الجمهورية الجزائرية الديمقراطية الشعبية

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة التعليم العالي و البحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة أبي بكر بلقايد – تلمسان –

Université Aboubakr Belkaïd – Tlemcen –

THESE

Présentée pour l'obtention du **grade** de **DOCTEUR EN SCIENCES**

En : Génie Biomédical

Spécialité : Génie Biomédical

Par : SETTOUTI Nesma

Sujet

Classification semi-supervisée des données médicales

Soutenue publiquement, le 02/06/2016, devant le jury composé de :

M. ABDERRAHIM Mohamed El Amine	MCA	Univ. Tlemcen	Président
M. CHIKH Mohamed El Amine	Professeur	Univ. Tlemcen	Directeur de thèse
M. BARRA Vincent	Professeur	Univ. Clermont Ferrand	Co- Directeur de thèse
M. ATMANI Baghdad	Professeur	Univ. Oran(Es-Senia)	Examinateur 1
M. EL BERRICHI Zakaria	Professeur	Univ. Sidi Bel Abbes	Examinateur 2
M. MESSADI Mohamed	MCA	Univ. Tlemcen	Examinateur 3

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Abou Bekr Belkaid
Faculté de Technologie
Département de Génie Biomédical
Laboratoire de Génie Biomédical GBM

THÈSE DE L'UNIVERSITÉ DE TLEMCEN

pour obtenir le grade de

DOCTORAT EN SCIENCES

Spécialité : **Génie Biomédical**

présentée et soutenue publiquement
par

Settouti Nesma

le 02 Juin 2016

Titre:

Classification Semi-Supervisée des données Médicales

Jury

Président du jury.	Dr. Abderrahim Mohammed El Amine,	MCA UABB Tlemcen
Examinateurs.	Pr. Atmani Baghdad,	UOES Oran
.	Pr. El-Berrichi Zakaria,	UDL Sidi Bel-Abbes
.	Dr. Messadi Mahamed,	MCA UABB Tlemcen
Directeur de thèse.	Pr. Chikh Mohamed Amine,	UABB Tlemcen
Co-Directeur de thèse.	Pr. Barra Vincent,	UBP Clermont-Ferrand

Je dédie ce travail à :

Mes grands parents,
Mes parents,
Mes frères et belles sœurs,
Mon neveu et nièces,
Mon mari,

Qu'ils trouvent ici l'expression de toute ma reconnaissance.

Remerciements

Une thèse est un effort collectif et elle est loin d'être un travail solitaire. En effet, je n'aurais jamais pu réaliser ce travail doctoral sans le soutien d'un grand nombre de personnes dont la générosité, la bonne humeur et l'intérêt manifestés à l'égard de ma recherche m'ont permis de progresser dans cette phase délicate. Je tiens donc ici à adresser mes remerciements à toutes les personnes qui ont participé de près ou de loin à cet ouvrage.

Je tiens tout d'abord à adresser mes remerciements les plus sincères à Professeur Mohamed Amine CHIKH pour avoir dirigé cette thèse et m'avoir permis de la réaliser dans les meilleures conditions. Je le remercie pour son soutien depuis mon Ingéniorat jusqu'à la fin de ma thèse à tous les niveaux. Son assistance et ses conseils m'ont chaque fois permis de rebondir dans les moments difficiles. Je le remercie vivement pour l'aide scientifique précieuse et pour tous les conseils donnés pendant la durée de cette thèse.

Mes remerciements s'adressent ensuite à Professeur Vincent BARRA qui a co-encadré cette thèse. Ses conseils avisés tout au long de mon séjour de finalisation de thèse au sein de son laboratoire LIMOS, Université Blaise Pascal, m'ont permis d'enrichir ce travail. J'aimerais également lui dire à quel point j'ai apprécié sa grande disponibilité et son respect sans faille des délais serrés de relecture des documents que je lui ai adressés.

Un grand merci à M. Mohammed El Amine ABDERRAHIM, Maitre de conférences à l'Université de Tlemcen, pour présider mon jury de thèse.

J'associe à ces remerciements M. Zakaria El-BERRICHI, Professeur à l'Université de Djilali El Yabes Sidi Belabes, M. Baghdad ATMANI, Professeur à l'Université d'Es Sénia Oran, et M. Mahamed MESSADI, Maitre de conférences à l'Université de Tlemcen pour avoir accepté d'examiner mon travail. J'imagine le travail que représente la lecture d'une thèse et les en remercie sincèrement.

Les différents membres du laboratoire Génie Biomédical GBM et tout particulièrement de l'équipe CREDOM qui ont également contribué à la réalisation de ces travaux, que ce soit au travers de longues discussions scientifiques ou bien grâce à l'ambiance chaleureuse et au très bon état d'esprit qu'ils entretiennent. J'ai une pensée particulière pour Mostafa El HABIB DAHO, Meryem SAIDI, Mohammed Amine LAZOUNI avec qui nous avons démarré nos activités de recherche.

Je tiens également à remercier Professeur Alexandre AUSSEM et plus particulièrement M. Haytham ELGHAZEL du laboratoire LIRIS, Claude Bernard Lyon 1, pour l'idée de ce sujet de thèse ainsi que le partage de leur savoir et connaissances dans ce domaine.

Mes derniers remerciements vont à Mohammed El Amine BECHAR, Khalida DOUIBI et Karima ENNAOUI pour leur soutien et pour toute l'aide qu'ils ont su m'apporter.

Résumé

En classification supervisée des données médicales, l'hypothèse de classement est apprise à partir d'un échantillon d'apprentissage généralement constitué de données étiquetées par un ou plusieurs médecins, experts du domaine. Néanmoins, face aux importants volumes de données disponibles actuellement, le coût de l'étiquetage des données devient très coûteux. Ainsi, il est impraticable, voire impossible d'étiqueter toutes les données disponibles. Cependant, sachant que la performance d'un classifieur est liée au nombre de données d'apprentissage, la principale question qui ressort est comment améliorer l'apprentissage d'un classifieur en ajoutant des données non étiquetées à l'ensemble d'apprentissage. La technique d'apprentissage issue de la réponse à cette question est appelée l'apprentissage semi-supervisé. Au cours de ces dernières années la classification semi-supervisée, qui fait usage des données non étiquetées pour améliorer la précision de l'hypothèse de classification ciblée, a connu un essor important dans le domaine de l'apprentissage artificiel.

Les méthodes d'ensemble comme approche de classification, nous offrent des taux erreurs minimales. Elles permettent de prendre naturellement en compte l'information apportée par les données non étiquetées dans l'apprentissage de la règle de classement. Ces algorithmes font appel de manière répétée à un apprenant de base pour produire différentes hypothèses ; au moment de la prédiction, ces hypothèses sont combinées au sein d'un vote. L'intérêt des techniques de combinaison a été établi par les faits que : quel que soit le mode de production des hypothèses et quelles que soient les modalités du vote final, l'erreur globale observée est plus faible que celle de n'importe quelle hypothèse impliquée dans le vote. Dans ces méthodes, l'importance de la diversité des hypothèses a été justifiée d'où l'intérêt de l'algorithme de la Forêt Aléatoire par son ensemble d'arbres de décision. Ces derniers ont la particularité d'être sensibles à l'ordre de présentation des données, cela a permis à la Forêt Aléatoire d'être une méthode très adaptée pour la tâche de classification des données réelles.

Dans cette thèse, nous nous appliquons sur la compréhension de cet algorithme des Forêts Aléatoires (RF) qui est considéré comme une technique de référence, compétitive avec la plupart des méthodes d'ensemble. De ce fait, avant d'aborder et d'élaborer notre approche en apprentissage semi-supervisé, nous consacrons toute une partie de cette thèse : à l'étude, l'optimisation et l'amélioration des performances de prédiction des RFs dans le contexte supervisé.

Par la suite, nous détaillons notre problématique majeure à savoir l'étiquetage automatique par apprentissage semi-supervisé. Nous introduisons de manière progressive le concept d'apprentissage semi-supervisé dans les méthodes d'ensemble, en commençant par les Forêts Aléatoires en apprentissage semi-supervisé, l'algorithme *co-Forest* et son application à la segmentation d'images médicales, et en dernier lieu, notre contribution proposée au problème d'annotation des données médiales à grande dimension (l'approche *Optim co-Forest*).

Dans la dernière partie de cette thèse, nous nous intéressons plus particulièrement à la sélection de variables en apprentissage supervisé et semi-supervisé. Le procédé de mesure de variables d'importance dans le paradigme des Forêts Aléatoires (RF) a eu une grande influence sur nos approches proposées. Afin d'améliorer l'efficacité de la sélection des ensembles de données à grande dimension, nous proposons notre approche d'évaluation de sélection de variables pertinentes en apprentissage semi-supervisé.

Nos divers algorithmes d'apprentissage supervisé et semi-supervisé ont été testés sur des données médicales artificielles et réelles et ont abouti à des résultats encourageants. Ces évaluations ont été enrichies par une discussion sur les avantages et les limites de chacune des méthodes développées.

Mots clés

Apprentissage semi-supervisé ; classification, méthodes d'ensemble ; Forêt Aléatoire ; co-Forest ; sélection de variables ; données à grande dimension ; données médicales.

Abstract

In supervised medical data classification, the classifier is learned from a training sample usually labeled by one or more physicians, experts in the field. Now, labeling, as well as the collection of the data itself, is a process which requires considerable human effort and is therefore very expensive. However, since we know that the larger the number of training samples, the better the performance of the classifier, making assumption that we have a correct classifier model, the issue becomes finding a way to improve supervised learning by adding unlabeled data. Training using both labeled and unlabeled data is called semi-supervised learning. In recent years, the semi-supervised classification, which used unlabeled data to improve the accuracy of the learned hypothesis, experienced significant growth, this in particular in the Machine Learning community.

The ensemble methods as classification approach offer a low generalization errors. It allow naturally to take into account the information provided by unlabeled data in the classification rule process. These algorithms used repeatedly learner hypothesis to generate different assumptions ; at the time of the prediction, these assumptions are combined in a vote. The interest of these ensemble methods has been established by the facts that : regardless of the production mode of hypotheses and whatever the terms of the final vote, the observed error of generalization is lower than any of the voters. In these methods, the importance of diversity of hypothesis has been demonstrated and that's why the algorithm Random Forest by its combination of decision trees that have the distinction of being sensitive to the order of presentation of data, remains the best candidate.

In the first part of the thesis, we focus on understanding the Random Forests (RF) algorithm which is considered as a reference technique, competitive with most existing ensemble methods. Therefore, before addressing and developing our approach in semi-supervised learning. We devote a whole part of this thesis to study, optimize and improve the RF prediction performance in the supervised learning.

Subsequently in the second part, we tackle the heart of our problem namely the automatic labeling in semi-supervised learning. To this end, we first adopt gradually the semi-supervised concept in ensemble methods, beginning with Random Forests in semi-supervised learning algorithm *co-Forest* and its application to semi-supervised segmentation of medical images, and finally, our contribution to the problem of the annotation of high dimensional medial datasets (*Optim co-Forest* approach).

In the last part of this thesis, we particularly interest in supervised and semi-supervised feature selection. The measure of feature importance in the paradigm of random forest (RF) had a great influence on our proposed approaches. To improve the efficiency of selection of high dimensional datasets, we evaluate our approach in selection of the most important and relevant features for a better learning and therefor, a better discrimination of noisy examples especially in the context of high dimensional dataset.

Experimental results on both artificial and real problems show the usefulness of our different semi-supervised algorithms. Depth analysis of the experimental results reveals the advantages and the limits of each method.

Keywords

Semi-supervised learning ; classification ; ensemble methods ; Random Forest ; co-Forest ; feature selection ; high dimensional dataset ; medical dataset.

ملخص

من خلال تصنيف منظم لبيانات طبية فإن فرضية الترتيب منشأة عادة من عينة التدريب المكونة من البيانات المعلّمة من طرف واحد أو أكثر من الأطباء والخبراء فى هذا المجال. ولكن أمام الكمية الهائلة من البيانات المتوفرة حالياً ، فإن تكلفة تعليم البيانات أصبحت جد باهظة ، وبالتالى غير ممكن بل مستحيلاً تعليم كل البيانات الموجودة. وعلماً أن أداء المصنف مرتبط مع عدد بيانات التدريب فإن سؤالاً رئيسياً يطرح: "كيف يمكن تحسين تدريب مصنف حين تضاف إليه بيانات غير معلّمة لمجموعة التدريب". التقنية المستنتجة من الجواب على هذا السؤال تسمى التعلم نصف المنظم. خلال السنوات الأخيرة فإن التعلم نصف المنظم الذى يستعمل البيانات غير المعلمة لتحسين دقة فرضية التصنيف المستهدف عرفت تقدماً ملحوظاً فى مجال التعليم الآلى كوسيلة للتصنيف ، تمنح لنا مناهج المجموعة (جماعات) نسباً ضئيلة من الأخطاء ، وتسمح لنا بأخذ بعين الاعتبار المعلومة المستنبطة من البيانات غير المعلّمة خلال تدريب قاعدة الترتيب وهذا بصفة طبيعية محضة.

هذه الخوارزميات تتطلب بصفة مكرّرة وجود قاعدة تعلُّم حتى يتسنى لها إنتاج مختلف الفرضيات ، وأثناء التنبؤ فإن هذه الافتراضات تكون ممزوجة ضمن التصويت. تأسست مصلحة مناهج المجموعة عن طريق الحقائق التالية بغض النظر عن نمط إنتاج افتراضيات ومهما كانت شروط التصويت النهائى فإن الخطأ العام الملحوظ أضعف من خطأ أي افتراضية مدرجة فى التصويت ضمن هذه النماذج تبرّرت أهمية تنويع الافتراضات ممَّ الحاجة الى خوارزميات الغابة العشوائية بمجموعة أشجار القرار. هذه الأخيرة تتميز بحاستها لترتيب تقديم البيانات ذلك ما سمح للغابة العشوائية أن تكون منهاج جد مناسب للقيام بتصنيف البيانات الحقيقية.

نركز فى هذه الأطروحة على فهم خوارزمية الغابة العشوائية (RF), التى تعتبر تقنية مرجعية منافسة لأغلبية المناهج الجماعية. لهذا قبل الشروع والقيام بمنهجية التعلّم نصف المنظم ، نكرس جزءًا كبيراً من الأطروحة ل: دراسة أمثلية وتحسين مهارات تنبؤ الغابات العشوائية(RFs) في مضمون منظم.

وبعد ذلك ، سنفصل اشكاليتنا الرئيسية التالية: التعلم الأتوماتيكى بالتدريب نصف المنظم ، ثم ندمج تدريجياً نظرية التعلم نصف المنظم داخل مناهج جماعية مبدئين بالغابات العشوائية التى هى فى طور التعلم نصف المنظم ، خوارزمية(co-forest), وتنفيذها لتقسيم صور طبية وفي الأخير مساهمتنا المقترحة لمشكل تنقيط البيانات الطبية ذات الحجم الكبير(optim-coforest).

فى الجزء الأخير من الأطروحة ، نهتم وبشكل خاص باختيار خصائص ضمن التدريب نصف المنظم قد أثرت عملية قياس خصائص ذات الأهمية فى نموذج الغابات العشوائية على نماذجنا المقترحة. لتحسين فعالية اختيار جماعة البيانات ذات الحجم الكبير نقترح طريقتنا لتقييم اختيار الخصائص الملائمة في التعلم نصف المنظم.

لقد اختُبرت مختلف خوارزمياتُنا بالتعلم المنظم ونصف المنظم على بيانات طبية مصطنعة وحقيقية وأدّت إلى نتائج مشجعة. إن تقويم كل هذا أثري بمناقشة حول مزايا وحدود النماذج الموسعة.

الكلمات المفتاحية

التعلم نصف المنظم ، التصنيف ، المناهج الجماعية ، الغابات العشوائية ،CO-FOREST ، اختيار الخصائص ، البيانات ذات الحجم الكبير ، البيانات الطبية.

Table des matières

Table des figures

Liste des tableaux

Glossaire

ACP : Analyse en Composantes Principales
APVs : Adjusted p-values
ASU Arizona State University
AUC : Area Under Curve
Bagging : Bootstrap Aggrigating
BS : Bagging Score
CART : Classification and Regression Trees
CDR : Cup-Disc Ratio
CIF : Conditional Inference Forest
CIT : Conditional Inference Tree
CLS : Constrained Laplacian Score
CYL : Simulated keratometric cylinder
DSE : Dossiers de Santé Electroniques
EM : Expectation-maximisation algorithm
EoC : Ensemble of Classifiers
FCART Fuzzy Classification and Regression Trees
FCM : Fuzzy C-means
fMRI : Functional Magnetic Resonance Imaging
Forest-RI : Forest with Random Input
FRF : Fuzzy Random Forest
FS : Feature Selection
GDI : Indice de Gini
IIG : Imprecise Info-Gain
Isomap : Isometric Feature Mapping
KC : Cas Kératocone
KCS : Cas suspects de Kératocône
LLE : Locally Linear Embedding
LSDF : Locality Sensitive Semi-supervised Feature Selection method
LVW : Las Vegas Wrapper
MCS : Multiple Slassifer Systems
MCLR : Monte Carlo Logic Regression
mRMR : Min-Redundancy, Max-relevance
OC : Optic Cup
OD : Optic Disc
Od : Œil droit
ONH : Tête du Nerf Optique
OOB Out-Of-Bag
Og : Œil gauche
OSI : Opposite Sector Index
PAC : Probably Approximately Correct
PERT : PErfect Random Tree

PIO : Pression intra-oculaire
PPV : Les plus proches voisins
RBF : Réseau à Base Radial
RCI : Roadway Characteristics and Inventory database
RF : Random Forest
RFE-SVM : Recursive Feature Elimination–Support Vector Machine
RGB : Red Green Blue
RNs : Réseaux de neurones
ROC : Receiver Operating Characteristic
RP : Random Patches
RSM : Random Subspaces Method
Se : Sensibilité
SEFR : Semi-supervised Ensemble learning guided Feature Ranking method
SETRED : self-training with data editing
SimK : Simulated keratometry
Sp : Specificité
sSELECT : sélection de variable par analyse spectrale
SSL : Semi-Supervised Learning
SubBag : Subspaces Bagging
SVM : Support Vector Machine
TC : Taux de Classification
UCI : University California Irvine
VIM : Variable Importance Measure

Première partie

I

Problématique et Objectifs

CHAPITRE 1

Introduction Générale

1 Motivations

Le stockage de l'information médicale et des données médicales ne cesse d'augmenter avec le développement technologique. La nécessité de développer et d'organiser de nouvelles façons à fournir : les informations de santé, les données et les connaissances ont été accompagnées par des avancées majeures dans les technologies de l'information et de la communication. Ces nouvelles technologies accélèrent l'échange et l'utilisation des données, informations et connaissances et éliminent les barrières géographiques et temporelles. Ces processus favorisent grandement le développement de l'informatique médicale.

L'informatique médicale existe maintenant depuis presque cinquante ans et elle a connu une croissance très rapide pendant cette dernière décennie. Malgré les avancés majeures de la science et de la technologie des soins de santé, cette discipline d'informatique médicale recèle le potentiel d'améliorer et de faciliter l'évolution constante de la masse de données toujours plus large concernant l'étiologie, la prévention et le traitement des maladies. Ce large champ d'intérêt couvre de nombreux thèmes de recherches multidisciplinaires avec un retentissement essentiel pour les soins des patients et le diagnostic des médecins.

La représentation et l'extraction de connaissances médicales occupent une place importante dans le paysage de la recherche de l'informatique médicale [1]. La nécessité de décrire sans ambiguïté les connaissances médicales dans des environnements cliniques, intrinsèquement caractérisées par des ambiguïtés terminologiques, diversité des données, a donné lieu à l'utilisation des systèmes d'aide au diagnostic au médecin [2].

Pour faire face à cet accroissement exponentiel des données à une échelle sans précédent, les techniques d'aide au diagnostic nécessitent des transformations et optimisations stratégiques majeures pour assurer l'exploitation, l'exploration et l'interprétabilité de ces masses de données. Elles nécessitent également une prise de conscience et une modification des pratiques des experts du domaine, pour lesquels ces évolutions constituent un défi pour minimiser les erreurs médicales.

2 Objectifs et Contexte de la thèse

Les quantités croissantes de données médicales produites annuellement comprennent une source inestimable de connaissances à découvrir, représenter et exploiter pour améliorer et renforcer le diagnostic médical. L'acquisition de ces dernières est devenue relativement simple et peu coûteuse, les ensembles de données sont de plus en plus importants, tant

en ce qui concerne le nombre de variables, qu'en terme de nombre d'instances. Cependant, ce n'est pas le cas pour les instances étiquetées. Habituellement, le coût d'obtention de ces labels est très élevé, pour cette raison, les données non étiquetées représentent la majorité des cas, surtout en comparaison avec la quantité de données étiquetées. L'utilisation de ces données nécessite un soin particulier, car plusieurs problèmes se posent à l'augmentation de la dimensionnalité et à l'absence d'étiquettes.

La fouille de données supervisée ou non supervisée, fournit les outils méthodologiques pour annoter ces données [3]. Les méthodes supervisées s'adressent généralement à la classification des données sur la base de connaissances préalables acquises par un apprentissage sur des données déjà annotées par l'expert, alors que les données dans les méthodes non supervisées (clustering) sont basées uniquement sur la similitude des instances de données sans apprentissage. Cette dernière pourrait être considérée comme avantageuse sur les méthodes supervisées. Cependant, les méthodes non supervisées, en général, nécessitent que le nombre de groupes cibles soit pré-spécifié par l'utilisateur, et les résultats ne soient pas associés avec des étiquettes de classe.

Dans ce contexte, le praticien dispose souvent d'un grand échantillon de données non étiquetées et d'un plus petit nombre étiqueté. Avec la disponibilité des données non étiquetées et la difficulté d'obtenir des étiquettes, les méthodes d'apprentissage semi-supervisé ont acquis une grande importance. Contrairement à l'apprentissage supervisé, l'apprentissage semi-supervisé vise les problèmes avec relativement peu de données étiquetées et une grande quantité de données non étiquetées.

L'approche semi-supervisée qui se situe à l'intersection entre l'approche supervisée et non supervisée, est alors une solution envisageable [4]. L'apprentissage semi-supervisé cherche à extraire une règle de décision ou de régression d'un ensemble d'apprentissage, avec une particularité : cet ensemble contient à la fois des objets étiquetés, et d'autres qui ne le sont pas (non étiquetés).

Dans cette thèse, nous nous focalisons sur l'amélioration des performances de la classification supervisée en utilisant les données non étiquetées (classification semi-supervisée). Nos principaux objectifs sont de répondre aux questions de recherche suivantes :

- *La connaissance seule des points avec labels est-elle suffisante pour construire une fonction de décision robuste capable de prédire correctement les étiquettes des points non étiquetés ?*
- *Comment juger de la pertinence d'un modèle à l'aide des données non étiquetées ?*
- *Comment améliorer les performances de ce dernier ?*
- *Dans le passage à l'échelle des données, ce prototype peut-il toujours être performant ?*
- *Le marquage manuel des régions d'intérêt des images médicales est une tâche fastidieuse pour l'expert médical. Comment améliorer l'annotation des images médicales au niveau du pixel pour automatiser les annotations d'une manière structurée ?*

3 Problématiques

L'apprentissage automatique traite généralement deux problèmes différents : l'apprentissage supervisé (classification) et l'apprentissage non supervisé (clustering). Cependant, au cours des dernières années, de nouveaux paradigmes ont émergé. Ils hybrident ces deux approches afin d'étendre l'applicabilité des algorithmes d'apprentissage automatique. Le premier paradigme qui a vu le jour est l'apprentissage semi-supervisé [5, 6], où l'ensemble d'apprentissage est composé de deux parties différentes. Dans la première partie, les étiquettes de classe sont spécifiées, alors que dans la seconde partie, seules les caractéristiques des données sont disponibles. L'importance de ces problèmes réside dans le fait que les données étiquetées sont souvent difficiles à obtenir, tandis que les non marquées sont facilement disponibles. Ainsi, l'utilisation des données non étiquetées

peut être un moyen d'améliorer les performances des algorithmes supervisés à faible coût supplémentaire. Les récents ouvrages [7, 8, 10] dans le domaine d'apprentissage automatique montrent une activité importante autour de cette question.

Plusieurs techniques ont été développées pour réaliser la tâche d'apprentissage semi-supervisé. Il existe principalement trois paradigmes [7] [8] qui abordent le problème de la combinaison des données labellisées et non labellisées et ce afin d'améliorer les performances de classification. Nous citons en particulier : l'apprentissage semi-supervisé, l'apprentissage transductif et l'apprentissage actif.

L'apprentissage semi-supervisé (SSL) renvoie à des méthodes qui exploitent soit les données non étiquetées par l'apprentissage supervisé ; ou bien les données étiquetées par l'apprentissage non supervisé.

L'apprentissage Transductif regroupe les méthodes qui tentent également d'exploiter les exemples non-étiquetés, mais en supposant que les exemples non étiquetés sont exactement les exemples de test.

L'apprentissage actif se réfère à des méthodes qui sélectionnent les exemples non étiquetés les plus importants. Un oracle peut être proposé pour l'étiquetage de ces instances, dont l'objectif est de minimiser l'étiquetage des données [9]. Parfois, il est appelé échantillonnage sélectif ou sélection d'échantillon.

Dans cette thèse, nous nous focalisons sur l'amélioration des performances de la classification supervisée en utilisant les données non étiquetées (SSL). Nous replaçons d'abord la classification semi-supervisée dans le cadre des problèmes de classification. Nous nous contentons à l'utilisation des méthodes d'ensemble comme méthodes de classification en apprentissage semi-supervisé.

Dans le cadre de données à grande dimension un autre problème surgit et concerne le nombre très important d'attributs. Un nombre élevé de d'attributs peut s'avérer pénalisant pour un traitement pertinent et efficace des données. D'un côté par les problèmes algorithmiques que cela peut entraîner (liés au coût calculatoire et à la capacité de stockage nécessaire), et d'autre part, parmi les variables existantes peuvent être non-pertinentes, inutiles et/ou redondantes perturbant ainsi le bon traitement des données. Or, il est très souvent difficile voire impossible de distinguer les variables pertinentes des variables non-pertinentes. L'intérêt d'application de techniques de sélection de variables est devenu essentiel pour améliorer les performances des classifieurs (Han et al. [10]) ; (Guyon & Elisseeff [11]) ; (Liu & Motoda, [12] [13]).

Les algorithmes de sélection d'attributs supervisé nécessitent la définition des labels de toutes les données. Par conséquent, la procédure de labéllisation réalisée par un expert humain peut s'avérer fastidieuse et coûteuse en temps de travail. C'est pour cette raison que, pour des applications réelles, on est généralement en présence de bases de données formées de nombreuses données non labellisées et avec peu de données labellisées. Ce contexte d'apprentissage est appelé semi-supervisé car l'analyste exploite à la fois les données non labellisées et les quelques données labellisées.

4 Contributions

La tâche d'apprentissage vise à apprendre à étiqueter des exemples pour lesquels il n'y a pas d'étiquette connue, soit parce qu'il est difficile et coûteux d'obtenir ces étiquettes, soit parce qu'il s'agit d'exemples non encore observés. Ainsi, pour entraîner un classifieur, nous disposons souvent d'une base d'exemples étiquetés, et d'une grande base d'exemples non étiquetés.

De ce fait, plusieurs travaux pionniers ont proposé des méthodes intéressantes [5] [14] [15] [16] [17] [18] [19] [6] montrant l'intérêt de ce type d'apprentissage. En effet, la plupart des travaux ont relevé un point très important, démontrant clairement que la prise en compte d'exemples non labellisés pouvait dégrader les performances par rapport à un apprentissage purement supervisé. Il est devenu donc clair qu'une analyse plus fondamentale est nécessaire afin de mieux comprendre les conditions qui permettent non seulement d'espérer, mais aussi de garantir des améliorations de performances. De cette vision est née notre approche d'apprentissage semi-supervisé.

Depuis la prolifération des bases de données partiellement étiquetées, la sélection de variables a connu un développement important dans le mode semi-supervisé. Nous proposons dans cette thèse d'aborder cette problématique avec une méthode d'ensemble à base de mesure d'importance de variables. Ce modèle est basé sur la classification semi-supervisée pour sélectionner les variables les plus pertinentes.

Notre contribution consiste à développer une méthode d'apprentissage semi-supervisé en présence d'un grand nombre de variables tout en fournissant une sélection de variables les plus pertinentes de manière intégrée. L'idée est de mettre en exécution les méthodes ensemblistes qui consistent à combiner plusieurs modèles pour produire une solution plus performante et plus robuste.

L'application de méthodes ensemblistes est préconisé dès qu'on veut dépasser le cap de réalisation de meilleures performances de prédiction. En effet, au lieu d'essayer d'optimiser une méthode ''en un seule fois'', les méthodes d'ensemble génèrent plusieurs règles de prédiction et mettent ensuite en commun leurs différentes réponses. L'heuristique de ces méthodes est qu'en générant beaucoup de prédicteurs, nous explorons massivement l'espace des solutions, et qu'en agrégeant toutes les prédictions, nous récupérons un prédicteur qui rend compte de toute cette exploration.

Nos contributions dans cette thèse sont résumées dans les trois points suivant :

1. Propositions et applications de différentes améliorations aux méthodes d'ensemble pour la sélection et la classification en apprentissage supervisé.
2. Propositions d'une approche de classification des données à grande échelle en apprentissage semi-supervisé par les méthodes d'ensemble.
3. Propositions d'une méthode de sélection de variables en apprentissage semi-supervisé sur les données à grande échelle.
4. Proposition d'une méthode d'annotation des images médicales au niveau du pixel pour la segmentation des régions d'intérêt de l'image.

5 Organisation du manuscrit

Ce manuscrit est structuré en deux grands volets répartis en cinq parties comme suit :

Partie 1 : Problématiques et Objectifs Introduit en deux chapitres, les problématiques qu'abordent cette thèse.

Le premier volet :

Classification en apprentissage supervisé et semi-supervisé par les méthodes d'ensemble.

Partie 2 : État de l'art et Contribution en apprentissage supervisé : Classification des données par approche ensembliste Cette partie relate en trois chapitres, les travaux des méthodes d'ensemble en apprentissage supervisé. Dans chacun des chapitres,

une amélioration est proposée et apportée aux méthodes d'ensemble, leur évaluation est réalisée sur les données médicales.

Partie 3 : État de l'art et Contribution en apprentissage semi-supervisé : Classification des données par approche ensembliste Les objectifs et la contribution apportée à la problématique de classification en semi-supervisé sont décrits en cette partie dans trois chapitres.

Le second volet :

Sélection de variables en apprentissage supervisé et semi-supervisé par les méthodes d'ensemble

Partie 4 : État de l'art et Contribution dans la sélection de variables en apprentissage supervisé et semi-supervisé par approche ensembliste Trois contributions dans la sélection de variables par mesure d'importance en apprentissage supervisé sont dépeintes dans les deux premiers chapitres englobant cette partie. Le chapitre 3 expose la sélection de variables par methode d'ensemble en apprentissage semi-supervisé.

Partie 5 : Conclusion et perspectives

La conclusion générale résume une synthèse des contributions apportées ainsi que les chemins définissant les perspectives possibles pour des travaux futurs.

CHAPITRE 2

Problématique générale

1 Données à grande échelle : propriétés

Dans les années à venir, on produira dans le monde plus de données de recherche que tout ce qui a été produit dans l'histoire de l'humanité [20]. Les nouvelles technologies, en progrès constant, vont permettre d'augmenter considérablement les vitesses et volumes de production des données. Ces données seront générées à partir d'appareils à très haut débit comme les séquenceurs, les simulations numériques de haute performance, les capteurs (e.g., environnementaux), les imageries scientifiques, les satellites. Les catégories les plus concernées par l'avènement des méthodes haut débit se situent principalement en bio-informatique et en bio-statistique [21].

Ces évolutions technologiques dans le numérique ont conduit durant les années 2000 à l'apparition de la notion de *Big Data* qui inverse le paradigme précédent dans lequel la production de données était gouvernée par la nature des problématiques [22]. A l'heure actuelle la question est donc plus de savoir comment organiser et analyser ces données que de les produire. Plus récemment, l'Open Data s'inscrit dans cette dynamique technologique.

Depuis, le phénomène, toujours en cours d'évolution, s'est étendu à l'ensemble des données, posant à tous des questions d'ordre technologique, stratégique, éthique et juridique.

L'évolution extrêmement rapide des technologies d'acquisition des données génère aussi dans certains domaines une obsolescence rapide des outils permettant d'en exploiter la production, dès lors que ces outils sont spécifiques d'une technologie particulière. Il n'en reste pas moins que les outils phares peuvent perdurer sur de très longues périodes sans être remplacés.

La majorité des données est désormais disponible sous forme numérique. Par contre, de nombreuses données nécessitent d'être structurées sous la forme de bases de données afin de pouvoir être pleinement utilisables et partageables. Du fait de l'accumulation et du manque de moyens humains, il est désormais de plus en plus fréquent que les données soient à la fois mal analysées et sous-employées [23].

La pluralité des applications rendues possibles mais aussi la diversité des données acquises et des technologies d'acquisition rendent nécessaire le développement de méthodologies appropriées et de nouveaux standards. Les problèmes soulevés par l'accroissement des volumes de données portent non seulement sur leur analyse statistique, mais aussi sur leur analyse « tout court » (extraction d'information, traitement algorithmique). Ces

traitements doivent pouvoir être adaptés, optimisés, de façon à pouvoir supporter le changement d'échelle. Par ailleurs, le volume des données change aussi la nature des questions qui peuvent être traitées, et pose des problèmes de modélisation en amont de tout traitement qu'il soit algorithmique ou statistique.

2 Données Annotées : propriétés

La richesse et la diversité des données biomédicales, permettent l'intégration translationnelle de multiples études bio-informatiques (translational bioinformatics research) [24], [25], [26], [22].

Cependant, les découvertes qui pourraient être réalisées par la fouille des données biomédicales sont limitées car la plupart des ressources publiques ne sont pas généralement décrites à l'aide de terminologies et d'ontologies.

La communauté biomédicale reconnaît d'ores et déjà l'importance des terminologies et des ontologies pour faciliter l'intégration de données et permettre de nouvelles découvertes [27]. Cependant, la variété des données est très importante et celles-ci sont rarement annotées par des terminologies claires et simples décrites dans des ontologies biomédicales.

Le plus souvent, les éléments d'une ressource (e.g., données expérimentales, diagnostics, maladies, échantillons, essais cliniques, images) ne sont pas toujours annotés, et quand c'est le cas, c'est avec des méta-données textuelles qui décrivent cet élément. Le problème est que ces descriptions textuelles sont rarement structurées. Il existe donc un challenge qui consiste à produire pour ces descriptions des annotations (ou labels, tags, étiquettes) qui faciliteraient la recherche et l'exploitation de ces données ainsi que leur intégration dans des systèmes d'aide au diagnostic au médecin [25], [28].

Habituellement, l'annotation peut être construite de trois manières principales : manuelle, automatique et semi-automatique.

- Dans l'annotation manuelle, des liens entre les données et les concepts sont fournis par des experts du domaine.
- Dans l'annotation automatique, des programmes spécialisés font l'analyse des données pour fournir de tels liens.
- Dans l'annotation semi-automatique, les programmes spécialisés suggèrent des liens entre les données et les concepts qui sont ensuite validés par des experts du domaine [29].

Pour expliquer plus clairement le concept d'annotation, citons pour exemple l'annotation et le séquençage des génomes : Le séquençage de l'ADN donne accès à une succession de nucléotides. Obtenir une carte complète et précise du génome d'un organisme se révèle être une tâche à la fois complexe et utile pour les scientifiques. Les généticiens et les biologistes moléculaires utilisent de multiples cartes pour explorer le terrain que constitue le génome : cartes de liaisons génétiques (mode de transmission des gènes dont les allèles déterminent les différences de phénotypes), cartes cytogénétiques (établies à partir de certaines positions caractéristiques visibles au microscope tels que les points de cassures de réarrangements chromosomiques) et des cartes de restriction comportant les sites de l'ADN sensibles aux enzymes de restriction [30].

Une fois la séquence assemblée, il s'agit de chercher dans ces successions de nucléotides des informations : c'est l'annotation. Celle-ci se déroule principalement en deux temps : l'annotation automatique et l'annotation manuelle.

L'annotation automatique consiste à développer et faire tourner des algorithmes sur les

séquences assemblées afin de reconnaître les gènes. C'est le travail des bio-informaticiens.

L'annotation manuelle est un travail long et fastidieux qui nécessite la mobilisation d'une communauté d'experts pour des groupes de protéines ou de fonctions biologiques. Il s'agit de parcourir une à une la séquence de gènes prédit par les bio-informaticiens et de décider si l'algorithme a bien prédit (validation le gène) ou à défaut (non validation du gène). Cette étape doit être répétée car une annotation peut toujours être améliorée. L'ensemble de ces informations est ensuite stocké dans des bases de données consultables.

Toujours dans le registre médical, un grand nombre d'images médicales sont générées quotidiennement dans les hôpitaux et les établissements médicaux, les besoins pour efficacement traiter, indexer, rechercher et récupérer ces images sont d'une importance primordiale. Étant donné l'augmentation des archives de contenus visuels disponibles, ces bases de grande taille motivent le développement de méthodes automatiques d'indexation, d'annotation et de requête. La quantité d'images possédant des annotations plus ou moins structurées qui ne cessent d'augmenter. Plusieurs techniques d'apprentissage automatique peuvent en bénéficier en estimant de manière plus robuste des modèles de prédiction de labels.

L'annotation d'images médicales peut être classée en deux catégories : l'annotation au niveau de l'image globale et celle au niveau pixel.
- L'annotation au niveau de l'image représente les méta-données de l'image, y compris où, quand, et comment l'image a été acquise.
- L'annotation au niveau pixel englobe des informations sur l'emplacement spatial d'un seul point ou une région d'intérêt (ROI : Region Of Interest). Elle peut être associée avec les marqueurs cliniques (par exemple, les tumeurs, les fractures osseuses, les caillots sanguins) présentés dans l'image. Les annotations peuvent être élaborées à partir d'observations ou de calculs. Les professionnels médicaux créent des annotations d'observations en associant un texte et/ou en représentant les majorations par des informations textuelles sur les ROI.

Les procédés traditionnels d'annotation, comme le marquage manuel des régions d'intérêt ainsi que l'indexation des images médicales sont des tâches fastidieuses pour l'expert médical. Pour améliorer l'annotation des images médicales au niveau pixelique et de l'image globale, de nouvelles technologies sont nécessaires pour automatiser les annotations d'une manière structurée et créer du contenu sémantique que les utilisateurs peuvent facilement rechercher.

Différents cadres d'applications comprennent le même problème d'annotation et nécessite l'intervention d'expert pour effectuer cette tâche. Exemples : dans le domaine de la reconnaissance vocale, l'enregistrement nous donne d'énormes quantités de données audio dont le coût est négligeable. Toutefois, l'étiquetage exige par la suite, quelqu'un pour l'écouter et le saisir. Pour l'intégration d'une nouvelle citation PubMed, le titre ainsi que le résumé de l'article correspondant sont indexés (grâce à des annotations manuelles) avec des termes de MeSH améliorant significativement la performance des recherches d'articles [22]. Des situations similaires sont valables pour la télédétection, la reconnaissance des visages, la détection des intrusions dans les réseaux informatiques [31], etc . . .

Toutefois, l'annotation des données biomédicales reste encore marginale. Ce n'est pas une pratique courante pour plusieurs raisons [22], [32] :

- Les annotations ont le plus souvent besoin d'être créées manuellement par des experts ou directement par les auteurs des données,
- La nécessité d'intervention d'experts du domaine ciblé, qui ne sont pas forcement disponible pour cette tâche,
- Le nombre de terminologies biomédicales disponibles est important. En outre, ces terminologies changent régulièrement et se chevauchent les unes les autres,

- Les annotations (labels, tags, étiquettes) sont généralement dans des formats différents et ne sont pas toujours accessibles aux utilisateurs,
- Les utilisateurs ne connaissent pas toujours les terminologies biomédicales pour faire les annotations eux-mêmes,
- L'annotation est souvent une tâche supplémentaire ennuyeuse et sans retour immédiat pour l'utilisateur.

3 Résumé

Face au rythme de croissance de ces volumes de données, l'annotation manuelle présente aujourd'hui un coût prohibitif. Dans cette thèse, nous nous intéressons aux approches produisant des annotations automatiques qui tentent d'apporter une réponse à ce problème [33]. Nous nous sommes focalisés sur les bases de données médicales et biologiques à moyenne et grande échelle dimensionnelle, ainsi qu'aux images médicales. Contrairement aux nombreuses bases généralistes (ne possèdent pas de connaissances à priori sur leur contenu) pour les bases bio-médicales, il est important de tenir compte de leur spécificité lors de l'élaboration d'algorithmes d'annotation automatique.

Il y a essentiellement deux types de méthodes utilisées pour l'annotation des données :

Les méthodes où une personne (expert) fournit des connaissances (linguistiques ou sur le domaine), et les méthodes dirigées par les données, où ces connaissances sont construites par apprentissage supervisé.

Ces deux types de méthodes ont certaines limitations [34], [35], [36] :

- Les méthodes à base de connaissances expertes sont simples à mettre en place mais coûteuses en temps pour ce qui est de la construction des connaissances. Elles ont aussi un potentiel de couverture réduit comparées aux méthodes statistiques.
- Les méthodes par apprentissage peuvent être très robustes si on dispose d'un bon nombre d'exemples d'entraînement et si les données de test sont du même type que les données d'apprentissage. Ces méthodes sont de ce fait dépendantes des données annotées, ressources qui ne sont pas toujours disponibles.

Il existe également des méthodes hybrides combinant ces deux techniques, nous nous intéressons à ces dernières et plus particulièrement à l'apprentissage semi-supervisé (SSL : semi supervised Learning). Avec la disponibilité des données non étiquetées et la difficulté d'obtenir des étiquettes, des méthodes d'apprentissage semi-supervisé ont acquis une grande importance. A la différence de l'apprentissage supervisé, l'apprentissage semi-supervisé vise les problèmes avec relativement peu de données étiquetées et une grande quantité de données non étiquetées.

L'apprentissage semi-supervisé est un des champs de recherche des plus intéressants pour une évolution du domaine de l'apprentissage artificiel au-delà du cadre supervisé à partir des données. Il peut également aider l'apprentissage humain (exemple l'aide au diagnostic médical) qui fonctionne fréquemment en mode supervisé.

Dans cette thèse, nous proposons de développer un modèle qui combine les données labellisées et non labellisées, qui sera plus précis que celui construit uniquement sur les données labellisées. Ainsi, un premier modèle sera fondé à partir des données labellisées et sera raffiné par la suite grâce aux données non labellisées. Nous nous focalisons sur l'amélioration des performances de la classification semi-supervisée en employant les méthodes d'ensemble.

En effet, la combinaison de différentes méthodes d'apprentissage/classification permet de tirer avantage de leurs forces tout en contournant leurs faiblesses. Aujourd'hui, force

est de constater que ce qu'on appelle maintenant systèmes multi-classifieurs (désignés souvent par l'acronyme MCS pour Multiple Classifier Systems) constitue une des voies les plus prometteuses de l'apprentissage automatique [8].

Parmi les systèmes multi-classifieurs émergent en particulier les Ensemble de Classifieurs, ou en anglais Ensemble of Classifiers (*EoC*). L'une des approches *EoC* les plus populaires et les plus efficaces consiste à combiner un ensemble de classifieurs de même type (par exemple un ensemble de réseaux de neurones, un ensemble d'arbres de décision, ou un ensemble de discriminants). Ces techniques ont fait l'objet de nombreux travaux et il en existe aujourd'hui un grand nombre capables de générer automatiquement des ensembles de classifieurs : Bagging [37], Boosting [38], Random Subspaces [39], ECOC [40], pour ne citer que les plus courantes. Ces méthodes ont pour objectif de créer de la diversité au sein d'un ensemble de classifieurs tout en cherchant à établir le meilleur consensus possible entre ces classifieurs.

De ce fait, l'approche ensembliste est la seule méthode qui permet la prise en compte rigoureuse recherchée tout en apportant des hypothèses supplémentaires sur les données non annotées. Dans ce cadre les deux principales contributions de ce travail sont le traitement des questions suivantes : *« Comment juger la pertinence d'un modèle à l'aide des données non étiquetées ? »* et *« Comment améliorer les performances de ce dernier ?».*

Dans le cadre d'application aux données à grande dimension, l'identification de sous-ensembles de variables pertinents parmi des milliers de variables potentiellement inutiles et superflues est un sujet de recherche ardu dans le domaine de reconnaissance de forme, ce qui a attiré énormément d'attention au cours des dernières années.

En apprentissage supervisé, les algorithmes de sélection de variables se basent uniquement sur des informations à partir de données étiquetées pour trouver les sous-ensembles de variables pertinents, à savoir, celles qui s'avèrent utiles pour la construction d'une hypothèse efficace. Un modèle de classification qui permet d'atteindre de bonnes solutions ou même meilleures avec un sous-ensemble restreint de caractéristiques [11], [41], [42].

Cependant, dans de nombreuses applications sur des données réelles, la quantité de données labellisées est très limitée et parait difficile de repérer et d'éliminer les variables redondantes et inutiles de l'ensemble des paramètres, surtout lorsqu'il s'agit de grandes dimensions. Cette situation se produit naturellement dans bon nombre d'applications dans le monde réel, où de grandes quantités de données peuvent être collectées à moindre coût et automatiquement. Mais l'étiquetage manuel des échantillons pose un problème de consommation de temps qu'on ne peut considérer comme acquis. Dans ce cas, les méthodes de sélection de variables non supervisées pourraient être envisagées pour exploiter l'information véhiculée par la grande quantité de données d'apprentissage non labellisées [43], [44], [45], [46].

La sélection de caractéristiques dans l'apprentissage non supervisé vise à trouver des sous-ensembles de variables pertinents qui produisent des regroupements "naturels" en regroupant les objets «similaires» en un ensemble basé sur une mesure de similarité.

De toute évidence, la combinaison des deux paradigmes (supervisé ou non) a permis l'émergence d'approches semi-supervisées sophistiquées pour la sélection de variables qui peuvent gérer à la fois les données étiquetées et non étiquetées. Le problème de la sélection de variables semi-supervisées a suscité beaucoup d'intérêt récemment et son efficacité a déjà été démontrée dans de nombreuses applications [47], [48], [49], [50].

Nous peuvons résumer le problème de *« malédiction de dimension »* par l'aphorisme de Liu et Motoda *"Less is more"* [12] qui met en exergue la nécessité de supprimer l'ensemble des portions non pertinentes des données de manière préalable à tout traitement si on désire en extraire des informations utiles et compréhensibles.

La sélection de variables en apprentissage semi-supervisé constitue une solution à ces problèmes. Ce processus vise en effet à la détermination d'un sous ensemble optimal (au sens d'un critère donné) de variables en tenant compte à la fois des données labellisées et non labellisées.

Notre seconde contribution consiste à proposer une méthode d'apprentissage semi-supervisée en présence d'un grand nombre de variables tout en fournissant une sélection de variables les plus pertinentes de manière intégrée. L'idée est de mettre en exécution les méthodes ensemblistes qui consistent à combiner plusieurs modèles pour produire une solution plus performante et plus robuste.

4 Démarche de cette thèse

Les méthodes d'ensemble constituent l'une des principales orientations actuelles de la recherche sur l'apprentissage par machine, elles ont été appliquées sur un large éventail de problèmes réels. Malgré l'absence d'une théorie unifiée sur des ensembles, il y a beaucoup de raisons théoriques pour combiner plusieurs apprenants, et une preuve empirique de l'efficacité de cette approche.

Le principe des méthodes d'ensemble est de construire une collection de prédicteurs et ensuite d'agréger l'ensemble de leurs décisions ; son objectif est d'être en mesure de trouver un ensemble d'hypothèses qui sont différentes dans leurs prises de décision afin qu'elles puissent se compléter mutuellement.

Parmi les approches issues des méthodes d'ensemble, la Forêt Aléatoire est l'un des derniers aboutissements de la recherche les plus efficaces pour l'apprentissage d'arbres de décision. Cette approche développées respectivement par Breiman 2001 [51], Amit et Geman 1997 [52] est consacrée à l'agrégation d'arbres randomisés. Elle génère un jeu d'arbres doublement perturbés au moyen d'une randomisation opérée à la fois au niveau de l'échantillon d'apprentissage et des partitions internes. La Forêt Aléatoire (Random Forest RF) est une technique de prévision d'ensemble réussie qui utilise le vote majoritaire ou une moyenne en fonction de la combinaison.

Au premier abord de cette thèse, une première partie (Figure 1) est consacrée à l'introduction du contexte et la problématique. Par la suite, nous nous sommes focalisés sur la compréhension de cet algorithme RF qui est considéré comme une technique de référence, compétitive avec la plupart des méthodes d'ensemble. Par ailleurs, malgré le succès des RFs, plusieurs travaux [53] [54] [55] [56] ont proposé l'amélioration de ce dernier.

Ainsi avant d'aborder et d'élaborer notre approche en apprentissage semi-supervisé. Nous avons consacré toute une partie de cette thèse « Partie II – Classification contexte supervisé » (Figure 1), à l'étude, l'optimisation et l'amélioration des performances de prédiction des RF's. Cette partie regroupe trois chapitres, où chaque chapitre expose une modification et correction à cette technique. En premier lieu, au niveau de l'agrégation des résultats (Forêt Aléatoire à vote pondéré [57]). En second lieu, au niveau de l'échantillonnage des données (Forêt à sous espaces aléatoires [58]) et en dernier lieu, au sein du classifieur lui même pour apporter plus d'interprétabilité à la forêt (Forêt Aléatoire Floue [59]).

Dans la troisième partie, la tâche d'étiquetage automatique par apprentissage semi-supervisé étant abordé. Pour ce faire, nous étudions le concept semi-supervisé dans les méthodes d'ensemble, par l'introduction de l'algorithme des Forêts Aléatoires en apprentissage semi-supervisé chapitre 1 (l'algorithme *co-Forest* [60]), l'annotation d'images médicales au niveau pixellique par la segmentation des régions d'intérêt en apprentissage semi-supervisé dans le chapitre 2. Dans le chapitre 3, nous abordons le problème d'an-

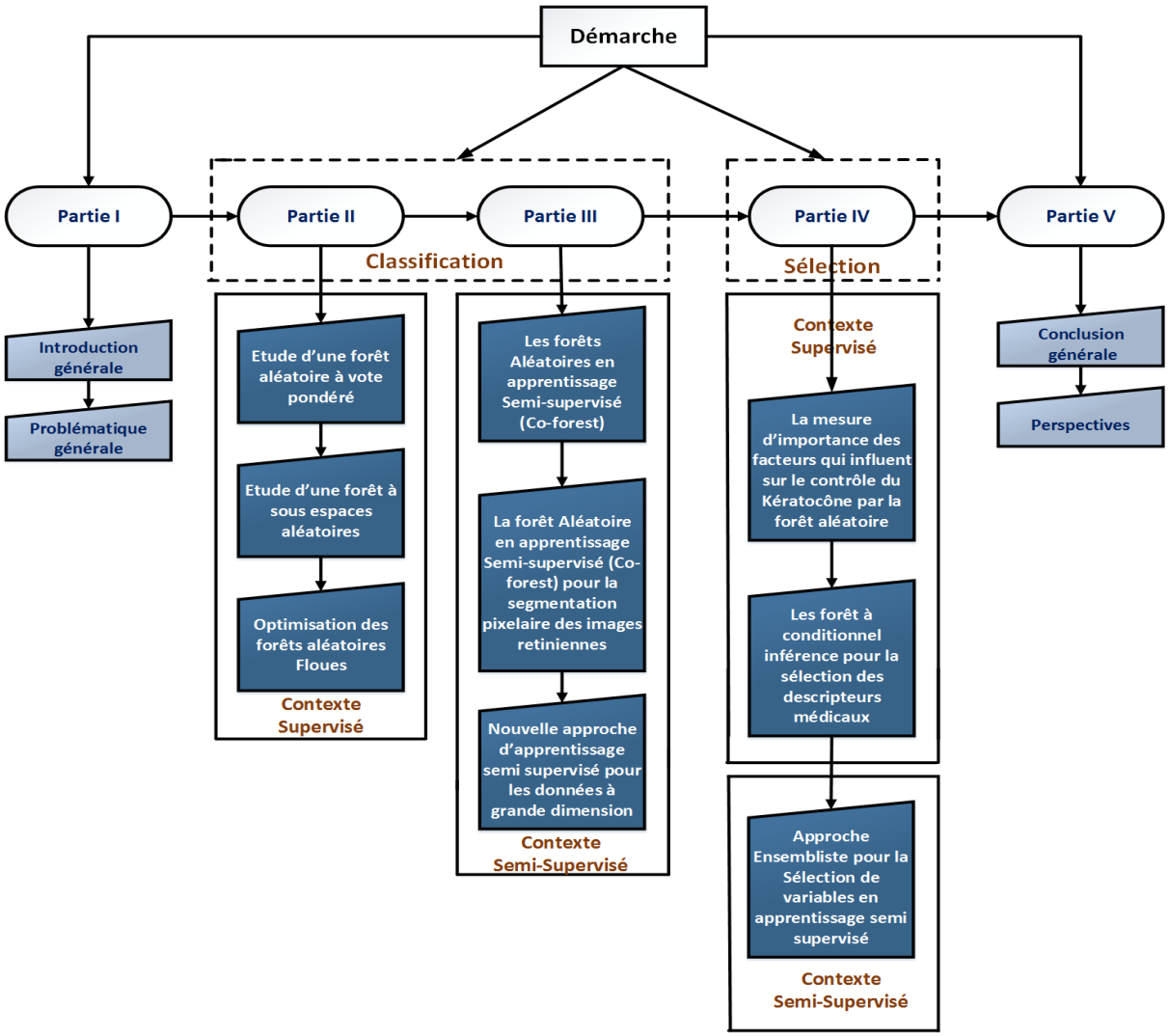


Figure 1 – Schéma représentatif des étapes de la démarche de cette thèse

notation sur les données à grande dimension (Approche Optim-coforest [61]).

Le deuxième volet de cette thèse se penche plus particulièrement sur la capacité de sélection de variables des forêts aléatoires. L'étude du principe de mesure des variables d'importance dans le paradigme des forêts aléatoires (RF) [51] a influencer et inspirer grandement nos travaux réalisés dans cette partie.

Dans la quatrième partie de cette thèse (Figure 1), nous étudions le potentiel de sélection des RF's en comparaison avec d'autres méthodes sur les données médicales chapitre 1 [62], [63] et [64], puis nous exploitons dans le chapitre 2 cette mesure sur des forêts à inférence conditionnelle [65]. Nous abordons par la suite dans le chapitre 3, notre seconde problématique qui est les données à grande dimension et la sélection en apprentissage semi-supervisé, en apportant une application de notre approche sur les données à grande dimension [66].

Finalement, nous terminons par la cinquième partie, avec une conclusion qui exposera une synthèse des contributions apportées ainsi que les pistes définissant des perspectives possibles pour de futurs travaux ainsi que les difficultés rencontrées lors de la réalisation de cette thèse.

Deuxième partie

II

État de l'art et Propositions en apprentissage supervisé : Classification des données par approche ensembliste

Introduction

Les Ensembles de Classifieurs (EoC) ou méthode ensembliste est une des approches multi- classifieurs les plus populaires et les plus efficaces. Elle consiste à combiner les décisions individuelles de plusieurs hypothèses $h_1, \cdots, h_T$ pour classer de nouveaux exemples.

L'heuristique de ces méthodes est qu'en générant beaucoup de prédicteurs, on explore grandement l'espace des solutions, et qu'en agrégeant toutes les prédictions, on récupère un prédicteur qui rend compte de toute cette exploration.

L'objectif visé est que le prédicteur final soit meilleur que chacun des prédicteurs individuels : même si les classifieurs individuels commettent des erreurs, il est peu probable qu'ils commettent les mêmes erreurs pour les mêmes entrées. D'où, surgit l'idée que les prédicteurs individuels doivent être différents les uns des autres : la majorité ne doit pas se tromper pour un même élément x.

Pour que cela soit possible, il faut également que les prédicteurs individuels soient relativement bons. Et là où un prédicteur se trompe, les autres doivent prendre le relais en renforçant l'apprentissage. Les conditions expliquant le succès de ces méthodes d'ensemble se résument ainsi :

– Les hypothèses construites ont un taux de succès meilleur que l'aléatoire.
– Les hypothèses présentent une certaine diversité.

L'émergence des méthodes de combinaison de plusieurs hypothèses répond aux limites des algorithmes induisant une seule hypothèse. En effet l'étude de Dietterich [67] désigne 3 limitations majeures à ces derniers :

Limitation statistique (variance) : si l'espace de recherche est grand proportionnellement au nombre d'exemples, plusieurs hypothèses de même performance peuvent être induites. L'algorithme est contraint d'en choisir une (surement pas la meilleure) (Figure 2), un simple vote peut réduire ce risque.

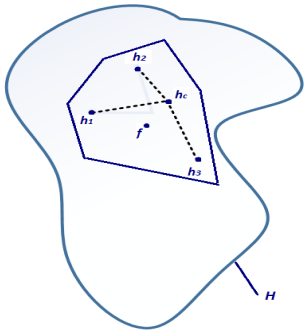


Figure 2 – Limitation statistique (variance)

Limitation de représentation (biais) : lorsque la famille d'hypothèses H ne contient pas de bonne approximation de f (Figure 3), une combinaison peut augmenter l'espace des fonctions possibles H.

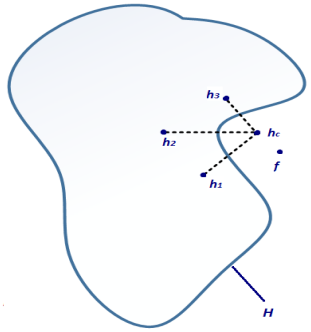


Figure 3 – Limitation de représentation (biais)

Limitation computationnelle : si l'usage d'heuristiques est nécessaire pour résoudre les problèmes algorithmiques, on aboutit souvent à des minima locaux (Figure 4). Une combinaison de minima locaux peut réduire le risque de choisir la mauvaise hypothèse.

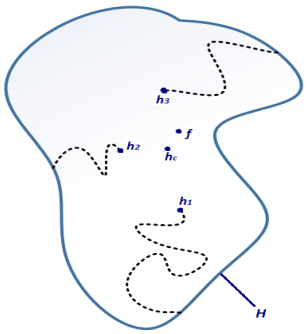


Figure 4 – Limitation computationnelle

Une classification possible des méthodes ensemblistes repose sur la nature des hypothèses de base [68] :

Méthodes ensemblistes hétérogènes Combinent un ensemble d'hypothèses $h_1, \cdots, h_T$ produites par des algorithmes différents $L_1, \cdots, L_T$ sur une même distribution D des exemples de LS.

Méthodes ensemblistes homogènes Associent un ensemble d'hypothèses $h_1, \cdots, h_T$ produites par un même algorithme d'apprentissage L (un ensemble de réseaux de neurones, un ensemble d'arbres de décision, ou un ensemble de discriminants). La diversité des hypothèses s'opère en modifiant la distribution de probabilité D_t des exemples utilisés pour construire h_t.

Il existe aujourd'hui un grand nombre de méthodes capables de générer automatiquement des ensembles de classifieurs : Bagging [37], Boosting [38], Random Subspaces [39]. . ., et chacune de ces procédures d'induction d'EoC dispose d'hyper-paramètres pour le contrôle de la construction des EoC.

Le principe de la méthode d'ensemble [67], est de construire une collection des prédicteurs, et puis agréger l'ensemble de leurs prédictions. L'heuristique de ces méthodes est qu'en générant beaucoup de prédicteurs, on explore grandement l'espace des solutions, et qu'en agrégeant toutes les prédictions, on récupère un prédicteur qui prend en considération toute cette exploration.

L'objectif est d'être en mesure de trouver un ensemble d'hypothèses qui sont différentes dans leurs prises de décision afin qu'elles puissent se compléter mutuellement. Pour que

cela soit possible, il faut également que les prédicteurs individuels soient relativement bons et différents les uns des autres. Le premier point est nécessaire, car agréger de mauvais prédicteurs ne pourra vraisemblablement pas donner un bon prédicteur. Le deuxième point est également naturel, car agréger des prédicteurs qui sont quasiment pareils donnera encore un prédicteur semblable et n'améliorera pas les prédictions.

Parmi les méthodes d'induction d'EoC, on trouve la méthode de la forêt aléatoire [51]. Cette méthode est un bagging amélioré au niveau des hyper-paramètres. Elle est basée sur la combinaison de classifieurs élémentaires de types arbres de décision. Individuellement, ces classifieurs ne sont pas efficaces, mais ils possèdent des propriétés intéressantes à exploiter au sein d'un EoC : ils sont particulièrement inconstants ; la spécificité des arbres utilisés dans les forêts aléatoires est que leur induction est perturbée d'un facteur aléatoire, et cela dans le but de générer de la diversité dans l'ensemble. C'est sur la base de ces deux éléments : utiliser des arbres de décision comme classifieurs élémentaires et faire intervenir l'aléatoire dans leur induction qu'a été introduit le formalisme de la forêt aléatoire.

L. Breiman a proposé un algorithme d'induction de la forêt aléatoire appelé Forest-RI [51], cet algorithme est considéré comme un algorithme de référence qui est compétitif avec les principales méthodes d'ensemble. Cependant, son utilisation pose jusqu'à aujourd'hui d'importantes difficultés.

Nous nous sommes intéressés dans cette partie à la classification des données médicales en exploitant certaines méthodes d'ensemble existantes, en particulier l'algorithme des forêts aléatoires. L'idée des méthodes d'ensemble est de combiner plusieurs classifieurs pour construire un meilleur modèle. Il existe de nombreuses méthodes qui génèrent automatiquement des ensembles de classifieurs. Certaines d'entre elles manipulent les exemples (comme dans Bootstrapping [69]), d'autres randomisent le choix des attributs (Random Subspaces Method [39]) et d'autres randomisent les deux exemples et attributs (Forêts aléatoires).

Parmi les algorithmes de génération de forêts aléatoires, la méthode Random Forest délivre un assemblage de plusieurs arbres de décisions, elle perd en intelligibilité ce qu'elle gagne en précision et se caractérise par sa double "randomisation" qui lui permet d'améliorer ses qualités de prédiction par rapport aux techniques plus simples comme le bagging.

Les forêts sont aléatoires et ont la particularité d'utiliser exclusivement des classifieurs élémentaires types d'arbres de décision. La raison principale est que ces classifieurs (arbres de décision) sont particulièrement adaptés pour une utilisation dans les méthodes d'ensemble, et aussi en raison de leur instabilité.

Par conséquent, des procédés pour générer la diversité dans ces ensembles sont parfois spécifiques à l'induction automatique des arbres de décision. La principale difficulté posée par diverses méthodes d'ensemble est leurs hyper-paramètres [70] qui permettent de contrôler la variété dans les données. Les Forêts aléatoires ne font pas exception à cette règle. Si nous prenons l'exemple de l'algorithme de référence pour l'induction des forêts aléatoires Forest-RI [51], deux paramètres principaux sont utilisés pour créer la diversité :

– Le nombre d'attributs sélectionnés de manière aléatoire à chaque nœud de l'arbre,
– Le nombre d'arbres induits dans la forêt.

D'autres méthodes comme les méthodes à sous espaces aléatoires (RSM) [39], Bagging à sous-espaces (SubBag) [71] et Random Patches (RP) [72], utilisent un autre paramètre qui représente la taille du sous-ensemble de variables sélectionnées. L'objectif de ces paramètres est de créer de la diversité dans l'ensemble (Bootstrapping, RSM, SubBag, RP, ...) et / ou de la diversité entre les classifieurs pendant l'étape de l'apprentissage (Random Forest, extra-trees, ...).

Un autre problème survient dans les méthodes d'ensemble qui est comment combiner la décision de chaque classifieur primaire dans l'ensemble. Dans les forêts aléatoires (RF), toute prévision est assurée par un vote à la majorité simple des classes d'arbres individuels. Cette méthode n'est pas toujours optimale puisque tous les arbres n'ont pas nécessairement les mêmes performances.

L'objectif de la partie 2 de cette thèse est de :

- Remplacer le vote de la majorité traditionnelle des forêts aléatoires par un vote pondéré [57],

- Soulever le problème de la diversité en développant une nouvelle méthode d'ensemble basée sur les forêts aléatoires et la méthode des sous-espaces aléatoires appelées Sub_RF (Forêt à sous espaces aléatoires [58]),

- Gagner en stabilité et en interprétabilité ce que l'on perd par le principe des ensembles de classifieurs. Nous introduisons la Forêt Aléatoire Floue (FRF) et nous proposons une optimisation de cette dernière [59].

CHAPITRE 1

Étude d'une Forêt Aléatoire à vote pondéré

1 Principe de la Forêt Aléatoire

La Forêt Aléatoire appelée RF (Random Forest) [51] est constituée d'un ensemble d'arbres simples de prévision, chacun étant capable de produire une réponse lorsqu'on lui présente un sous-ensemble de données. Cette méthode est un bagging amélioré au niveau des hyper-paramètres.

Le Bagging est une méthode d'ensemble introduite par Breiman [37]. Le mot Bagging est la contraction des mots Bootstrap et Aggregating. Le Bootstrap est un principe de ré-échantillonnage statistique [69] traditionnellement utilisé pour l'estimation de grandeurs ou de propriétés statistiques.

Bootstrapping Un bootstrap d'un ensemble T est l'ensemble obtenu en tirant $|T|$ fois des éléments de T uniformément au hasard et avec remise. Le bootstrapping d'un ensemble d'entraînement T produit un nouvel ensemble T' qui présente en moyenne $1 - e^{-1} \approx 63\%$ instances uniques différentes de T quand $|T| >> 1$ [73] (la création de plusieurs bags à partir d'une base de donnée (voir exemple illustré Figure 5)).

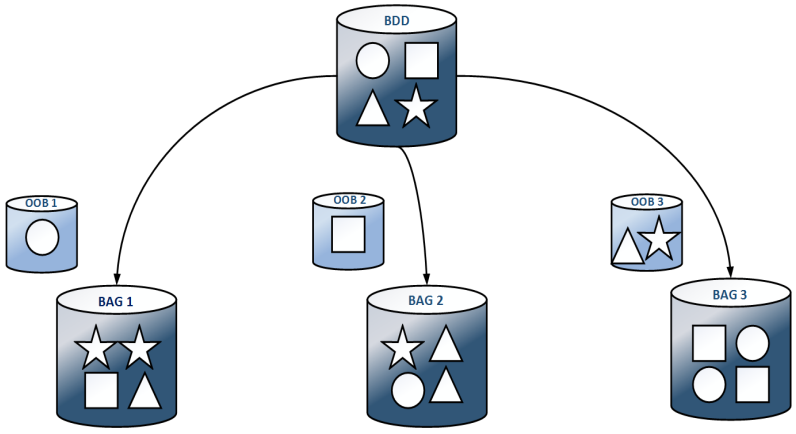


Figure 5 – Schéma représentatif du principe du bootstrapping

Agrégation Le principe est de produire plusieurs bootstraps $T'_1, \ldots, T'_m$, chaque bootstrap T'_i étant utilisé pour entraîner un prédicteur t_i (penser ici à un arbre de régression,

mais la technique s'applique à n'importe quelle famille de prédicteurs). Étant donné une instance (x, y), on fait régresser chaque arbre, ce qui nous donne un ensemble de valeurs $y_1, \ldots, y_m$ prédites. Celles-ci sont alors agrégées en calculant leurs moyennes.

$$y = \frac{1}{m} \sum_i y_i$$

En résumé, l'idée du Bagging est d'utiliser plusieurs ensembles de données ré-échantillonnées à partir de l'ensemble des données observées et ce à l'aide d'un tirage aléatoire avec remise (Figure 5). Ainsi chaque classifieur élémentaire de l'ensemble sera entraîné sur un des échantillons bootstrap de sorte qu'il soit tous entraîné sur un ensemble d'apprentissage différent. L'agrégation de ces classifieurs permet d'obtenir un prédicteur plus performant.

1.1 Forêt à entrées aléatoires "Forest-RI"

Une Forêt Aléatoire est un prédicteur constitué d'un ensemble de classifieurs élémentaires de type arbres de décision CART [74]. L. Breiman [51] propose une amélioration du bagging avec un algorithme d'induction de forêts aléatoires (Forest-RI — pour Random Forest - Random Input —) qui utilise le principe de randomisation "Random Feature Selection" proposé par Amit et Geman [52]. L'algorithme 1 Random Forest-RI appartient à la famille la plus large des forêts aléatoires défini par Breiman [51].

Algorithm 1 Pseudo code de l'algorithme Random Forest-RI

1: **Entrées :** L'ensemble d'apprentissage L, Nombre d'arbres N.
2: **Sortie :** *Ensemble d'arbres* E
3: **Processus :**
4: **for** $i = 1 \rightarrow N$ **do**
5: $T^i \leftarrow BootstrapSample(T)$
6: $C^i \leftarrow ConstructTree(T^i)$ où à chaque noeud :
- Sélection aléatoire de $K = \sqrt{M}$ Variables à partir de l'ensemble d'attributs M
- Sélection de la variable la plus informative K en utilisant l'index de Gini
- Création d'un nœud fils en utilisant cette variable

7: $E \leftarrow E \cup \{C^i\}$
8: **end for**
9: Retourner E

L'induction des arbres se fait sans élagage et selon l'algorithme CART, toutefois, au niveau de chaque nœud, la sélection de la meilleure partition, basée sur l'index de Gini, s'effectue uniquement sur un sous-ensemble d'attributs de taille préfixée (généralement égale à la racine carrée du nombre total d'attributs) sélectionnés aléatoirement depuis l'espace originel des caractéristiques [75].
Pour la problématique de classification, la réponse prend forme d'une classe qui associe un ensemble (classe) de valeurs indépendantes (prédicteur) à une des catégories présentes dans la variable indépendante. En utilisant les ensembles d'arbres on obtient une amélioration significative de la prévision (c'est-à-dire une meilleure tendance à prévoir les nouvelles données), par rapport aux techniques classiques (par exemple CART) [76]. La réponse de chaque arbre dépend du sous-ensemble de données choisi indépendamment (avec remplacement) et avec la même distribution pour tous les arbres de la forêt. La prédiction globale de la forêt aléatoire est calculée en prenant la majorité des votes de chacun de ses arbres.

2 Objectifs

La Forêt Aléatoire (Random Forest RF) est une technique de prévision d'ensemble réussie qui utilise le vote majoritaire ou une moyenne en fonction de la combinaison. Cependant, il est clair que chaque arbre dans une forêt aléatoire peut avoir une contribution différente

au traitement d'une certaine instance.

Plusieurs travaux [53–56] ont proposé des améliorations de la forêt aléatoire, en se basant sur le fait qu'individuellement, chaque classifieur donne de faibles prédictions. L'idée principale est de renforcer chaque classifieur sans exclure la variété entre eux et de diminuer la variance sans perdre sa force de prédiction. D'autres travaux [53] [56] ont affirmé que l'amélioration de chaque classifieur individuellement est insuffisante. Ils ont proposé alors de changer la méthode d'agrégation classique par d'autres techniques.

Dans ce travail [57, 77], nous apportons aussi nos améliorations aux Forêts Aléatoires, et nous démontrons que les performances de prédiction des RF's peuvent encore être améliorées avec le remplacement de l'indice de GINI par d'autres indices (twoing ou deviance). Nos expérimentations démontrent également que l'agrégation des prédictions des arbres par vote pondéré donne de meilleurs résultats que l'agrégation classique avec vote majoritaire.

Ce chapitre est organisé comme suit : la première section concerne l'état de l'art, nous exposons les différentes techniques utilisées dans la littérature pour améliorer ces méthodes. La seconde section concerne les expérimentations ; dans cette section nous présentons l'approche proposée, les bases de données utilisées, les expérimentations réalisées, les résultats obtenus avec leurs interprétations. Et dans la dernière section, une conclusion générale et des perspectives pour ce travail sont proposées.

3 État de l'art

Plusieurs travaux prouvent qu'un modèle de classification induisant une seule hypothèse possède des problèmes [37] [51]. Ils ont donc proposé de combiner chacun de ces classifieurs individuels faibles pour former un unique système de classification appelé Ensemble de Classifieurs (comme les forêts aléatoires).

A ce jour plusieurs études se sont intéressées aux forêts aléatoires, plus particulièrement à l'amélioration de l'algorithme CART au niveau du critère de segmentation, ainsi qu'à l'étape de l'agrégation des classifieurs pour obtenir un meilleur classifieur final. Nous présentons dans cette section l'état de l'art des améliorations des forêts aléatoires ainsi que nos propositions dans ce sens.

3.1 Choix de la variable

La construction d'un arbre de décision passe par plusieurs étapes ; la première étape est de choisir une variable de segmentation qui maximise un critère donné. Nous citons par la suite les travaux qui améliorent cette étape en utilisant d'autres critères de segmentation.

Robnik-Sikonja dans [53] étudie certaines possibilités d'augmenter ou de diminuer la force de corrélation des arbres dans la forêt. La sélection aléatoire des attributs rend les arbres individuels plutôt faibles. Le premier objectif était de renforcer individuellement les arbres sans perdre la variété entre eux, et d'augmenter la variance sans décroitre sa force de prédiction.

L'étude expérimentale de Robnik-Sikonja dans [53], repose sur deux forêts aléatoires variantes, dont la première est une méthode classique en utilisant l'indice de Gini [74], alors que la seconde est améliorée en utilisant les cinq critères suivants : (Gini, Gain de ration [78], MDL [79], RelieF [80] et Myopic RelieF [81]).

Les travaux d'Abellan & Masegosa [54] sont des études expérimentales utilisant différents arbres de décision comme classifieurs par la méthode Bagging [37]. Le but de ces études est de déterminer le meilleur critère de partitionnement parmi quatre critères

(Info-gain [82], ratio Info-Gain [78], l'indice de Gini, et imprécise Info-Gain [83]). Ils ont prouvé que la meilleure forêt Aléatoire est celle qu'utilise le critère Imprécise Info-Gain (IIG).

Une autre étude expérimentale a été faite par Andres Cano et al. [55]. Les auteurs ont utilisé deux approches pour construire les arbres de décision. Une approche bayésienne avec un nombre de nœuds de partitionnement $K = 1$ (K le nombre de caractéristiques à sélectionner aléatoirement à chaque nœud), et une forêt aléatoire classique dont le nombre K est variable. Pour les deux approches, quatre groupes d'arbres différents 10, 50, 100 et 200 ont été évalués sur 23 bases de données du dépôt d'UCI [84]. Après l'utilisation d'une décomposition bias-variance de l'erreur, les auteurs ont confirmé que le nombre K affecte directement la performance de la forêt aléatoire. En effet, pour une diminution de cette valeur K la variance est réduite tandis que le bias augmente [85] et vice versa. Cette tendance est rompue avec l'injection supplémentaire d'aléatoire dans le critère de division.

Evanthia et al. [56] utilisent trois types de forêts aléatoires : la forêt aléatoire classique (indice de Gini), deux autres forêts aléatoires améliorées l'une avec ReliefF pour la segmentation, et l'autre avec plusieurs critères d'évaluation. La motivation derrière ces méthodes d'induction de forêt est de favoriser simultanément l'augmentation des performances individuelles des arbres et celle de la diversité. Pour leurs expérimentations ils ont modélisé la base de donnés alzheimer pré-traitée par fMRI (functional magnetic resonance imaging) [86, 87].

3.2 Le mécanisme de vote

Construire une collection de prédicteurs n'est pas une condition suffisante pour avoir une bonne classification, une autre étape est très importante pour la construction d'une forêt aléatoire, c'est l'agrégation de l'ensemble des prédictions. L'objectif visé est que ce prédicteur final soit meilleur que chacun des prédicteurs individuels. Dans l'étape suivante nous présentons les travaux qui améliorent cette étape.

Dans une forêt aléatoire classique, l'agrégation est représentée par un vote majoritaire des prédictions des classifieurs individuels. Plusieurs travaux ont mis au point des alternatives au vote majoritaire par l'application d'autres mécanismes de vote.

Robnik-Sikonja propose dans [53] d'étudier une amélioration du système de vote majoritaire classique des forêts aléatoires, dans le but d'obtenir des classifieurs finaux plus performants. L'idée proposée est d'établir une sélection d'arbres de décision qui participent au vote final sur les données "similaires" en terme de voisinage. Plus simplement, pour chaque exemple de la base de test, ce système détermine parmi les données d'apprentissage, celles qui lui ressemblent le plus. Il décide alors de ne faire confiance qu'aux arbres de décision qui savent le mieux classer ces données dites "similaires". Ce système de vote majoritaire s'appuie sur la procédure utilisée par Breiman dans [51] pour mesurer la similarité.

Dans [56] Evanitha et al., présentent des améliorations au système du vote ordinaire. Il s'agit de proposer six diversités de l'algorithme de vote pondéré. Le premier algorithme [53] est le même que celui utilisé par Breiman dans [51]. Ce principe est basé sur les performances individuelles des arbres sur les donnés similaires. Le deuxième algorithme [88], le troisième [89] et le cinquième algorithme [90] sont basés sur des métriques entre les données. Le quatrième [91] et le sixième [92] utilisent la pondération des arbres selon leurs taux de classification. Evanitha et al. effectuent un certain nombre d'expérimentations sur la base de données Alzheimer pour étudier l'évolution des performances des forêts aléatoires lorsqu'elles implémentent le vote pondéré ou le vote majoritaire classique.

3.3 Contributions/propositions

Dans ce travail, nous proposons de remplacer le vote majoritaire par le vote pondéré suivant la performance locale de chaque arbre. Ce choix est justifié par le fait que le vote classique dans son principe donne un même poids à chaque décision de chaque arbre ce qui est préjudiciable pour nombre d'exemples. Ce système de vote dépend aussi du choix d'une majorité de classifieurs qui donnent une même classe pour la prédiction d'un exemple donné, alors que les arbres n'ont pas les mêmes performances. Nous mettons en œuvre également une deuxième modification au niveau du choix de variable de division en remplaçant l'indice de Gini par ses diversités à savoir Towing et Deviance.

4 Méthodes

4.1 Indice d'évaluation Gini

Ce critère a été introduit par Breiman et al. [51] dans la méthode d'induction d'arbres CART (*Classification And Regression Tree*) [74]. C'est également le critère que Breiman a décidé d'utiliser dans ses algorithmes d'induction de forêt aléatoire pour mesurer l'impureté du nœud.

En classification, on cherche à diminuer l'indice de Gini, et donc à augmenter l'homogénéité des nœuds obtenus (un nœud étant parfaitement homogène s'il ne contient que des observations de la même classe). Il existe plusieurs critères selon lesquels l'impureté de nœud est minimisée dans un problème de classification. Les trois métriques de Gini communément utilisés incluent : Gdi, Deviance et Twoing.

La métrique Gdi

L'indice GDI est appliqué par défaut dans l'algorithme de CART, il est défini par :

$$\text{Gdi} = 1 - \sum_{i=1}^{k} P^2(i)$$

Où p est la proportion d'observation de la classe i qui atteint le nœud.

Un nœud avec juste une classe (un nœud pur) a l'indice Gini égal à zéro ; autrement dit l'indice de Gini est positif.

La métrique Twoing

Constatant que l'indice de Gini n'est pas efficace lorsque le nombre de classes est élevé, Breiman a proposé dans [74] la règle *Twoing* qui fonctionne pour les arbres binaires, où le nombre de nœud est égal à deux et la partition T se divise en deux noeuds, t_G et t_D.

Conçu pour les problèmes multi-classes, cette approche se base sur la séparation entre les classes plutôt que la diversité du nœud. Dans un problème multi-classes chaque division est traitée comme un problème binaire. Les divisions qui tiennent des ensembles de classes liées sont préférées. L'approche offre l'avantage de révéler des similarités entre les classes et peut être aussi appliquée aux classes ordonnées.

$$Twoing = p(t_G)p(t_D)(\sum_{i=1}^{k} |p_i(t_G) - p_i(t_D)|)$$

La règle de Twoing n'est pas une mesure de pureté d'un nœud, mais une mesure de différence pour décider comment diviser un nœud.

En dénotant $p_i(t_G)$ par la fraction des membres de classe i dans le nœud gauche fils après la division, et $p_i(t_D)$ dénote la fraction des membres de classe i dans le nœud fils juste après la division. $p(t_G)$ et $p(t_D)$ sont les fractions d'observations qui se divisent respectivement vers la gauche et la droite. Si l'expression est grande, la division donne des nœuds fils purs. De même si l'expression est petite, la division fait ressortir chaque nœud fils semblable l'un à l'autre, et par la suite semblable au nœud parent. De ce fait, la division n'augmente pas la pureté du nœud.

La métrique de Déviance

Aussi appelée la trans-entropie ou la mesure de déviance d'impureté, est utilisée pour calculer l'impureté de nœud. Un nœud pur a la déviance zéro ; autrement, la déviance est positive, définit par :

$$Deviance = -\sum_{i=1}^{k} p(i) \log p(i)$$

4.2 L'agrégation par vote

Une forêt aléatoire est l'agrégation d'une collection d'arbres aléatoires. Cette méthode revient à faire la moyenne des arbres en régression, et à faire un vote majoritaire en classification.

Le vote majoritaire

Pour chaque instance $(x; y)$ et forêt aléatoire, l'apprentissage de chaque arbre de la forêt est lancé pour aboutir un ensemble de valeurs de classes prédîtes. Celles-ci sont alors agrégées par un vote majoritaire entre les classifieurs (Algorithme 2).

Algorithm 2 L'agrégation par vote majoritaire

1: Entrainer l'ensemble des classifieurs H sur une X donnée pour prédire les classes C_j
2: Calculer $V_{t,j}$ le vote pour la classe retournée par chaque classifieur

$$V_{t,j}(X) = \begin{cases} 1 & \text{si } h_t \text{ donne la classe j} \\ 0 & \text{sinon.} \end{cases}$$

3: Calculer le vote total obtenu par chaque classe $V_j = \sum_{t=1}(V_{t,j})$
4: Choisir la classe qui a la plus grande valeur des votes V_j comme une classe finale pour la donnée X

Le vote pondéré

Dans ce travail, nous proposons d'étudier une amélioration du système de vote majoritaire classique des forêts aléatoires, dans le but d'obtenir des classifieurs finaux plus performants.

L'idée est de pondérer les décisions (la classe attribuée) de chaque arbre de la forêt par sa performance locale. La performance locale de chaque arbre est tout simplement son taux de bonne classification par ses OOB (Out Of Bag [1]). En finalité la somme de toutes les pondérations pour chaque classe est réalisée et l'élément à classer prend la valeur de celle qui a le plus grand poids (Algorithme 3).

1. Un échantillon bootstrap S est, par exemple, obtenu en tirant aléatoirement n observations avec remise dans l'échantillon d'apprentissage S, chaque observation ayant une probabilité $1/n$ d'être tirée, les éléments restants sont les éléments OOB.

Algorithm 3 L'agrégation par vote pondéré

1: **Entrées :** X : Une instance à classer, T : ensemble de classifieurs, OOB de chaque classifieur.
2: **Sortie :** Classe C_x
3: **Processus :**
4: Calculer la performance $Perform_t$ de chaque classifieur «t» de «T»

$$Perform_t = \text{Taux de classification } (OOB_t)$$

5: Calculer $Score_{c,x}$ le score de chaque classe «c» obtenu par les «T» classifieurs pour l'instance «X».

$$Score_{c,x} = \sum_{t=1}^{T} (Test_{t,c}(x) * Perform_t)$$

$$Test_{t,c}(x) = \begin{cases} 1 \text{ si l'arbre «}t\text{» donne la classe «}c\text{» pour l'instance «}X\text{»} \\ 0 \text{ sinon.} \end{cases}$$

6: La classe c_x avec le score le plus élevé est choisie comme classe finale

$$C_x = Argmax(Score_{c,x})$$

5 Résultats et interprétation

Dans ce chapitre, nous allons présenter l'impact de chaque indice de division ainsi que chaque type de vote sur une forêt aléatoire. La pondération de chaque arbre est calculée selon son taux de bonne classification de ses éléments OOB. Dans ce qui suit nous présentons les résultats obtenus par les différentes combinaisons. Les expérimentations ont été réalisées en deux temps. Premièrement, mettre en place le nombre d'arbres pour la construction de la forêt Aléatoire, et en second lieu, séléctionner l'indice de division et le type de vote optimal pour améliorer les performances de classification de la forêt aléatoire.

5.1 Description des bases de données médicales

Afin d'évaluer les différentes méthodes, nous avons fait appel à quatre bases de données médicales du répertoire UCI Machine Learning [84] résumées dans le tableau (Table 1).

La base de données Breast Cancer La base de données du cancer du sein dénommée « Wisconsin Breast Cancer Database » a été collectée à l'Université du Wisconsin. Elle contient les informations médicales de 699 cas cliniques relatifs au cancer du sein classés comme bénin ou malin : 458 patientes (soit 65%) sont des cas bénins et 241 patientes (soit 35%) sont des cas malins. La base de données contient neuf attributs qui représentent des cas cliniques, et l'attribut de la classe binaire.

La base de données Bupa liver disorder La base de données Liver disorders a été dénotée par Richard S. Forsyth dans une recherche médicale de la compagnie de soins médicaux internationaux BUPA. Elle englobe une étude médicale sur 345 individus des maladies du déséquilibre de foie (200 malades, 145 non malades). La base de données contient six attributs, les cinq premiers sont des analyses de sang. Ils sont supposés être sensibles aux désordres du foie qui pourraient survenir suite à la consommation excessive d'alcool.

La base de données Haberman L'ensemble de données concerne une étude qui a été accomplie entre 1958 et 1970 à l'Université « Chicago's Billings Hospital » sur la

survivance de patients qui avaient subi la chirurgie pour le cancer du sein. La base de données contient 306 patients (225 ont survécu et 81 ont succombé à la mort). Elle contient trois attributs et un quatrième pour la classe.

La base de données Pima La base de données Pima Indians Diabetes a été réalisée sur une étude de 768 femmes Indiennes Pima (500 non diabétiques 268 Diabétiques), Ces mêmes femmes, qui ont stoppé leur migrations en Arizona, États Unis, et ont adopté un mode de vie occidentalisé, développent un diabète dans presque 50% des cas. Le diagnostic est une valeur binaire variable « classe» qui permet de savoir si le patient montre des signes de diabète selon les critères de l'organisation Mondiale de la Santé.

Bases de données	# Instances	# Variables	# Classes
Breast	699	9	2
Bupa Liver	345	6	2
Habermann	306	3	2
Pima	768	8	2

Table 1 – Paramètres des bases de données utilisées

5.2 Choix de la taille de la forêt

Pour choisir le nombre d'arbres optimal dans une forêt, nous avons effectué plusieurs tests avec, à chaque fois, une forêt de taille différente. Les résultats (Figure 6) indiquent qu'à partir de 100 arbres le taux d'erreur reste plus ou moins stable. Dans ce qui suit on va utiliser 100 arbres pour toutes les expérimentations.

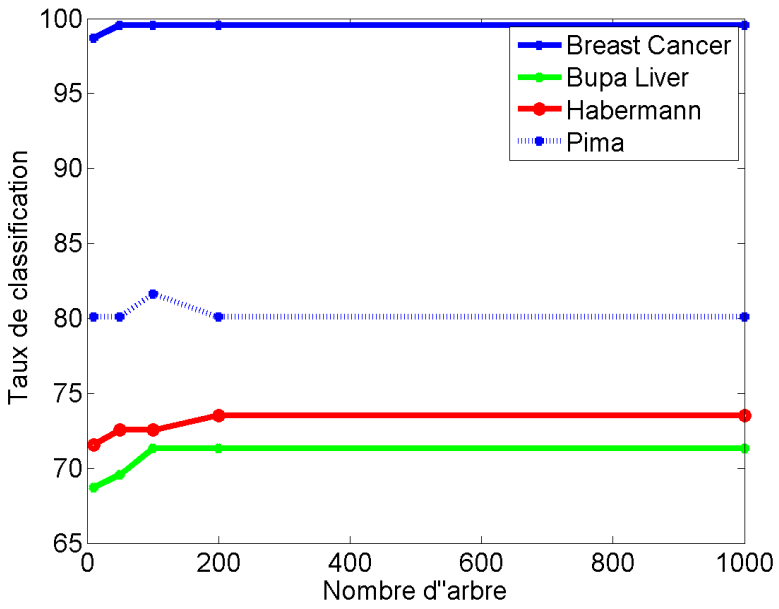


Figure 6 – Taux de classification par rapport aux nombres d'arbres.

5.3 Choix de l'indice de division et le type de vote

Nous comparons dans cette partie les performances (taux de classification) de six différentes forêts aléatoires. Dans le but d'améliorer la performance des forêts nous avons remplacé dans trois forêts le vote majoritaire par le vote pondéré, en utilisant les mêmes critères pour les trois forêts (indice de Gini (Gdi), Twoing et Deviance). Le tableau (Table

2) contient les résultats des différentes forêts. Nous constatons que le vote pondéré donne de meilleurs résultats par rapport au vote majoritaire classique, aussi que le critère Déviance donne la plupart du temps de meilleurs résultats par rapport à l'indice de Gini (Gdi) et Twoing.

Bases de données	**Vote majoritaire**			**Vote pondéré**		
	GDI	TWOING	DEVIANCE	GDI	TWOING	DEVIANCE
Breast cancer	**99.335**	99.1416	**99.5708**	**99.514**	99.1416	**99.5708**
Bupa Liver	70.7218	69.7321	**71.4348**	**72.1739**	71.3043	**72.1739**
Habermann	**72.8406**	71.9167	72.5490	72.5490	72.5490	**73.521**
Pima	81.0546	81.9117	**82.0313**	82.0313	81.6406	**82.3**

Table 2 – Les performances des forêts aléatoires utilisant l'indice de Gini et ses deux variantes

Afin de mieux évaluer les performances réalisées, nous avons développé deux forêts aléatoires différentes, la première est une forêt aléatoire classique avec l'indice de Gini (Gdi) et le vote majoritaire pour agréger les 100 classifieurs. La deuxième forêt est améliorée en remplaçant Gini par Deviance et un vote pondéré pour l'agrégation. Une validation croisée égale à 10 est effectuée pour chaque forêt afin de comparer les performances de sensibilité, spécificité et taux de classification des deux méthodes sur les quatre bases de données (voir Table 3).

Bases de données	**RF classique**			**RF améliorée**		
	TC%	Se%	Sp%	TC%	Se%	Sp%
Breast cancer	99.28	98.15	99.62	99.51	99.14	99.81
Bupa Liver	70.72	86.33	42.54	72.17	87.38	44.71
Haberman	72,84	41,6	82,98	73,52	43,6	82,85
Pima	81.05	61.56	90.40	82.3	64.81	90.69

Table 3 – Les performances de la forêt améliorée pour l'ensemble de données médicales

Breast Cancer Les deux forêts aléatoires donnent des valeurs de sensibilité et spécificité très élevées. Ces résultats montrent que les forêts aléatoires ont effectuées un bon apprentissage pour les données négatives et positives, donc elles réalisent une très bonne détection des patients bénins et malins.

En comparant les résultats des deux forêts nous remarquons que les résultats sont presque identiques vu la nature de la base de données. Notant que que plusieurs travaux n'utilisant qu'un mono-classifieur comme les réseaux de neurones ont obtenu avec cette base 100% de bonne classification.

Bupa liver La sensibilité des deux approches est très élevée, cela signifie que les deux forêts ont réalisées un bon apprentissage pour les données positives. Par contre, la spécificité des deux méthodes est faible d'où la détection des patients non malades est moins pertinente.
En comparant les deux forêts aléatoires, nous remarquons que la forêt améliorée fournit la plupart du temps de meilleurs résultats que la forêt classique (8 cas sur 10).

Haberman La sensibilité des deux méthodes est faible, donc on obtient une mauvaise détection des patients décédés depuis cinq ans. Une très bonne valeur de la spécificité des deux forêts nous donne une bonne détection des patients qui ont survécu durant les cinq ans.

Pima Les résultats de cette expérimentation montrent que la spécificité des deux forêts est très élevée prouvant que les forêts aléatoires ont réalisées un bon apprentissage sur les données négatives. Nous pouvons en conclure que lorsqu'un patient est non diabétique la méthode RF le détecte avec succès.

En comparant les résultats obtenus, notre constat immédiat, est que la forêt aléatoire améliorée fournit la plupart du temps de meilleurs résultats que la forêt aléatoire classique (7 cas sur 10).

Pour la comparaison des deux forêts aléatoires classique et améliorée, nous constatons que la deuxième méthode donne la plupart du temps de meilleurs résultats par rapport à la forêt classique.

6 Conclusion

Dans ce chapitre, nous proposons une étude comparative des performances de classification reliée au domaine médical du modèle ensembliste forêt Aléatoire (Random Forest).

Pour cela, nous avons procédé dans un premier temps à une analyse de travaux effectués dans le domaine. Ceci nous a permis de mettre en évidence plusieurs avantages ainsi que certaines limites des forêts aléatoires employées dans le cadre de la classification des données médicales. Nous avons constaté en conséquence que les classifieurs déjà proposés à base des forêts aléatoires font preuve de bonnes performances mais peuvent être encore améliorés pour apporter plus de précisions aux résultats.

Nous avons en premier lieu, fait appel à la forêt aléatoire en appliquant l'indice de Gini et le vote majoritaire. En second lieu, nous avons procédé au développement de plusieurs variances du même classifieur en utilisant l'indice de Déviance et Twoing rule au niveau du choix de la variable de division des nœuds des arbres. Et en dernier lieu, nous avons appliqué le vote pondéré comme méthode d'agrégation de l'ensemble d'arbres.

Le taux de classification obtenu avec notre méthode est parmi les meilleurs résultats obtenus jusqu'à aujourd'hui pour la classification de ces bases de données. Les résultats obtenus sont très compétitifs par rapport aux autres versions des forêts aléatoires.

CHAPITRE 2

Étude d'une Forêt à Sous espaces Aléatoires

1 Objectifs

L'intérêt principal de ce travail est donc d'étudier les performances d'une version modifiée des forêts aléatoires que nous appelons Forêt à Sous espaces Aléatoires *Sub_RF* (Subspaces Random Forests). La méthode proposée [58, 77] pour l'induction des arbres, tente d'améliorer la précision en apportant plus de diversité au sein des ensembles de classifieurs.

Pour cela, ce chapitre a été réparti comme suit : Un état de l'art des travaux du domaine est exposé dans la section 2. Par la suite dans la section 3, nous présentons le principe des méthodes utilisées et nous détaillons notre approche d'induction d'arbres. Nos résultats sont présentés et discutés dans la section 4 suivante. Enfin, une synthèse générale qui met en évidence les principales propriétés de notre technique ainsi que quelques perspectives sont proposées dans la conclusion.

2 État de l'art

Il existe différentes possibilités pour la génération des ensembles de classifieurs. La première possibilité consiste à manipuler les données comme dans le Bagging [37] avec l'utilisation du principe de Bootstrap ou dans le RSM (Random Subspace Methods [39]) où nous appliquons uniquement un sous-ensemble aléatoire de l'espace des caractéristiques.

Une deuxième possibilité d'induire des ensembles de classifieurs est de faire appel a différents algorithmes de classification (ou un algorithme avec différents paramètres) tout en formant chacun sur le même ensemble de données (ici nous parlons des méthodes d'ensemble hétérogènes) [93–96].

Une troisième possibilité consiste à combiner les deux premières méthodes en randomisant les données (Bootstraping [1] par exemple) et l'algorithme d'apprentissage (en utilisant différents algorithmes ou bien un seul avec différents paramètres).

Dans cette optique, Breiman a proposé les Forêts Aléatoires (Random Forests) [51]. L'algorithme des Forêts Aléatoires est l'une des réalisations les plus populaires de la

1. Un échantillon bootstrap O est, par exemple, obtenu en tirant aléatoirement n observations avec remise dans l'échantillon d'apprentissage O, chaque observation ayant une probabilité $1/n$ d'être tirée.

recherche consacrée à l'agrégation d'arbres aléatoires. Depuis leur proposition, plusieurs chercheurs ont tenté d'améliorer les forêts aléatoires par la modification du mécanisme de vote ou la technique d'induction des arbres.

L'algorithme PERT (pour "PErfect Random Tree" les arbres aléatoires parfaits) en est un exemple, il a été proposé par Cutler et Guohua (2001) [97]. Dans cet algorithme, pour la division de chaque nœud non-terminal, deux exemples de différentes classes sont d'abord tirés au hasard parmi l'ensemble d'apprentissage. Ensuite, un attribut aléatoire est sélectionné et le point de découpage est aléatoirement et uniformément établi entre les valeurs de cet attribut pour les deux exemples pris au hasard.

Geurts et al. (2006) [85] ont mis en place un autre algorithme intégrant plus d'aléatoire au sein de l'arbre de décision connu sous le nom d'Extra-trees (Pour Extremely Randomized trees) qui est similaire à PERT et qui combine la randomisation des attributs de la méthode RSM avec une sélection totalement aléatoire du point de découpage.

Dans un autre travail, Panov et al. [71] ont combiné les Sous-espaces aléatoires et le Bagging pour construire de meilleurs ensembles. L'idée du SubBag (pour Subspaces Bagging) consiste à sélectionner des échantillons aléatoires avec remplacement à partir de l'ensemble originel (comme dans le Bagging) et pour chaque échantillon, seulement 75% des attributs sont choisis au hasard (en utilisant le RSM [39]). Ainsi, les sous-bases obtenues sont plus randomisées par rapport à celles générées par le Bootsrping.

Dans leur papier [71], Panov et al. ont prouvé que cette approche donne des performances comparables à celles des forêts aléatoires, avec l'avantage supplémentaire d'être applicable à n'importe quel algorithme d'apprentissage sans la nécessité de le randomiser.

Louppes et Geurts ont proposé dans [72] un travail similaire au SubBag appelé Random Patches où chaque modèle de l'ensemble est construit à partir d'un sous ensemble de données et de variables obtenues par tirage aléatoire de l'ensemble originel des données. Leur méthode est très efficace pour l'application des grandes bases de données, car elle réduit considérablement l'espace mémoire et le temps d'exécution.

3 Propositions

Dans ce travail, une version modifiée de RF appelée *Sub_RF* (pour Subspaces Random Forests) est proposée. Notre procédure exploite le principe de SubBag [71] pour la mise en place de l'algorithme des forêts aléatoires.

3.1 Les sous-espaces Aléatoires

La Méthode de Sous-espaces Aléatoires ou bien RSM (pour Random Subspaces Method) a été proposée par Ho [39]. Cette méthode est assez similaire au Bagging, mais il ne s'agit pas de manipulation des données, mais plutôt les caractéristiques par un échantillonnage sans remise.

L'idée de base est de former chaque classifieur sur un sous-espace d'attributs tirés aléatoirement de la base originelle. Chaque sous-espace aléatoire a la même dimension P, avec $P < M$, où M est la dimension de l'espace originel de descripteurs. Dans [39], Ho a montré, pour le paramètre P, que les meilleurs résultats sont généralement obtenus par $P \approx M/2$ caractéristiques. Ho a également montré que le RSM est applicable à tout type de classifieur.

3.2 Forêt à Sous espaces Aléatoires (*Sub_RF*)

La méthode proposée permet la création d'un ensemble de classifieurs en utilisant le procédé du SubBag [71] pour la génération des échantillons d'apprentissage tandis que les classifieurs sont des arbres de décisions générés en utilisant l'algorithme Forest-RI [51].

Cette méthode d'induction d'arbres a été nommée *Sub_RF* (pour Subspaces Random Forest) (Algorithme 4).

Algorithm 4 Pseudo code de l'algorithme *Sub_RF*

Entrées : L'ensemble d'apprentissage L, Nombre d'arbres N, Taille du sous-espace S.
Sortie : *Ensemble d'arbres*
Processus :
for $i = 1 \rightarrow N$ **do**
  $T^i \leftarrow BootstrapSample(T)$
  $T^i \leftarrow SelectRandomSubSpaces(T^i, S)$
  $C^i \leftarrow ConstructRF\_tree(T^i)$
  $E \leftarrow E \cup \{C^i\}$
end for
Retourner E

Cet algorithme ajoute un autre niveau de randomisation à la méthode du SubBag grâce à la fonction *ConstructRF_tree*() qui permet de créer des arbres à l'aide du principe des forêts aléatoires. Pendant le processus d'apprentissage des arbres, à chaque nœud, la sélection de la meilleure répartition sur la base de l'indice de Gini est effectuée non pas sur l'ensemble des attributs M mais sur un sous-ensemble sélectionné de façon aléatoire (racine de M selon l'algorithme Forest-RI).

4 Résultats et interprétations

Notre approche a été testée sur cinq bases de données du dépôt d'UCI Machine Learning Repository [84]. Les bases de données qui ont été utilisées dans nos expériences sont décrites dans le tableau (Table 15).

Bases de données	# Instances	# Variables	# Classes
Breast	699	9	2
Ecoli	366	7	8
Isolet	7797	617	26
Liver	345	6	2
Pima	768	8	2

Table 4 – Paramètres des bases de données d'expérimentations

Dans nos expériences, quatre algorithmes différents sont mis en œuvre à savoir PERT [97], Random Forest [51], SubBag [71] et notre méthode proposée *Sub_RF*. L'objectif est de visualiser et d'étudier l'évolution de l'erreur de chaque technique.

Tout d'abord, chaque base de données a été divisée en deux sous-ensembles de données, une pour l'apprentissage et l'autre pour le test en utilisant la validation croisée (5-fold cross validation). La division de la base a été effectuée par tirage au hasard de la base initiale en respectant la distribution des classes. Comme il a été déjà expliqué, notre méthode utilise le Bootstraping pour générer les sous-bases, ainsi nous aurons en moyenne, pour chaque échantillon bootstrap 63,2% des exemples uniques à partir de l'ensemble d'origine, le reste étant des doubles (première randomisation).

Pour nos expérimentations, les paramètres recommandés pour chaque algorithme seront utilisés. Dans [71], Panov et Dzeroski ont suggéré d'exploiter 75% de l'espace des attributs (à l'aide de la méthode RSM) pour chaque Bag (seconde randomisation). Pour le paramètre K de l'algorithme Random Forest, plusieurs travaux de la littérature ont montré qu'un nombre d'attributs égal à $\sqrt{M}$ (M est la taille de tout l'espace des attributs) est un bon compromis pour produire une forêt performante [51] [98].

Pour le choix de la taille de l'ensemble, nous avons développé plusieurs forêts aléatoires avec différents nombre d'arbres : 10, 50, 100, 200, 500 et 1000. Les résultats indiquent que, pour plus de 100 arbres, le taux d'erreur reste plus ou moins stable (Figure 7).

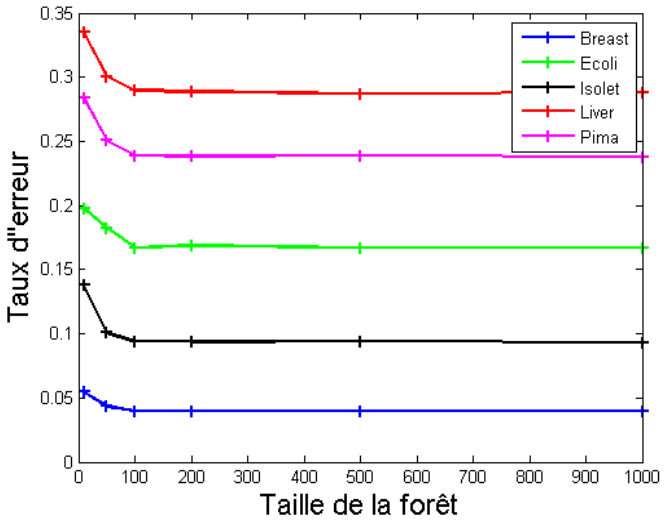


Figure 7 – Taux d'erreur de RF avec les différents nombres d'arbres

Dans ce qui suit, la taille de chaque forêt est fixée à 100 arbres. Dans un second temps, nous avons effectué une comparaison entre RF, SubBag et notre méthode proposée *Sub_RF* afin d'évaluer la performance de cette nouvelle technique pour générer des classifieurs à base d'arbres.

Bases de données	PERT	RF	SubBag	*Sub_RF*
Breast	0,0478	0,04	**0,03**	0,0403
Ecoli	0,2555	0,1672	0,1551	**0,1499**
Isolet	0,1982	0,0899	0,0808	**0,0781**
Liver	0,2816	0,2812	0,2603	**0,2501**
Pima	0,2541	0,2387	0,2311	**0,2283**

Table 5 – Taux d'erreurs des différents algorithmes

Les résultats (Table 5) montrent que *Sub_RF* donne de meilleurs performances par rapport aux autres méthodes, en raison de l'augmentation de sa randomisation. Cela peut être expliqué par le fait que, contrairement au RF, les arbres du *Sub_RF* sont très différents car ils n'utilisent pas tous les attributs et, contrairement à PERT, ils choisissent la meilleure variable. Sur la Base Breast Cancer, SubBag donne de meilleures performances que Sub_RF, cela peut s'expliquer par leur ressemblance dans la première et la seconde étape de randomisation. A partir de ces résultats, nous pensons que *Sub_RF* fournit globalement le meilleur compromis en terme de randomisation dans le cadre de la génération des forêts aléatoires.

5 Conclusion

Dans ce chapitre, une nouvelle méthode de génération d'arbres appelée Subspaces Random Forest (Sub_RF) qui utilise le Bagging, la méthode des Sous-espaces aléatoires et les Forêts Aléatoires a été essentiellement proposée. Cette méthode s'est, en fait, avérée efficace par rapport à la forêt aléatoire classique.

Pour cette raison, nous avons testé expérimentalement notre approche sur cinq bases de données du répertoire de l'UCI. Les résultats montrent que notre approche suggérée est compétitive aux méthodes existantes dans l'état de l'art.

Il reste plusieurs questions et limites de notre approche auxquelles nous souhaitons répondre à l'avenir. Tout d'abord, nous voulons renforcer d'avantage ces résultats avec une analyse théorique ainsi que tester notre algorithme sur de grandes bases de données. Nous aimerions également tester certaines techniques de sélection ensemble afin de ne garder que les meilleurs arbres de la forêt vu que les arbres générés sont très différents.

CHAPITRE 3

Optimisation des Forêts Aléatoires Floues

1 Objectifs

Pour certains domaines d'application, il est essentiel de produire des procédures de classification compréhensibles par l'utilisateur. C'est en particulier le cas pour l'aide au diagnostic médical où le médecin doit pouvoir interpréter les raisons du diagnostic automatique.

A ce titre plusieurs travaux sont effectués afin de développer des outils d'aide au diagnostic et de classification des maladies interprétables et transparents. Les arbres de décision répondent à ces contraintes car ils représentent graphiquement un ensemble de règles et sont aisément interprétables mais leur instabilité a conduit à l'élaboration de méthodes comme le Bagging (pour Bootstrap Aggregating) [37].

La Forêt Aléatoire «Random Forest» de Breiman [51] nous permet dans une certaine mesure de remédier à ce problème. Malgré la puissance de performance de ce dernier, la notion de stabilité n'est pas forcément présente. L'idée proposée est d'intégrer le flou pour retrouver la stabilité grâce aux arbres flous.

Dans ce travail [59], nous traitons l'extraction de la connaissance à partir des données, en appliquant les forêts aléatoires floues (Fuzzy Random Forest FRF) qui combinent la robustesse des arbres de décision, la puissance du caractère aléatoire qui augmente la diversité des arbres dans la forêt ainsi que la flexibilité de la logique floue. Les FRF's ont la spécificité de contrôler des données imparfaites, de réduire le taux d'erreurs et de mettre en évidence plus de robustesse et plus d'interprétabilité.

De ce fait, nous nous focalisons sur l'étude des forêts aléatoires floues, et plus particulièrement leur l'amélioration par la méthode C-moyennes flous (FCM : Fuzzy C-Means) et ce afin d'optimiser l'apprentissage structurel des paramètres flous. Cette technique permet de réduire le nombre de sous-ensembles flous et de minimiser le nombre de règles pour une connaissance plus ciblée.

Ce chapitre est composé de 5 sections décrites comme suit : en premier lieu, un état de l'art sur les forêts aléatoires floues est réalisé dans ce contexte. En second lieu, nous présentons les méthodes et l'approche proposée. Par la suite, nous exposons les résultats et interprétations des algorithmes implémentés avec une synthèse sur les différentes techniques. En dernier lieu, seront présentées une conclusion générale et les perspectives futures pour ce travail .

2 Univers flou

Nous ne pouvons aborder la forêt floue sans tout d'abord présenter les bases de la théorie floue.

Dans la vie quotidienne nous appliquons implicitement la logique floue, car les informations ne sont pas toujours précises. Autrement, chaque personne peut se trouver dans des situations où elle utilise des informations incomplètes, après raisonnement, elle prend des décisions. Il a été nécessaire de créer une logique (Logique floue) qui admet des valeurs de vérité en dehors de l'ensemble binaire pour pouvoir tenir compte et manipuler ce genre d'information [99].

2.1 Le problème

Ces imperfections peuvent être distinguées en trois classes :

- *Imprécision :* désigne les connaissances qui ne sont pas perçues ou clairement définies. Par exemple : La température de la chambre est très élevée [100].
- *Incertitude :* désigne les connaissances dont la validité est sujette à questions. Par exemple : Je crois que la température est de 30 [100].
- *Incomplète :* du fait d'une rupture dans la transmission des données [101].

2.2 La logique floue

La logique floue est une extension de la logique classique (appelée aussi la logique booléenne), Lotfi Zadeh est le fondateur de la théorie des ensembles flous qui est définie comme : "une collection telle que l'appartenance d'un élément quelconque à cette collection peut prendre toutes les valeurs entre 0 et 1" [102].

La logique est basée sur deux concepts principaux :
- Ensembles et variables flous et opérateurs associés.
- Prise de décision à partir d'une base de règles : Si. . . Alors

2.3 Notion d'ensemble et sous ensemble flou

Ensemble flou

Soit X une collection d'objets notés x. Un ensemble flou A de X est caractérisé par une fonction d'appartenance μ_A. La valeur μ_A, pour x dans X, est comprise entre 0 et 1. Elle définit le degré d'appartenance de l'objet x à l'ensemble flou A [101].

Fonction d'appartenance

Il existe plusieurs formes classiques de fonctions d'appartenance floues (comme trapézoïdale, triangulaire ou gaussienne,. . .) qui modélisent des variables linguistiques (grand, moyen ou jeune,. . .). Elle permet de décrire une appartenance floue à une classe [103].

Sous ensemble flou

La logique floue permet de caractériser une appartenance graduelle à un sous ensemble, appelé sous ensemble flou [104].

- Dans l'approche classique :

$$\text{Si } \mu_A \text{ est la fonction d'appartenance de l'ensemble A.}$$
$$\forall x \in X \quad \mu_A(x) = 0 \quad \text{si } x \notin X$$
$$\mu_A(x) = 1 \quad \text{si } x \in X$$

- **Dans l'approche floue :**
– Un élément peut appartenir plus ou moins fortement à cette classe.
– Un sous-ensemble flou A d'un référentiel X est caractérisé par une fonction d'appartenance μ_A :

Si μ_A est la fonction d'appartenance de l'ensemble flou A.
$$\forall x \in X \qquad \mu_A \in [1,0]$$

Un ensemble flou est déterminé par sa fonction d'appartenance (degré d'appartenance ou valeur de vérité) [104].

3 Forêt Aléatoire Floue (Fuzzy Random Forest)

3.1 Le principe des forêts d'arbre de décision flou

Les forêts aléatoires floues (FRF) sont un ensemble d'arbres de décision floue. Les arbres de décision floue généralisent les arbres de décision classiques et sont beaucoup mieux adaptés pour le traitement de données numériques continues, et cela sans introduire des seuils de décision classiques [105]. Le parcours de la racine à une feuille constitue une règle floue. A chaque étape de construction d'un nœud de l'arbre, le choix de l'attribut est très important. Chaque attribut n'a pas le même pouvoir discriminant relatif aux classes.

En sélectionnant les attributs possédant un fort pouvoir discriminant, l'arbre de décision sera moins profond et plus efficace. La mesure du pouvoir discriminant d'un attribut se fait grâce à une mesure de discrimination. La stratégie de partitionnement détermine la façon de découper la base d'apprentissage durant le développement de l'arbre ; et le critère d'arrêt définit les choix d'arrêt du développement de l'arbre.

L'objectif est de définir à chaque nœud l'attribut permettant de déterminer au mieux la classe des éléments de la base d'apprentissage.

3.2 Construire des partitions floues

L'élément important pour la prise en compte de données imprécises (l'information imparfaite) est l'existence d'une partition floue sur l'univers des valeurs d'attribut numérique – symbolique [106], elle est en général donnée par l'expert du domaine.

Pour cela, il faut en premier un système d'apprentissage inductif autonome doté d'une méthode automatique de construction de partition floue pour gérer et traiter les données numériques. Par la suite des données numériques – symboliques pour construire les arbres de décisions floues.

3.3 Les caractéristiques d'un arbre de décision flou

Les arbres de décision flous généralisent les arbres de décision classiques et sont mieux adaptés pour le traitement des données numériques / symboliques dont le parcours se fait de la racine à la feuille et constitue une règle floue [105] pour obtenir la réponse de l'arbre. Ils sont aussi construits avec des seuils fixés pour chaque attribut numérique qui seront considérés comme des valeurs floues lors de l'utilisation de l'arbre général.

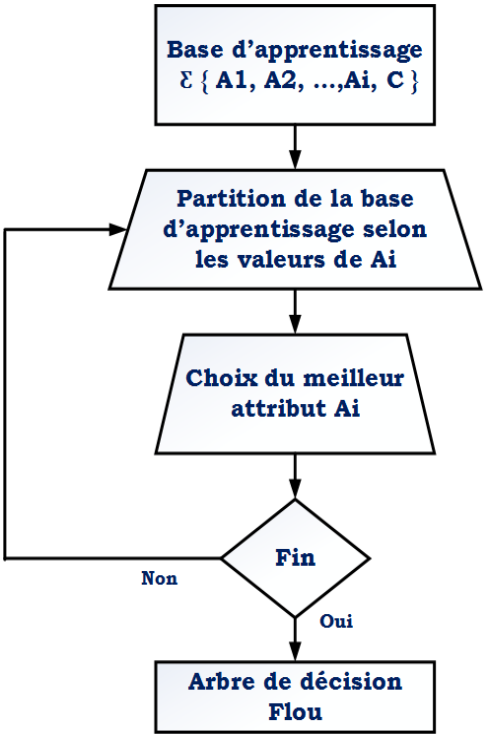
La construction d'arbre de décision par la stratégie TDIDT (Top Down Induction of Decision Tree) se fait par les étapes suivantes :

ξ : La base d'apprentissage composée des exemples définis par des attributs A_i et une classe C.

Choisir un attribut A_i pour partitionner la base d'apprentissage ξ. Un nœud est créé dans l'arbre portant un test sur la valeur A_i s'effectue généralement en décomposant la base en sous base. Chaque sous base est composée d'exemples possédant une valeur identique pour A.

Si tous les exemples de la base possèdent la même classe, le processus de l'arbre peut s'arrêter sinon il reprend la partition (au choix du meilleur attribut).

Dans le cadre flou on repartit les exemples de toutes les sous bases en leurs affectant un degré d'appartenance.



3.4 Comparaison entre l'arbre de décision classique et l'arbre de décision flou

Arbre classique (booléen)	Arbre flou
Le processus de classification suit le premier chemin valide.	Tous les chemins de l'arbre sont évalués lors du processus de classification.
Le critère de sélection pour diviser les données lors de la construction de l'arbre n'est pas toujours approprié.	Meilleur pouvoir de généralisation entre l'ensemble d'apprentissage et l'ensemble de test.
L'arbre est sensible au bruit dans l'ensemble d'apprentissage.	L'utilisation de degrés d'appartenance flous permet un traitement robuste face au bruit.
Le processus de décision dépend des valeurs seuils.	L'usage de valeurs linguistique élimine le problème des valeurs seuils.

Table 6 – La comparaison entre l'arbre de décision classique et flou [107]

4 État de l'art

En passant en revue les différents travaux réalisés au préalable sur la Forêt Aléatoire Floue, une seule équipe dirigée par le Professeur P. Bonissone traite le problème des données imparfaites en utilisant des systèmes multi-classifieurs basés sur des forêts d'arbres de décisions floues. De ce fait, nous citons les différents travaux réalisés par cette équipe.

Dans leur papier «A Fuzzy Random Forest : Fundamental for Design and Construction» [108] P. Bonissone et al. ont effectué des tests sur différentes bases en comparant les performances de l'algorithme FRF (Fuzzy Random Forest) tout en appliquant des arbres de type Fuzzy ID3 avec les algorithmes de Bayes, C4.5, Random forest et ADABOOST.

Sur une autre étude, Bonissone et al. «Combination Methods in a Fuzzy Random Forest » [109] ont mis en œuvre deux autres méthodes de combinaisons : méthode minimum (chaque classe est sélectionnée par la valeur minimum de toutes les feuilles atteintes dans l'arbre) et la méthode maximum (chaque classe est sélectionnée par la valeur maximum de toutes les feuilles atteintes dans l'arbre) avec les 2 techniques FRF et FID [1]. D'après les résultats comparatifs, les auteurs ont constaté que FRF est nettement meilleure que FID.

Toujours dans la continuité de l'amélioration du principe de FRF, Bossinone et al. dans «Weighted in a Fuzzy Random Forest» [110] ont combiné une méthode pondérée en utilisant une inférence suivant 2 stratégies :

- SM 1 : en combinant l'information des feuilles atteintes dans chaque arbre pour obtenir la décision de chaque arbre, en appliquant le même procédé de combinaison aux autres pour produire la décision globale de la forêt.
- SM2 : en combinant l'information des feuilles atteintes de tous les arbres pour réaliser la décision globale de la forêt.

Les résultats obtenus démontrent que l'implémentation d'inférence pondérée avec la deuxième stratégie est mieux adaptée qu'avec la première.

Dans un papier plus général, Bonissone et al. «A fuzzy Radom forest» [111] ont récapitulé toutes les différentes stratégies et ont dressé un bilan comparatif entre elles. Les méthodes de combinaison utilisées sont :

- La méthode pondérée qui dépend explicitement des données.
- La méthode non pondérée qui dépend des données implicites.

Les résultats recueillis montrent de meilleures performances (jusqu'à 65%) pour les méthodes de combinaison pondérées par rapport aux méthodes non pondérées.

Dans leur dernier papier intitulé «Towards the learning from low quality data in a fuzzy random forest ensemble» [112], ils proposent une étude comparative entre la forêt floue avec des séparateurs flous non uniformes/uniformes.

- La séparation floue non-uniforme (reconnaissance de forme de type non symétrique).
- La séparation floue uniforme. (Reconnaissance de forme de type symétrique, homogène).

Leurs résultats prouvent clairement que FRF avec les séparations floues non-uniformes aboutissent à de meilleurs résultats que les séparations floues uniformes.

Tout récemment le chercheur C. Marcela [113] a présenté une étude sur l'influence du nombre d'arbres pour la construction de la forêt d'arbre de décision flou. Chacune de ces forêts a été employée pour classifier tous les exemples de l'ensemble d'essai, en appliquant 3 graphiques :

- Classique : correspond au taux d'erreur en employant un arbre de décision flou classique de type fuzzy CART.
- Zadeh : correspond au taux obtenu en employant les T-conormes de Zadeh (les opérateurs minimum et maximum) en classifiant des exemples.
- Lukasiewicz : correspond au taux d'erreur obtenu en employant les t-normes de Lukasiewicz en classifiant les exemples.

1. Le FID est un programme qui produit un arbre de décision basé sur la logique floue en permettant la gestion des données imparfaites, en particulier des étiquettes linguistiques

5 Propositions

La FRF accroit la robustesse des arbres de décision flous. La puissance du caractère aléatoire augmente la diversité des arbres dans la forêt, et la flexibilité de la logique floue. Tout ces points permettent de : contrôler des données imparfaites, réduire le taux d'erreur et mettre en évidence plus de robustesse et plus d'interprétabilité.

Dans le cadre de notre travail, nous nous intéressons à l'étude des forêts aléatoires floues optimisées pour les doter d'une meilleure capacité de prise en compte des données imparfaites (numériques, imprécises, incertaines, ou incomplètes), et améliorer leur mise en œuvre dans un raisonnement flou. De ce fait, nous proposons la construction de forêts d'arbres de décision flous (de types Fuzzy CART) pour la classification de l'ensemble des données, en employant la logique floue de Sugeno [114].

6 Base de données

6.1 Pima Diabetes

La base Pima Indian Diabetes (PID) d'Arizona [84] est constituée de 768 femmes dont 268 sont diabétiques et 500 non diabétiques. Chaque cas est formé de 9 attributs, le 9 ème représente la classe du patient (Table 7).

Attributs	**Description de l'attribut**	**Moyenne**	**Écart type**
Npreg	Nombre de grossesses	3.845	3.37
Glu	Concentration du glucose plasmatique	120.895	31.973
BP	Tension artérielle diastolique, (mm Hg)	69.105	19.356
Skin	épaisseur de pli de peau du triceps, (mm)	20.536	15.952
Insu	Dose d'insuline, (mu U/ml)	79.799	115.244
Bmi	Indice de masse corporelle, (poids kg/taille m)	31.993	7.884
Ped	Fonction de pédigrée de diabète (l'hérédité)	0.472	0.331
Age	âge (Année)	33.241	11.76

Table 7 – Description des attributs de la base de données Pima Diabetes

6.2 Liver Disorder

Liver Disorder (Bupa) [84] est une base de données sur les troubles hépatiques collectée par Liver disorder Medical Research Ltd, elle contient 345 exemples de sexe masculin (200 non malades et 145 malades), sont définis par les 7 attributs dont le dernier représente la classe (Table 8).

Attributs	**Description de l'attribut**	**Moyenne**	**Écart type**
mcv	Volume globulaire moyen	90.159	4.448
alkphos	Alcalines phosphatase	69.87	18.348
SGPT	Aminotransferase alanine	30.406	19.512
SGOT	Aspartate aminotransferase	24.643	10.064
gammagt	Gamma-glutamyl transpeptidase	38.248	39.255
drinks	Nombre de boissons d'une demi – pinte l'équivalent de boissons alcoolisées bues par jour	3.455	3.338

Table 8 – Description de la base de données Liver disorder Bupa

7 Expérimentations

7.1 Les forêts aléatoires Classiques (RF)

Pour les deux bases de données nous avons réparti l'ensemble des données en 2/3 pour l'apprentissage et 1/3 pour le test, tout en gardant une distribution des classes équilibrées, les résultats obtenus sont les suivants :

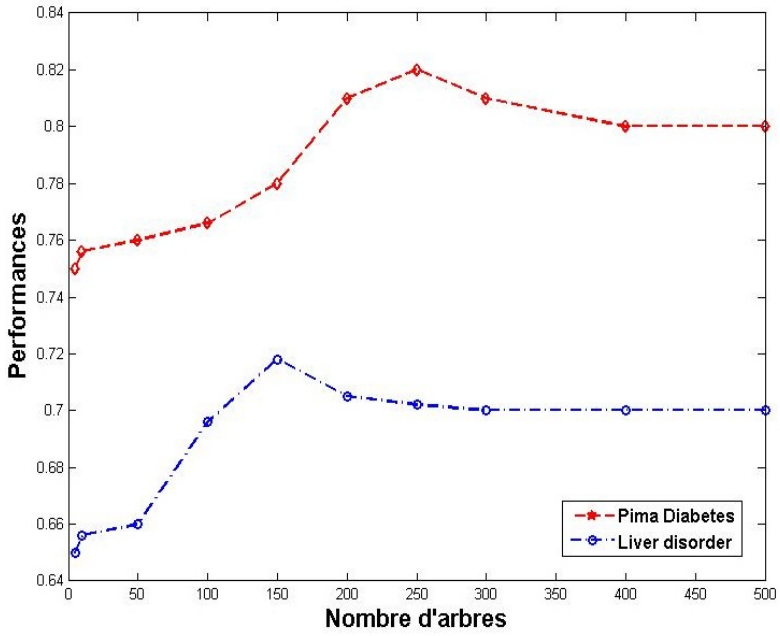


Figure 8 – Performances de classification en fonction du nombre d'arbres pour la base Pima Diabetes et Liver disorder Bupa

- Les performances de la forêt aléatoire (RF) (Figure 8) sur la base Pima Diabetes ont donné les meilleurs résultats au niveau de 200 arbres avec un taux de classification de 82.5%, mais la stabilité est plus présente à partir de 400 arbres.
- Pour la base Liver Disorder le taux de classification est de 72% pour 150 arbres et en augmentant le nombre d'arbres on remarque plus ou moins une stabilité à partir de 250 arbres.

Dans les méthodes d'ensemble et plus particulièrement la forêt aléatoire RF, l'aspect de stabilité des performances étant le plus important. A partir des résultats réalisés nous avons constaté une certaine stabilité qui est présente avec un nombre d'arbres important sur des bases de données à petite dimension. Notre but est d'implémenter l'arbre classique avec le flou afin de voir si la stabilité sera plus présente avec moins de classifieur.

7.2 Forêt Aléatoire Floue avec l'arbre Fuzzy CART (FRF-FCART)

Dans la construction du FRF nous intégrons la logique floue au niveau des nœuds pour chaque arbre en utilisant l'équation 3.1 de l'indice flou de Gini [106]. Cet indice peut s'étendre pour la prise en compte de valeurs floues de la même façon que l'entropie de Shannon.

L'indice flou de Gini $H_G(V/U)$ est défini par :

$$H_G(V/U) = \sum_{i=1}^{m} p * (U_i).G_G(V/U_i), \tag{3.1}$$

avec : $G_G(V/U_i) = 1 - \sum_{j=1}^{m} p * (V_j/U_i)^2$

Les performances de la forêt aléatoire floue (Figure 9) sur la base Pima Diabetes ont donné de meilleurs résultats au niveau de 30 arbres avec un taux de classification de 76% avec une stabilité bien présente sur les différents nombres d'arbres. Pour la base Liver Disorder Bupa, le taux de classification est de 66% et sa stabilité est présente à partir de 100 arbres.

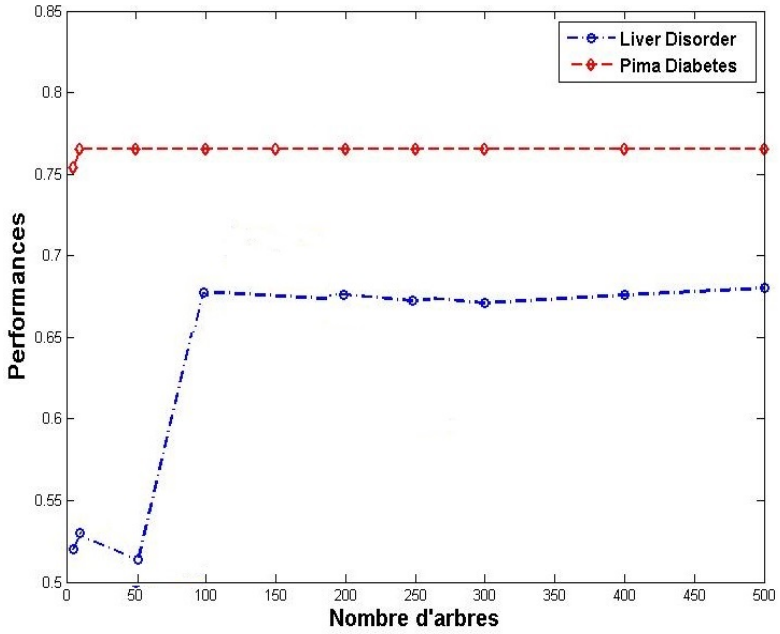


Figure 9 – Performances de classification en fonction du nombre d'arbres pour la base Pima et Bupa

Nous remarquons que Fuzzy CART donne des résultats assez médiocres avec un temps d'exécution important bien que la stabilité recherchée est présente.

La connaissance obtenue

Nous distinguons dans les figures suivantes les fonctions d'appartenances pour les variables de la base de données Pima Diabetes et Liver Disorder Bupa.

La base Pima Diabetes Le classifieur Fuzzy CART génère une base de connaissances de 60 règles de classification. Dans la Figure 10, la partition floue de chaque paramètre est faite de manière automatique.

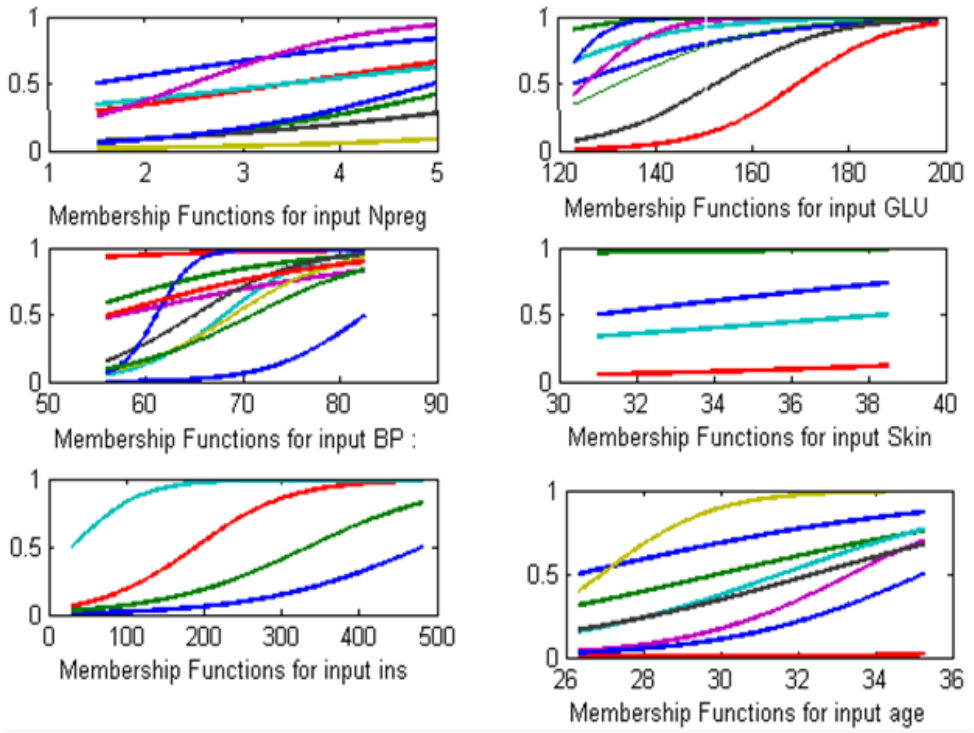


Figure 10 – Les fonctions d'appartenance de la base de données Pima avec Fuzzy CART

La base Liver Disorder Le classifieur Fuzzy CART génère une base de connaissances de 35 règles de classification.

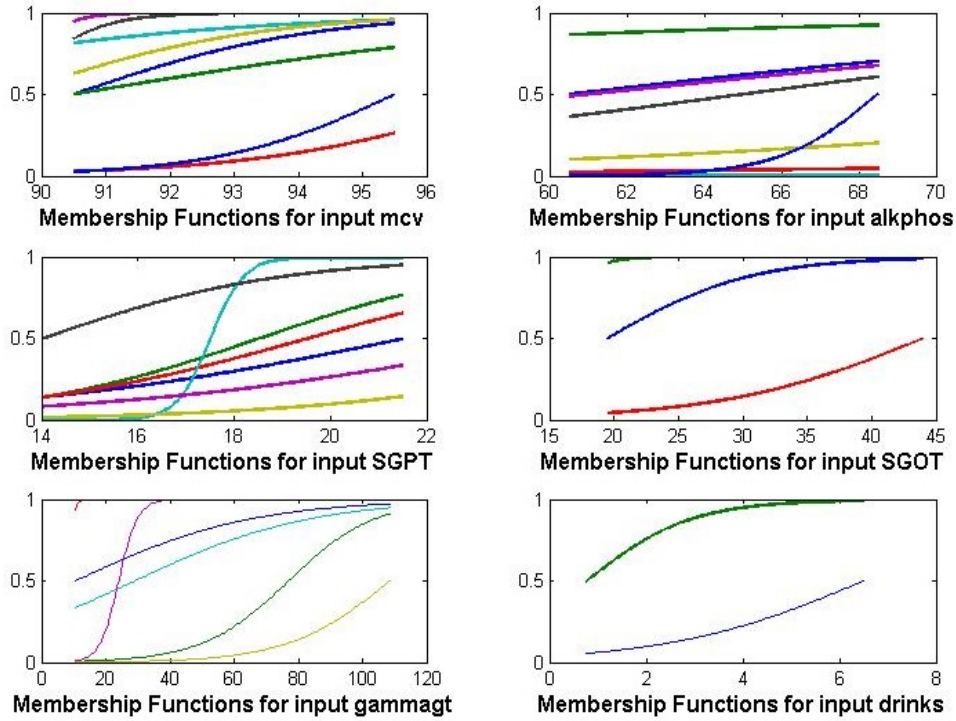


Figure 11 – Les fonctions d'appartenance de la base de données Liver Disorder avec Fuzzy CART

Dans le même cas que la base de données PIMA, nous retrouvons une partition floue de

manière automatique passant par exemple de 8 fonctions d'appartenance pour le paramètre MCV à 2 pour le paramètre DRINKS (Figure 11).

L'inconvénient majeur de cette approche est le fait de solliciter toutes les règles de la base de connaissance (60 règles pour Pima et 35 règles pour Bupa) pour chaque exemple présenté à l'entrée du classifieur, ce qui complique la tâche du médecin.

L'application des arbres de type Fuzzy CART dans les forêts aléatoires floues démontre la stabilité recherchée, mais en contre partie nous a fait perdre en performances et en interprétabilité avec un temps d'exécution très grand. La connaissance obtenue est incohérente et cela peut s'expliquer par les points suivants :

- Incomplétude des partitions floues : exemple deux voisins de sous-ensembles flous dans une partition floue qui ne se chevauche pas.
- Indiscernabilité des partitions floues : exemple les fonctions d'appartenance de deux sous-ensembles flous sont tellement semblables que la partition floue est indiscernable.
- Inconsistance des règles floues : exemple Les fonctions d'appartenance perdent leurs sens prescrits physiques.
- Trop de sous-ensembles flous.

Pour remédier à tous ces points, nous proposons l'algorithme d'optimisation Fuzzy C-Means [115, 116] qui a déjà montré son intérêt d'application pour l'optimisation de l'apprentissage structurel des paramètres flous [117]. Il va également nous permettre d'augmenter la distinction des partitions floues et de réduire le nombre de sous-ensembles flous.

7.3 Forêt Aléatoire Floue avec l'arbre modifié Fuzzy C-Means CART (FRF-FCM-FCART)

Principe de Fuzzy C-Means (FCM) dans Fuzzy CART

FCM est un algorithme de clustering qui permet de faire une répartition des fonctions d'appartenance au niveau des nœuds. Son principe de regroupement permet d'améliorer les performances de chaque arbre et cibler la connaissance. Il est utilisé dans la classification non supervisée. Dans notre cas d'application le nombre de clusters va être égal à 2, car nous traitons un problème de classification binaire et le paramètre de fuzzification m est fixé à 2 de manière heuristique.

L'approche FCM- Fuzzy CART

L'optimisation du Fuzzy CART se fait comme suit (Figure 12), FCM tente de partitionner les données numériques dans des clusters. L'appartenance d'un point de données à un cluster spécifique est exprimée par la valeur d'appartenance de ce point à ce cluster. La valeur d'appartenance est calculée par la minimisation d'une fonction objective de FCM, qui recherche l'appartenance réalisant le moins d'erreur.

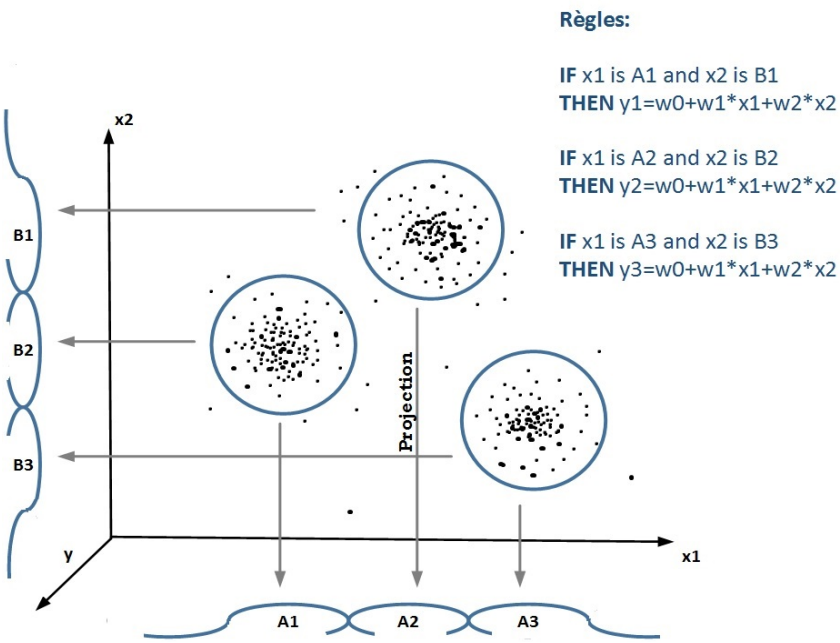


Figure 12 – Schéma représentatif de la répartition des clusters en fonctions d'appartenance avec FCM

Nous allons effectuer la même répartition des bases de données faite précédemment pour l'expérimentation des Forêts Aléatoires Floues avec l'arbre modifié Fuzzy C-Means CART (FRF-FCM-FCART), les résultats obtenus sont les suivants :

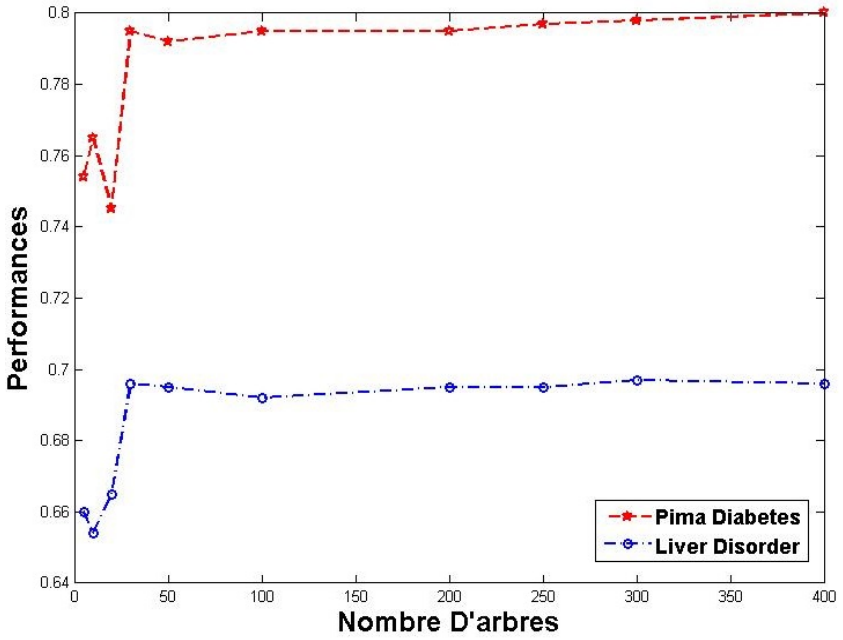


Figure 13 – Erreur de classification en fonction du nombre d'arbres pour la base Pima et Bupa

Les performances de la forêt aléatoire Floue en fonction du nombre d'arbres (Figure 13), ont réalisé les meilleurs résultats au niveau de 30 arbres sur la base PIMA Diabetes avec un taux de classification de 79%, avec une stabilité quasi constante à partir des 30 arbres. Quant à la base Liver Disorder Bupa le meilleur taux de classification est de 69% et la stabilité commence à partir de 30 arbres.

La Connaissance obtenue

La base Pima Diabetes Le classifieur FCM- Fuzzy CART génère une base de connaissances de 2 règles de classification (Figure 14) :

Règle 1 : *If (Npreg is petit) and (Glu is grand) and (Bp is grand) and (Skin is grand) and (PED is grand) and (Age is grand) then (class is malade).*

Règle 2 : *If (Npreg is grand) and (Glu is petit) and (Bp is petit) and (Skin is petit) and (PED is petit) and (Age is petit) then (class is non malade).*

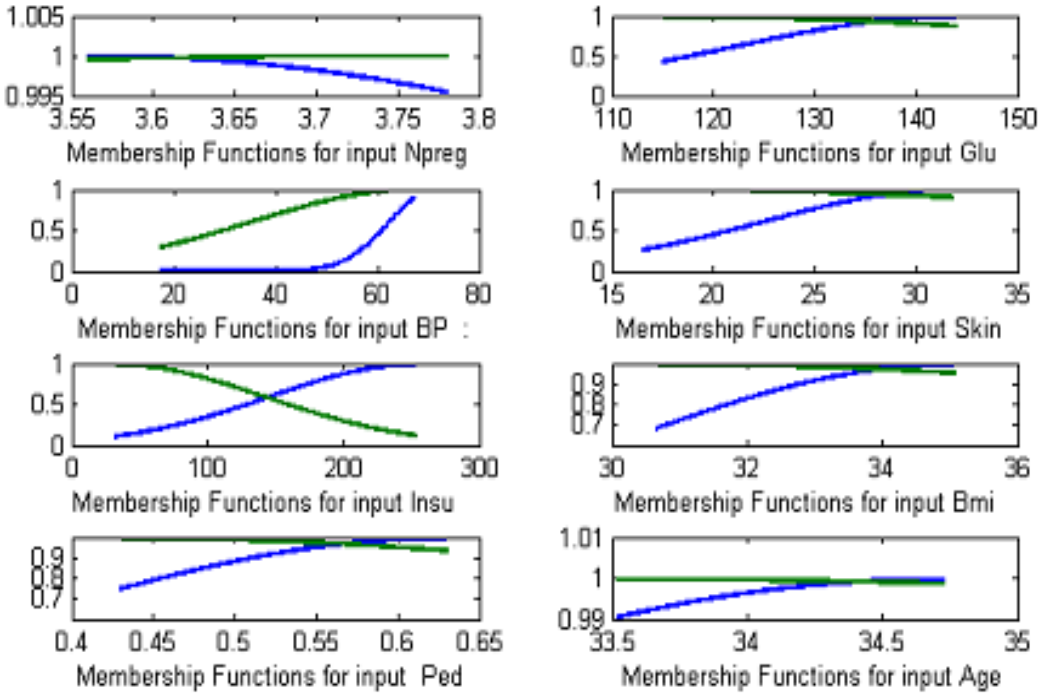


Figure 14 – Les fonctions d'appartenance de la base de données Pima Diabetes avec FCM

La base Liver Disorder Avec le classifieur FCM- Fuzzy CART, une base de connaissances de 2 règles de classification est réalisée (Figure 15) :

Règle 1 : *If (mcv is grand) and (alkphos is grand) and (SGPT is grand1) and (SGOT is petit) and (gammagt is grand) and (drinks is grand) then (class is malade).*

Règle 2 : *If (mcv is petit) and (alkphos is petit) and (SGPT is petit) and (SGOT is petit) and (gammagt is petit) and (drinks is petit) then (calss is non malade).*

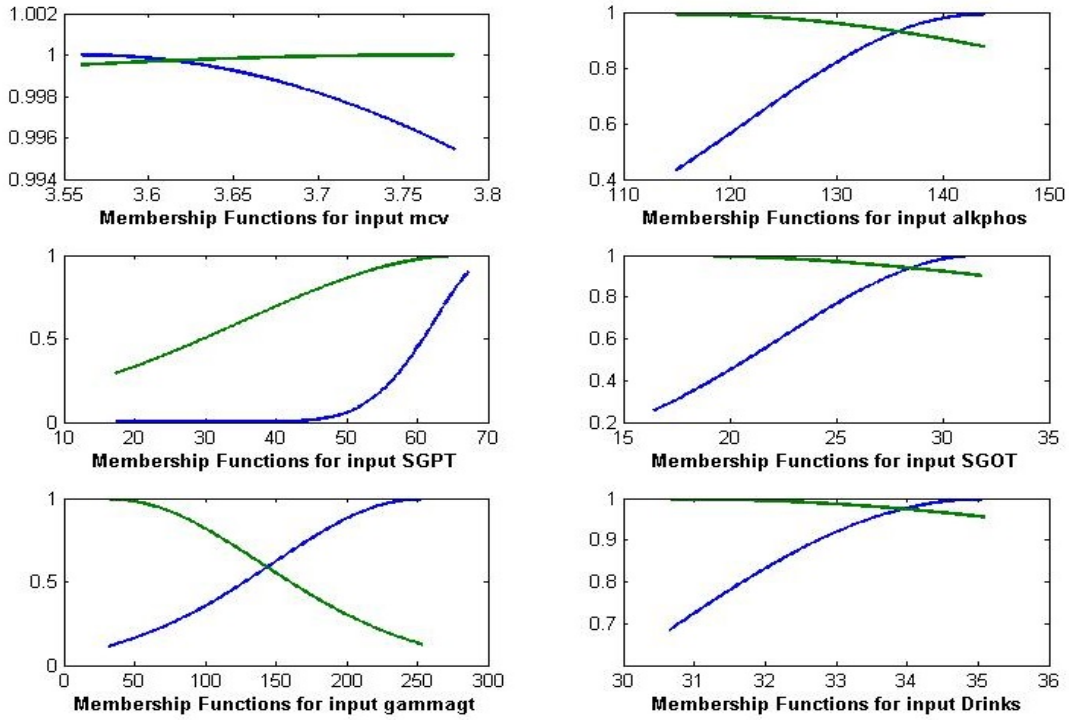


Figure 15 – Les fonctions d'appartenance de la base de données Liver Disorder Bupa avec FCM

Dans cette partie, chaque attribut d'entrée a deux fonctions d'appartenance. Nous remarquons que l'application de fuzzy c-means a réduit considérablement le nombre de règles, celui-ci a été minimisé de 60 règles à 2 pour Pima Diabetes et de 35 à 2 pour Liver Disorder Bupa, ce qui diminue la complexité de la base de connaissances de manière significative.

7.4 Discussion

Dans le cadre de ce travail, nous nous sommes intéressés à la construction d'une forêt d'arbres de décision floue (de types Fuzzy CART) pour la classification de données médicales. Les résultats obtenus sont non satisfaisants, en terme de temps d'exécution et taux de classification, d'où l'initiative d'améliorer cette approche par l'algorithme d'optimisation Fuzzy C-means. Il permet une meilleure répartition des données et ainsi une régularisation des contraintes qui s'appliquent sur les paramètres des fonctions d'appartenance floues. Cette approche nous a permis de réduire le nombre de sous-ensembles flous tout en minimisant le nombre de règles pour une connaissance plus ciblée.

Cet algorithme réduit le nombre de sous-ensembles flous et minimise le nombre de règles pour une connaissance ciblée. Delà donc automatiser et optimiser la structure et les paramètres des fonctions d'appartenance.

Des résultats obtenus, nous pouvons dire que les arbres de décision flous constituent une technique importante dans tout système pour la modélisation de la connaissance dans l'aide au diagnostic médical.

8 Conclusion

Ce chapitre présente un modèle de classification de forêts aléatoires floues, pour cela nous avons appliqué deux critères pour évaluer la méthode proposée. Le premier est la

stabilité des performances du classifieur, le second est le bon taux de classification.

L'algorithme de Fuzzy C-Means (FCM) a permis d'optimiser l'arbre Fuzzy CART en réduisant son temps d'exécution mais tout en gardant une très bonne stabilité et une bonne reconnaissance.

Avec l'hybridation de ces méthodes, une extraction de connaissances par le système d'inférence flou a pu être faite passant de 60 règles avec le Fuzzy CART à 2 avec Fuzzy FCM. La connaissance réalisée est plus fiable, précise et suffisamment simple pour être comprise, le tout en améliorant les performances avec un bon taux de classification pour les deux bases de données.

Nous prévoyons dans les perspectives de ce travail, son adaptation à un problème multi-classes ; aussi ordonnancer les connaissances par ordre de pertinence et ainsi faire un élagage des arbres les moins informatifs (sélection de classifieurs).

Conclusion

Les méthodes de classification ont pour but d'identifier les classes auxquelles appartiennent des objets à partir de certains traits descriptifs. Elles conviennent en particulier au problème de la prise de décision automatisée. La procédure de classification sera extraite automatiquement à partir d'un ensemble d'exemples, ce dernier consiste en la description d'un cas avec la classification correspondante.

Un système d'apprentissage doit alors, à partir de cet ensemble d'exemples, extraire une procédure de classification, il s'agit en effet d'extraire une règle générale à partir des données observées. La tâche générée tentera d'identifier les classes auxquelles appartiennent des objets à partir de certains traits descriptifs et avoir une meilleure prédiction pour classifier les nouveaux exemples.

Les méthodes d'ensemble constituent l'une des principales orientations actuelles de la recherche sur l'apprentissage par machine, elles ont été appliquées à un large éventail de problèmes réels. Malgré l'absence d'une théorie unifiée sur des ensembles de données, il y a beaucoup de raisons théoriques pour combiner plusieurs apprenants, et la preuve empirique confirme l'efficacité de cette approche.

Parmi les algorithmes issus des méthodes d'ensemble, les forêts aléatoires ("Random Forest" RF) [51] sont l'une des derniers aboutissements de la recherche et les plus efficaces pour l'apprentissage d'arbres de décision. RF permet l'agrégation d'arbres randomisés tout en synthétisant les approches développées respectivement par Breiman 1996 [37], Amit et Geman 1997 [52].

Les performances d'une forêt d'arbres dépendent de la qualité des exemples qui la compose et de leur indépendance. Aussi les forêts aléatoires sont fondées sur des arbres non élagués afin de réduire l'erreur de biais ; en outre, le processus aléatoire permet d'assurer une faible corrélation entre les arbres pour sauvegarder leur diversité.

Nous nous sommes intéressés dans ce travail à étudier et améliorer les performances du modèle ensembliste Forêt Aléatoire sur une tâche de classification reliée au domaine médical. À cet effet, nous avons procédé en premier temps à une analyse de travaux effectués dans le domaine qui a permis de mettre en évidence plusieurs avantages ainsi que certaines limites des forêts aléatoires employées dans le cadre de la classification des données médicales. Nous avons constaté en conséquence que les classifieurs déjà proposés à base des forêts aléatoires font preuve de bonnes performances mais peuvent être encore améliorés pour apporter plus de précisions aux résultats.

Nous avons en premier lieu, fait appel à la forêt aléatoire en utilisant l'indice de Gini et le vote majoritaire, puis nous avons procédé au développement de plusieurs variances du même classifieur en utilisant l'indice de Déviance et Twoing rule au niveau du choix de

la variable de division des nœuds des arbres et enfin nous avons appliqué le vote pondéré comme méthode d'agrégation de l'ensemble d'arbres.

En second lieu, une nouvelle méthode de génération d'arbres appelée Subspaces Random Forest (Sub_RF) qui regroupe les principes du Bagging, la méthode des Sous-espaces aléatoires et les Forêts Aléatoires a été essentiellement proposée. Cette méthode s'est, en effet, avérée efficace par rapport à la forêt aléatoire classique.

En dernier lieu, un modèle de classification des forêts aléatoires floues est proposé. Spécifiquement, nous nous sommes plus intéresses à deux critères pour évaluer cette approche. Le premier est la stabilité des performances du classifieur, le second est la bonne reconnaissance. L'algorithme de Fuzzy C-Means (FCM) a permis d'optimiser l'arbre Fuzzy CART en réduisant son temps d'exécution mais tout en gardant une très bonne stabilité et un bonne reconnaissance. Avec l'hybridation de ces méthodes une extraction de connaissances par le système d'inférence flou a pu être faite passant de 60 règles avec le Fuzzy CART à 2 avec Fuzzy FCM. La connaissance réalisée est plus fiable, précise et suffisamment simple pour être comprise, le tout en améliorant les performances avec un bon taux de classification pour les deux bases de données.

Il reste plusieurs questions et limites à nos approches auxquelles nous souhaitons répondre à l'avenir. Nous prévoyons dans les perspectives de ce travail, de renforcer ces résultats avec une analyse théorique ainsi que tester nos algorithmes sur de grandes bases de données. Aussi l'adaptation aux problèmes multi-classes et multi-labels [118].

Troisième partie

III

État de l'art et Propositions en apprentissage semi-supervisé : Classification des données par approche ensembliste

Introduction

Les méthodes modernes d'acquisition automatique permettent d'obtenir de nombreuses variables sur de nombreux individus pour un faible coût. Toutefois, la variable d'intérêt est souvent plus difficile à repérer que les autres. Ceci est particulièrement vrai dans les problèmes de prédiction. Dans ce cas, il est préférable d'apprendre une règle qui permet de prédire la variable d'intérêt étant donné la disponibilité des variables obtenues à moindre coût.

Dans ce contexte, le praticien dispose souvent d'un grand échantillon de données non étiquetées et d'un plus petit échantillon de données étiquetées. Exemple, avec le développement rapide d'Internet, il est facile d'obtenir des milliards de pages Web à partir de serveurs Web. Cependant, la classification de ces pages web dans des classes est une tâche longue et difficile. Aussi dans le domaine de la reconnaissance vocale, l'enregistrement nous donne d'énormes quantités de données audio dont le coût est négligeable. Toutefois, l'étiquetage exige quelqu'un pour l'écouter et le saisir par la suite. Des situations similaires sont valables pour la télédétection, la reconnaissance des visages, l'imagerie médicale, la recherche d'images par contenu [15] et la détection des intrusions dans les réseaux informatiques [31].

Vu la disponibilité des données non étiquetées et la difficulté de leur annotation, les méthodes d'apprentissage semi-supervisé ont acquis une grande importance. A la différence de l'apprentissage supervisé, l'apprentissage semi-supervisé vise les problèmes avec relativement peu de données étiquetées et une grande quantité de données non étiquetées. La question qui se pose est alors de savoir si la seule connaissance des points avec labels est suffisante pour construire une fonction de décision capable de prédire correctement les étiquettes des points non étiquetés. Différentes approches proposent de déduire des points non étiquetés des informations supplémentaires et de les inclure dans le problème d'apprentissage.

Plusieurs sortes de techniques ont été développées pour réaliser la tâche d'apprentissage semi-supervisé. Il existe principalement trois paradigmes [7] [8] qui abordent le problème de la combinaison des données labellisées et non labellisées afin d'améliorer les performances. De ce fait, nous citons en bref ces catégories : apprentissage semi-supervisé, apprentissage transductif et l'apprentissage actif.

L'Apprentissage semi-supervisé (SSL) renvoie à des méthodes qui tentent d'exploiter soit les données non étiquetées pour l'apprentissage supervisé où les exemples non étiquetés sont différents des exemples de test ; soit d'exploiter les données étiquetées pour l'apprentissage non supervisé.

L'Apprentissage Transductif regroupe les méthodes qui tentent également d'exploiter les exemples non étiquetés, mais en supposant que les exemples non étiquetés sont exactement les exemples de test.

L'Apprentissage actif se réfère à des méthodes qui sélectionnent les exemples non étiquetés les plus importants, et un oracle peut être proposé pour l'étiquetage de ces instances, dont l'objectif est de minimiser l'étiquetage des données [9]. Parfois, il est appelé échantillonnage sélectif ou sélection d'échantillon.

Dans cette partie, nous nous focalisons sur l'amélioration des performances de la classification supervisée en employant les données non étiquetées (SSL). Nous mettons en place d'abord l'approche semi-supervisée dans le cadre des problèmes de classification, orientée vers l'application des méthodes d'ensemble en classification semi-supervisée.

L'apprentissage semi-supervisé fait ressortir l'hypothèse que les données non étiquetées vont améliorer l'apprentissage. Mais, cela n'est pas toujours vrai [17]. Chapelle et al. [7] distinguent trois hypothèses que les données doivent satisfaire pour pouvoir profiter des exemples non-étiquetés.

- L'hypothèse de régularité : si deux points x_1, x_2, sont proches dans une région à haute densité, alors les sorties correspondantes doivent être y_1, y_2,
- L'hypothèse de classe : si les points sont dans le même cluster donc ils sont susceptibles d'être de la même classe,
- L'hypothèse de sous-espace : les données (de grande dimension) se situent généralement sur une variété de basse dimension.

La première hypothèse est une condition que les données doivent remplir pour permettre une généralisation à partir d'un petit nombre d'étiquettes. En effet, les données pour lesquelles deux exemples très proches, situés dans des régions denses de l'espace d'apprentissage, et qui n'ont pas plus de chance de posséder des étiquettes communes, deviendront des données dont la nature empêche un apprentissage semi-supervisé. Les deux autres hypothèses spécifient des conditions que les données peuvent remplir et sous lesquelles il est possible d'établir une généralisation à partir d'un petit nombre d'exemples étiquetés : soit les données sont regroupées en classes homogènes, soit en des sous-espaces denses et homogènes. La différence entre ces deux hypothèses concerne essentiellement l'espace dans lequel est mesuré la distance entre deux exemples.

Parmi les méthodes qui profitent de ces hypothèses, Zhu [17] distingue plusieurs familles (Figure 16) la plupart de ces méthodes supposent une structure sous-jacente qui corrèle les données non étiquetées avec une étiquette de classe et les rend par conséquent instructives.

Les paramètres d'apprentissage semi-supervisé peuvent être divisés en mono et multi-vue (Figure 16). Dans l'apprentissage à vue unique, comme EM [119] et l'auto-apprentissage [120], l'algorithme reçoit un ensemble unique de caractéristiques qui sont utilisées pour l'apprentissage. En apprentissage à multi-vue, tels que Co-training [5] et Co-EM [121], l'algorithme d'apprentissage reçoit deux ou plusieurs ensembles de caractéristiques (vue multiple). Ces deux vues font appels aux multiples classifieurs de type homogène (tri-Training [16], Co-Forest [18], Co-training by Committee [122]) ou hétérogène (statistical co-learning [14], Democratic co-learning [15]). Cette catégorie de méthodes a un lien direct avec les méthodes ensemblistes puisqu'elle fait appel à plusieurs classifieurs pour gérer les exemples non-étiquetés.

Dans la partie précédente de cette thèse, nous avons démontré de manière empirique l'efficacité d'application des approches d'ensemble (de manière plus spécifique celle de l'algorithme des forêts Aléatoires). De ce fait, l'approche ensembliste est la seule méthode qui permet la prise en compte rigoureuse recherchée tout en apportant des hypo-

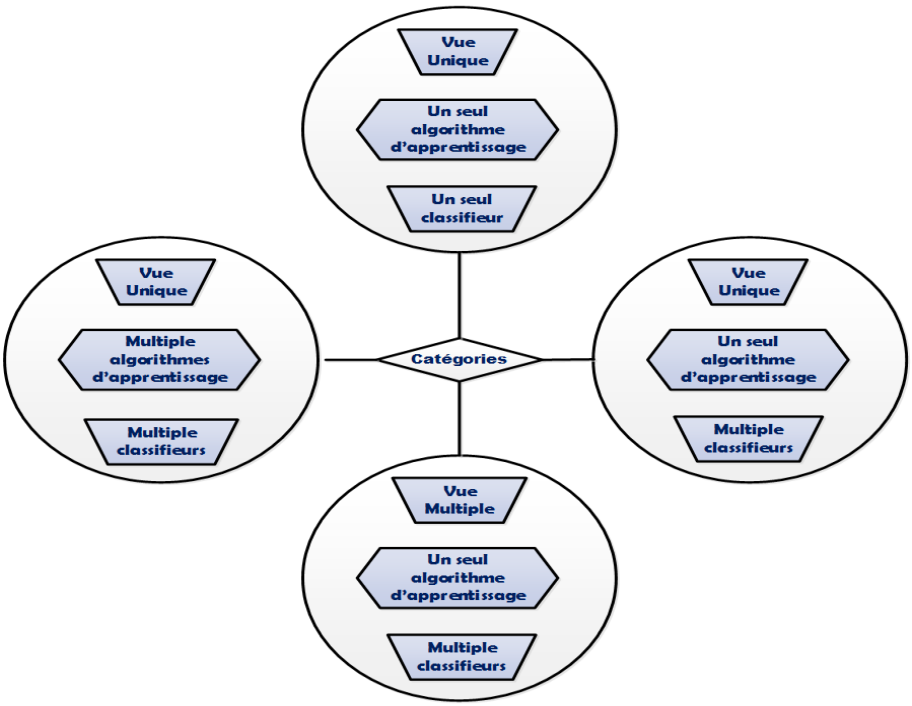


Figure 16 – Les différentes approches pour l'apprentissage semi-supervisé SSL

thèses supplémentaires sur les données non classées. Dans ce cadre nos deux principales contributions à ce travail sont le traitement des questions suivantes : *« Comment juger la pertinence d'un modèle à l'aide des données non étiquetées ? »* et *« Comment améliorer les performances de ce dernier ? ».*

Les données non étiquetées peuvent dans un certain nombre de situations dégrader la règle de classement. Selon Cozman and Cohen [123], ceci ne peut être le cas que si le modèle postulé est mal spécifié. En effet, si le modèle utilisé est bien spécifié l'information apportée par ces dernières est efficacement prise en compte par le modèle, et conduit à une amélioration de la règle de classement supervisé.

Dans cette troisième partie de la thèse, nous cherchons à répondre à la question suivante *« Le modèle utilisé est-il pertinent ? »*. Pour cela, nous partons de l'idée que lorsque différentes méthodes peuvent être utilisées pour estimer un même paramètre il est intéressant de les comparer [124]. Nous comparons alors les estimations supervisées et semi-supervisées des paramètres dans différents domaines applicatifs.

Nous proposons alors la mise en place d'expérimentations qui permettent de détecter, si les paramètres estimés de ces différentes façons sont suffisamment proches compte tenu de l'hypothèse que le modèle spécifie. Ces expérimentations concèdent à détecter les situations où le semi-supervisé est susceptible d'améliorer les performances du supervisé.

Cette troisimèe partie de la thèse se répartie en trois contributions principales suivantes :

– Extrapolation du principe des forêts aléatoires en apprentissage semi-supervisé : application sur données biologiques à grande dimension [60].

– Proposition d'approche de segmentation semi-supervisée des images médicales par classification pixellique semi-supervisée : application sur les images réelles rétiniennes.

– Présentation d'une nouvelle approche de classification semi-supervisée optimisée pour une application à grande échelle de données [61].

CHAPITRE 1

Les Forêts Aléatoires en Apprentissage Semi-Supervisé (*co-Forest*)

1 Objectifs

En apprentissage supervisé, les algorithmes infèrent un modèle de prédiction à partir de données préalablement étiquetées. Cependant, l'étiquetage est un processus long et coûteux qui nécessite souvent l'intervention d'un expert. Cette phase contraste avec une acquisition automatique des données. Ce n'est alors pas habituel de se retrouver avec un volume important de données dont seulement une petite partie a pu être étiquetée. Par exemple, en recherche d'images par le contenu, l'utilisateur souhaite étiqueter le minimum d'images pour fouiller une base aussi grande que possible. Dans ce contexte, l'apprentissage semi-supervisé (SSL) intègre les données non-étiquetées dans la mise en place du modèle de prédiction. En ce sens, l'apprentissage semi-supervisé est à mi-chemin entre l'apprentissage supervisé et non-supervisé : il cherche à exploiter les données non-étiquetées pour apprendre la relation entre les exemples et leur étiquette.

L'apprentissage semi-supervisé a été largement appliqué dans de nombreux domaines tel que le diagnostic médical, la reconnaissance des formes . . .etc. Les méthodes d'apprentissage semi-supervisées font appel aux données non étiquetées en plus des données étiquetées. Leur utilisation offre une meilleure classification des ensembles de données de grande dimension, dont seulement un petit nombre d'exemples étiquetés est disponible. Les méthodes d'ensemble sont considérées comme une solution efficace pour résoudre le problème de dimensionnalité et permettent d'améliorer la robustesse et la capacité de généralisation des apprenants individuels. Dans ce chapitre, nous nous intéressons plus particulièrement à l'algorithme d'ensemble *Random Forest* en semi-supervisé nommé *co-Forest* pour la classification des données biologiques à grande dimension. L'algorithme est évalué sur sa capacité à prédire correctement l'étiquette des exemples non étiquetés, ainsi que sa robustesse lorsque le nombre d'exemples étiquetés disponibles est restreint.

Pour ce faire, ce chapitre va être réparti comme suit : dans la section 2, nous présentons un état de l'art des approches d'ensemble semi-supervisées existantes en expliquant le principe de chacune. Dans la section 3, nous détaillons notre proposition. Nos résultats sont présentés et discutés dans la section suivante. Enfin, une synthèse générale qui met en évidence les principales propriétés de cette technique ainsi que ses points forts.

2 État de l'art

Le premier à avoir vu le jour dans la catégorie des techniques d'apprentissage ensemblistes en semi-supervisé est l'algorithme *co-Training*, proposé par Blum et Mitchell [5] pour la classification semi-supervisée des pages web.

Le *co-Training*, suppose que les variables sont naturellement partitionnées en deux ensembles $x = (x_1; x_2)$. Par exemple, pour les pages Web on considère l'ensemble des liens hypertextes et l'ensemble du contenu, sous les hypothèses suivantes :

1. Chaque composant est suffisant pour la classification,
2. Les composants sont indépendants conditionnellement à la classe,

Blum et Mitchell (1998) démontrent des garanties de type Probably Approximately Correct (PAC) (Valiant [125]) sur l'apprentissage en présence de données étiquetées et non étiquetées (Algorithme 5).

Algorithm 5 *co-Training* pseudo code pour la classification de documents

1: **Entrée :** une collection initiale de documents étiquetés
2: **Sortie :** Deux classifieurs, $C1$ et $C2$, qui prédisent l'étiquette des nouveaux documents.
3: **Répéter jusqu'à ce qu'il n'y est plus de document sans étiquette.**
- Construire le classifieur $C1$ en utilisant la partie x_1 de chaque document
- Construire le classifieur $C2$ en employant la partie x_2 de chaque document.
- Pour chaque classe k, ajouter à la collection de documents étiquetés le document non-étiqueté classé dans la classe k par le classifeur $C1$ avec la plus forte probabilité.
- Pour chaque classe k, ajouter à la collection de documents étiquetés le document non-étiqueté classé dans la classe k par le classifeur $C2$ avec la plus forte probabilité.
4: Ces prédictions peuvent ensuite être combinées.

Les auteurs ont également montré que l'indépendance des deux sous-ensembles d'attributs est une condition nécessaire pour que cet algorithme du *co-Training* améliore la prédiction. En pratique, cependant, il n'est pas toujours possible d'obtenir deux sous-ensembles d'attributs indépendants par rapport à l'étiquette, ce qui rend le *co-Training* difficilement généralisable.

Pour contourner cette difficulté, Goldman and Zhou [14] proposent d'adapter cette stratégie à un ensemble de classifieurs hétérogènes, appelé *statistical co-learning*. Leur méthode suit la procédure mise en place par Blum and Mitchell [5], même s'ils sont obligés de contrôler la qualité des exemples nouvellement étiquetés avant de les ajouter définitivement à l'apprentissage.

Grâce à l'aide de trois classificateurs au lieu de deux, Zhou et al. [16] ont proposé l'algorithme d'apprentissage *tri-Training*, qui nécessite des sous-ensembles d'attributs ni suffisants ni redondants et pas d'algorithmes d'apprentissage supervisé spéciaux qui pourraient diviser l'espace d'exemple dans un ensemble de classes d'équivalence. L'algorithme *tri-Training* réalise la prédiction par un vote majoritaire plutôt que par un classifieur combiné ou un stacking.

D'autre part, une amélioration de l'algorithme *co-Training* a vu le jour sous le nom de *co-Forest*, qui étend le paradigme du *co-Training* en appliquant *Random Forest* [51]. Il a été introduit par Li et Zhou [18] dans l'application à la détection de micro calcifications pour le diagnostic du cancer du sein. *co-Forest* utilise $N \geq 3$ classifieurs au lieu des 3 dans *Tri training*. Les $N - 1$ classifieurs sont employés pour la détermination des

exemples de confiance, appelés Ensemble de concomitance = $H_i = H_{N-1}$. La confiance d'un exemple non étiqueté peut être simplement estimée par le degré de confiance sur l'étiquetage, à savoir le nombre de classifieurs qui sont d'accords sur l'étiquette assignée par H_i.

Les approches proposées par [5], [15], [16] (ainsi que d'autres telles que celles de [126]) montrent l'avantage à utiliser plusieurs classifieurs. Cependant, l'apprentissage de ces classifieurs implique de prédire les exemples non-étiquetés avant de les appliquer. De ce fait, par sa mesure de confiance l'algorithme de Li et Zhou [18], propose le meilleur compromis à l'approche semi-supervisée.

Pour bien mettre en évidence ce principe, dans ce chapitre, nous nous intéressons à l'application de l'algorithme *co-Forest* pour la classification des données biologiques à grande dimension en apprentissage semi-supervisé.

3 Proposition

Dans ce travail nous nous focalisons sur l'application de l'algorithme semi-supervisé *co-Forest* sur les données biologiques transcriptomes (Figure 17), où le nombre de variables p, peut atteindre des dizaines voire des centaines de milliers. Dans un même temps, pour beaucoup d'applications, le nombre d'observations n, se trouve réduit à quelques dizaines.

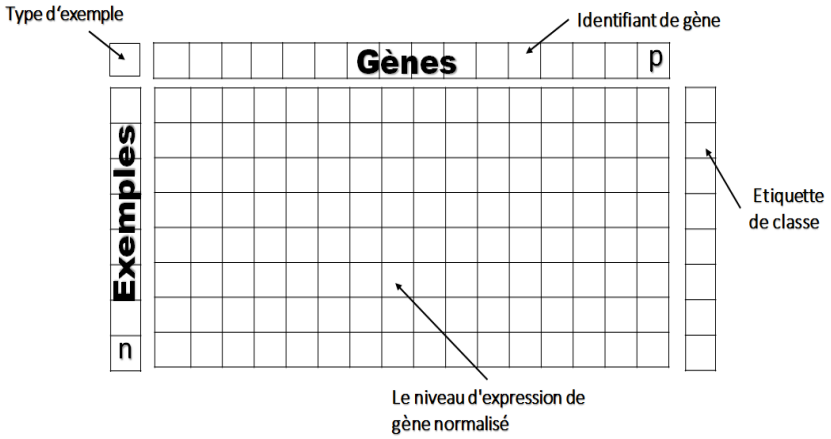


Figure 17 – Structure des bases biologiques

Le domaine typique de telles situations est le domaine médical où l'on peut maintenant faire énormément de mesures sur un individu donné (mesures d'expression de gènes par exemple), mais le nombre d'individus sur lequel on fait l'expérience est réduit (dans le cas d'étude d'une maladie, le nombre de porteurs de la maladie qui participent à une étude est souvent très limité).

4 Principe de l'algorithme *co-Forest*

L'algorithme *co-Forest* a été proposé par Li and Zhou [18], il étend le paradigme *co-Training* [5] par la méthode d'ensemble Forêts aléatoires (*Random Forest*) [51].

Algorithm 6 Pseudo code de l'algorithme *co-Forest*

Entrées : L'ensemble d'exemples labellisés L, ensemble d'exemples non labellisés U, seuil de confiance θ, nombre d'arbres N
Sorties : Ensemble d'arbres h_i,
Processus :
Construction de la forêt aléatoire à N arbres
for $i = 1 \rightarrow N$ **do**
  $\hat{e}_{i,0} \leftarrow 0.5$
  $W_{i,0} \leftarrow 0$
end for
$t \leftarrow 0$
Répéter jusqu'à ce qu'il n'y ait aucuns changements dans les arbres de la forêt aléatoire.
$t \leftarrow t + 1$
for $i = 1 \rightarrow N$ **do**
  $\hat{e}_{i,t} \leftarrow EstimateError\ (H_i, L)$
  $L_{i,t} \leftarrow \phi$
  if $(\hat{e}_{i,t} < \hat{e}_{i,t-1})$ **then**
    $U_{i,t} \leftarrow Subsample\ (\mathrm{U},\ \frac{\hat{e}_{i,t-1}.W_{i,t-1}}{\hat{e}_{i,t}})$
    for $x_u \epsilon\ U_{i,t}$ **do**
      if $Confidence\ (H_i, x_u) > \theta$ **then**
        $L'_{i,t} \leftarrow L_{i,t} \bigcup (x_u, H_i(x_u))$
        $W_{i,t} \leftarrow W_{i,t} +$ ***Confidence*** (H_i, x_u)
      end if
    end for
  end if
end for
for $i = 1 \rightarrow N$ **do**
  if $(\hat{e}_{i,t}.W_{i,t} < \hat{e}_{i,t-1}.W_{i,t-1})$ **then**
    $h_i \leftarrow LearnRandomTree\ (L \bigcup (L'_{i,t}))$
  end if
end for
fin de Répéter
Retourner h_i
Vote Majoritaire $H^*(x) \leftarrow argmax_{y \epsilon label} \sum_{i:h_i(x)=y} 1$

Nous désignons par L et U l'ensemble des données étiquetées et non étiquetées respectivement. En *co-Training*, deux classifieurs sont formés à partir de L, puis chacun d'eux sélectionne les exemples les plus confiants dans U pour la phase de labellisation, à partir de leur propre fonction de classement ou par séparation hyperplan. Ainsi, une partie importante de *co-Training* réside dans la façon d'estimer la confiance de la prévision, en d'autres termes, *comment estimer ou obtenir la confiance d'un exemple non labellisé ?*.

Dans *co-Forest*, un ensemble de N classifieurs désignés comme H^* est utilisé dans *co-Training* au lieu de deux classifieurs. De cette façon, nous pouvons estimer efficacement la confiance de chaque classifieur. Si nous voulons considérer l'exemple labellisé le plus confiant par un des classifieurs h_i $(i = 1, 2, \ldots, N)$ de l'ensemble H^*, nous employons l'ensemble de concomitance de h_i noté par H_i qui regroupe tous les autres classifieurs à l'exception de h_i. Par conséquent, la confiance de l'étiquetage peut être calculée comme le degré d'accord sur l'étiquetage, c'est à dire le nombre de classifieurs qui s'accordent sur le label attribué par H_i. L'idée générale de *co-Forest* est de former tout d'abord un ensemble de classifieurs par les données étiquetées L et d'affiner chaque classifieur avec des données non marquées jugées les plus confiants par son ensemble de concomitance.

Plus précisément, dans chaque itération d'apprentissage autour de *co-Forest*, l'ensemble de concomitance H_i permettra de tester chaque exemple dans U. Pour un exemple non

labellisé xu, si le nombre de classifieurs qui s'accordent sur une étiquette particulière dépasse un seuil θ prédéfini, cette nouvelle étiquette lui sera affectée et par la suite il sera copié dans le nouveau ensemble L'. A l'itération suivante, L' est employé pour le raffinage de h_i. Nous noterons que les exemples non étiquetés ne sont pas supprimés de U, de sorte qu'ils peuvent être sélectionnés par d'autres $H_j(j \neq i)$ dans les itérations suivantes (Algorithme 5).

Un problème qui peut affecter la performance globale de *co-Forest* est que toutes les données non étiquetées dont la confiance est au-dessus de θ seront ajoutées à L_i, ce qui rend L_i assez grand dans l'avenir. Mais dans le cas où un classifieur ne peut représenter la distribution sous-jacente ; une énorme quantité de données étiquetées deviendra nuisible à la performance au lieu d'augmenter la précision de la prédiction.

Ce phénomène a été découvert dans plusieurs algorithmes d'apprentissage semi-supervisé. Inspiré par Nigam et al. [121], *co-Forest* reprend aussi le principe en assignant un poids à chaque exemple non marqué. Un exemple est pondéré en fonction de la confiance prédictive d'un ensemble concomitant. Cette approche permet de réduire l'influence de θ, même si θ est petit, les exemples qui ont une faible confiance prédictive peuvent être limités.

$$\frac{\hat{e}_{i,t}}{\hat{e}_{i,t-1}} < \frac{w_{i,t-1}}{w_{i,t}} < 1 \tag{1.1}$$

Selon Li et Zhou [18], dans les deux itérations ($(t-1)$ avec $t > 1$), la condition (eq. 1.1) pour l'actualisation de l'ensemble d'apprentissage devrait être satisfaite pour améliorer de manière itérative la capacité de généralisation.

En supposant que le classificateur h_i est reconstruit (phase ré-apprentissage) sur l'ensemble de données $L_i U L'_{i,t}$ dans la t i ème itération. Et que $\hat{e}_{i,t}$ désigne le taux d'erreur de H^* sur $L'_{i,t}$ aussi que $\hat{e}_{i,t} w_{i,t}$ est la moyenne pondérée des exemples mal labellisés par H^*.

Les hypothèses que $\hat{e}_{i,t} < \hat{e}_{i,t-1}$ et $w_{i,t-1} < w_{i,t}$, $w_{i,t} < \frac{\hat{e}_{i,t-1} w_{i,t-1}}{\hat{e}_{i,t}}$ doivent être réunis par l'obtention du sous-échantillon de U et ainsi satisfaire la condition en (eq. 1.1).

En résumé, le principe de l'algorithme *co-Forest*, consiste en N arbres aléatoires qui sont d'abord formés à partir d'un ensemble d'apprentissage bootstrap [1] de l'ensemble labellisé L pour créer une forêt aléatoire. Ensuite, à chaque itération, chaque arbre aléatoire sera affiné avec les exemples nouvellement marqués par son ensemble de concomitance, où la confiance de l'exemple labellisé dépasse un certain seuil θ. Cette méthode permettra de réduire les chances qu'un arbre dans une forêt aléatoire soit biaisé lorsque nous utilisons les données non étiquetées.

5 Expérimentations et Résultats

Nous comparons les performances de l'algorithme *co-Forest* à ceux de *Random Forest* classique, afin d'évaluer les situations où le semi-supervisé est susceptible d'améliorer les performances du supervisé. Le tableau (Table 9) décrit l'ensemble de données biologiques appliquées pour l'expérimentation en termes de nombre d'exemples, de gènes (attributs) et de classes.

1. Un échantillon bootstrap L est, par exemple, obtenu en tirant aléatoirement n observations avec remise dans l'échantillon d'apprentissage L_n, chaque observation ayant une probabilité 1/n d'être tirée [69].

Base de données	♯ Gènes	♯ Exemples	♯ classes	Références
Colon	2000	62	2	[127]
Prostate	12533	102	2	[128]
Childhood tumors	9945	23	2	[129]
Leukemia	7129	38	2	[130]
Breast Cancer	12625	24	2	[131]
Dataset-C	7129	60	2	[132]

Table 9 – Caractéristiques des bases d'expérimentation

Les données d'apprentissage sont aléatoirement divisées en deux ensembles : L labellisé et non labellisé U fixés par un taux (μ), qui peut être calculé par la taille de U sur la taille de $L \cup U$. Afin de simuler différentes quantités de données non étiquetées, quatre différents taux de « non-labéllisation » $\mu = 20\%$, 40%, 60% et 80%, sont étudiés. Notons que les distributions de classe pour L et U sont maintenues similaires à celle de l'ensemble originel de données.

Dans ces expérimentations, nous reprenons les mêmes paramètres fixés dans [18], avec une valeur de N fixée à 6 arbres, le seuil de confiance θ à 0,75, soit un exemple non labellisé est considéré comme étant de confiance si plus de 3/4 des arbres s'accordent sur le label donné.

Pour estimer l'erreur sur chaque jeu de données, nous avons prédéterminé un ensemble d'exemples étiquetés. Pour chaque ensemble, l'algorithme est évalué sur sa capacité à prédire correctement l'étiquette des exemples non étiquetés. Les exemples étiquetés ont été choisis aléatoirement, avec pour seule condition celle d'avoir au moins un exemple de chaque classe présent dans chaque ensemble.

μ	Techniques	Colon	Prostate	Childhood Tumors	Leukemia	Breast Cancer	Dataset-C	Moyenne
80 %	Random Forest	0,45	0,36	0,80	0,48	0,53	0,14	0,41
	co-Forest	0,41	0,34	0,77	0,48	0,52	0,09	0,36
	Improv.	3,64%	1,91%	3,08%	0,82%	1,38%	5,50%	2,26%
60 %	Random Forest	0,41	0,23	0,57	0,24	0,43	0,06	0,26
	co-Forest	0,36	0,22	0,45	0,21	0,41	0,05	0,23
	Improv.	4,98%	0,40%	11,67%	2,78%	2,32%	0,22%	3,60%
40 %	Random Forest	0,35	0,20	0,27	0,16	0,40	0,02	0,17
	co-Forest	0,33	0,15	0,21	0,13	0,36	0,02	0,15
	Improv.	2,20%	4,64%	5,95%	2,17%	4,24%	0,11%	2,05%
20 %	Random Forest	0,38	0,21	0,44	0,10	0,26	0,01	0,15
	co-Forest	0,35	0,20	0,43	0,09	0,23	0,01	0,14
	Improv.	2,78%	0,75%	1,00%	0,97%	2,57%	0,08%	1,50%
0 %	Random Forest	0,36	0,20	0,34	0,06	0,25	0,01	0,12
	co-Forest	0,35	0,20	0,30	0,06	0,26	0,01	0,12
	Improv.	0,29%	0,91%	3,98%	0,09%	-0,60%	0,02%	0,78%

Table 10 – L'erreur Moyenne des algorithmes comparés aux différents taux μ

L'algorithme *co-Forest* a permis d'améliorer les performances de *Random Forest*. Cela peut être observé aux différents pourcentages de non labéllisation μ (Figure 18), en particulier lorsque le taux est élevé (les données étiquetées sont très faibles).

Afin de comparer la performance d'amélioration de *co-Forest* sur *Random Forest*, les améliorations *"improv."* sont moyennées sur l'ensemble des jeux de données dans tous les exemples non labellisés, et ainsi une amélioration de la performance globale est obtenue.

Nous avons enregistré (Table 10) un pourcentage d'amélioration *Improv.* de 3.64% sur la base Colon avec seulement 12 exemples étiquetés. Des performances très intéressantes sont réalisées sur la base Childhood tumors où *co-Forest* dépasse largement *Random Forest* par un taux d'amélioration égal à 11%, avec seulement 9 échantillons étiquetés (Figure 18).

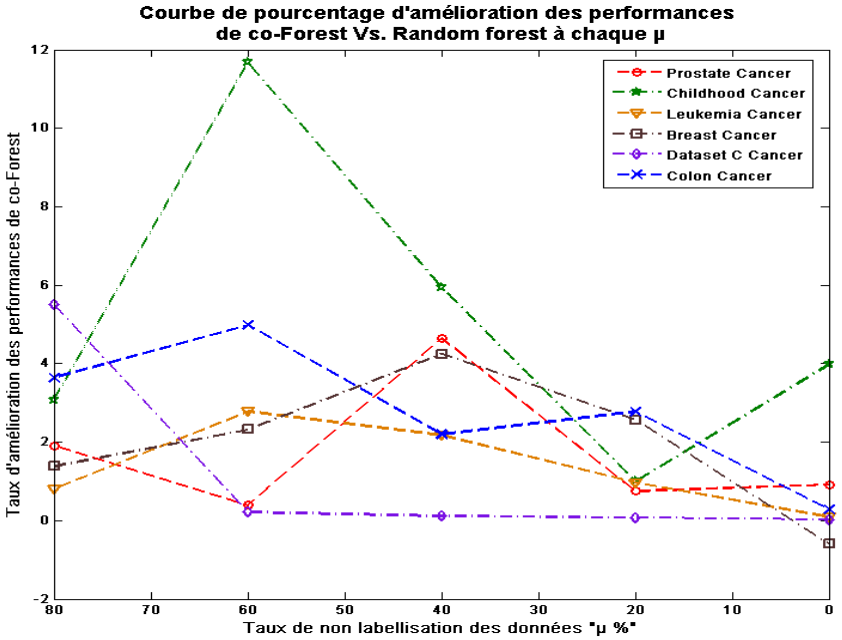


Figure 18 – Courbe de performances d'amélioration de co-Forest Vs. Random forest à chaque μ

L'amélioration moyenne la plus élevée de *co-Forest* sur l'algorithme supervisé *Random forest* est de 3,6% sur l'ensemble des données biologiques, avec un taux de non-labéllisation μ égale à 60% (voir Table 10). Enfin, pour chaque jeu de données, l'algorithme *co-Forest* a montré sa robustesse quand le nombre d'exemples étiquetés disponibles diminue (voir Figure 19).

6 Conclusion

Dans ce travail l'algorithme de *co-Forest* est présenté. Il fait appel aux échantillons non étiquetés pour améliorer les performances d'apprentissage à partir des échantillons labellisés. Il étendant le paradigme de *co-Training*, en exploitant la puissance de la forêt aléatoire. *co-Forest* permet la sélection des échantillons non labellisés les plus confiants par la participation de tout ces classifieurs. Des expérimentations menées sur des ensembles de données biologiques prouvent l'efficacité de *co-Forest* et confirme ainsi sa capacité de mesure en améliorant la performance de l'hypothèse apprise sur une petite quantité d'échantillons labellisés tout en exploitant les échantillons non labellisés.

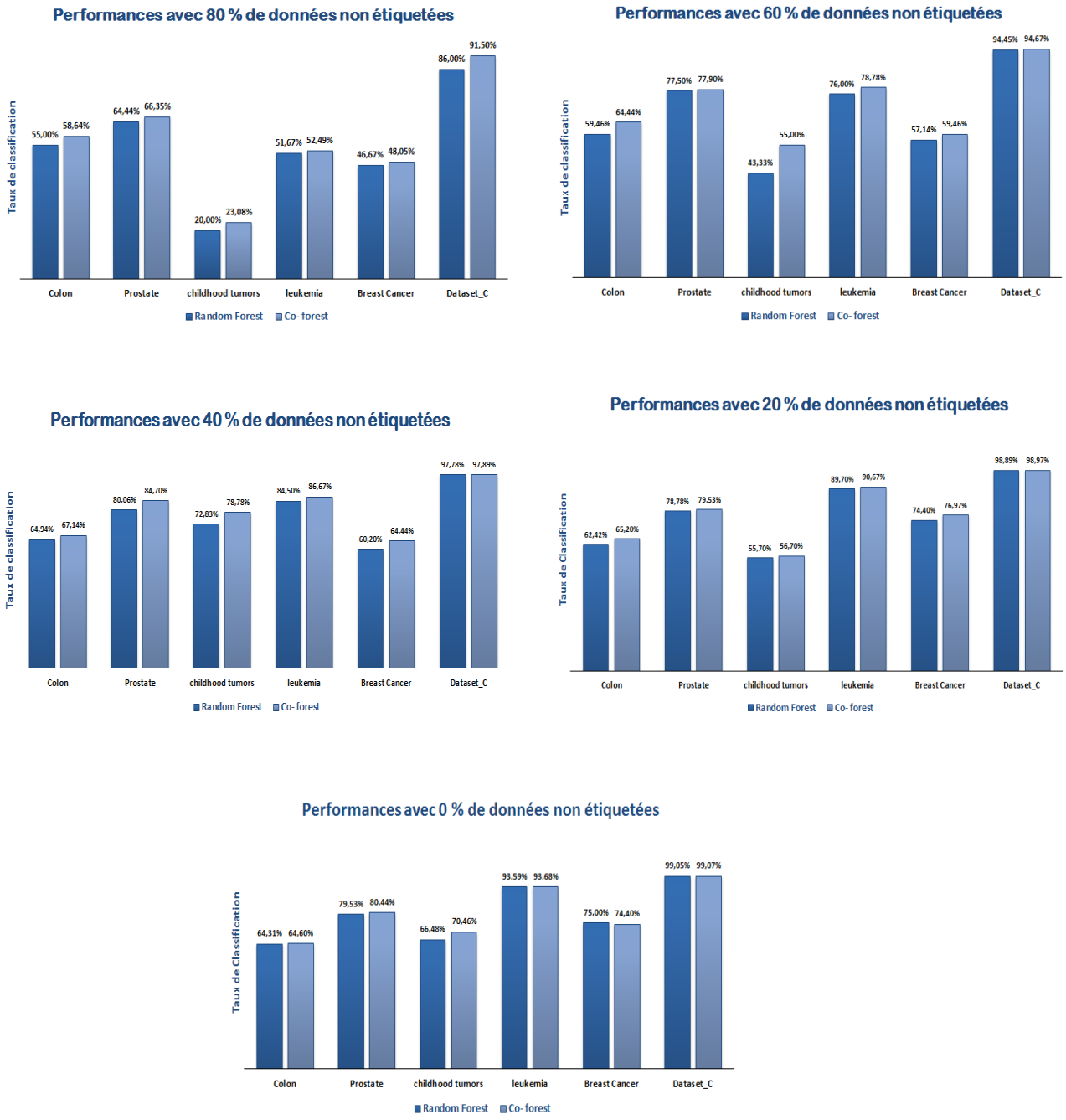


Figure 19 – Histogrammes de performances pour chaque degré de non labéllisation μ

CHAPITRE 2

Les Forêts Aléatoires en Apprentissage Semi-Supervisé (*Co-forest*) pour la segmentation des images rétiniennes

1 Contexte

La majorité des applications dans le domaine d'aide au diagnostic médical nécessitent l'acquisition de données d'imagerie de natures diverses : radiologies, examens scanner ou IRM, imagerie échographique, vidéo, . . .etc. Une tâche fondamentale dans le traitement de ces données est la segmentation, c'est-à-dire l'extraction des structures d'intérêt dans les images, en 2D ou en 3D. Ces informations servent notamment de base à la visualisation des organes, la classification des objets, la génération de modèles de simulation, ou des mesures surfaciques ou volumiques.

Dans le cadre de ce travail, nous nous intéressons à l'annotation des images au niveau super-pixellique par l'approche de segmentation semi-automatique. Ce procédé nécessite une interaction plus ou moins importante de l'expert. Ce type de traitement est utile soit pour traiter directement les données, soit pour définir un résultat de référence qui pourra être appliqué pour l'évaluation de méthodes de segmentation automatiques.

La segmentation automatique d'image vise à l'extraction automatisée d'objets caractérisés par un contour. Elle a pour but de rassembler des pixels entre eux suivant des critères prédéfinis, le plus souvent les niveaux de gris ou la texture. Les pixels sont ainsi regroupés en régions, qui constituent une partition de l'image. Néanmoins, cette tâche reste difficile à réaliser surtout dans les cas de figure où les bords d'un objet sont manquants et/ou il y a un faible contraste entre les régions d'intérêt (ROI) et le fond.

Dans cette partie de thèse, nous nous focalisons sur la segmentation semi automatiques des images fond d'œil de la rétine pour le suivi médical de la maladie de Glaucome. Dans les images rétiniennes, la papille, ou disque optique (« disc » anglais), ou tête du nerf optique, est le lieu de rassemblement des fibres optiques à l'entrée du nerf optique (Figure 20). C'est aussi la structure anatomique oculaire qui se dégrade progressivement par le glaucome. Cette altération s'exprime par l'apparition d'une excavation (« cup » Figure 20), ou par l'élargissement d'une excavation constitutionnelle physiologique. Elle est visible à l'examen du fond d'œil, et peut précéder de plusieurs années l'apparition des désordres péri-métriques. Le dépistage du glaucome s'effectue à l'aide de plusieurs examens comme la mesure du champ visuel, celle de la pression intraoculaire ou encore le fond d'œil qui permet d'analyser l'état du nerf optique.
Actuellement, la recherche tente d'identifier plus précisément les patients présentant des facteurs à risques. L'objectif étant une prise en charge toujours plus précoce de la

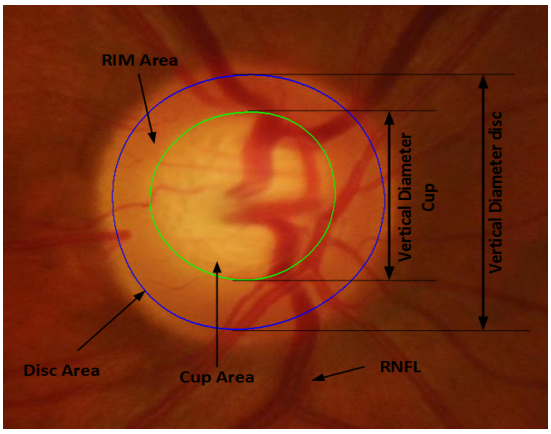


Figure 20 – Mesure du rapport cup/disque Optique

maladie afin de retarder le développement de déficits visuels. Les recherches récentes sur la détection automatique du glaucome visent à détecter le cup et le disque optique pour calculer le rapport cup/disque (C/D : la valeur du diamètre du cup divisée sur le diamètre du disque Figure 20).

Plusieurs études médicales [133, 134], ont démontré que la forme du disque physiologique de la papille d'une population normale est généralement arrondie ou de forme ovale horizontale. Par ailleurs, la perte d'épaisseur des fibres nerveuses rétiniennes péri-papillaires (RNFL) se produit au niveau des pôles supérieurs et inférieurs du disque. La forme du cup est agrandit verticalement plus qu'horizontalement. Par conséquent, le rapport vertical cup sur disque (VC/D) apporte une information clinique plus pertinente pour le suivi du glaucome.

2 Objectifs

Dans le but de mesurer automatiquement le rapport cup/disque (VC/D), de nombreuses approches ont été proposées dans la littérature pour la segmentation d'images du fond d'œil afin d'extraire la région du disque optique. Tous ces travaux peuvent être repartis au sein des familles d'approches comme : les méthodes de segmentation, les techniques d'extraction de caractéristiques et les méthodes de classification qui comprennent les algorithmes supervisés et non supervisés. La segmentation supervisée par des méthodes telles que les réseaux de neurones et les machines à vastes marges (SVM) conduisent à une grande précision, mais ils nécessitent une grande quantité de données étiquetées pour leur apprentissage, ce qui est difficile, coûteux et lent à acquérir dans les applications réelles. D'autre part, les méthodes d'apprentissage non supervisé tels que K-means et C-Moyens floues (FCM) suppriment les coûts d'étiquetage mais ont un rendement inférieur par rapport aux méthodes supervisées.

Pour résoudre ces problèmes, nous proposons une approche semi-supervisée pour la segmentation des régions cup et disque pour le calcul du rapport (VC/D). Plusieurs algorithmes semi-supervisés tels que l'auto-apprentissage [135], Co-training [5], Expectation Maximization (EM) [136] et, récemment, la méthode d'ensemble Co-Forest [18, 137], ont été développés, mais aucun d'entre eux n'a été utilisés pour segmentation semi-supervisée.

Plusieurs approches ont été développées afin d'automatiser le plus possible le processus de segmentation cup et disque optique des images rétiniennes. La majorité des méthodes existantes sont basées sur la modélisation et caractérisation du contour [138–141]. Toutefois, des études récentes [142, 143] indiquent que le procédé de segmentation par classification pixellique est potentiellement très utile et intéressant. Néanmoins, la grille

de pixels est une représentation non naturelle des scènes visuelles. Elle est plutôt un "artefact" d'un processus de formation d'image numérique. Ainsi, il serait plus naturel, et sans doute plus efficace, de travailler avec des entités perceptuellement significatives obtenues à partir d'un processus de regroupement de bas niveau.

De ce fait, les algorithmes de regroupement des pixels ou de sur-segmentation de l'image connus sous le nom de "super-pixels", peuvent être utilisés pour remplacer la structure rigide de la grille de pixels [144]. Un superpixel est un correctif d'image, qui permet de mieux aligner les bords d'intensité qu'un patch rectangulaire. Il capture la redondance dans l'image et réduit la complexité du traitement subséquent. En effet, le super-pixel peut causer une accélération substantielle dans le temps de traitement, puisque le nombre de super-pixels d'une image varie de 25 à 2500, contrairement aux centaines de milliers de pixels dans une image [145].

Les super-pixels sont devenus des blocs de construction clés pour de nombreux algorithmes de vision par ordinateur ; La construction des super-pixels sur l'image centrale, nous permet de travailler localement sur l'information extraite. Les avantages de cette approche peuvent être résumés dans les points suivants [144] :

1. Les super-pixels adhèrent au mieux aux contours de l'image.
2. Dans le pré-traitement, les super-pixels sont rapides au calcul et simple à utiliser.
3. Pour la segmentation, les super-pixels augmentent à la fois la vitesse de traitement et améliorent la qualité des résultats.

Dans cette partie de la thèse, nous proposons une méthode de segmentation semi-supervisée par la classification de super-pixels. Pour ce faire, la méthode proposée SP3S (*Super-Pixel for Semi-Supervised Segmentation*) comporte deux étapes principales. Dans la première étape, la forêt aléatoire en apprentissage semi-supervisé « *co-Forest* » est formée uniquement sur 5% de la base d'image où les super-pixels sont annotés par un expert ophtalmologique. Dans la seconde étape, les super-pixels non labellisés sont impliqués pour le renforcement de l'apprentissage du classifieur afin de mieux discriminer les régions cup et disque des images rétiniennes. Pour calculer le rapport VC/D, un modèle de forme géométrique actif est utilisé pour dresser le contour du cup et disque optique.

Pour ce faire, ce chapitre est organisé comme suit : Dans la section 2, nous introduisons les travaux connexes pour la rétine fond segmentation pour le calcul du C/D. L'algorithme *co-Forest* est également présenté dans section 3. L'approche proposée sera détaillée dans la section 4. La section 5 présente les résultats expérimentaux. Enfin, dans la section 6 nous concluons par l'apport de cette approche pour le suivi de la maladie du glaucome.

3 État de l'art du domaine

Plusieurs approches ont été développées afin d'automatiser le plus possible le processus de segmentation des images rétiniennes. Dans la littérature, un certain nombre d'études ont été rapportées sur la segmentation automatique du disque optique (OD) [138–141, 146, 147] ; d'autres études ont également été signalées qui s'intéressent qu'à la segmentation du cup otique (OC) [142, 148], et enfin d'autres travaux ont relevé le défi de la segmentation commune du cup et disque optique [143, 149, 150]. Tous ces travaux peuvent être divisés en familles d'approches comme : l'approche contour, les algorithmes morphologique, les méthodes de classification des pixels, etc . . .

Klein et Walter [146] proposent un procédé basé sur l'opération morphologique afin d'extraire le disque optique ; ils ont appliqué la ligne de partage des eaux à l'image gradient. Lalonde et al. [138] quant à eux, ils ont fait l'extraction du disque optique (OD) à l'aide du Filtre de Canny, et le contour du disque optique a été déterminé par un modèle circulaire.

Un autre travail [147] propose la segmentation du disque optique à l'aide des motifs de l'opérateur local binaire (LBP), où une égalisation d'histogramme est effectuée pour améliorer la qualité de l'image d'entrée avant l'application de la méthode de LBP. D'autres opérateurs morphologiques et de filtrage sont appliqués pour filtrer les artefacts et enlever le bruit de l'image segmentée.

Dans l'approche de Merickel et al. [140], le chemin optimal correspondant à la frontière du disque a été déterminé sur la base de la fonction de coût, la texture et celle des informations a priori du contour. Xu et al. [141] ont utilisé la technique de modèle déformable qui comprend la détermination du contour du disque à travers la minimisation de la fonction d'énergie définie par l'intensité de l'image, l'image gradient et la douceur de frontière. Li et Chutatape [139] ont fait appel à l'analyse en composantes principales pour localiser le disque optique. Un modèle de forme actif modifié est proposé dans la détection de la forme du disque optique, un système de coordonnées du fond d'œil est établi afin de fournir une meilleure description des caractéristiques dans les images rétiniennes. Wong et al. [151] ont opté pour la technique level-set suivie par un ajustement géométrique en ellipse afin de lisser les contours du disque optique. Lowell et al. [152], ont également proposé un modèle adaptatif pour positionner le disque optique, et un modèle de contour déformable pour le segmenter. Ce dernier utilise un modèle elliptique global et un modèle déformable local.

Mittapalli et Kande dans [148] présentent deux méthodes de segmentation pour la détection du disque (OD) et cup (OC) optique. La première en appliquant les motifs binaires locaux [153]. La deuxième méthode pour la détection du cup optique est basée sur la méthode de regroupement SWFCM (spatially weighted fuzzy c means clustering) [154] utilisant les informations des propriétés structurelles et les niveaux de gris de la région claire.

En contrepartie des approches qui ont largement porté sur l'estimation du rapport C/D, Joshi et al. [149] proposent une évaluation plus complète du disque optique et effectuent la détection du glaucome utilisant plusieurs paramètres du disque. La méthode présentée pour l'évaluation du glaucome est présentée sous la forme de deux méthodes de segmentation pour OD et OC. Une nouvelle méthode de segmentation de l'OD est proposée, elle intègre l'information locale de l'image autour de chaque point d'intérêt dans l'espace de fonction multi-dimensionnelle afin de fournir de la robustesse contre les variations trouvées dans et autour de la région d'OD. En parallèle, une stratégie multi-étapes est mise en œuvre pour obtenir un sous-ensemble fiable de vessel-bends appelé r-bends suivis par un raccord pour dériver le contour du cup optique.

Abramoff et al. [143] ont adopté une méthode de classification pixellique en faisant appel à l'analyse de la fonctionnalité de l'algorithme du plus proche voisin. Le résultat final de leur approche comprenait la classification de chaque pixel à chaque région (classe) disque, cup, ou fond.

Dans l'approche de Cheng et al. [142], la segmentation du cup et disque optique des images rétiniennes est déterminée par la méthode de classification de super-pixels. Le processus proposé comprend : une étape de sur-segmentation par l'algorithme de classification itératif linéaire (SLIC) [144] afin de diviser l'image en super-pixels. Le processus suit une phase d'extraction de caractéristiques pour calculer les caractéristiques de chaque superpixel. Enfin une classification avec la méthode SVM est réalisé pour déterminer chaque super-pixels et d'estimer le contours cup et disque. Dans le même registre, Xu et al. [155] reprennent le même cadre d'application des super-pixels sauf qu'ils adoptent une approche non supervisée. Muramatsu et al. [156] comparent deux méthodes de classification pixelliques différentes qui emploient le modèle de contour actif (ACM), à savoir : C moyens Flou (FCM) réseau de neurones (ANN) pour la segmentation des régions cup et disque optique.

Dans le but de combler le déficit de méthodes rentables, sensibles et précises pour dépister le glaucome, ce travail traite le problème du suivi du glaucome par approche de classification de super-pixels en apprentissage semi-supervisé. Contrairement aux méthodes précédentes de segmentation des contours avec des données non marquées, notre méthode SP3S, met en place la forêt aléatoire en apprentissage semi-supervisé pour la segmentation des images rétiniennes à partir de toutes les données disponibles (étiquetées et non étiquetées). Notre précédé effectue une classification pixellique sur une représentation super-pixellique afin de réduire la complexité de l'image dans la tâche de traitement. Dans ce qui suit, les détails de notre approche sont présentés pour la mesure automatique du rapport C/D afin d'établir un diagnostic rapide, fiable et efficace du glaucome.

4 L'approche proposée

L'objectif de ce travail est la segmentation automatique du cup et disque optique pour la reconnaissance du glaucome. Pour ce faire ; nous proposons une approche basée essentiellement sur une classification semi-supervisée des super-pixels de l'image. L'intervention d'un ophtalmologue expert est importante dans l'identification des régions cups et disques d'images du fond d'œil de la rétine. L'algorithme proposé est illustré dans la Figure 21.

En premier lieu, un pré-traitement de sur-segmentation basé sur une technique de super-pixels est mise en place : l'image est divisée en petites régions homogènes en couleur, appelées super-pixels, desquelles sont extraites la composante spatiale.

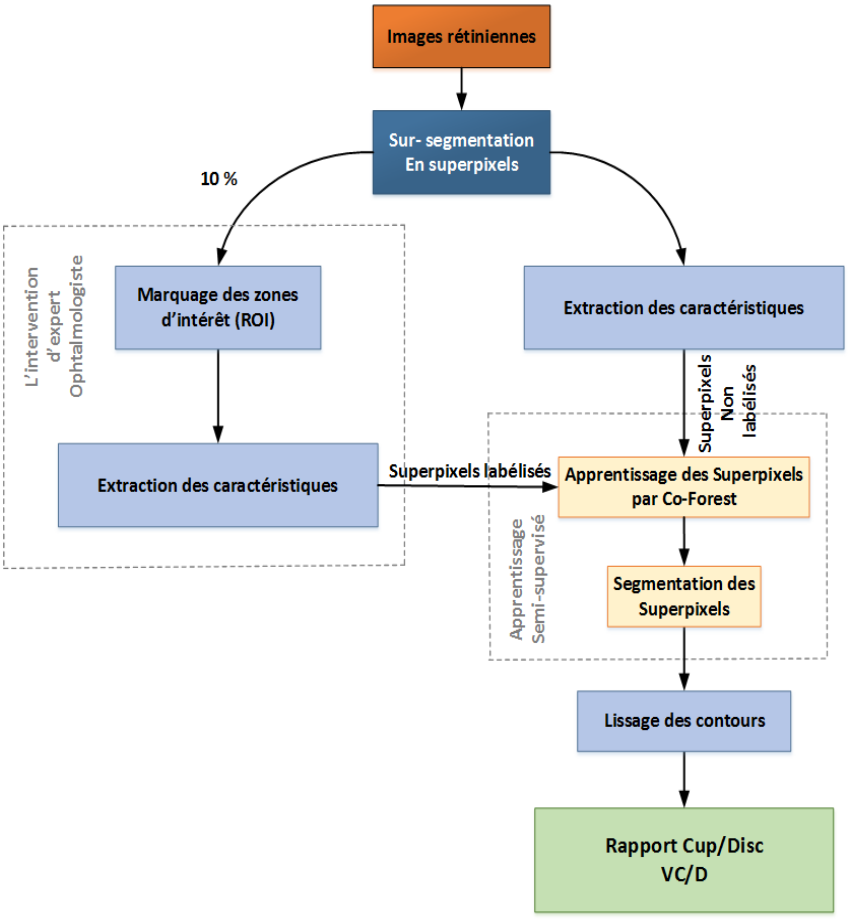


Figure 21 – Schéma représentant le processus SP3S de segmentation automatique du cup et disque optique

Par la suite, intervient l'expert pour dresser les régions d'intérêt pour les 10 % des

images sur-segmentées de la base. L'annotation se fait pour les régions : cup, disque et fond. Notre contribution est l'application de l'algorithme semi-supervisé *co-Forest* pour la classification des super-pixels afin de réaliser une segmentation semi-automatique (annotation pixellique) des régions cups et disques. Dans la phase finale, les contours cup et disque optique sont lissés par un modèle déformable (AGSM) de façon à mieux estimer leurs limites pour le calcul du rapport VC/D. Les détails de chaque blocs sont abordés dans les sections suivantes.

4.1 Sur-segmentation

La sur-segmentation est un pré-traitement qui subdivise une image en régions compactes et uniformes composées de pixels qui possèdent les mêmes propriétés du point de vue de la colorimétrie, de l'intensité, etc. Ces régions, appelées super-pixels, sont des agrégats de pixels spatialement cohérents qui préservent l'homogénéité et l'information contenue dans l'image (Figure 22). Les frontières des super-pixels ont la propriété d'adhérer aux contours des objets présents dans l'image. Sur-segmenter l'image offre la possibilité de manipuler des sous-régions cohérentes de l'image plutôt que des pixels. De ce fait, en tant que pré-traitement, la sur-segmentation réduit significativement la complexité de l'image.

L'algorithme de super-pixels

Les techniques dites de super-pixels sont de plus en plus populaires et couramment utilisés dans le domaine du traitement d'image et de la vision par ordinateur, notamment pour la segmentation [157], l'estimation de bruit gaussien [158], etc . . .. Les approches à base de super-pixels sont devenues des alternatives crédibles à l'utilisation d'une grille régulière et rigide de pixels. Plusieurs algorithmes de sur-segmentation sont recensés dans la littérature [159–161]. Ces approches permettent des traitements plus rapides, requérant une capacité mémoire moindre, en offrant la possibilité de ne calculer les descripteurs de texture qu'a l'échelle des super-pixels plutôt qu'à celui, plus dense, des pixels.

Les techniques de super-pixels reposent sur le regroupement de pixels voisins partageant des caractéristiques similaires (texture, contour, couleur, etc . . .) au sein de régions polygonales. Par conséquent, les techniques de super-pixels produisent une image sur-segmentée représentant une carte compacte du contenu initial de l'image [144]. Dans ce travail, nous utilisons l'algorithme SLIC (Simple Linear Iterative Clustering) introduit par Achanta et al. [144], il a été démontré que l'algorithme SLIC surpasse tous les algorithmes concurrents tant qu'en point de vue de la qualité de la segmentation qu'en point de vue de la vitesse d'exécution [144]. Il produit des super-pixels compacts et uniformes en couleur sans requérir l'ajustement de multiples paramètres.

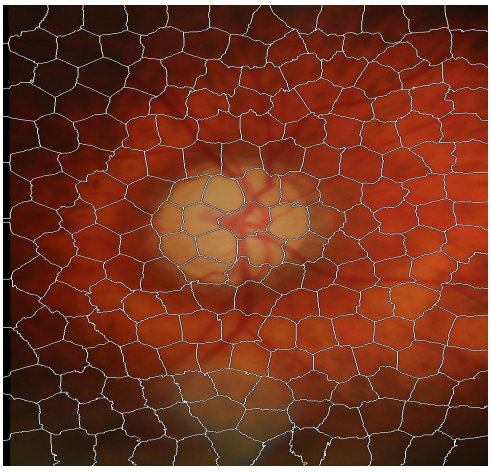


Figure 22 – Sur-segmentation obtenue à l'aide de l'algorithme SLIC

Cette méthode récente, permet de construire des super-pixels réguliers en surface. Tout d'abord, les centres des super-pixels sont initialisés sur une grille régulière, espacée de S pixels avec $S = \sqrt{\frac{N}{K}}$, où N est le nombre total de pixels dans l'image et K le nombre de super-pixels souhaités. Ils peuvent être éventuellement déplacés afin d'éviter de se trouver sur un contour de l'image. Cette méthode est itérative et comprend deux étapes :

1. L'assignation des pixels à un centre C_k suivant un critère d'appartenance,
2. La mise à jour des centres.

Cette approche tente de minimiser dans l'étape 1, le critère d'appartenance correspondant à une distance entre C_k et le pixel courant p définie par :

$$D_s(C_k; p) = d_{lab}(C_k; p) + \frac{m}{S} d_{xy}(C_k; p) \tag{2.1}$$

où d_{lab} est la distance calorimétrique et d_{xy} est la différence entre les positions dans l'image courante, telles que :

$$d_{lab}(C_k; p) = \sqrt{(l_k - l_i)^2 + (a_k - a_i)^2 + (b_k - b_i)^2}$$

$$d_{xy}(C_k; p) = \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2}$$

Ainsi, les paramètres sont le nombre approximatif de super-pixels K et leur compacité m. D'après Achanta et al. [144], $m \in [0 - 20]$ lorsque nous travaillons sur l'espace couleur CIELAB. Le terme m joue un rôle de pondération entre la couleur et la position. Quand $c = 0$ les super-pixels peuvent être très souples et adhèrent aux contours de l'image et quand $m = 20$ ils se rapprochent d'une forme régulière. Les auteurs proposent de fixer m à 10 car cette valeur permet d'obtenir des performances supérieures à celles réalisées par Felzenszwalb et Huttenlocher [160]. Pour chaque centre est défini une zone de recherche d'appartenance de taille (2S × 2S) et centrée sur C_k. Seuls les pixels p appartenant à cette zone sont parcourus comme le montre la Figure 23.

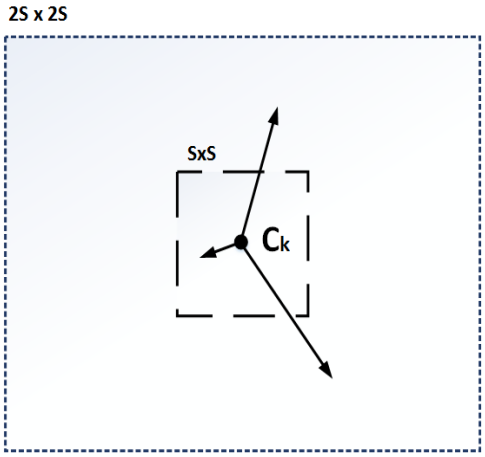


Figure 23 – Zone de recherche de pixels similaires au centre C_k de référence

4.2 Méthodes d'extraction des caractéristiques

Les espaces couleurs

Différents espaces couleurs ont été utilisés dans la segmentation par classification pixellique, mais beaucoup d'entre eux partagent des caractéristiques similaires. Par conséquent, dans ce travail, nous nous limitons à quatre espaces couleurs qui sont les plus représentatifs et couramment utilisés dans le domaine de traitement d'image [162].

Types de variables	**Forme**
RGB	$R(i,j)$ $G(i,j)$ $B(i,j)$
LUV	$L = 116(\frac{Y}{Y_n})^{1/3} - 16 \quad \text{Si } \frac{Y}{Y_n} > 0.008856$ $= 903.3(\frac{Y}{Y_n}) \quad \text{Si } \frac{Y}{Y_n} \leq 0.008856$ $U = 13L(U' - U'_n)$ $V = 13L(V - V'_n)$
HSV	$\text{H} = \frac{G-B}{(Max-Min)} \quad \text{Si} \quad \text{R} = \text{Max}$ $= \frac{B-R}{(Max-Min)}+2 \quad \text{Si} \quad \text{G} = \text{Max}$ $= \frac{R-G}{(Max-Min)}+4 \quad \text{Si} \quad \text{B} = \text{Max}$ $\text{S} = \frac{Max(R,G,B)-Min(R,G,B)}{Max(R,G,B)}$ $V = Max(R,G,B)$
YUV	$Y = 0.2989R + 0.5866G + 0.1145B$ $U = 0.5647(B-Y) = -0.1687R - 0.3312G + 0.5B$ $V = 0.7132(R-Y) = 0.5R - 0.4183G - 0.0817B$

Table 11 – Tableau des paramètres de caractérisation

L'espace RGB est un système additif qui décompose les couleurs en trois quantités des trois couleurs primaires : le rouge, le vert et le bleu (Table 11). C'est le système le plus employé dans les images couleurs et les moniteurs. Le modèle RVB utilise le système de coordonnées cartésiennes. Le point (1, 1, 1) représente le blanc, le point (0, 0, 0) représente le noir et la diagonale représente les niveaux de gris.

L'intention de l'espace couleur LUV consiste à produire un espace de couleur plus linéaire que les autres espaces couleurs existants (Table 11). Linéaire perceptuelle signifie qu'un changement de la même quantité dans une valeur de couleur doit produire une variation de la même importance visuelle.

L'espace HSV est un modèle de représentation dit "naturel", c'est-à-dire proche de la perception physiologique de la couleur par l'œil humain. Il consiste à décomposer la couleur selon des critères physiologiques (Table 11).

L'espace YUV est destiné principalement à la vidéo analogique, ce modèle de représentation est utilisé dans les standards vidéo PAL et SECAM. La luminance est représentée par Y, tandis que les chrominances U et V sont issues de la transformation de l'espace RGB (Table 11).

Analyse spatiale

La localisation spatiale d'un pixel peut fournir des informations de l'emplacement et de l'orientation par rapport au centre d'intérêt. Dans [163], Giannakeas et al. ont utilisé la distance Euclidienne et la distance Manhattan pour mesurer la similarité entre la position de pixel candidat et le centre d'intérêt (figure 24). Plus la distance est minime plus le pixel candidat à une grande probabilité d'appartenir à la même classe du centre.

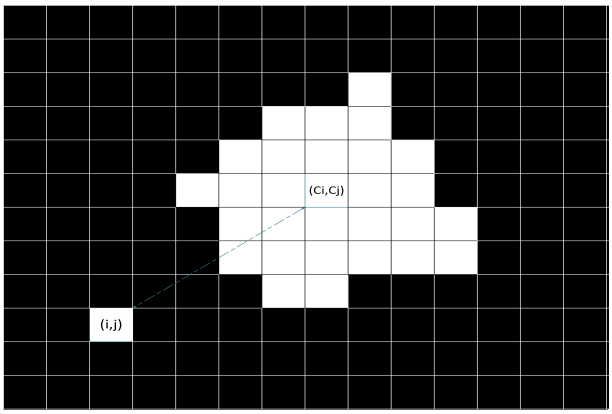


Figure 24 – Procédure d'extraction des caractéristiques spatiales

4.3 La Forêt Aléatoire en apprentissage semi-supervisé "*co-Forest*"

L'algorithme *co-Forest* repose sur le paradigme du *co-Training* [5], où deux classifieurs sont d'abord formés à partir de L, puis chacun d'eux choisit les exemples les plus confiants en U de son point de vue. Il met par la suite à jour les autres classifieurs avec ces exemples nouvellement étiquetés. Un des aspects les plus importants dans *co-Training* est d'estimer la confiance d'un exemple donné non étiqueté.

Dans *co-Training* standard, l'estimation de la confiance profite directement à partir de deux sous-ensembles d'attributs suffisants et redondants, où la confiance d'étiquetage d'un classifieur pourrait être considérée comme sa confiance pour un exemple non étiqueté. Lorsque la condition des deux sous-attributs suffisants et redondants n'est pas présente, la validation croisée est appliquée à chaque itération d'apprentissage afin d'estimer la confiance pour les données non étiquetées [14]. L'inefficace estimation de la confiance réduit considérablement l'applicabilité de l'étendue de l'algorithme *co-Training* dans des applications telles que le diagnostic assisté par ordinateur.

Cependant, si un ensemble de classifieurs N, qui est désigné par H^*, est utilisé dans le *co-Training* au lieu de deux classifieurs, la confiance peut être estimée de manière efficace. Lors de la détermination des exemples les plus confiants étiquetés pour un classifieur de l'ensemble de $H_i (i = 1, \ldots, N)$, tous les classifieurs sont utilisés sauf h_i. Ces classifieurs forment un nouvel ensemble, qui est appelé l'ensemble de concomitance de H^*, noté par H_i. Notons que H_i diffère de H^* seulement par l'absence de h_i. La confiance pour un exemple sans étiquette peut être simplement estimée par le degré d'accords sur l'étiquetage, c'est à dire le nombre de classifieurs qui sont d'accord sur l'étiquette assignée par H_i. En utilisant la méthode *Co-forest* [18], l'algorithme construit un ensemble de classifieurs sur L, puis affine chaque classifieur avec des exemples nouvellement étiquetés choisis par son ensemble de concomitance.

Le fonctionnement de *Co-forest* peut se résumer par les étapes suivantes :

Étape 1 *Co-forest* lance l'apprentissage des H^* sur des bootstrap [1] de L Figure 25.

1. Un échantillon bootstrap L est, par exemple, obtenu en tirant aléatoirement n observations avec remise dans l'échantillon d'apprentissage L_n, chaque observation ayant une probabilité 1/n d'être tirée.

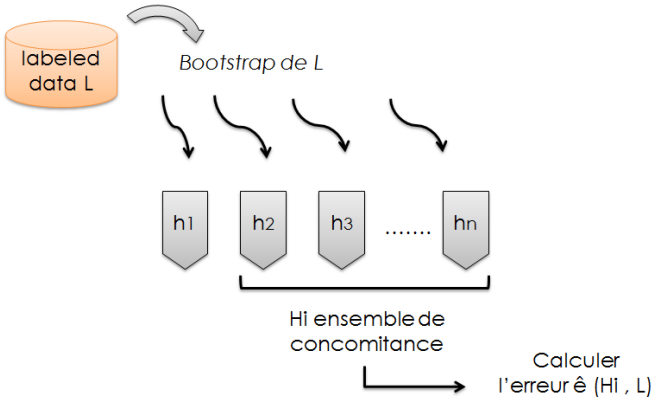


Figure 25 – Apprentissage des arbres sur les données labellisées L

Étape 2 l'ensemble de concomitance examine chaque exemple de U

Si le nombre de votant sur une étiquette de classe pour xu est d'accord $> \theta$ **Alors** xu est labellisé et copié dans un nouveau ensemble L'

Remarque : Nous pourrons être confronté à une situation où $L' \geq U$ cela affecte les performances de h_i

Solution : introduire un poids par la prédiction de confidence par l'ensemble de concomitance (Figure 26).

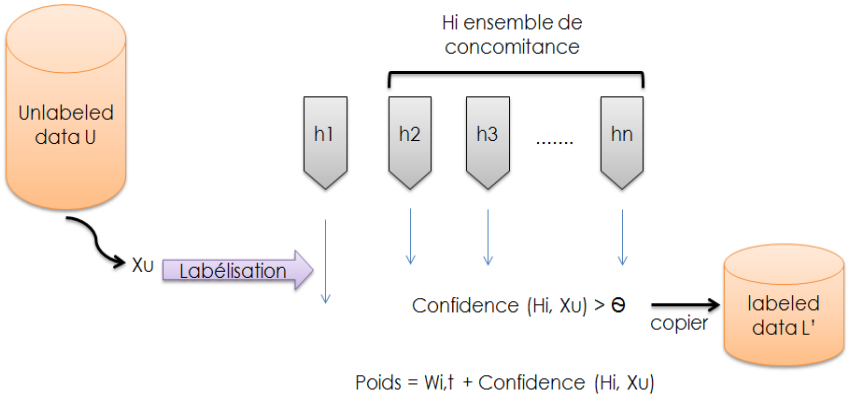


Figure 26 – Labellisation des données non labellisées U par l'ensemble de concomitance

Étape 3 Chaque arbre aléatoire est raffiné avec des exemples nouvellement marqués $L \cup L'$ ensuite sélectionnés par son ensemble de concomitance sous la condition suivante :

$$e_{i,t}.W_{i,t} < e_{i,t-1}.W_{i,t-1}$$

Où : $W = \sum w_{ij}$ et w_{ij} : la confidence prédictive de H_i sur x_i dans L' Figure 27.

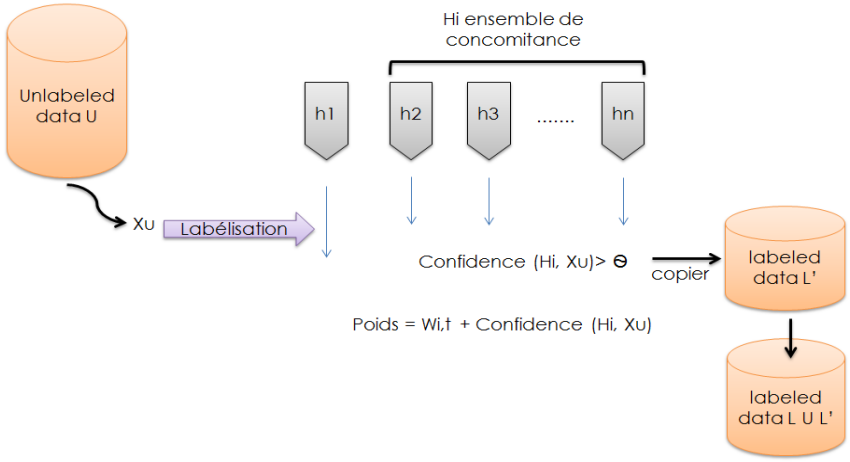


Figure 27 – Ré-apprentisage par les exemples nouvellement marqués L ⋓ L'

Pour que le succès de cette méthode d'ensemble soit présent, il faut que deux conditions soient satisfaites :

– Chaque prédicteur individuel doit être relativement bon.
– Les prédicteurs individuels doivent être différents les uns des autres.

En plus simple, il faut que les prédicteurs individuels soient de bons classifieurs. Et là où un prédicteur se trompe, les autres doivent prendre le relais.

Afin de maintenir la diversité dans *Co-forest*, l'idée est d'appliquer les forêts aléatoires. Elles permettent d'injecter l'aléatoire dans son principe d'apprentissage, pour maintenir cette condition. Les auteurs de *Co-forest* ont fixé un seuil pour la labéllisation des *U* où seulement les *U* dont *le total de poids* $< \frac{e_{i,t-1}.W_{i,t-1}}{e_{i,t}}$ seront sélectionnés (voir Figure 28).

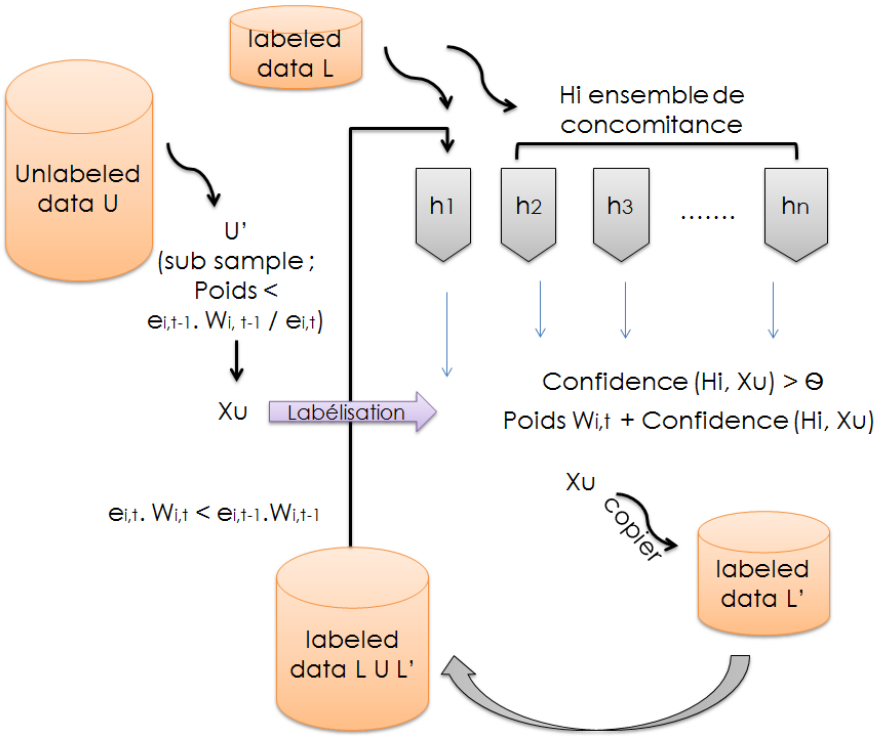


Figure 28 – Schéma de principe de l'algorithme *Co-forest*

4.4 Modèle géométrique déformable

Nous avons mis en œuvre un algorithme de modèle déformable basé sur la technique AGSM (Active Geometric Shape Model) Wang et Boyer [164]. Les modèles déformables sont des méthodes permettant de localiser les frontières d'un objet qui peut être représenté par une équation paramétrique. L'idée est de modeler de manière itérative les paramètres de forme de l'objet selon le champ de force, et ce afin de trouver les paramètres optimaux. Pour ajuster un modèle déformable de type snakes [165], ASM (Active Shape Model) [166], etc . . ., les points du modèle vont être déplacés le long du champ de force dans chaque itération. Un bon champ de force doit respecter les gradients de l'image et être lissé pour assurer une large gamme de capture. Dans le modèle AGSM de Wang et Boyer, ces derniers utilisent le Gradient Vector Flow (GVF) (eq. 2.2) pour minimiser une énergie fonctionnelle (eq. 2.3) (Où f représente l'image lissée).

$$\mathbf{v}(x,y) = [u(x,y), v(x,y)] \tag{2.2}$$

$$\xi = \iint (\mu(u_x^2 + u_y^2 + v_x^2 + u_y^2) + \parallel \nabla f \parallel^2 \parallel \mathbf{v} - \nabla f \parallel^2) dxdy \tag{2.3}$$

Où $f(x,y)$ est le négatif de l'énergie externe dérivée de $I(x,y)$ l'image en niveaux de gris, qui est considérée comme une fonction de variables continues de position (X,Y) (eg.2.4). Ce champ de force peut être résolu en utilisant le calcul des variations, et mesuré par itération numérique [167].

$$f(x,y) = -E_{ext}(x,y) = -G_\sigma(x,y) * I(x,y) \tag{2.4}$$

Le principe de base d'AGSM est d'associer chaque paramètres : de position, de taille, les paramètres de forme et ou les paramètres d'orientation, avec une force, puis régler le paramètre en fonction de cette force ou orientation.

5 Base de données

RIM-ONE Release 3 est la troisième version de la base de données RIM-ONE [168]. C'est une base de données d'image rétinienne du fond d'œil, qui porte exclusivement sur la segmentation du nerf optique (ONH). La base de données est composée de 159 images du fond d'œil de taille 1424 x 1072, ces images sont segmentées de manière manuelle par deux différents experts en ophtalmologie pour générer les images réalité terrain. La segmentation moyenne est également disponible comme segmentation de référence.

Ces images du fond d'œil ont été capturées à partir de différentes sources médicales dans trois hôpitaux espagnols (Hospital Universitario de Canarias, Hôpital Clinico San Carlos et l'hôpital Universitario Miguel Servet). La compilation d'images provenant de différentes sources médicales garantissent l'acquisition d'un ensemble d'images représentant et hétérogène. Les images sont classées en différentes catégories :

- 85 images de patients non-glaucomateux,
- 29 images de patients avec un glaucome modéré,
- 6 images de patients atteints de glaucome sévère,
- 39 images de patients glaucomateux.

6 Résultats et expérimentations

6.1 Expérimentations

Dans la présente étude, afin de réaliser la sur-segmentation des images rétiniennes, nous fixons la compacité des super-pixels par l'algorithme SLIC à $m = 10$. Il a été démontré

par Achanta et al. [144], que cette valeur offre un bon équilibre entre la similitude des couleurs et la proximité spatiale pour le traitement en super-pixels des images.

La première série d'expérimentations a été effectuée pour évaluer la performance selon différents nombre de super-pixels K allant de 500, 1000, 1500 et 2000. Le tableau (Table 12) résume les taux de classification pour chaque K. Les résultats montrent que $K = 500$ est remarquable pour cette application malgré qu'il y a une légère amélioration lorsque k atteint 1000 à 2000. Sachant qu'un petit nombre de K nécessite moins de temps de calculs (i7-4790 CPU @ 3.60 GHz, 16Go RAM, MATLAB R2011b), par conséquent nous adoptons $K = 500$ avec un filtre de 15×15 lors des essais ultérieurs.

Nombre de super-pixels K	500	1000	1500	2000
Taux de classification	90,16	90,81	90,74	90,85
Temps d'exécution (sec)	1207,37	2813,57	3524,40	4812,87

Table 12 – Tableau de performances selon différents nombre de super-pixels K

Pour la suite de notre expérimentation, nous avons sélectionné 10% de la base de données (15 images) pour réaliser l'apprentissage. L'expert ophtalmologiste interviendra dans l'étiquetage de ces quinze images par marquage des zones d'intérêt (ROI), permettant ainsi une meilleure perception des régions cup et disc. Dans la partie d'apprentissage de la forêt aléatoire semi-supervisée, un nombre d'arbre égale à 100 étant choisi. L'évaluation est réalisée par une validation croisée égale à 5. Les détails des paramètres d'expérimentation sont résumés dans le tableau (Table 13).

Base étiquetée	15 images
Base d'apprentissage	64 images
Base de test	80 images
Degré de confiance	75%
Nombre d'arbres	100
Validation croisée	5

Table 13 – Paramètres de classification

6.2 Résultats

Une analyse quantitative est effectuée pour évaluer la performance globale de notre méthode de segmentation. Cette évaluation est basée sur la similitude entre les régions et les contours détectés par notre méthode en comparaison avec ceux des experts individuels et leur moyenne. Pour estimer le chevauchement entre les régions segmentées et celles vérité terrain, les mesures de précision et du rappel des super-pixels sont calculées.

$$Precision = \frac{VP}{VP + FP} \qquad Rappel = \frac{VP}{VP + FN}$$

Avec : VP : Vrai Positif : nombre de super-pixels positifs classés positifs. FP : Faux Positif : nombre de super-pixels négatifs classés positifs. FN : Faux Négatif : nombre de super-pixels positifs classés négatifs.

Pour mieux apprécier la qualité des résultats de notre méthode, nous faisons appel à une mesure de performance appelée F-score (F) qui est la moyenne harmonique de la précision et du rappel. Elle est définie comme suit :

$$F = 2 \ \frac{Precision \ . \ Rappel}{Precision + Rappel}$$

La valeur de F-score se situe entre 0-1, une valeur F-score élevée nous renseigne sur la pertinence de la technique.

	Cup optique			Disque optique		
	Expert 1	**Expert 2**	**Moyenne**	**Expert 1**	**Expert 2**	**Moyenne**
Cas Glaucomateux	0,8475	0,8563	0,8643	0,8982	0,8965	0,8989
Cas Suspects	0,8400	0,8614	0,8660	0,8756	0,8898	0,8909
Cas Normaux	0,8871	0,9033	0,9282	0,9149	0,9169	0,9178

Table 14 – Évaluation F-score des régions cups et disques optiques

Afin d'évaluer la performance globale sur l'ensemble des images, la moyenne F-score est calculée sur les 3 catégories d'images : cas normal, cas glaucome et cas suspect (Table 14). De ce tableau, nous pouvons observer la mesure F score de notre méthode SP3S pour les régions disque et cup optique en comparaison avec la segmentation des deux experts et leur moyenne.

Une des raisons du grand intérêt porté à cette maladie du Glaucome réside dans sa multiformité où chaque patient développe différentes formes de glaucomes à différents stades d'évolution. Les traitements sont donc adaptés à chaque cas et il n'existe pas de traitement standard. Chaque expert ophtalmologiste jugera de manière qualitative la taille et forme du cup et disque optique pour évaluer la progression ou stade du glaucome.

La comparaison entre les segmentations expert 1 et 2 ainsi que leur moyenne (Table 14), montre une correspondance plus grande (F score élevé) entre les contours de segmentation de notre approche SP3S avec celle de la moyenne des experts ophtalmologistes. En effet, ceci se reflète dans les cas dit difficile à diagnostiquer (Figure 29), où nous sommes confrontés à des marquages de régions différent et a des visions propres à chaque expert ophtalmologiste. Établir une moyenne de ces segmentations permet de faire un compromis entre les différents avis pour une mesure qui approche la réalité. Par notre méthode de segmentation semi-automatique, notre algorithme permet de prendre en considération les différents avis d'experts ophtalmologistes et fournir le meilleur consensus par rapport à chacune de ces estimations.

L'application du modèle déformable AGSM a permis le tracé de courbe déformable qui épouse la forme des régions cup et disque. Cette capacité de se mouvoir automatiquement vers les données recherchées est représentée géométriquement et est guidée par une loi d'évolution régissant ses déformations. Cette dynamique est basée sur la notion d'énergie interne et externe, le but étant de minimiser l'énergie totale présente le long de la courbe. Des contraintes permettent de conserver une courbe lisse avec des points équidistants tout en laissant un certain champ libre pour les déformations. L'énergie interne correspond à la morphologie et aux caractéristiques de la courbe (courbure, longueur, etc.). L'énergie externe provient de l'image, les critères sont variables (présence de bords marqués, bruit, etc.). En effet, le modèle AGSM ellipsoïdal [164] a permis de dresser la segmentation en superpixels par classification semi-supervisée en des formes approchantes à celles dessinées par les experts, réalisant un contour quasi-identique à celui de la moyenne des experts.

6.3 Discussion

Pour plus d'exhaustivité, nous sélectionnons 6 images dites difficiles a diagnostiquer dans le cas des patients glaucomateux, suspects et normaux (Figure 29). Nous discutons les performances et la qualité de segmentation de notre approche en mode semi-supervisé par l'algorithme *co-Forest* avec les marquages des deux experts et leur moyenne.

Six images représentant deux cas de chaque catégorie : glaucome (a,b), normal (c,d), suspect (e,f) sont affichées dans la Figure 29). Elles sont décrites comme difficiles a diagnostiquer par les experts ophtalmologistes à cause de la multiformité du glaucome dans les cas suspects et glaucomateux, aussi par leur faible qualité visuelle (qualité de l'image). En effet, nous pouvons clairement noter qu'à partir des images (a) cas glauco-

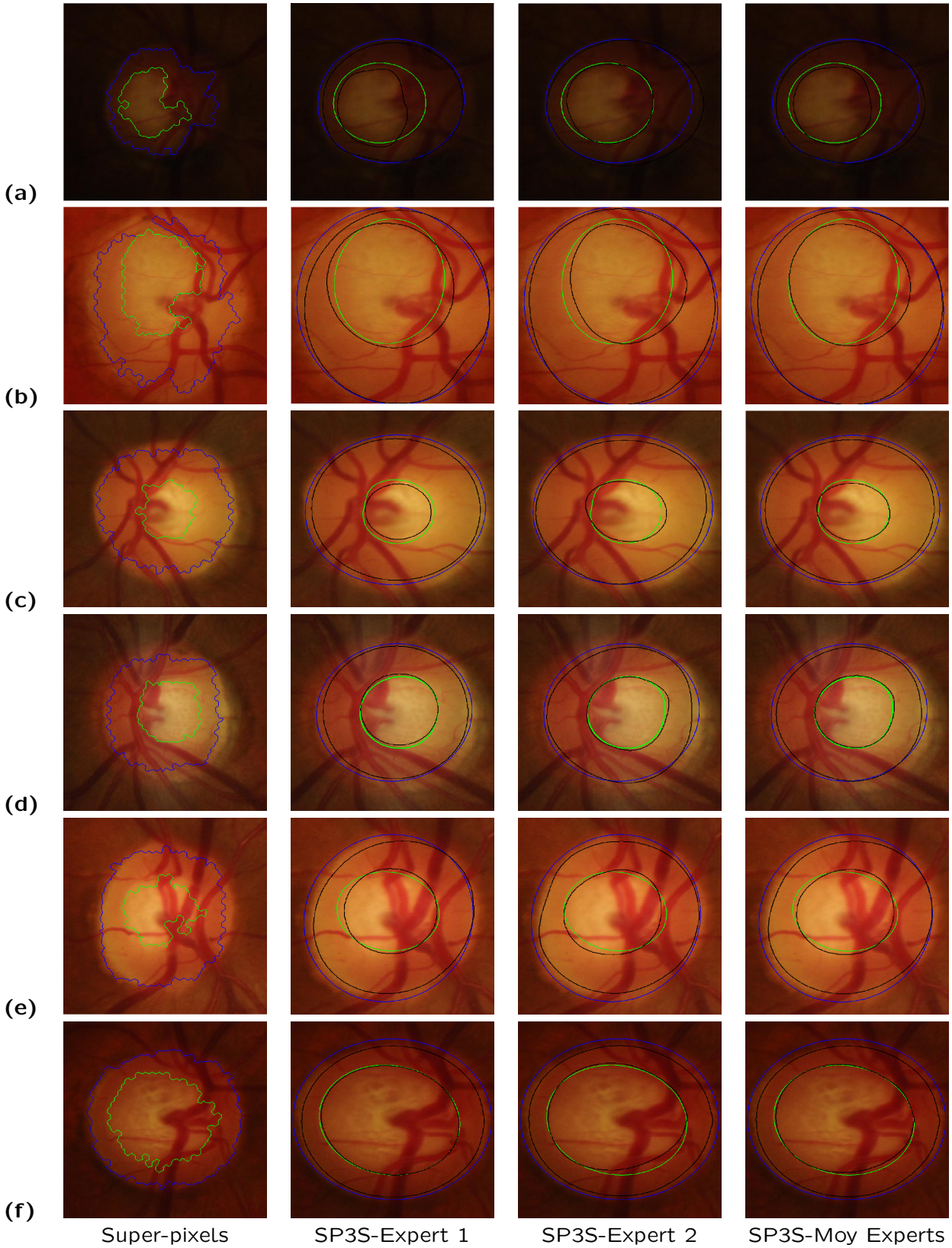


Figure 29 – Exemple de segmentation semi-supervisée des cas rétiniens glaucomateux (a,b), normaux (c,d), suspects (e,f). La segmentation des experts est représentée par un contour noir, la segmentation par SP3S avec un contour vert (cup) et un contour bleu (disque).

mateux et (b) cas suspect, une divergence du point de vue des experts dans le marquage des régions disque et cup optique.

La mesure F-score calculée dans les 3 catégories des six images indique que notre approche SP3S de segmentation semi-automatique par apprentissage semi supervisé réalise des performances très satisfaisantes en comparaison aux annotations des experts.

Pour discuter ces résultats, nous prenons comme exemple l'image (a) cas glaucomateux d'une qualité d'image très sombre où la visibilité n'est pas assez nette. Notre méthode a permis une concordance de marquage avec l'expert 1 sur la région disque optique avec un score égale à 0,9636. Pour la segmentation du cup optique elle coïncide mieux avec celle de l'expert 2 par un score atteignant les 0,9575. Et ainsi le meilleur compromis a été établi par la moyenne des experts avec pour le cup 0,95215 et le disque 0,9588. Nous pouvons déduire de cet exemple que SP3S se rapproche le mieux des avis moyennés des deux expert.

Dans sa globalité, ce travail, nous ouvre une multitude de pistes de recherche pour la segmentation automatique des images. L'idée d'extrapoler la segmentation de région par classification super-pixellique au contexte d'apprentissage semi-supervisé par l'approche *co-Forest*, nous a permis d'exploiter les données non-étiquetées dans la mise en place du modèle de prédiction ensembliste. En ce sens, les données non-étiquetées ont renforcé la reconnaissance des régions d'intérêts pour un rapport Cup/Disque calculé approximant celui des experts ophtalmologistes.

Une comparaison des rapports VC/D entre la moyenne de marquage des experts et celle de notre approche SP3S sur les trois catégories d'images : Glaucome, suspect et normal sont représentées dans les Figures 30, 31, 32. Les différences de mesures des ratios VC/D de la moyenne des ophtalmologistes en comparaison avec notre méthode de segmentation par l'algorithme *co-Forest* tendent à montrer de petites variations dans les 3 cas inclus, ayant une mesure presque semblable à celle de la moyenne des experts.

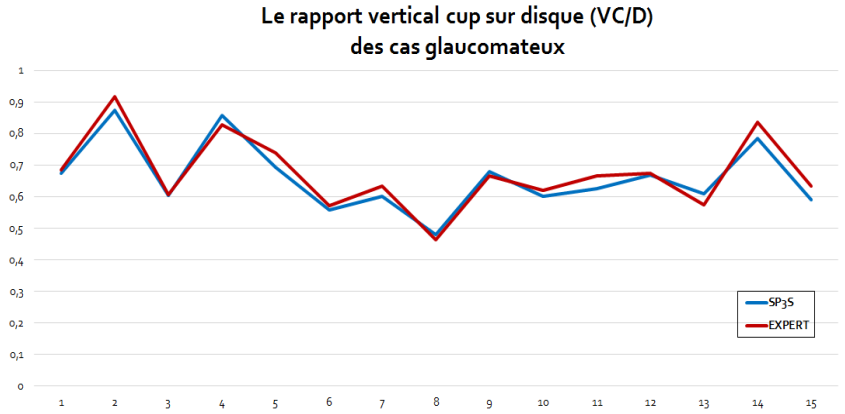


Figure 30 – Comparaison du rapport Cup/Disc mesuré par un ophtalmologiste et notre approche proposée.

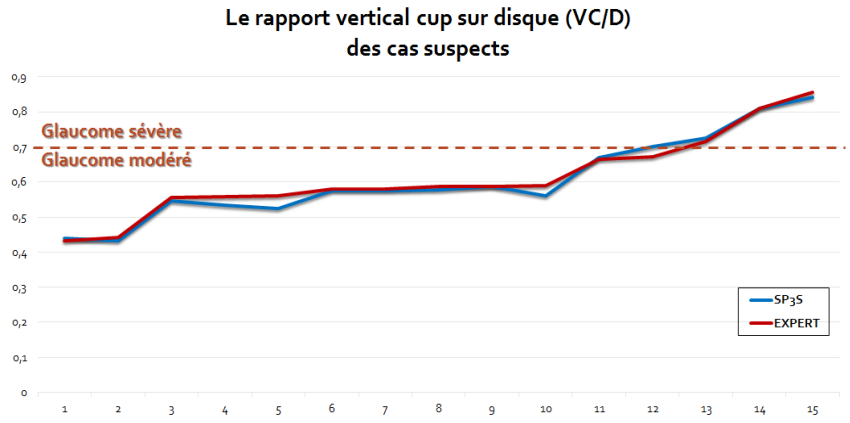


Figure 31 – Comparaison du rapport Cup/Disc mesuré par un ophtalmologiste et notre approche proposée. Les cas 1-12 sont des cas de glaucome modéré, et 13-15 cas sont des cas de glaucome sévère.

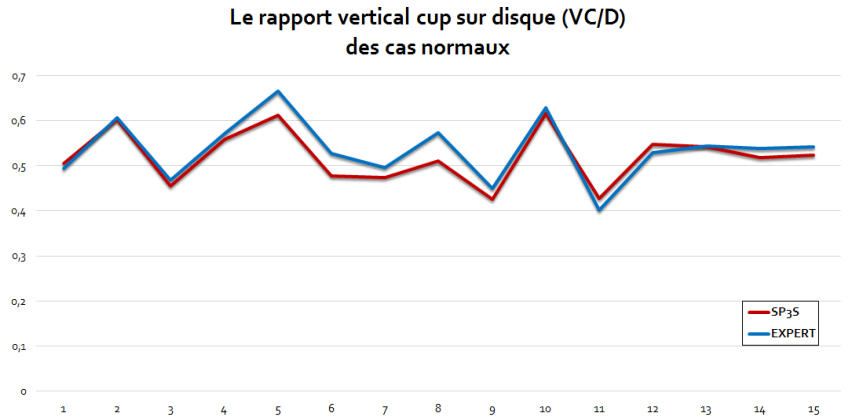


Figure 32 – Comparaison du rapport Cup/Disc mesuré par un ophtalmologiste et notre approche proposée.

7 Conclusion

Le seul moyen de prévenir le glaucome est de le dépister précocement, pour pouvoir le traiter le plus tôt possible. Plus le diagnostic est précoce, plus les chances de traitement et de prévention de la cécité sont efficaces. Cette maladie peut être évaluée en surveillant la pression intra-oculaire (PIO), par le champ visuel et l'aspect de la papille optique (rapport Cup/Disque).

Dans ce travail, nous avons présenté une méthode de segmentation des régions cups et disques SP3S (*Super-Pixels for Semi-Supervised Segmentation*) et ce afin de réaliser un suivi médical concret de la maladie du glaucome. De ce fait, nous avons proposé une méthode de segmentation semi-automatique des régions Cup et Disque des images rétiniennes par classification super-pixellique en apprentissage semi-supervisé. L'objectif visé est d'impliquer l'expert à l'apprentissage de notre modèle pour une meilleure discrimination des régions d'intérêt.

Les résultats obtenus sont très prometteurs et encourageants, indiquant une grande capacité de reconnaissance et de segmentation des régions ciblées. Notre processus peut

être utilisé comme une étape d'examen préliminaire dans le diagnostic automatique du glaucome en particulier dans les programmes de dépistage.

Enfin, même si nous considérons l'approche SP3S comme un pas en avant dans l'étude de la segmentation semi-automatique, nous pouvons objecter le faible nombre d'images disponibles pour chacun des cas étudiés. Une étude statistique représentative de la performance humaine pour cette tâche nécessiterait un échantillonnage beaucoup plus large de la population, en faisant intervenir des facteurs comme ceux environnementaux, l'âge ou l'hérédité. Notre méthode pourrait être appliquée directement dans un tel cadre sans augmenter le temps de calcul, puisqu'elle fait appel à une sur-segmentation par super-pixels qui accélèrent le processus classique initialement réalisé par des pixels.

CHAPITRE 3

Nouvelle approche d'apprentissage semi-supervisé pour les données à grande dimension

1 Objectifs

L'apprentissage semi-supervisé est un des champs de recherche les plus intéressants pour une évolution du domaine de l'apprentissage artificiel au-delà du cadre supervisé des données. Il peut également aider l'apprentissage humain (exemple : l'aide au diagnostic médical) qui fonctionne fréquemment en mode supervisé. Dans ce chapitre nous proposons une nouvelle approche en apprentissage semi-supervisé appelée *Optim co-Forest*.

L'algorithme *Optim co-Forest* combine à la fois le ré-échantillonnage de données (Bagging [37]) et est secondé de deux stratégies de sélection : la première implique la sélection de sous ensembles de paramètres aléatoires inspirée de l'approche *Rel-RASCO* [6] afin de générer l'ensemble des classifieurs suivant le principe de *co-Forest* [18]. Cela permettra de préserver la diversité des classifieurs et ainsi donc avoir une meilleure discrimination des différentes classes. La seconde stratégie est une extension de la mesure d'importance des *forêts aléatoires RF* [51], elle fait appel à un assortiment d'ensemble de données marquées et non marquées, pour mesurer la pertinence des variables. Un classement de toutes les variables est finalement réalisé par rapport à leurs pertinences dans tous les classifieurs semi- supervisés.

Des expérimentations sur des ensembles de données de l'UCI [84] vérifient l'efficacité de *Optim co-Forest*. Des applications sur des ensembles de données à grande dimension confirment la puissance de notre méthode dans le passage à l'échelle grâce à ses approches de sélections adoptées.

Ce chapitre introduit l'algorithme *Optim co-Forest*, ce qui nous mène à organiser ce papier comme suit : un état de l'art des techniques d'apprentissage ensemblistes en semi-supervisé. Une revue de quelques méthodes d'ensemble dans le domaine semi-supervisé est effectuée. Nous présentons en détail l'évolution de ces dernières ainsi que leurs avantages et leurs limites. Nous exposons ensuite dans la section 3, le processus général de notre approche proposée et ses différentes étapes. Nous validons notre algorithme et les choix que nous avons réalisés par une phase d'expérimentation. Nous montrons la capacité de notre méthode à améliorer les performances de classification par une comparaison avec les méthodes représentatives de la littérature. Finalement, nous terminerons par une conclusion présentant une synthèse des contributions apportées ainsi que les pistes définissant des perspectives possibles pour de futurs travaux et aussi les difficultés rencontrées lors de la réalisation de ce travail.

2 Les techniques d'apprentissage ensemblistes en semi-supervisé

Les méthodes d'apprentissage semi-supervisé utilisent des données non marquées, afin de modifier ou hiérarchiser les hypothèses obtenues à partir de données étiquetées seules.

L'auto-apprentissage "*Self-Training*" est une technique couramment utilisée en apprentissage semi-supervisé. Dans l'auto-apprentissage, un classifieur est d'abord généré avec la petite quantité de données étiquetées. Il est ensuite utilisé pour classer les données non étiquetées. Les points non labellisés les plus confiants, avec leurs étiquettes prédites, sont ajoutés à l'ensemble d'apprentissage. Le classifieur est reconstruit sur l'ensemble de ces nouvelles données. A noter que le classifieur utilise ses propres prédictions pour le ré-apprentissage. La procédure est également appelée auto-enseignement (self-learning). Un inconvénient majeur de cette approche réside dans le calcul d'erreur de classification, qui est de ce fait renforcé tout au long du processus de ré-apprentissage.

D'autres approches ont été élaborées par la suite, elles tentent de remédier à cet inconvénient en «éliminant» les points non marqués si la confiance de prédiction descend en dessous d'un seuil. L'algorithme du *co-Training* (Blum et Mitchell [5]) (Mitchell [169]) étant le premier à proposer ce procedé, il suppose que :

– Les caractéristiques peuvent être divisées en deux ensembles ;
– Chaque sous-ensemble de variables est suffisant pour former un bon classifieur ;
– Les deux ensembles sont conditionnellement indépendants en fonction de la classe.

Initialement deux classifieurs sont générés séparément avec les données étiquetées sur les deux sous-ensembles respectivement. Chaque classifieur classe par la suite les données non marquées, et « alimente » l'autre classifieur avec les exemples les plus confiants nouvellement labellisés. Chaque apprenant est reconstruit avec les exemples d'apprentissage supplémentaires fournis par l'autre classifieur, et par la suite le processus est répété.

En *co-Training*, les deux classifieurs (ou hypothèses) doivent être d'accord sur les données non étiquetées ainsi que les données étiquetées.

Nous avons besoin d'hypothèses qui assurent que les sous-ensembles de caractéristiques sont suffisamment bons, de façon a ce que nous pouvons faire confiance aux étiquettes prédites par chaque apprenant sur U. De ce fait, nous avons besoin que les sous-ensembles de variables soient conditionnellement indépendants afin que les exemples de données de haute confiance d'un classifieur soient des échantillons iid (Independent and identically distributed) pour l'autre classifieur. Zhu [17] schématise l'hypothèse dans la Figure 33. L'algorithme *co-Training* présente des conditions strictes sur l'ensemble de vue des données. On peut se demander si ces conditions peuvent être réalisées dans des applications réelles. Goldman et Zhou [14] utilisent deux apprenants de type différent mais sur le même ensemble de caractéristiques. Ils font appels aux exemples de données U de haute confiance identifiés avec un ensemble de tests statistiques pour relancer l'apprentissage de l'autre et vice versa.

Chawla et Karakoulas [170] ont réalisé des études empiriques sur cette version de *co-Training* et les ont comparé avec plusieurs autres méthodes, en particulier pour le cas où les données étiquetées et non étiquetées ne suivent pas la même distribution. Plus tard, Zhou et Goldman [15] ont proposé une approche à une seule vue avec multiples classifieurs appelée *Democratic Co-Learning*.

Zhou et Li [16] proposent *tri-Training* qui utilise trois classifeurs. Si deux d'entre eux sont d'accord sur la classification d'un point non marqué, ce dernier est utilisé pour l'apprentissage du troisième classifieur. Cette approche évite ainsi la nécessité de mesurer explicitement la confiance de l'étiquette de chaque apprenant. Il peut être appliqué à des

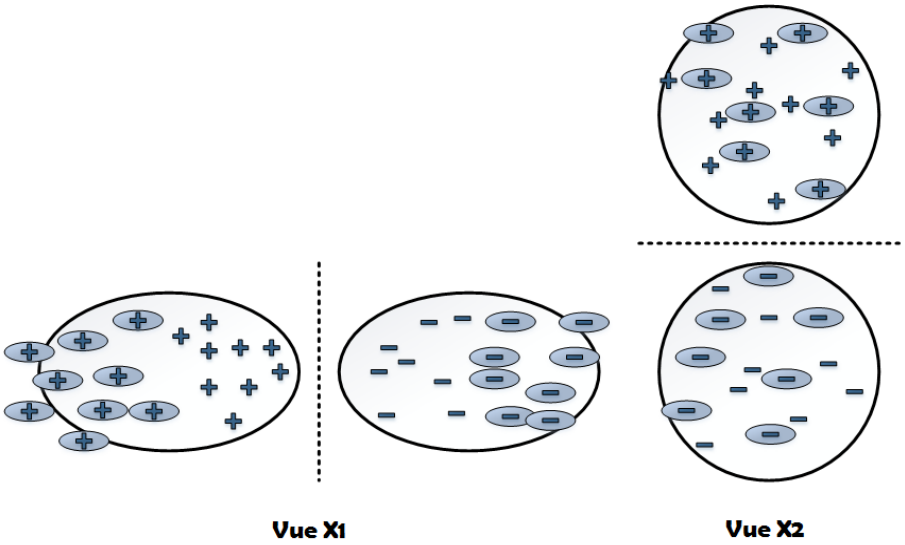


Figure 33 – Illustration de l'hypothèse conditionnelle indépendante de *co-Training* sur l'espace partagé de caractéristiques. Avec cette hypothèse les points de données les plus confiants en vue X_1, représentés par des étiquettes encerclées, seront dispersés au hasard en vue X_2. Ceci est avantageux si elles doivent servir pour le ré-apprentissage du classifieur en vue X_2.

ensembles de données sans points de vue différents, ou à différents types de classifieurs. Balcan et al. [171] assouplissent l'hypothèse d'indépendance conditionnelle avec un état d'expansion beaucoup plus faible, et justifient la procédure itérative de *co-Training*. Ando et Zhang [172] proposent un modèle à deux avis qui assouplit également l'hypothèse d'indépendance conditionnelle.

D'autres part, une amélioration de l'algorithme *co-Training* a vu le jour sous le nom de *RASCO* (Random Subspace Method for Co-training) [173]. *RASCO* obtient les différentes répartitions des caractéristiques en utilisant la méthode de sous-espace aléatoire RSM [39]. Cette modification a apporté une réduction d'erreur d'apprentissage importante en comparaison avec les algorithmes *co-Training* traditionnel et *tri-Training*. Dans son principe *RASCO* utilise les répartitions aléatoires afin de former différents classifieurs. Les échantillons de données non étiquetées seront étiquetés et ajoutés à l'ensemble d'apprentissage basé sur la combinaison des décisions des classifieurs qui sont construit sur les différentes répartitions de variables.

Cependant, dans le cas où on est en présence de beaucoup de variables non pertinentes, *RASCO* peut sélectionner des sous-espaces de caractéristiques qui ne conviennent pas pour une bonne classification. Dans [6], Yaslan et al. proposent d'appliquer des sous-espaces de caractéristiques aléatoires pertinentes pour l'apprentissage de *co-Training*. L'algorithme dénommé $Rel - RASCO$, produit des sous-espaces aléatoires pertinents à l'aide de calcul de pertinence parmi des dizaines de variables. La pertinence d'une variable est obtenue par le calcul de l'information mutuelle entre la variable et son étiquette de classe.

Dans un autre volet, un nouvel algorithme qui étend le paradigme du *co-Training* en appliquant *Random Forest* [51] a été introduit par [18] dans l'application à la détection des micro-calcifications pour le diagnostic du cancer du sein. Cet algorithme nommé *co-Forest* utilise $N \geq 3$ classifieurs au lieu des 3 dans *Tri training*. Les $N - 1$ classifieurs sont employés pour la détermination des exemples de confiance, appelés Ensemble de concomitance $= H_i = H_{N-1}$. La confiance d'un exemple non étiqueté peut être simplement estimée par le degré d'accord sur l'étiquetage, à savoir le nombre de classifieurs

qui sont d'accord sur l'étiquette assignée par H_i.

Les approches proposées par [5], [15], [16] (ainsi que d'autres parmi lesquelles [126]) montrent l'avantage à utiliser plusieurs classifieurs. Cependant, faire l'apprentisssage de ces classifieurs implique de prédire les exemples non-étiquetés avant de les employer. De ce fait l'algorithme de Li et Zhou, propose le meilleur compromis à l'approche semi-supervisée.

Dernièrement, Deng et Guo dans leur papier [137], apportent une amélioration de *co-Forest* avec des résultats très intéressants en intégrant une méthode de filtrage adaptative « Adaptive DATA Editing », l'algorithme étant appelé *ADE-Co-Forest*. *ADE-Co-Forest* emploie l'approche d'édition de données *RemoveOnly* [174] pour identifier et éliminer les exemples «suspects» bruités mal étiquetés au sein du sous-ensemble d'apprentissage nouvellement labellisé.

Le principe de *RemoveOnly* est comme suit : pour chaque exemple x nouvellement labellisé en L_i, il sélectionne les k plus proches voisins de x de l'ensemble d'apprentissage $L \bigcup Li$ selon la règle du plus proche voisin. Il vérifie ensuite, si au moins k' voisins des k plus proches voisins de x détiennent la même étiquette de classe avec x, sinon, l'exemple concerné x est identifié comme un «suspect» mal labellisé et retiré de L_i.

La première constatation déduite par rapport à ces approches d'édition de données (filtrage) c'est qu'elles présentent des avantages au niveau de leur efficacité calculatoire et de leur robustesse face au sur-apprentissage. Mais elles ne tiennent pas en considération les interactions entre les éléments, et tendent à sélectionner les exemples comportant des informations redondantes plutôt que complémentaires. De plus, ces méthodes ne tiennent absolument pas compte des choix faits pour la méthode de classification. Au contraire elles appauvrissent l'ensemble de ré-apprentissage au lieu de l'enrichir avec des éléments nouveaux pour renforcer l'apprentissage.

De ce fait, dans ce travail nous nous intéressons aussi à l'amélioration de *co-Forest*, mais contrairement à *ADE-Co-Forest* [137] qui se base sur le calcul de distance pour la suppression des éléments bruités, nous proposons une approche intégrée (Embedded) de sélection d'attributs pertinents dans le processus de ré-apprentissage des classifieurs. Cette approche permettra de garder tous les éléments nouvellement labellisés mais avec un classement de pertinence au niveau des variables de reconstruction des arbres de décision de la forêt aléatoire.

3 Notre approche proposée « L'algorithme *Optim co-Forest* »

L'algorithme présenté est une amélioration de la méthode *co-Forest* de Li et Zhou [18] pour la classification semi-supervisée (voir chapitre 1 et 2). Nous proposons d'optimiser l'approche en intégrant une sélection d'attributs dans le processus de ré-apprentissage en plus d'apporter quelques corrections aux limites observées dans *co-Forest* et reprises dans *ADE-Co-Forest* ; mais avant d'aborder les étapes de notre approche, nous présentons d'abord le principe de la dernière optimisation de l'algorithme *co-Forest* nommé *ADE-Co-Forest*.

3.1 L'algorithme *ADE co-Forest*

Deng et Guo [137], ont mis en évidence un problème qui peut affecté *co-forest* ainsi que d'autres algorithmes de type *co-Training*, à savoir, que les exemples non étiquetés peuvent être mal étiquetés et intégrés dans le processus d'apprentissage. Cela est dû au nombre limité d'exemples initialement labellisés qui génèrent habituellement des classifieurs faibles, manquant de précision et de diversité. Dans leur article [137], Deng et Guo

proposent un nouvel algorithme qui combine *co-forest* avec une technique d'édition de données adaptative nommée *ADE co-Forest*.

ADE co-Forest exploite une technique spécifique d'édition de données (appelée aussi de vérification ou de filtrage) afin d'identifier et éliminer éventuellement les exemples mal étiquetés à travers les itérations de $co - labeling$. De plus, il emploie une stratégie d'adaptation afin d'enclencher ou non l'opération d'édition en fonction des différents cas de figure. La stratégie d'adaptation combine cinq théorèmes ou pré-conditions, toutes assurent une réduction itérative de l'erreur de classification et une augmentation dans l'échelle des nouveaux exemples d'apprentissage dans le cadre de la théorie de l'apprentissage PAC [125].

Dans *ADE co-Forest* la technique d'édition de données *RemoveOnly* (Algorithme 7) [174] est employée pour identifier les données mal étiquetées.

Algorithm 7 Édition de donnée *Remove only*

1: **Entrée :** soit $S = X$
2: **Sortie :** S
3: **Pour chaque** $x_i \in X$
- Trouver les k plus proches voisins de x_i dans $(X - x_i)$
- **Si** aucune étiquette de classe n'est commune avec au moins k' voisins
- **Alors** retirer x_i de S

Son principe est que l'étiquette de chaque instance non marquée n'est pas seulement déterminée par de multiples classifieurs, mais aussi par la règle du plus proche voisin. Si l'étiquette est compatible avec celles des données sélectionnées par un minimum de k' proches voisins, les données d'instance non labellisées avec la plus grande confiance, sont ajoutées dans l'ensemble d'apprentissage. Sinon, elles sont refusées et écartées de l'ensemble du ré-apprentissage.

3.2 L'algorithme *Optim co-Forest*

Dans le cadre d'une application sur des ensembles de données à grande dimension, les caractéristiques peuvent être corrélées, redondantes ou peut-être même bruitées et donc non pertinentes. Dans ce cas de figure *co-Forest* peut sélectionner ces variables et de ce fait apprendre sur des données erronées résultant des performances de classification individuelles médiocres. Cet inconvénient peut être évité par la sélection de variables les plus pertinentes, ce qui implique l'utilisation d'algorithme de sélection intelligente à l'instar de la sélection aléatoire.

Le procédé de mesure des variables d'importance dans le paradigme des forêts aléatoires (RF) [51], a eu une grande influence sur notre approche proposée. Dans cette étude, nous montrons que ces idées sont également applicables à la sélection de variables en semi-supervisé. De ce fait, nous proposons une méthode d'évaluation des variables d'importance en semi-supervisé basée sur le principe de *co-Forest* appelée *Optim co-Forest*.

L'algorithme classe les caractéristiques à travers un framework composé de méthodes d'ensemble, dans lequel la pertinence d'un élément est évaluée par son exactitude prédictive en faisant appel aux données étiquetées et non étiquetées.

Dans *co-Forest*, la mesure d'importance de variable ne peut être estimée qu'à partir des échantillons OOB (Out of Bag) puisque l'échantillon bootstrap utilisé pour l'apprentissage de chaque arbre aléatoire est modifié après première itération. Les données OOB sont toutes étiquetées. Toutefois, étant donné la quantité très réduite de données marquées, la diversité des données OOB n'est pas assez suffisante. Les estimations OOB

sont biaisées car elles dépendent de trop peu de données.

Nous citons aussi deux conditions nécessaires pour que le succès d'une méthode d'ensemble soit présent, il faut que chaque prédicteur individuel soit relativement bon ; et que les prédicteurs individuels soient différents les uns des autres. Ainsi, pour maintenir la diversité entre les membres du comité dans *co-Forest*, nous avons mis en œuvre deux stratégies.

Optim co-Forest combine à la fois le ré-échantillonnage de données (Bagging [37]) et deux stratégies de sélection. La première implique la sélection intelligente d'un sous ensemble de paramètres aléatoires inspirée de l'approche *Rel-RASCO* [6] et ce afin de générer l'ensemble des classifieurs suivant le principe de *co-Forest*. Cela permettra de conserver la diversité des classifieurs ainsi que leur capacité à produire la meilleure discrimination pour chaque classe.

Une fois que chaque membre de l'ensemble est obtenu, la seconde stratégie appliquée consiste à une extension de la mesure d'importance de RF [51]. Elle fait appel à un assortiment d'ensemble de données marquées et non marquées, pour mesurer la pertinence des variables. Un classement de toutes les variables est finalement réalisé par rapport à leurs pertinences dans tous les classifieurs semi-supervisés.

La combinaison de ces deux stratégies dans la construction de l'ensemble des classifieurs en semi-supervisé mène à l'exploration d'un plus grand espace de solutions et delà récupérer un prédicteur qui rend compte de toute cette exploration.

Dans ce chapitre, nous proposons d'établir des sous-espaces aléatoires de caractéristiques pertinentes pour *co-Forest*. L'algorithme proposé, *Optim co-Forest* (Algorithme 8) produit des sous-espaces aléatoires pertinents à l'aide du score de pertinence des caractéristiques, obtenus en calculant l'information mutuelle entre les caractéristiques et les étiquettes de classe. Afin de maintenir aussi la diversité (l'aspect aléatoire), chaque variable d'un sous-espace est choisie en fonction des probabilités proportionnelles à la pertinence des scores de caractéristiques.

Dans notre proposition *Optim Co-forest* (Algorithme 8 step 2), la première phase consiste à construire la forêt aléatoire classique proposée par Breiman [51]. Cette méthode met en place l'approche du Bagging [37] avec un algorithme d'induction aléatoire des forêts (Forest-RI : Random forest - random Input). Cette approche utilise le principe de randomisation "Random Feature Selection" proposée par Amit et Geman [52] pour générer un ensemble d'arbres doublement perturbé en pratiquant une randomisation fonctionnant à la fois sur l'échantillon d'apprentissage et les partitions internes de l'arbre de décision.

Chaque arbre est donc généré en premier abord à partir d'un sous-échantillon (un échantillon bootstrap [175]) de l'ensemble d'apprentissage, semblable aux techniques du Bagging [37]. Par la suite, l'arbre est construit suivant la méthodologie CART [74] à la différence qu'à chaque nœud, la sélection de la meilleure répartition basée sur l'indice de Gini est effectuée non pas sur l'ensemble des attributs M mais sur un sous-ensemble choisi au hasard de celui-ci. La taille K de ce sous-ensemble est établie entre ($1 \leq K \leq M$) [75].

Pour définir le paramètre K de l'algorithme des Forêts aléatoires, plusieurs travaux dans la littérature [51,176], ont montré qu'un nombre d'attributs égal à $\sqrt{M}$ est un bon compromis pour produire une forêt efficace.

Dans *Optim Co-forest* (Algorithme 8 step 2), nous introduisons l'approche des sous-espaces pertinents (Algorithme 9) sur les K sous-ensembles d'attributs sélectionnés au hasard (Figure 34). Pour ce faire, nous nous sommes basés sur le principe de génération des sous-ensembles de l'algorithme *Rel-RESCO* [6] car il présente la caractéristique de sélectionner les attributs de façon aléatoire et cela à l'aide d'un calcul de pertinence.

Algorithm 8 Pseudo code de l'algorithme Optimized Co-Forest

Entrées : L'ensemble d'exemples labellisés L, ensemble d'exemples non labellisés U, seuil de confiance θ, nombre d'arbres $NTree$
Sorties : Ensemble d'arbres h_i,
Processus :

Construction de la forêt aléatoire à $NTree$ arbres avec les O ***relevent random subspaces***.
$I \leftarrow$ mesure d'importance des variables par H
for $i = 1 \rightarrow NTree$ **do**
  $\hat{e}_{i,0} \leftarrow 0.5$
  $W_{i,0} \leftarrow 0$
end for
$t \leftarrow 0$
Répéter jusqu'à ce qu'il n'y ai aucun changement dans les arbres de la forêt aléatoire.
$t \leftarrow t + 1$
for $i = 1 \rightarrow NTree$ **do**
  $\hat{e}_{i,t} \leftarrow EstimateError\ (H_i, OOB_L)$
  $L^1_{i,t} \leftarrow \phi$
  if $(\hat{e}_{i,t} < \hat{e}_{i,t-1})$ **then**
    $U^1_{i,t} \leftarrow Subsample\ (\mathrm{U},\ \frac{\hat{e}_{i,t-1}.W_{i,t-1}}{\hat{e}_{i,t}})$
    $U^2_{i,t} \leftarrow Subsample\ (\mathrm{U},\ \frac{\hat{e}_{i,t-1}.W_{i,t-1}}{\hat{e}_{i,t}})$
    for $x_u \epsilon\ U^1_{i,t}$ **do**
      if $Confidence\ (H_i, x_u) > \theta$ **then**
        $L^1_{i,t} \leftarrow L^1_{i,t} \bigcup (x_u, H_i(x_u))$
        $W_{i,t} \leftarrow W_{i,t} +$ ***Confidence*** (H_i, x_u)
      end if
    end for
  end if
end for
for $i = 1 \rightarrow NTree$ **do**
  if $(\hat{e}_{i,t}.W_{i,t} < \hat{e}_{i,t-1}.W_{i,t-1})$ **then**
    $h_i \leftarrow LearnRandomTree\ (L \bigcup (L^1_{i,t}, I))$
  end if
end for
for $i = 1 \rightarrow Ntree$ **do**
  for $feature\ f$ **do**
    $I_{i,t}(f) \leftarrow$ **Measure of Importance** $(f, (OOB_{i,t} \bigcup U^2_{i,t}))$
  end for
end for
for $feature\ f$ **do**
  $I(f) \leftarrow \frac{\sum I_{i,t}(f)}{NTree}$
end for
fin de Répéter
Vote Majoritaire $H^*(x) \leftarrow argmax_{y \epsilon label} \sum_{i:h_i(x)=y} 1$

Cette approche permet d'apporter la diversité aux classifieurs tout en produisant autant de sous-espaces que nécessaire.
Pour cela, nous supposons que nous disposons d'un problème de classification à C classes. Les entrées $x \in R^2$ sont de dimension K. Les étiquettes l sont représentées à l'aide d'un codage $\in 1, \ldots, C$. Il comporte un ensemble de données étiquetées L qui se compose de N échantillons ; et un ensemble de données non marquées U avec seulement des entrées.

Notre objectif est de sélectionner O sous-espaces de caractéristiques $S_1, ..., S_O$ (Algorithme 8 step 2), (Figure 34), de sorte que nous pouvons former des classifieurs sur chacun de ces sous-espaces et de combiner leurs sorties afin de produire une bonne classification sur les nouveaux exemples.

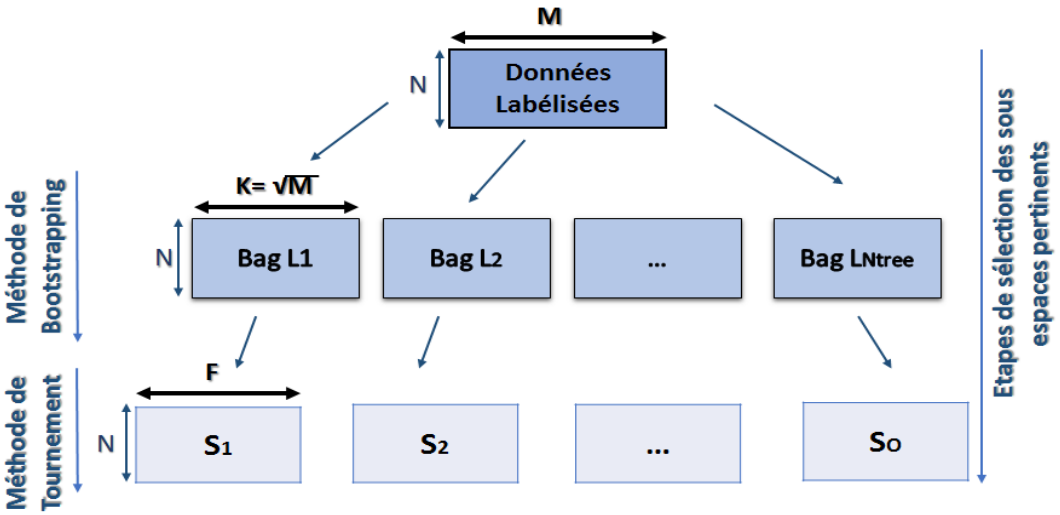


Figure 34 – Procédure de sélection des sous-espaces d'attributs pertinents

Lors de la production de chaque sous-espace de caractéristiques, la fonction *Relevent* sélectionne chaque variable selon son score de pertinence réalisé à partir de l'information mutuelle entre la fonction et les étiquettes de classe (Algorithme 9).

Algorithm 9 L'algorithme *Relevant Random Subspace*

1: **Entrées :** discrétiser L, $Ntree$: nombre d'arbres, F nombre de variables à sélectionner dans chaque bag, labels l.
2: **Sortie :** S_O
3: **For** $O = 1$ **to** $Ntree$ **Do**
4: Rel = ***score relevance*** (Bag_{L_O}, l) % calcul d'information mutuelle entre les caractéristiques et le vecteur classe l
- **For** $m = 1$ **to** F **Do**
- $S_O \leftarrow$***tournement*** (Rel,m)
- **end**
5: **end**

Si nous notons V_j, $j = \{1, 2, .., K\}$ le vecteur de caractéristiques à N dimensions pour la j i ème variable. La pertinence Rel (V_j) d'une caractéristique V_j, à savoir, l'information mutuelle MI (V_j, l), entre V_j et les classes ciblées l peut être évaluée comme suit :

$$Rel(V_j) = MI(V_j, l) = \sum_{n,c} p(V_{n,c}, l_{n,c}) log \frac{p(V_{n,c}, l_{n,c})}{p(V_{n,c})p(l_{n,c})} \tag{3.1}$$

Où $V_{i,j}$ représente la j i ème variable et $l_{n,c}$ c'est l'étiquette c pour le n ème échantillon d'apprentissage.

Afin d'être en mesure de calculer les probabilités de l'équation (3.1), nous devons passer par une phase de discrétisation des variables. De ce fait, nous reprendrons le principe appliqué dans *Rel-RASCO* [6]. Où 10 sous-ensembles de tailles égales seront placés entre les valeurs minimales et maximales observées pour une variable V_j dans l'ensemble d'apprentissage étiqueté. Nous approximons les probabilités par la moyenne du nombre d'échantillons dans chaque sous-ensemble.

Nous générons dans un premier lieu O sous-espaces $S_1, \ldots, S_O$, chacun contenant $m > 0$ variables. Dans un second lieu, nous appliquons la méthode de sélection par *tournement* [177] entre des paires de caractéristiques individuelles (avec taille de tournement fixée à 2) pour mettre en place les sous-espaces de caractéristiques (Figure 35).

La sélection par *tournement* est effectuée comme suit : Deux variables sont sélectionnées au hasard à partir des K caractéristiques disponibles. Parmi ces deux variables, celle qui présente un score de pertinence supérieur est ajoutée au sous-ensemble de caractéristiques sélectionnées. La variable sélectionnée est extraite de l'ensemble des caractéristiques disponibles et la procédure est répétée jusqu'à ce que l'ensemble de caractéristiques sélectionnées m arrive au nombre F requis de variables.

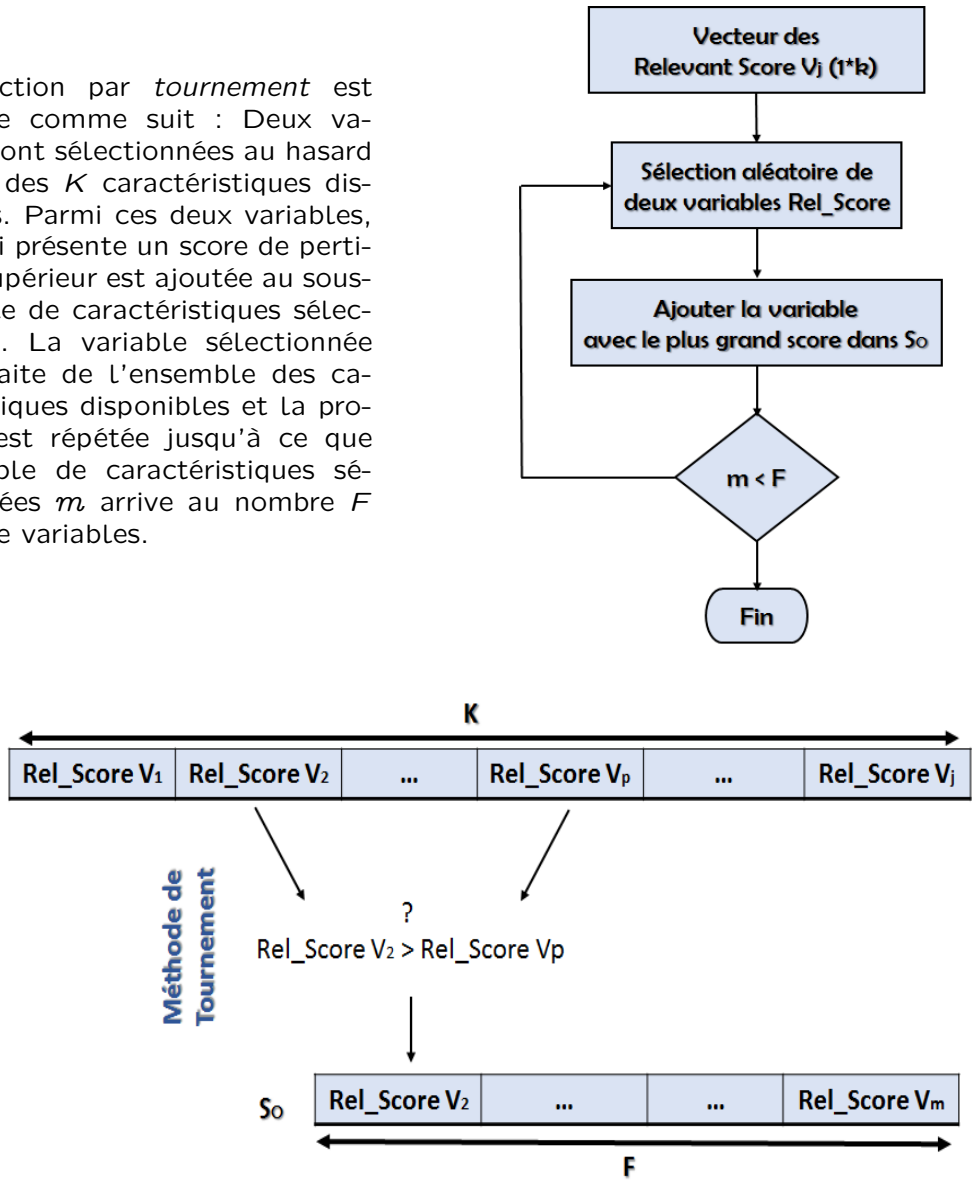


Figure 35 – Principe de sélection par *tournement*

Notre algorithme *Optim co-Forest* procède par la suite à la construction de son comité par l'apprentissage des classifieurs sur les O sous-espaces d'ensembles de caractéristiques pertinents en se référant au principe de *co-Forest*.

Nous apporterons des modifications sur le calcul d'erreur $e_{i,t}$ de *co-Forest* classique qui est évalué avec précision uniquement lors de la première itération par l'estimation d'erreur out-of-bag. Aux itérations suivantes, elle a tendance à être sous-estimée et erronée car elle dépend de l'ensemble d'apprentissage.

Dans *Optim co-Forest* l'erreur est calculée à tous les niveaux sur les éléments *Out Of Bag* (Algorithme 8 step 12) pour assurer que l'erreur réalisée est non biaisée et ainsi plus précise.

La seconde stratégie de sélection que nous avons élaborée dans *Optim co-Forest* consiste à une mesure d'importance des variables pertinentes à partir des ensembles *Out Of Bag*. Ces derniers ne sont pas utilisés pour la construction du modèle correspondant, cela permettra d'aboutir à une estimation non biaisée (Algorithme 8 steps 30-36).

En premier lieu, nous construisons les ensembles *Out Of Bag* « en dehors du sac » pour chaque classifieur, nous sélectionnons les cas bien prédits (classés) des exemples étiquetés OOB_l et nouvellement étiquetés dans un sous ensemble qu'on appellera U^2 (Algorithme 8 step 16).

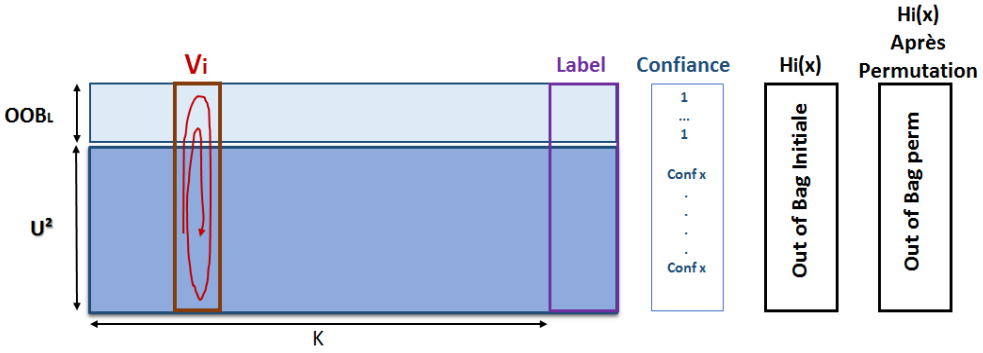


Figure 36 – Principe de mesure d'importance des variables

En second lieu, nous reprenons la confiance de chaque exemple sélectionné (Algorithme 10). Notons que la confiance des exemples étiquetés est mise à 1, si l'étiquette de classe donnée par h_i correspond à l'étiquette réelle. Pour les exemples non étiquetés, leurs confiances seront calculées en fonction du degré d'accord sur l'étiquette parmi les membres de concomitance H^*.

Algorithm 10 La fonction *Confidence*

1: **Entrée :** un exemple non labellisé x, un comité H_x ou x est out of bag, nombre de classes C
2: **Sorties :** *conf (x)* et *label(x)*
3: Appliquer H_x pour générer la probabilité de distribution de classe de x tel que $P(x) = \{p_c(x) : \ c = 1 \ldots, C\}$
4: *conf (x)*$= max_{1 \leq c \leq C} P(x)$
5: *label(x)*$= argmax_{1 \leq c \leq C} P(x)$

En dernier lieu, les valeurs de chaque variable V_i sont permutées au hasard, et h_i est appliqué pour prédire la classe de ce nouveau modèle *Out Of Bag Perm*. La procédure est répétée pour chaque variable $V \in V_1 \ldots V_i$ (voir Figure 36).

En fin de procédure, est calculée la somme des confiances des exemples pour lesquels l'étiquette prédite dans *Out Of Bag Perm* diffère de la première étiquette initiale dans *Out Of Bag*, . Cette dernière valeur est moyennée sur $NTree$ (la taille du comité). La valeur ainsi obtenue est prise comme l'importance de la variable V. L'idée clé de

notre approche est l'utilisation de la confiance de l'étiquette dans l'évaluation de mesure d'importance des variables. Ainsi, les exemples non étiquetés jouent un rôle important dans l'évaluation variable d'importance.

3.3 Les avantages de notre approche

L'intérêt de notre approche se résume sur les différents avantages apportés par rapport aux algorithmes *co-Forest*, *ADE-Co-Forest*, et *Random Forest* de l'état de l'art.

Premièrement, *Optim co-Forest* surpasse RF lorsque l'ensemble d'apprentissage labellisé disponible est faible. La forêt aléatoire RF s'appuie sur les données d'apprentissage disponibles pour encourager la diversité. Donc, si la taille de l'ensemble d'apprentissage est petite comme c'est le cas en semi-supervisé, alors la diversité parmi les membres de l'ensemble sera limitée. Par conséquent, l'erreur d'ensemble sera faible. *Optim co-Forest* ajoute progressivement des exemples nouvellement labellisés à l'ensemble d'apprentissage en construisant des sous-espaces de variables pertinentes. De ce fait, *Optim co-Forest* peut améliorer la diversité et l'erreur moyenne de l'ensemble des membres construits par RF [51], *co-Forest* [18], *ADE co-Forest* [137] et améliore aussi le principe de mesure de variable d'importance.

Deuxièmement, comme *Optim co-Forest* fait appel à une méthode de création d'ensemble diversifiée, la mesure de variable d'importance sur la base d'un ensemble de classifieurs (calcul de confiance) est plus précise que l'utilisation d'un seul classifieur.

Troisièmement, nous noterons également que la mesure de variable d'importance s'effectue de manière différente dans notre approche que celle des *forêts aléatoires RF* ainsi que *co-Forest* [18] et *ADE-Co-Forest* [137]. En *co-Forest*, la mesure de variable d'importance ne peut être estimée qu'à partir des échantillons OOB puisque l'échantillon bootstrap utilisé pour former chaque arbre aléatoire est rejeté après la première itération. Toutefois, étant donné que la quantité de données labellisées est très faible, la diversité des données *out of bag* est insuffisante. Les estimations *out of bag* sont donc biaisées car elles dépendent d'un nombre de données restreint.

4 Expérimentations et résultats

Le même benchmark de données sur lequel Li et Zhou [18] ainsi que Deng et Guo [137] ont fait leurs expériences pour évaluer les performances de leurs algorithmes respectifs *co-Forest* et *ADE-Co-Forest* est utilisé dans cette étude. Ce benchemark est composé de 10 bases de données du dépôt d'UCI [84] (Blake et al., 1998), les informations détaillées de ces ensembles de données sont indiquées dans Table 15.

Bases	*#instances*	*#variables*	*#classes*
Bupa	345	7	2
Colic	368	22	2
Diabetes	768	8	2
Hepatitis	155	19	2
Hypothyroid	3163	25	2
Ionosphere	351	34	2
kr-vs-kp	3196	36	2
Sonar	208	60	2
Vote	435	16	2
Wpbc	194	33	2

Table 15 – Description des bases d'expérimentation

Pour chaque ensemble de données, une validation croisée égale à dix (10 cross validation) est effectuée pour l'évaluation. Les données d'apprentissage sont aléatoirement divisées

en deux ensembles : L labellisé et non labellisé U fixé par un taux (μ), qui est calculé par la taille de U sur la taille de $L \cup U$. Afin de simuler les différentes quantités de données non étiquetées, quatre différents taux de « non-labellisation » $\mu = 20\%$, 40%, 60% et 80%, sont étudiés. Aussi nous prenons note que la distribution de classe pour L et U est maintenue similaire à celle de l'ensemble originel.

Dans les expérimentations suivantes, la valeur de N est fixée à 6 arbres. Le seuil de confiance θ est fixé à 0,75, à savoir, un exemple nouvellement labellisé est considéré comme étant de confiance si plus de 3/4 des arbres sont d'accord sur son étiquette assignée.

Pour estimer l'erreur sur chaque jeu de données, nous avons prédéterminé un ensemble d'exemples étiquetés. Pour chaque ensemble, l'algorithme est évalué sur sa capacité à prédire correctement l'étiquette des exemples non labellisés. Les exemples étiquetés ont été choisis aléatoirement, avec pour seule contrainte la présence d'au moins un exemple de chaque classe pour chaque ensemble.

Afin d'évaluer les performances d'*Optim co-Forest* avec celles de : *co-Forest*, *Random Forest* et *ADE-co-Forest*. Ces dernières sont moyennées (Average) sur l'ensemble des jeux de données dans tous les exemples non labellisés, et une amélioration de la performance globale est réalisée.

Pour chaque ensemble de données avec un taux μ de « non labellisation » spécifique, la validation croisée est répétée dix fois, et les résultats sont moyennés et enregistrés. Les tableaux 16, 17, 18 et 19 résument les taux d'erreurs moyens des classifieurs appris sur les différents taux μ.

Afin de mieux évaluer les résultats obtenus pour chaque algorithme, nous faisons appel aux tests non paramétriques. Nous adoptons plus particulièrement la méthodologie de test de Friedman post-hoc proposée par Demsar [178] pour la comparaison de divers algorithmes sur plusieurs jeux de données [179].

	Optim Co-forest		*Random Forest (RF)*		*Co-forest*		*ADE-Co-forest*	
Bases	Mean±Std	T(s)	Mean±Std	T(s)	Mean±Std	T(s)	Mean±Std	T(s)
Bupa	0.675 ± 0.0408	2.035	0.464 ± 0.0596	0.103	0.670 ± 0.0530	1.944	0.663 ± 0.0428	1.954
Colic	0.843 ± 0.0073	5.814	0.570 ± 0.0200	0.108	0.841 ± 0.0167	4.643	0.840 ± 0.0218	5.546
Diabetes	0.742 ± 0.0215	4.633	0.372 ± 0.0084	0.134	0.740 ± 0.0268	3.393	0.742 ± 0.0257	4.334
Hepatitis	0.835 ± 0.0208	0.921	0.653 ± 0.0409	0.081	0.832 ± 0.0434	0.663	0.841 ± 0.0667	0.886
Hypothyroid	0.990 ± 0.0044	36.153	0.500 ± 0.0054	0.307	0.989 ± 0.0043	16.284	0.990 ± 0.0048	34.398
Ionosphere	0.931 ± 0.0149	1.973	0.481 ± 0.0287	0.094	0.929 ± 0.0221	1.461	0.928 ± 0.0300	1.671
kr-vs-kp	0.987 ± 0.0529	0.785	0.477 ± 0.1805	0.046	0.983 ± 0.0686	0.453	0.986 ± 0.0610	0.681
Sonar	0.799 ± 0.0283	1.548	0.526 ± 0.0611	0.097	0.797 ± 0.0189	1.106	0.796 ± 0.0664	1.218
Vote	0.958 ± 0.0073	0.980	0.567 ± 0.0274	0.085	0.954 ± 0.0277	0.913	0.956 ± 0.0289	0.882
Wpbc	0.793 ± 0.0073	1.237	0.263 ± 0.0173	0.088	0.788 ± 0.0328	1.247	0.796 ± 0.0311	1.074
Average rank	1.3		4		2.6		2.1	

Table 16 – La moyenne des performances des algorithmes comparés avec un taux de non labellisation des données $\mu = 80\%$

	Optim Co-forest		*Random Forest (RF)*		*Co-forest*		*ADE-Co-forest*	
Bases	Mean±Std	T(s)	Mean±Std	T(s)	Mean±Std	T(s)	Mean±Std	T(s)
Bupa	0.674 ± 0.0169	2.154	0.642 ± 0.0420	0.082	0.660 ± 0.0290	0.935	0.662 ± 0.0586	2.096
Colic	0.842 ± 0.0559	2.915	0.821 ± 0.0117	0.091	0.838 ± 0.0170	3.045	0.84 ± 0.0241	2.894
Diabetes	0.75 ± 0.0290	3.365	0.730 ± 0.0353	0.104	0.743 ± 0.0250	2.584	0.749 ± 0.0239	3.207
Hepatitis	0.828 ± 0.0159	0.622	0.811 ± 0.0631	0.069	0.828 ± 0.0750	0.462	0.831 ± 0.0253	0.536
Hypothyroid	0.988 ± 0.0000	26.453	0.987 ± 0.0031	0.286	0.988 ± 0.0059	12.767	0.988 ± 0.0040	25.835
Ionosphere	0.921 ± 0.0061	1.246	0.909 ± 0.0300	0.076	0.921 ± 0.0496	1.161	0.92 ± 0.0668	1.193
kr-vs-kp	0.983 ± 0.0039	2.012	0.976 ± 0.0359	0.095	0.980 ± 0.0282	0.952	0.981 ± 0.0177	1.827
Sonar	0.785 ± 0.0256	0.598	0.738 ± 0.0915	0.072	0.777 ± 0.0625	0.726	0.782 ± 0.0769	0.550
Vote	0.955 ± 0.0485	0.494	0.945 ± 0.0283	0.074	0.951 ± 0.0396	0.472	0.957 ± 0.0544	0.446
Wpbc	0.753 ± 0.0011	0.601	0.689 ± 0.0380	0.071	0.746 ± 0.0407	0.621	0.754 ± 0.0420	0.597
Average rank	1.5		4		2.7		1.8	

Table 17 – La moyenne des performances des algorithmes comparés avec un taux de non labellisation des données $\mu = 60\%$

	Optim Co-forest		*Random Forest (RF)*		*Co-forest*		*ADE-Co-forest*	
Bases	Mean±Std	T(s)	Mean±Std	T(s)	Mean±Std	T(s)	Mean±Std	T(s)
Bupa	0.65 ± 0.0061	2.183	0.626 ± 0.0420	0.082	0.645 ± 0.0290	0.935	0.651 ± 0.0586	2.096
Colic	0.839 ± 0.0061	2.921	0.812 ± 0.0117	0.091	0.829 ± 0.0170	3.045	0.832 ± 0.0241	2.894
Diabetes	0.739 ± 0.0043	3.256	0.733 ± 0.0353	0.104	0.741 ± 0.0250	2.584	0.74 ± 0.0239	3.207
Hepatitis	0.833 ± 0.0163	0.554	0.805 ± 0.0631	0.069	0.818 ± 0.0750	0.462	0.825 ± 0.0253	0.536
Hypothyroid	0.988 ± 0.0010	26.946	0.986 ± 0.0031	0.286	0.987 ± 0.0059	12.767	0.987 ± 0.0040	25.835
Ionosphere	0.931 ± 0.0467	1.354	0.898 ± 0.0300	0.076	0.921 ± 0.0496	1.161	0.923 ± 0.0668	1.193
kr-vs-kp	0.982 ± 0.0107	2.014	0.968 ± 0.0092	0.097	0.976 ± 0.0231	0.105	0.978 ± 0.0294	1.986
Sonar	0.778 ± 0.0199	0.659	0.721 ± 0.0915	0.072	0.757 ± 0.0625	0.726	0.764 ± 0.0769	0.550
Vote	0.957 ± 0.0045	0.499	0.945 ± 0.0283	0.074	0.953 ± 0.0396	0.472	0.955 ± 0.0544	0.446
Wpbc	0.75 ± 0.0334	0.603	0.718 ± 0.0380	0.071	0.724 ± 0.0407	0.621	0.735 ± 0.0420	0.597
Average rank	1.3		4		2.75		1.95	

Table 18 – La moyenne des performances des algorithmes comparés avec un taux de non labellisation des données $\mu = 40\%$

	Optim Co-forest		*Random Forest (RF)*		*Co-forest*		*ADE-Co-forest*	
Bases	Mean±Std	T(s)	Mean±Std	T(s)	Mean±Std	T(s)	Mean±Std	T(s)
Bupa	0.623 ± 0.0142	1.136	0.604 ± 0.0098	0.077	0.615 ± 0.0613	0.985	0.62 ± 0.0323	1.127
Colic	0.831 ± 0.0028	2.298	0.793 ± 0.0640	0.078	0.822 ± 0.0794	2.058	0.821 ± 0.0820	2.233
Diabetes	0.75 ± 0.0147	2.301	0.721 ± 0.0129	0.086	0.737 ± 0.0272	1.983	0.745 ± 0.0219	2.297
Hepatitis	0.825 ± 0.0114	0.256	0.792 ± 0.0286	0.077	0.813 ± 0.0363	0.392	0.82 ± 0.0481	0.229
Hypothyroid	0.986 ± 0.0119	18.023	0.983 ± 0.0129	0.251	0.984 ± 0.0137	8.845	0.985 ± 0.0035	17.626
Ionosphere	0.912 ± 0.0323	0.773	0.869 ± 0.0549	0.071	0.908 ± 0.0195	0.730	0.911 ± 0.0430	0.715
kr-vs-kp	0.971 ± 0.0123	1.126	0.95 ± 0.0133	0.065	0.966 ± 0.0347	0.923	0.9691 ± 0.0283	1.025
Sonar	0.754 ± 0.0097	0.432	0.678 ± 0.0831	0.062	0.705 ± 0.0664	0.522	0.747 ± 0.0831	0.402
Vote	0.953 ± 0.0118	0.487	0.94 ± 0.0477	0.070	0.944 ± 0.0559	0.614	0.949 ± 0.0424	0.457
Wpbc	0.751 ± 5.8514E-17	0.389	0.697 ± 0.0462	0.063	0.75 ± 0.0395	0.466	0.757 ± 0.0666	0.340
Average rank	1.1		4		2.9		2	

Table 19 – L'erreur Moyenne des algorithmes comparés avec un taux de non labellisation des données $\mu = 20\%$

Les Tableaux 16, 17, 18 et 19 démontrent que, par rapport aux taux de performances moyennes des *Forêts Aléatoires* en fonction des différents taux μ, les trois dérivés de *co-Forest* en apprentissage semi-supervisé sont en mesure d'exploiter les données non marquées et ce afin d'améliorer l'hypothèse initiale apprise par RF tout en tablant seulement les données marquées.

Pour chaque taux μ de non labellisation, *Optim co-Forest* a prouvé la plus grande amélioration, tandis que *co-Forest* a fait ressortir la plus petite performance des trois. En outre, les pairwised t-tests à travers les taux d'erreurs de chaque ensemble de données, indiquent que le succès de *Optim co-Forest* sur le classifieur *Random Forest* est maximal pour les 10 ensembles de données, tandis que les améliorations sur *co-Forest* et *ADE-Co-Forest* sont significatives sur plus de la moitié des ensembles de données.
Optim-co-Forest nous a permis d'améliorer dans certains cas de manière significative et dans d'autres cas de manière minime les performances de *Random Forest*, *co-Forest* et *ADE-Co-Forest*. Cela peut être constaté aux différents taux de « non-labellisation » μ, en particulier lorsque le taux est élevé, c'est à dire, les données étiquetées sont très limitées.

Les résultats réalisés varient sur différents ensembles de données. Par exemple, les performances d'*Optim co-Forest* sur $Sonar$, $kr-vs-kp$ et $Ionosphere$ sont tout à fait remarquables. Par contre *Optim co-Forest* donne des résultats équivalents ou presque médiocres sur $Diabetes$ et $Bupa$. Ceci peut être expliqué par le nombre d'attributs dans leur ensemble de données.

En effet, chaque arbre dans la forêt aléatoire est construit en utilisant le meilleur attribut entre plusieurs attributs choisis au hasard pour la répartition interne à chaque nœud. Plus le nombre d'attributs est petit dans les sous-ensembles, plus forte sont les possibilités que certains attributs soient sélectionnés. Par conséquent, il y a plus de chance pour que la répartition interne de l'arbre soit la même. Ainsi, les arbres dans la forêt aléatoire générés avec un nombre d'attributs minime pourraient être moins diversifiés par rapport à ceux qui sont formés sur plusieurs attributs.

Nous pouvons remarquer dans le tableau (Table 15), que l'ensemble de données *Diabetes* a seulement 8 attributs et celui de *Bupa* n'a que 6 attributs tandis que l'*Hepatite* et *Wpbc* ont un nombre plus important d'attributs. Ceci peut expliquer les performances réalisées sur le *Diabetes* et *Bupa* qui tendent à être inférieures à celles de *Ionosphere* et *Sonar* (voir Tables 16, 17, 18 et 19).

Lors de l'évaluation des différentes approches, nous avons aussi besoin de choisir un algorithme qui fonctionne rapidement et utilise les ressources informatiques disponibles de manière efficace. Nous allons, en effet, évaluer l'efficacité des algorithmes comparés par la mesure du temps d'exécution moyen de chacun. Tables 16, 17, 18 et 19) indiquent le temps d'exécution T(s) sous le même support matériel d'exécution (CPU @ 2.60GHz i7-4720HQ). Nous remarquons qu'*Optim co-Forest* enregistre légèrement plus de temps d'exécution que les trois autres approches, les deux stratégies appliquées pour optimiser *co-Forest* explique la légère différence dans l'application de petites bases de données.

Afin de mieux évaluer les résultats obtenus pour chaque algorithme, nous adoptons dans cette étude la méthodologie post-hoc du test de Friedman proposé par Demsar [178] pour la comparaison de plusieurs algorithmes sur plusieurs jeux de données.

4.1 Comparaison par tests non paramétriques

Un test non paramétrique est un test dont le modèle ne précise pas les conditions que doivent remplir les paramètres de la population dont a été extrait l'échantillon. Cependant certaines conditions d'application doivent être vérifiées. Les échantillons considérés doivent être aléatoires (lorsque tous les individus ont la même probabilité de faire partie de l'échantillon) et simples (tous les individus qui doivent former l'échantillon sont prélevés indépendamment les uns des autres) [180], et éventuellement indépendants les uns des autres (emploi de tables de nombres aléatoires). Les variables aléatoires prises en considération sont généralement supposées continues.

Avantages des tests non paramétriques

- Leur emploi se justifie lorsque les conditions d'applications des autres méthodes ne sont pas satisfaites, même après d'éventuelles transformation de variables.
- Les probabilités des résultats de la plupart des tests non paramétriques sont des probabilités exactes quelle que soit la forme de la distribution de la population dont est tiré l'échantillon.
- Pour des échantillons de taille très faible jusqu'à $N = 6$, la seule possibilité est l'utilisation d'un test non paramétrique, sauf si la nature exacte de la distribution de la population est précisément connue. Ceci permet une diminution du coût ou du temps nécessaire à la collecte des informations.
- Il existe des tests non paramétriques permettant de traiter des échantillons composés à partir d'observations provenant de populations différentes. De telles données ne peuvent être traitées par les tests paramétriques sans faire des hypothèses irréalistes.
- Seuls des tests non paramétriques existant permettent le traitement de données qualitatives : soit exprimées en rangs ou en plus ou moins (échelle ordinale), soit nominales.
- Les tests non paramétriques sont plus facile à apprendre et à appliquer que les tests paramétriques. Leur relative simplicité résulte souvent du remplacement des valeurs observées soit par des variables alternatives, indiquant l'appartenance à l'une ou à l'autre classe d'observation, soit par les rangs, c'est-à-dire les numéros d'ordre des valeurs observées rangées par ordre croissant. C'est ainsi que la médiane est généralement préférée à la moyenne, comme paramètre de position [181].

Désavantages des tests non paramétriques

- Les tests paramétriques, quand leurs conditions sont remplies, sont les plus puissants que les tests non paramétriques.

– Un second inconvénient réside dans la difficulté à trouver la description des tests et de leurs tables de valeurs significatives.

4.2 Le test post-hoc de Friedman

Le test de Friedman est un test non-paramétrique utilisé pour comparer les observations répétées sur les mêmes sujets. Il est également appelé une analyse non paramétrique en blocs aléatoires de variance. Le test de Friedman utilise le test statistique chi-square avec $a - 1$ degrés de liberté, où a est le nombre de mesures répétées. Lorsque le $p - value$ de ce test est faible (inférieure à 0,05) nous sommes dans le cas de figure de rejeter l'hypothèse nulle.

Tout d'abord, nous avons utilisé le test de Friedman non-paramétrique pour évaluer le rejet des hypothèses lorsqu'elles sont évaluées avec les mêmes conditions au même niveau. Il classe les algorithmes pour chaque ensemble de données séparément, le meilleur algorithme obtient le rang inférieur, par exemple dans notre cas avec 4 classifieurs, il sera égal à 1, le deuxième rang à 2 etc.

Ensuite, le test de Friedman compare les rangs moyens des algorithmes et calcule la statistique Friedman. Dans notre cas, avec 4 algorithmes et 10 ensembles de données, Friedman statistique (répartis selon chi-square avec 3 degrés de liberté) : est de 27,72 pour $\mu = 20$ %, 24,33 pour $\mu = 40$ % , 22,68 pour $\mu = 60$ % et 23.16 pour $\mu = 80$ %. P-valeur calculée par Friedman Test : 0.000004, 0.000021, 0.000047 et 0.000037 respectivement, de sorte que l'hypothèse nulle est rejetée à un niveau élevé d'importance pour tous les μ.

Si une différence statistiquement significative de la performance est détectée, cela signifie que certaines hypothèses à l'expérimentation ont des répartitions différentes les unes des autres, par conséquent, notre prochaine étape sera d'essayer de savoir quelles paires d'algorithmes sont significativement différentes des autres. De ce fait, nous procédons au test post-hoc.

Demsar a montré dans son étude [178], que la puissance du test post-hoc est beaucoup plus grande lorsque tous les classifieurs sont comparés seulement à un classifieur de commande et non entre eux. Par conséquent, le test de comparaison ne devra pas être fait par paires. Lorsque tous les classifieurs sont comparés avec un classifieur de contrôle, nous pouvons utiliser l'une des procédures générales pour contrôler l'erreur dans les tests d'hypothèses multiples, telles que le test de Bonferroni. Cette méthode est généralement conservatrice, plus souple et facile à utiliser [178].

Le test statistique pour comparer l'i-ème et j-ème classifieur à l'aide de ces méthodes est :

$$z = \frac{(R_i - R_j)}{\sqrt{\frac{k(k+1)}{6N}}}$$

Avec R_j est le rang moyen des algorithmes $R_j = \frac{1}{N}\sum_i r_i^j$. r_i^j le rang de la $j-$ème du $k$ème algorithmes sur la $i - th$ de la N-ème base de données.

La valeur z est utilisée pour trouver la probabilité correspondante de la table de distribution normale [178], qui est ensuite comparée à une valeur appropriée de α (Table A1 dans [182]). Les tests diffèrent dans leur façon de régler la valeur de α pour compenser les comparaisons multiples.

La p-value fournit des informations pour savoir si un test d'hypothèse statistique est significatif ou non. Selon Garcia et al. [183], quand p-value est considérée dans une comparaison multiple, il reflète la probabilité d'erreur d'une certaine comparaison, mais il ne prend pas en compte les comparaisons restantes appartenant à la famille. De ce fait, ils

recommandent l'utilisation de valeurs p ajustées (APVs : adjusted p-values) en raison du fait qu'ils fournissent plus d'informations dans une analyse statistique.

Dans ce chapitre, nous avons mesuré les performances de chaque classifieur au moyen de l'exactitude des données d'essai en utilisant 5 répétitions de 10 validations croisées, ainsi chaque résultat appartenant à l'échantillon analysé par les tests statistiques représente en fait la moyenne de 50 exécutions de l'algorithme en question. Les valeurs de p ajustées obtenues en appliquant la méthode *post hoc* sur les résultats de la procédure Friedman sont résumées dans le tableau (Table 20).

μ	i	Algorithmes	$z = (R_0 - R_i)/SE$	unadjusted p
20%	3	RF	5.022947	0.000001
	2	Co-forest	3.117691	0.001823
	1	ADE-Co-forest	1.558846	0.119033
40%	3	RF	4.676537	0.000003
	2	Co-forest	2.511474	0.012023
	1	ADE-Co-forest	1.125833	0.260236
60%	3	RF	4.330127	0.000015
	2	Co-forest	2.078461	0.037667
	1	ADE-Co-forest	0.519615	0.603332
80%	3	RF	4.676537	0.000003
	2	Co-forest	2.251666	0.024343
	1	ADE-Co-forest	1.385641	0.165857

Table 20 – Le tableau de comparaison Post Hoc FRIEDMAN pour $\alpha = 0.05$

La procédure de Bonferroni-Dunn rejette les hypothèses qui ont une $p - valeur$ non ajustée inférieure à 0.016667. Dans le cas des taux de non labellisitation $\mu = 20$ et 40 %, sur les bases de données à petites dimensions, le test a montré que *Random Forest* et *Co-forest* sont significativement différents d'*Optim Co-forest*, et *ADE-Co-forest* et *Optim Co-forest* très semblables. Par contre, au taux 60 et 80 %, *Optim Co-forest*, *ADE-Co-forêt* et *co-Forest* sont trop semblables, le test rejette seulement la *Forêt Aléatoire*.

4.3 Passage à l'échelle : Application sur les bases à grande dimension

Comme indiqué précédemment dans la présentation de l'algorithme *Optim co-Forest*, notre approche proposée étant plus adaptée aux ensembles de données à grande dimension, où les caractéristiques peuvent être corrélées, redondantes ou peut-être même bruitées et donc non pertinentes. Dans ce cas de figure *Optim co-Forest* permet de sélectionner de manière intelligente des variables plus pertinentes, ce qui implique des performances plus intéressantes. De ce fait, nous allons confirmer cette hypothèse par les tests suivants.

Bases	*#instances*	*#variables*	*#classe*
Arcene	200	10000	2
BaseHock	1993	4862	2
CNAE-9	1080	856	9
Leukemia	73	7129	2
Madelon	2598	500	2
Musk	476	166	2
Ovarian	54	1536	2
PCMAC	1943	3289	2
Relathe	1427	4322	2
Toxicology	171	5748	4

Table 21 – Description des bases d'expérimentation à grande dimension

Nous avons sélectionné un ensemble de 10 bases de données à grande dimension du dépôt d'ASU [184] et d'UCI [84], leurs caractéristiques sont résumées dans Table 21.

Afin d'étudier l'efficacité d'*Optim co-Forest* sur les données à grande dimension en apprentissage semi-supervisé, les performances de *RF*, *co-Forest ADE-Co-Forest* sont en outre analysées.

Les résultats obtenus ont été réalisés avec les paramètres suivants : le nombre d'arbres N égal à 100 arbres, le seuil de confiance θ toujours égal à 0,75. La méthode de validation croisée avec 10 divisions étant appliquée, les tests suivent la même stratégie de simulation que précédemment avec différentes quantités de données non étiquetées μ « taux de non-labellisation » variant de 20%, 40%, 60% et 80%, avec une distribution de classe pour L et U similaire à celle de l'ensemble de données originel.

Il peut être observé à partir des tableaux 22, 23, 24 et 25 qu' *Optim co-Forest* est en mesure d'améliorer la performance de l'hypothèse apprise à partir de données non étiquetées pour différents taux de non labellisation. Le taux d'erreur moyen sur les quatre taux μ des différents tableaux a été réduit respectivement.

	Optim Co-forest		*Random Forest (RF)*		*Co-forest*		*ADE-Co-forest*	
Bases	Mean±Std	T(s)	Mean±Std	T(s)	Mean±Std	T(s)	Mean±Std	T(s)
Arcene	0.5118 ± 0.0777	3.1632	0.4647 ± 0.0503	0.1371	0.5176 ± 0.0369	1.3980	0.4706 ± 0.0716	2.8093
BaseHock	0.8547 ± 0.0359	996.2300	0.5700 ± 0.0116	1.3317	0.6243 ± 0.0124	98.2694	0.6201 ± 0.0642	956.8076
CNAE-9	0.6889 ± 0.0282	26.2365	0.3722 ± 0.0449	0.5091	0.4922 ± 0.0528	14.4724	0.4778 ± 0.0654	24.0344
Leukemia	0.7231 ± 0.0114	0.5313	0.6538 ± 0.0583	0.0942	0.6538 ± 0.0798	0.2811	0.6769 ± 0.0570	0.4775
Madelon	0.5621 ± 0.0158	85.2317	0.5002 ± 0.0126	0.5012	0.5007 ± 0.0309	14.3252	0.5005 ± 0.0148	81.2428
Musk	0.5100 ± 0.0100	1.9855	0.4812 ± 0.0565	0.1601	0.5112 ± 0.0518	1.5900	0.4975 ± 0.0617	1.6966
Ovarian	0.6000 ± 0.0031	0.3564	0.4778 ± 0.0609	0.0961	0.5000 ± 0.0497	0.0998	0.4778 ± 0.0248	0.2320
PCMAC	0.8040 ± 0.0568	1235.6480	0.5260 ± 0.0346	2.1784	0.5473 ± 0.0476	62.2557	0.5488 ± 0.0598	1098.6757
Relathe	0.6587 ± 0.0389	276.5900	0.5673 ± 0.0190	1.5407	0.6059 ± 0.0438	69.3345	0.5908 ± 0.0512	252.3587
Toxicology	0.3053 ± 0.0130	1.4632	0.2632 ± 0.0517	0.1706	0.2667 ± 0.0505	0.3165	0.2667 ± 0.0147	1.2165
Average rank	1.2		3.9		2.1		2.8	

Table 22 – La moyenne des performances des algorithmes comparés sur des ensembles de données à grande dimension avec un taux de non labellisation $\mu = 80\%$

	Optim Co-forest		*Random Forest (RF)*		*Co-forest*		*ADE-Co-forest*	
Bases	Mean±Std	T(s)	Mean±Std	T(s)	Mean±Std	T(s)	Mean±Std	T(s)
Arcene	0.5165 ± 0.0256	10.7892	0.4412 ± 0.0411	0.3282	0.4824 ± 0.0483	1.8689	0.4912 ± 0.0305	9.2982
BaseHock	0.8718 ± 0.0111	1965.3546	0.5471 ± 0.0153	4.7914	0.6321 ± 0.0063	149.3320	0.6351 ± 0.0195	1871.3423
CNAE-9	0.7189 ± 0.0618	61.2355	0.3567 ± 0.0191	0.7362	0.4633 ± 0.0341	24.6943	0.4878 ± 0.0388	59.4440
Leukemia	0.7385 ± 0.0150	1.5688	0.6538 ± 0.0834	0.1373	0.6611 ± 0.0632	0.5043	0.6741 ± 0.0608	1.0460
Madelon	0.5095 ± 0.0147	160.0256	0.5000 ± 0.0279	1.1366	0.5042 ± 0.0208	21.4946	0.5090 ± 0.0163	158.7042
Musk	0.4587 ± 0.0163	2.8955	0.4375 ± 0.0476	0.1277	0.4450 ± 0.0307	2.3389	0.4463 ± 0.0162	2.4255
Ovarian	0.7289 ± 0.0932	0.8623	0.4556 ± 0.1204	0.1407	0.5089 ± 0.0633	0.4353	0.5137 ± 0.1009	0.5000
PCMAC	0.7954 ± 0.0280	1112.0135	0.5214 ± 0.0348	4.2671	0.5501 ± 0.0176	54.0222	0.5498 ± 0.0219	1002.7259
Relathe	0.6516 ± 0.0253	402.8965	0.5572 ± 0.0125	2.8730	0.5954 ± 0.0280	47.6561	0.5820 ± 0.0306	390.1300
Toxicology	0.3158 ± 0.0480	3.4786	0.2667 ± 0.0419	0.2570	0.2772 ± 0.0563	1.6305	0.2667 ± 0.0419	3.2478
Average rank	1		3.95		2.7		2.35	

Table 23 – La moyenne des performances des algorithmes comparés sur des ensembles de données à grande dimension avec un taux de non labellisation $\mu = 60\%$

Nous noterons également que *ADE-Co-Forest* a des performances nettement inférieures à celles de *co-Forest* lors du passage à l'échelle sur les différents μ. Dans un autre cas de figure *Optim co-Forest* réalise des performances considérables même dans les conditions les plus extrêmes avec un taux de non labellisation μ=80% ; ce qui nous ramène en citant la base *Leukemia* par exemple, a seulement 15 individus labellisés avec 7129 caractéristiques et pourtant une amélioration moyenne de 9.939% sur l'erreur moyenne est réalisée.

	Optim Co-forest		*Random Forest (RF)*		*Co-forest*		*ADE-Co-forest*	
Bases	Mean±Std	T(s)	Mean±Std	T(s)	Mean±Std	T(s)	Mean±Std	T(s)
Arcene	0.6453 ± 0.0297	12.1003	0.5076 ± 0.0460	0.4207	0.5877 ± 0.0403	2.2795	0.6043 ± 0.0197	11.4293
BaseHock	0.8092 ± 0.0133	7025.6955	0.5862 ± 0.0195	8.0011	0.7051 ± 0.0068	110.3242	0.6698 ± 0.0152	6866.1659
CNAE-9	0.6303 ± 0.0294	87.2365	0.3878 ± 0.0370	1.4157	0.5239 ± 0.0159	24.4810	0.4264 ± 0.0146	85.6367
Leukemia	0.7231 ± 0.0498	1.9632	0.6538 ± 0.0885	0.1528	0.6654 ± 0.0918	0.5478	0.6538 ± 0.0501	1.6467
Madelon	0.5635 ± 0.0218	183.2152	0.5016 ± 0.0174	1.4692	0.5089 ± 0.0280	24.0347	0.5066 ± 0.0189	171.2990
Musk	0.6469 ± 0.0346	3.9965	0.4781 ± 0.0591	0.1535	0.6231 ± 0.0214	1.8201	0.5725 ± 0.0347	3.5546
Ovarian	0.7111 ± 0.0778	0.5123	0.4944 ± 0.0843	0.1362	0.5944 ± 0.0556	0.3709	0.5722 ± 0.0633	0.4694
PCMAC	0.6435 ± 0.0204	6324.5980	0.5367 ± 0.0150	6.9202	0.5857 ± 0.0158	90.4668	0.5678 ± 0.0182	6182.8384
Relathe	0.7623 ± 0.0334	1121.3587	0.5723 ± 0.0179	5.5956	0.6551 ± 0.0135	47.3533	0.6166 ± 0.0387	1074.2383
Toxicology	0.4228 ± 0.0509	5.4237	0.3649 ± 0.0771	0.3383	0.4000 ± 0.0603	1.8807	0.3877 ± 0.0400	5.0429
Average rank	1		3.95		2.1		2.95	

Table 24 – La moyenne des performances des algorithmes comparés sur des ensembles de données à grande dimension avec un taux de non labellisation $\mu = 40\%$

	Optim Co-forest		*Random Forest (RF)*		*Co-forest*		*ADE-Co-forest*	
Bases	Mean±Std	T(s)	Mean±Std	T(s)	Mean±Std	T(s)	Mean±Std	T(s)
Arcene	0.5809 ± 0.0497	26.8956	0.4706 ± 0.0524	0.5852	0.5574 ± 0.0682	2.8265	0.5603 ± 0.0472	25.4365
BaseHock	0.8020 ± 0.0167	1533.2555	0.5775 ± 0.0162	6.3920	0.6985 ± 0.0086	39.5137	0.661 ± 0.0115	1496.3231
CNAE-9	0.5006 ± 0.0213	122.3255	0.3281 ± 0.0145	1.7349	0.4550 ± 0.0352	353.5539	0.3872 ± 0.0201	120.8174
Leukemia	0.7615 ± 0.099	1.7524	0.6538 ± 0.0699	0.1562	0.6692 ± 0.0926	0.4856	0.6577 ± 0.0632	1.7313
Madelon	0.5470 ± 0.0302	210.2354	0.5008 ± 0.0139	2.2302	0.5106 ± 0.0371	337.9970	0.5088 ± 0.0458	202.7113
Musk	0.5969 ± 0.0515	4.7859	0.4763 ± 0.0421	0.1903	0.5825 ± 0.0428	2.8359	0.53 ± 0.0456	4.5300
Ovarian	0.7056 ± 0.1049	0.6234	0.4722 ± 0.0304	0.1232	0.5611 ± 0.1009	0.4204	0.5389 ± 0.1204	0.5704
PCMAC	0.7287 ± 0.0295	1678.9560	0.5341 ± 0.0258	8.5303	0.5804 ± 0.0159	49.0181	0.5624 ± 0.0150	1654.5269
Relathe	0.7216 ± 0.0171	963.2540	0.5581 ± 0.0240	7.4081	0.6365 ± 0.0182	51.7928	0.599 ± 0.0149	943.0689
Toxicology	0.4351 ± 0.0574	5.2364	0.3789 ± 0.0771	0.3383	0.3947 ± 0.0603	1.8807	0.3807 ± 0.0400	5.0429
Average rank	1		4		2.1		2.9	

Table 25 – La moyenne des performances des algorithmes comparés sur des ensembles de données à grande dimension avec un taux de non labellisation $\mu = 20\%$

Le principe de sélection adapté dans *Optim co-Forest* avec la mesure de variables d'importance a permis d'apporter des améliorations notables, le distinguant comme un excellent candidat dans la classification des données à grande dimension en apprentissage semi-supervisé.

La mesure d'importance des variables adaptée par *Optim Co-Forest* a apportée des améliorations significatives, le distinguant comme un excellent candidat dans la classification semi-supervisé de données à grande dimension. En outre, en terme de temps de calcul T (s) (tableaux 22, 23, 24 et 25), on peut voir clairement que, dans le cadre de données à grand nombre d'attributs, *Optim Co-Forest* est gourmand en temps d'exécution, ceci est relié à la complexité temporelle des deux stratégies de sélections adoptées.

La comparaison Friedman post hoc pour les quatre algorithmes avec les P-valeurs ajustées sur les résultats de la procédure Friedman sont résumés dans Table 26.

μ	i	Algorithmes	$z = (R_0 - R_i)/SE$	unadjusted p
20%	3	RF	5.196152	0
	2	ADE-Co-forest	3.290897	0.000999
	1	Co-forest	1.905256	0.056747
40%	3	RF	5.10955	0
	2	ADE-Co-forest	3.377499	0.000731
	1	Co-forest	1.905256	0.056747
60%	3	RF	5.10955	0
	2	Co-forest	2.944486	0.003235
	1	ADE-Co-forest	2.338269	0.019373
80%	3	RF	4.676537	0.000003
	2	ADE-Co-forest	2.771281	0.005584
	1	Co-forest	1.558846	0.119033

Table 26 – Le tableau de comparaison Post Hoc FRIEDMAN pour $\alpha = 0.05$

Pour les ensembles de données à grande dimension, la procédure de Bonferroni-Dunn rejette les hypothèses qui ont une $p-valeur$ non ajustée inférieure à 0.016667, à chaque niveau de non labélisation μ. Ainsi, le test a montré qu'à $\mu = 20$, 40 et 80 %, *Random forest* et *ADE-Co-forest* sont significativement différents de *Optim Co-forest*. Parcontre, une grande similarité est présente entre *Optim Co-forest* et *Co-forest*.

Par conséquent, au taux 60 %, *Optim Co-forest* et *ADE-Co-forest* sont très semblables, le test rejette les algorithmes *Random forest* et *Co-forest*. Ceci peut être expliqué par l'approche d'édition de données appliquée par *ADE-Co-forest* qui lui permet d'éliminer les données bruitées et d'outrepasser les performances de *Co-forest* à un certain nombre de données labellisées disponibles.

Afin de soutenir davantage ces comparaisons, nous avons comparé, sur chaque ensemble de données l'hégémonie de *Optim Co-forest* avec chaque paire de méthodes. Nous utilisons Pairwise t-test (avec $t = 0,05$) sur les performances des tableaux 22, 23, 24 et 25 pour respectivement $\mu = 80$, 60, 40 et 20 %.

Les résultats de ces comparaisons par paires sont représentés dans le tableau (Table 27) en termes des statuts «Win | Tie | Loss» [185] ; les trois valeurs dans chaque cellule

indiquent combien de fois *Optim Co-forest* est nettement mieux / pas significativement différent / plus médiocre que l'autre méthode sur les 10 ensembles de données à grande dimension.

Algorithmes	*Random forest*			*Co-forest*			*ADE-Co-forest*		
Paired t-test (with $t = 0.05$)	**Win**	**Tie**	**Loss**	**Win**	**Tie**	**Loss**	**Win**	**Tie**	**Loss**
$\mu\% = 20$	10	0	0	7	3	0	9	1	0
$\mu\% = 40$	10	0	0	7	3	0	9	1	0
$\mu\% = 60$	10	0	0	7	2	1	7	3	0
$\mu\% = 80$	10	0	0	8	2	0	9	1	0

Table 27 – Le tableau de comparaison Pairwise t-test d'*Optim Co-forest* avec les 3 autres classifieurs

Il peut être clairement vu du tableau de comparaison Pairwise t-test (Table 27), qu'*Optim Co-forest* surpasse considérablement *Co-forest* et *ADE-Co-forest* dans le cadre d'application des données à grande dimension.

5 Conclusion

La tâche d'apprentissage vise à savoir étiqueter des exemples pour lesquels l'étiquette est inconnue, soit parce qu'il est difficile et couteux d'obtenir ces étiquettes, soit parce qu'il s'agit d'exemples non encore observés. Ainsi, pour entrainer un classifieur, on dispose souvent d'une base d'exemples étiquetés, et d'une énorme masse d'exemples non étiquetés.

De ce fait, plusieurs travaux pionniers ont proposé des méthodes attrayantes [5] [14] [15] [16] [17] [18] [19] [6] montrant l'intérêt de ce type d'apprentissage. Néanmoins, la plupart des travaux ont relevé un point très important, c'est que la prise en compte d'exemples non labellisés pouvaient dégrader les performances par rapport à un apprentissage purement supervisé. Il est devenu donc clair qu'une analyse plus fondamentale est nécessaire pour de mieux comprendre les conditions qui permettent non seulement d'espérer, mais garantir des améliorations de performances. De cette vision est née notre approche *Optim co-Forest*.

Dans ce chapitre, l'algorithme *Optim co-Forest* est proposé, il a la faculté de tirer profit des échantillons non marqués pour améliorer les performances du système formé à partir des échantillons marqués. En étendant le paradigme *co-Forest*, *Optim co-Forest* exploite la puissance de *Random Forest*, une méthode d'ensemble bien connue, pour s'attaquer au problème de la sélection des échantillons non labellisés les plus confiants pour le ré-apprentissage de la forêt.

Optim co-Forest exploite également deux stratégies de sélection : la première concerne la sélection de sous-ensemble de paramètres aléatoires qui nous permettra de garder la diversité des classifieurs. La seconde méthode est la mesure d'importance de variables pour mesurer la pertinence de ces variables. La combinaison de ces deux stratégies dans la construction de l'ensemble des classifieurs en semi-supervisé conduit à l'exploration d'un plus grand espace de solutions et par conséquent avoir un prédicteur plus compétitif et adéquat aux espaces à grande dimension.

Des expérimentations sur des ensembles de données de l'UCI de petite et grande dimension vérifient l'efficacité de *Optim co-Forest*. Elles montrent clairement que le principe d'*Optim co-Forest* est plus intéressant et performant et se distingue dans le cas de données à grande dimension.

Conclusion

En classification et plus généralement en apprentissage, on distingue classiquement deux approches : le cas supervisé et le cas non supervisé. On peut combiner celles-ci dans un apprentissage semi-supervisé. Dans ce cadre, on enrichit un ensemble de données non étiquetées par un certain nombre d'exemples étiquetés. Ces derniers, en proportion généralement faible, servent à guider les algorithmes de classification via leur paramétrage et leurs conditions d'initialisation.

Typiquement, ce type de problème se présente lorsqu'évaluer la classe d'un exemple est coûteux. Comme exemple lorsqu'on souhaite déterminer si un patient donné est atteint ou non d'une maladie qu'on ne peut détecter que par des moyens modernes onéreux.

D'un autre point de vue, nous avons étudié dans la partie précédente (Partie 2) les avantages et la capacité prédictive des méthodes d'ensemble, plus spécifiquement des forêts aléatoires. Sur le même principe, nous avons proposé dans ce travail de reprendre l'algorithme *Random Forest* dans un contexte semi supervisé sur des applications à grande échelle de données. De ce fait, nous avons mis aux points trois contributions où nous avons illustré les résultats théoriques avec des données synthétiques et réelles :

Dans un premier temps, nous avons étudié la méthode d'apprentissage *co-Forest* sur des données biologiques pour l'application classique d'apprentissage semi supervisé des forêts aléatoires, et comparé les résultats obtenus avec ceux obtenus par la même approche en apprentissage supervisé, démontrant ainsi la capacité de *co-Forest* à améliorer les performances en exploitant les échantillons non labellisés

En second lieu, nous avons cherché par la suite à étendre la méthode *co-Forest* pour en faire un outil de segmentation d'images médicales par approche de classification pixellique. L'approche expérimentale est fournie et testée sur des images rétiniennes réelles. Le principe repose sur une approche de segmentation semi-supervisée par la classification de super-pixels des régions cup et disque pour le calcul du rapport Cup/disque. Ce procédé nous a permis la reconnaissance automatique par apprentissage semi-supervisé des régions Discs et Cups pour la mesure du rapport CDR afin de mesurer la progression du glaucome. Pour ce faire, la méthode proposée SP3S (*Super-Pixel for Semi-Supervised Segmentation*) comporte deux étapes principales. Dans la première, la forêt aléatoire en apprentissage semi-supervisé « *co-Forest* » est formée uniquement sur 10% des images super-pixels annotées par un expert ophtalmologiste. Dans la seconde étape, les super-pixels non labellisés sont impliqués pour le renforcement de l'apprentissage du classifieur afin de mieux discriminer les régions cup et disque des images rétiniennes. Pour calculer le rapport VC/D, un modèle de forme géométrique actif est utilisé pour dresser le contour du cup et disque. Les résultats réalisés ont démontré des erreurs minimes comparées à ceux des experts du domaine.

En dernier lieu, nous avons proposé un nouvel algorithme de classification semi-supervisé *Optim co-Forest* qui améliore les méthodes décrites dans la littérature pour une application à moyenne et grande échelle de données. Cette méthode a démontré sa faculté de tirer profit d'échantillons non marqués pour améliorer les performances du système formé à partir d'échantillons marqués. En étendant le paradigme de *co-Forest*, *Optim co-Forest* exploite la puissance de *Random Forest*, pour s'attaquer au problème de la sélection des échantillons non labellisés les plus confiants pour le ré-apprentissage de la forêt en classification semi-supervisée.

Enfin l'étude que nous avons mené dans le cadre de l'évaluation d'algorithmes semi-supervisés nous démontre clairement la supériorité des méthodes semi-supervisées dans le contexte de données partiellement étiquetées. Nous avons pu observer que la combinaison des données étiquetées et non étiquetées permettait une évaluation plus globale, objective et quantitative des méthodes supervisées seules afin d'améliorer les performances de classification.

Ces applications montrent également, que les méthodes proposées passent à l'échelle de manière très intéressante malgré le volume de données. L'algorithme développé en dernier chapitre de cette partie peut être appliqué pour la sélection de caractéristiques. Cette piste sera exploitée dans la partie suivante de cette thèse.

Quatrième partie

IV

État de l'art et Propositions en apprentissage supervisé et semi-supervisé : Sélection de variables par approche ensembliste

Introduction

Les avancées technologiques ont facilité l'acquisition et le recueil de nombreuses données, notamment dans le domaine médical lors des routines cliniques. Ces données peuvent être utilisées comme support de décision médicale, conduisant aux développements d'outils capables de les analyser et de les traiter.

Dans la littérature, nous trouvons régulièrement la notion d'aide au diagnostic, ces systèmes sont même considérés comme étant essentiels dans beaucoup de disciplines, ils reposent sur des techniques issues de l'intelligence artificielle. Cependant, le problème majeur que rencontre l'application de ces approches réside dans la haute dimension des données. Ces problèmes désignent les situations où nous disposons de peu d'observations alors que le nombre de variables explicatives est très grand. Cette situation est de plus en plus fréquente dans diverses applications, en particulier celles liées aux domaine médical et biologique.

Aujourd'hui, la difficulté réside non seulement dans l'obtention des données pertinentes mais également dans leurs analyses. L'objectif consiste à développer des méthodes d'analyse permettant d'extraire un maximum d'informations à partir des données récoltées par les experts ou appareils médicaux. Ces faits ont fait émerger un grand nombre de questions, *"Quelle est la bonne procédure de sélection à utiliser ?"*, *"Comment parvenir à une technique complètement explicite, simple à implémenter et rapide aux calculs ?"*.

La sélection de données contribue au renforcement de l'aide au diagnostic médical, le degré et le taux de progression des variables pertinentes mesurées de façon répétitive sur chaque sujet permettent de quantifier la sévérité de la maladie et la susceptibilité de sa progression. Ceci est usuellement intéressant, sur les plans cliniques et scientifiques, d'aider l'expert dans ses prises de décisions dans un laps de temps rapide.

Les méthodes de réduction de la dimension peuvent être divisées en deux grandes catégories : l'extraction de variables et la sélection de variables [13, 186].

– **Les méthodes d'extraction de variables** consistent à transformer l'ensemble de variables de départ en un nouvel ensemble de variables, généralement plus petit, tout en conservant autant que possible la structure originale des données. Nous pouvons distinguer les méthodes linéaires non-supervisées comme l'Analyse en Composantes Principales (ACP) [187], les méthodes linéaires supervisées comme l'Analyse Factorielle Discriminante (AFD) [188], les méthodes non linéaires non supervisées comme l'ACP à noyau [189], Locally Linear Embedding (LLE) [190], Isometric FeatureMapping (Isomap) [191] et les méthodes non linéaires supervisées comme l'Analyse Factorielle Discriminante à noyau...

Le principal inconvénient de ces méthodes est leur temps de calcul. Une méthode d'extraction de variables nécessite le calcul des attributs initiaux pour ensuite extraire les attributs pertinents. Ces derniers sont obtenus en combinant, linéairement ou non, les variables initiales. Un autre inconvénient des méthodes d'extraction est qu'elles imposent un effort important à l'utilisateur pour interpréter et comprendre la nouvelle représentation des données : il est donc difficile de donner une interprétation sémantique des attributs extraits, car ces derniers sont une combinaison des attributs initiaux.

– **Les méthodes de sélection de variables** [186] aident à choisir un sous-ensemble pertinent de variables à partir de l'ensemble original de caractéristiques selon un critère de performance. Ces méthodes [192] permettent alors de caractériser plus rapidement les données et sont donc utilisées pour des applications où les coûts en temps de calcul doivent être minimisés. La sélection de variables ne modifie pas la représentation originale des données : les attributs sélectionnés gardent leur sémantique de départ et peuvent alors être interprétés plus facilement par l'utilisateur.

Problématique

Durant ces dernières années, la sélection de variables est devenue l'objet qui attire l'attention de nombreux chercheurs. Cette sélection permet d'identifier et d'éliminer les variables qui pénalisent les performances d'un modèle complexe dans la mesure où elles peuvent être bruitées, redondantes ou non pertinentes. De plus, la mise en évidence des variables pertinentes facilite l'interprétation et la compréhension des aspects médicaux et biologiques ; ainsi, elle permet d'améliorer la performance de prédiction des méthodes de classification et de passer outre le fléau de la haute dimension de ces données (phénomène connu sous le nom en anglais de *the curse of dimentionality*).

Le problème spécifique de la sélection de variables nécessite une approche particulière puisque le nombre de variables est très largement supérieur vis-à-vis du nombre d'échantillons (expériences ou observations). Dans la littérature d'apprentissage artificiel trois approches sont envisagées. Elles relèvent des méthodes de type filtre, enveloppe "*wrapper*" ou embarqué "*Embedded*".

Les méthodes *wrapper* et *Embedded* sélectionnent de façon implicite les variables où la sélection se fait lors du processus d'apprentissage. Ces deux approches sont caractérisées par la pertinence des attributs sélectionnés mais demande un temps de calcul long à l'opposé de la méthode filtre. L'approche filtre est couramment utilisée à ce jour pour analyser les données biologiques, elle consiste à parcourir la sélection des variables avant le processus de l'apprentissage et ne conserve que les caractéristiques informatives.

Les approches de sélection de variables

La sélection de variables est d'un intérêt particulier et crucial pour toute base de données dont le nombre de variables ou d'observations est très grand. Or, avec le développement des outils informatiques qui permettent de stocker et de traiter toujours mieux ces données, ce type de situations est fréquemment rencontré, ceci explique l'intérêt actuel porté au thème de la sélection de variables qui permet d'extraire des variables pertinentes et de réduire la dimension de l'espace originel.

Dans l'apprentissage automatique, la sélection de variables est un dispositif crucial où le but est d'isoler le sous-ensemble de caractéristiques permettant d'expliquer efficacement les valeurs de la variable cible, ainsi trois approches sont généralement citées dans la littérature :

– Approche Filtre (*filter* approach),

– Approche enveloppe (*wrapper* approach),
– Approche embarquée (*Embedded* approach).

Approche filtre

L'approche filtre sélectionne un sous-ensemble de variables en pré-traitement des données d'un modèle (l'étape d'analyse des données). Le processus de sélection est indépendant du processus de classification (Figure 37) [193].

Figure 37 – Principe de l'approche filtre

Un de ses avantages est d'être complètement indépendante du modèle de données construit. Elle propose un sous-ensemble de variables satisfaisant pour expliquer la structure cachée des données.

Ce principe est aussi adapté dans la sélection de variables non supervisées (Géurif [194], Mitra et al. [46], Bennani et Géurif [195]). De plus les procédures filtres sont généralement moins gourmandes en temps de calcul puisqu'elles évitent les exécutions répétitives des algorithmes d'apprentissage sur différents sous-ensemble de variables.

En revanche, leur inconvénient majeur est qu'elles ignorent l'impact des sous-ensembles choisis sur les performances de l'algorithme d'apprentissage.

Approche enveloppe

Les techniques enveloppe "*wrapper*" ont été introduites par John et al. en 1994 [196]. Leur principe est de générer des sous-ensembles candidats et de les évaluer grâce à un algorithme de classification. Cette évaluation est faite par un calcul d'un score, par exemple un score d'un ensemble sera un compromis entre le nombre de variables éliminées et le taux de performance de classification sur un fichier de test.

L'appel à l'algorithme de classification est fait à chaque évaluation (c'est-à-dire, à chaque sélection d'une variable, le taux de classification est calculé pour juger la pertinence d'une caractéristique) et cela par un mécanisme de validation croisée qui est fréquemment utilisé.

Le principe de *wrapper* est de générer un sous-ensemble bien adapté à l'algorithme de classification (Figure 38). Les taux de reconnaissance sont élevés car la sélection prend en compte le biais intrinsèque de l'algorithme de classification. Un autre avantage est sa simplicité conceptuelle ; il n'y a pas besoin de comprendre comment l'induction est affectée par la sélection des variables, il suffit de générer et de tester.

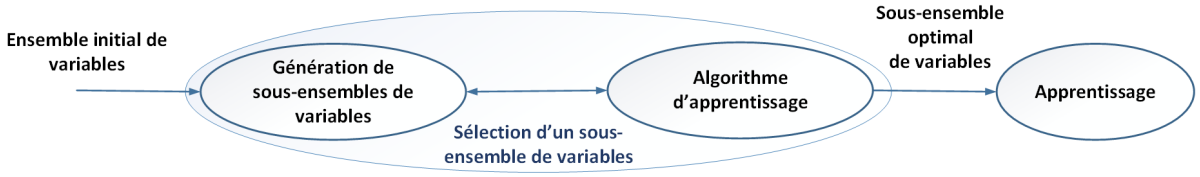


Figure 38 – Principe de l'approche enveloppe *wrapper*.

Cependant, trois raisons font que les *wrapper*s ne constituent pas une solution parfaite :

- D'abord, elles n'apportent pas vraiment de justification théorique à la sélection, de ce fait, elles ne permettent pas de comprendre les relations de dépendances conditionnelles existantes entre les variables.
- D'autre part la procédure de sélection est spécifique à un algorithme de classification particulier et les sous-ensembles trouvés ne sont pas forcément valides si nous changeons de méthode d'induction.
- Finalement, l'inconvénient principal de la méthode réside dans les calculs qui deviennent de plus en plus longs, voir irréalisables quand le nombre de variables est très grand.

Approche embarquée

Les méthodes embarquées "*Embedded*" intègrent directement la sélection dans le processus d'apprentissage, les arbres de décision [74] sont l'illustration la plus emblématique. Cette catégorie regroupe toutes techniques qui évaluent l'importance d'une variable en cohérence avec le critère utilisé pour évaluer la pertinence globale du modèle.

La sélection de variables

La sélection de variables (*Feature Selection* "FS") est un domaine très actif depuis ces dernières années. Sa particularité s'inscrit dans le cadre de la fouille de données. En effet, la fouille de données de très grande dimension devient un enjeu crucial pour des applications tel que le génie génétique, les processus industriels complexes, etc ....

Il s'agit en fait de résumer et d'extraire intelligemment de la connaissance à partir des données brutes. L'intérêt de la sélection de variables est regroupé dans les points suivants :

- Lorsque le nombre de variables est vraiment trop grand, l'apprentissage consomme un très grand temps. Par contre, la sélection peut réduire l'espace des caractéristiques et permettre l'exécution dans un temps convenable.

- D'un autre point de vue, créer un classifieur revient à créer un modèle pour les données. Or une attente légitime pour un modèle est d'être le plus simple possible (paradigme d'Occam's razor [197]). La réduction de la dimension de l'espace de caractéristiques permet alors de réduire le nombre de paramètres nécessaires à la description de ce modèle.

- Elle améliore la performance de la classification : sa vitesse et son pouvoir de généralisation.

- Elle augmente la compréhensibilité des données.

La sélection de données consiste à choisir un sous-ensemble optimal de variables pertinentes, à partir d'un ensemble de variables originelles. Elle comporte trois éléments essentiels : une mesure de pertinence, une procédure de recherche et un critère d'arrêt.

Mesure de pertinence

La mesure de pertinence associée aux méthodes de sélection de variables sont souvent basées sur des heuristiques calculant l'importance individuelle de chaque variable dans le modèle obtenu. Ces heuristiques sont de différentes natures : statistique, probabiliste, information mutuelle, mesure d'indépendance ou bien la vraisemblance entre les variables.

Procédure de recherche

Trouver une solution optimale suppose une recherche exhaustive parmi les $(2_n - 1)$ sous-ensembles de variables possibles. Bien que des méthodes de recherche efficaces comme celle de Branch et Bound ont été proposées, elles s'appuient sur une propriété de mono tonicité de la mesure de pertinence qui est en pratique difficile à assurer (Bennani [198] ; Bennani [199]). Une recherche exhaustive est dès lors inapplicable même pour un nombre de variables de l'ordre de quelques dizaines. En pratique, nous utilisons des approches sous-optimales comme les algorithmes de recherche gloutonne ou les méthodes de recherche aléatoire.

Les stratégies gloutonnes les plus utilisées sont les méthodes séquentielles dont font partie la méthode de sélection en avant (*forward selection*) (Dy et Brodley [200] ; Raftery et Dean [201]), la méthode d'élimination en arrière (*backward elimination*) (Guérif et Bennani [202]) et les méthodes bidirectionnelles comme la méthode stepwise ou celle proposée par Sorg-Madsen et al. [203] qui combine une approche par sélection filtre à une approche symbiose par élimination (*wrapper* ou *Embedded*).

La méthode de sélection en avant débute avec un ensemble vide et progresse en ajoutant une à une les variables les plus intéressantes. A l'inverse, la méthode d'élimination en arrière commence par l'ensemble de toutes les variables dont les moins pertinentes sont supprimées tour à tour. Les méthodes bidirectionnelles combinent ces deux modes de recherche. Les algorithmes génétiques font partie des méthodes de recherche aléatoire qui sont parfois utilisées. Il est reproché généralement à la méthode de sélection en avant, de ne pas prendre en compte le problème de la pertinence mutuelle.

Critère d'arrêt

Le nombre de variables à sélectionner fait généralement partie des inconnues du problème et il doit être déterminé automatiquement à l'aide d'un critère d'arrêt de la procédure de sélection de variables. Ce problème se rapporte à celui de la sélection de modèles sachant que de nombreux auteurs utilisent soit des critères de maximum de vraisemblance ([200] ; [201]) soit des critères de séparabilité des classes ([200] ; [202]).

Il convient de noter que les critères de séparabilité utilisés dans [200], [202] ne sont plus utilisables lorsque le nombre de variables est important car leur évaluation fait intervenir soit l'inversion soit le calcul du déterminant des matrices de covariance.

D'autres auteurs comme Sorg-Madsen et al. [203] utilisent une combinaison de la mesure de pertinence et de la procédure de recherche. Par contre, pour la sélection d'un sous-ensemble de variables, l'attribution d'un score pour chaque variable est la technique la plus optimale et plusieurs solutions ont été proposées dans la littérature.

Une première solution est d'attribuer un score à chaque variable indépendamment des autres, et de faire la somme de ces scores. Pour évaluer une variable, l'idée est de déterminer la corrélation de la variable avec la classe. Cependant, Elisseeff et Guyon [204] ont démontré par des exemples simples que cette approche nommée *feature ranking* pose des problèmes dans le cas général. En effet, cette approche n'élimine pas les variables redondantes, d'autre part, il est possible que des variables peu corrélées avec la classe deviennent utiles lorsque nous les repositionnons dans le contexte des autres variables.

Une seconde solution est d'évaluer un sous-ensemble dans sa globalité. Une méthode plus spécifique au problème est décrite dans Koller et Sahami [205] où les auteurs proposent d'éliminer une variable si elle possède une couverture markovienne, c'est-à-dire si elle est indépendante de la classe en connaissance des autres variables. Il existe aussi d'autres solutions avec un principe intermédiaire entre *feature ranking* et *subset ranking* basé sur une idée de Ghiselli [206] démontrant de bons résultats dans le cadre de la CFS

(*Correlation based Feature Selection*) par M.Hall [207].

La corrélation ou la dépendance entre deux variables peut être définie de plusieurs façons. L'application du coefficient de corrélation statistique comme dans [207] est trop restrictive, car elle ne capture que les dépendances linéaires. Le test de Fisher à un score important indique donc que les moyennes des deux classes sont significativement différentes.

En revanche, un test d'indépendance statistique peut être utilisé comme celui du *Chi-2*, *t-statistic*, *F-statistic* [208], [209], [210]. Aussi, il est possible d'utiliser la notion d'information mutuelle (MI) qui est fondée sur un calcul probabiliste et d'entropie de Shannon [211] ainsi que la méthode mRMR (*minimum Redondance Maximum Relevance*) qui vise deux objectifs en parallèle : prendre les variables pertinentes et éliminer les variables redondantes.

Dans les parties précédentes de cette thèse, nous avons mis en avant les avantages et les inconvénients de la méthode *Forêt Aléatoire*, montrant que cette méthode délivre un assemblage de plusieurs arbres de décisions, elle perd en intelligibilité ce qu'elle gagne en précision, aussi elle se caractérise par sa double "randomisation" qui lui permet d'améliorer ses qualités de prédiction par rapport aux techniques plus simples comme le bagging. Cette implémentation présente l'avantage de ne dépendre que d'un nombre très limité de paramètres pour s'exécuter, ce qui rend leur exploration plus riche. Les deux principales caractéristiques de la méthode sont sa capacité de sélection des variables d'importance et le nombre d'arbres formant la forêt globale de prédiction.

Dans la quatrième partie de cette thèse, nous proposons d'étudier et d'exploiter le procédé de mesure des variables d'importance du paradigme des forêts aléatoires (RF) [51].

De ce fait, la partie 4 s'articule autour des trois études et applications suivantes :

- Étude des capacités de sélection des Forêts Aléatoires RF's pour l'extraction des descripteurs médicaux pertinents [64], [62], [63] et [212].

- Proposition de mesure d'importance des variables par les forêts à inférence conditionnelle [65].

- Application de l'approche de classification semi supervisée optimisée à grande échelle de données [66] pour la sélection de variables en semi supervisé.

CHAPITRE 1

La Mesure d'importance des facteurs qui influent sur le contrôle du Kératocône par la Forêt aléatoire

1 Objectifs

Le but de cette recherche est d'identifier les facteurs importants qui influent sur le Kératocône en appliquant l'approche de sélection de variables. Le Kératocône est une maladie déformante de la cornée qui perd progressivement sa forme normalement sphérique pour prendre localement la forme d'un cône de plus en plus cambré. Cette déformation progressive, non inflammatoire, entraîne une myopie et un astigmatisme. L'évolution peut se faire avec l'apparition d'un astigmatisme irrégulier pouvant s'accompagner d'une baisse de la fonction visuelle.

Le diagnostic du Kératocône peut s'avérer difficile à poser, surtout au stade peu avancé, puisque les symptômes associés à cette maladie peuvent aussi être associés à d'autres troubles oculaires. Seuls des tests spécifiques peuvent révéler la présence de la maladie.

L'hétérogénéité et l'imperfection de détection du Kératocône constituent la difficulté majeure rencontrée lors du diagnostic. Les experts doivent faire face à cette difficulté au niveau de la modélisation des connaissances ainsi qu'au niveau des mécanismes de raisonnement.

Dans ce chapitre, nous présentons une méthode de sélection basée sur l'algorithme des *forêts aléatoires*. En plus d'être très performantes en prédiction, les forêts aléatoires calculent un indice d'importance des variables. La phase de sélection de variables appliquée par l'approche des *forêts aléatoires* de type embarqué (embedded) est comparée avec les méthodes classiques des catégories filtre, enveloppe et embarqué comme suit : Maximum de pertience et minimum de redondance, Relief, LSW, SVM-RFE, par la suite une classification est réalisée afin d'identifier les facteurs importants qui influent sur l'identification du Kératocône.

Dans ce travail [64], nous utilisons une base de données expérimentale réelle composée de 492 échantillons et 62 attributs, obtenus à partir de 246 sujets enregistrés entre 2010 et 2014. La collecte de données locale a été réalisée au niveau de la clinique LAZOUNI Tlemcen.

De ce fait, nous ciblons deux objectifs distincts en sélection de variables. En effet, nous distinguons le but d'interprétation du but de prédiction. Notre méthode de sélection de variables tente de satisfaire ces deux objectifs. Pour cela nous allons répartir ce travail

comme suit, dans la deuxième section, un état de l'art sur la classification du Kératocône et les différentes techniques exploitées y sont présentés. La troisième section concerne la description de notre propre base de données avec un exemple de lecture d'une carte de topographie cornéenne Orbascan II. La quatrième section présente les techniques de sélection employées en comparaison avec la mesure d'importance des forêts aléatoires. Vient par la suite, la section résultats et interprétations présentant aussi les avantages et inconvénients de chacune des approches de sélection employées. En dernier lieu, une conclusion est apporté qui résume les avantages à l'application de cette approche par rapport aux méthodes classiques.

2 État de l'art des travaux sur la détection automatique du Kératocône

Dans la littérature plusieurs travaux se sont intéressés à la détection automatique du Kératocône. Durant ces dix dernières années, nous notons une grande tendance des chercheurs appartenant à des disciplines diverses qui s'intéressent à la détection automatique du Kératocône et cela a conduit à l'apparition d'un grand nombre de publications. Un dénombrement de telles publications reflète la fertilité et la vitalité de ce champ.

Dans ce chapitre nous nous sommes investis à collecter, synthétiser, analyser et classifier les différentes techniques de classification du Kératocône présentes dans la littérature. Cette tâche a été très ardue vu le nombre énorme et croissant des publications mais aussi la multitude des domaines d'application et la grande diversité des solutions originales traitant le problème de la maladie du Kératocône. En effet, nous citerons par la suite uniquement les recherches réalisées sur les cartes de l'Orbscan II , que nous classerons par ordre chronologique de contribution. L'ensemble des publications et les différentes approches et méthodes sont synthétisées comme suit :

Les premiers à s'être attaqués à ce problème par approche intelligente sont Smolek et Klyce 1997 [213], ils se sont intéressés à la détection automatique du Kératocône par une approche neuronale. Le réseau de neurones pour le dépistage du Kératocône a été conçu pour détecter la présence d'un Kératocône (KC) ou suspects de Kératocône (KCS). De ces expérimentations, les auteurs ont conclu que les réseaux de neurones ont permis une bonne distinction des KC de KCS de la topographie.

Par la suite, Accardo et Pensiero [214] ont traité l'identification du début du Kératocône, ils ont émis la question de savoir si un dépistage du Kératocône doit être effectué en tenant compte des deux yeux d'un même sujet ou chaque œil séparément. Pour cela, ils ont utilisé l'approche neuronale par plusieurs combinaisons du nombre d'entrées, nœuds cachés et de sortie. Les auteurs ont prouvé que de meilleurs résultats sont réalisés par les paramètres des deux yeux d'un même sujet.

L. A. Carvalho [215] a proposé de développer un réseau de neurones artificiel qui pourrait classer les types spécifiques de formes de la cornée en utilisant des coefficients de Zernike comme entrée pour améliorer la précision de détection du Kératocône. L'auteur a collecté une base de données à partir d'un système Eyesys 2000 de 80 cornées classées en cinq catégories : (1) normales, (2) l'astigmatisme à-la- règle, (3) l'astigmatisme contre- la - règle, (4) le Kératocône, (5) post- kératomileusis assisté par laser in situ. Il a montré qu'en utilisant un système de représentation de surface de la cornée simple et bien enrichi en s'appuyant sur les informations de l'élévation de la cornée, peut alors générer des paramètres d'entrée simples qui sont indépendants de la définition de courbure et donc efficaces.

Dans [216], Mahmoud et al. présentent une simulation des indices topographiques utilisés pour la détection et l'évaluation du Kératocône et cela pour permettre leur application à des cartes acquises à partir de plusieurs machines topographiques. Les expérimentations

de ce travail par les indices simulés étaient significativement corrélées avec les indices natifs correspondants aux 3 appareils topographiques : un Tomey TMS- 1, un EyeMapAlcon et Keratron topographe. Les auteurs sont parvenus à la déduction que tous les indices simulés étaient significativement corrélés avec les indices natifs correspondants.

Une étude comparative a été proposée par Z. Schlegel, [217] entre la partie antérieure et la partie postérieure des cartes d'élévation de la cornée et entre le cas normal et le cas suspect. La méthode utilisée est de calculer les corrélations de l'élévation des paramètres entre les surfaces antérieures et postérieures. La technique des réseaux de neurones a été appliquée. Les résultats ont permis de montrer que la corrélation calculée entre la tonicité conique et la partie antérieure et postérieure de la surface de la cornée était meilleure dans le cas Kératocône suspect que celui normal. Par contre, l'asphéricité conique et la courbure apicale sont moins corrélées dans le cas Kératocône suspect que celui normal.

La technique de régression logistique a été appliquée par deSanctis et al. [218] pour soutenir des points de référence identifiés par l'analyse des courbes ROC, et pour vérifier la validité du modèle. Les courbes (ROC) des caractéristiques ont été réalisées pour déterminer l'exactitude prédictive globale de l'essai et identifier les points de coupures optimales de l'élévation de la cornée postérieure pour maximiser la sensibilité et la spécificité de la discrimination Kératocône et Kératocône infraclinique des cornées normales. La Courbe ROC a montré une grande précision globale prédictive de l'élévation postérieure à la fois pour le Kératocône et le Kératocône infraclinique (aire sous la courbe 0,99 et 0,93 respectivement). L'étude réalisée a clairement indiqué que l'élévation de la cornée postérieure permet une discrimination très efficace du Kératocône par rapport aux cornées normales.

Afin d'identifier le Kératocône à partir des cartes ORBSCAN II, Souza et al. [219] implémentent les approches : les séparateurs à vaste marge, les perceptrons multi-couche et réseau à base radial. Plusieurs paramètres ont été utilisés comme référence pour le dépistage. La validation croisée a été appliquée pour l'apprentissage et le test des classifieurs.L'ensemble des résultats de l'apprentissage sur les cartes de l'Orbscan II par les techniques : SVM, le RNS et RBF montrent clairement une bonne détection du Kératocône.

Gatinel et. Saad [220], proposent la prévention de l'ectasie cornéenne par une nouvelle méthode de détection du Kératocône fruste. La méthode statistique d'analyse discriminante a été appliquée afin de séparer les observations en les classant grâce à une fonction score de diverses variables. Pour discriminer entre les cornées saines de celles atteintes de Kératocône fruste, les auteurs ont utilisé la courbe ROC.

Un nouveau test diagnostic a été établi par Smadja et al. [221], en utilisant la puissance des techniques d'analyse discriminante. Le travail est basé sur la classification du kératocone infraclinique par la méthode des arbres de décision CART. Pour cette méthode l'auteur a introduit 55 paramètres de la cornée antérieure et postérieure. Il s'est basé pour l'évaluation des performances de l'algorithme sur les paramètres de la courbure d'élévation tachymétrie, et les paramètres de front d'onde. Les résultats avec l'arbre de décision ont permis la discrimination entre le cas normal et le cas Kératocône avec une sensibilité de 100% et une spécificité de 99,5%.

3 Contribution

Dans ce travail nous nous intéressons plus particulièrement à la capacité de sélection de variables dans les forêts aléatoires pour l'identification des facteurs influents à la détection du Kératocône. Les Forêts Aléatoires peuvent être appliqué sur des données à la fois nominales et continues. Elle a la capacité de s'adapter aux problématiques où le nombre de catégories pour les variables est trop grand pour permettre l'utilisation de

techniques classiques comme les Arbres de Décision, la Régression Logit, etc ....
De plus, sa faculté de pouvoir faire de la prévision sur un sous-ensemble aléatoire de prédicteurs apporte une solution adéquate pour résoudre les problématiques où le nombre de variables à étudier dépasse un certain seuil de compréhension.

Le Kératocône constitue actuellement un champ primordial de recherches, il y a d'importantes discordances dans les estimations du nombre de personnes atteintes de Kératocône, mais beaucoup d'études estiment que la prévalence est comprise entre une personne sur 2000 et une sur 500 [222, 223]. Il correspond à une déformation de la cornée vers l'avant, les modifications visuelles sont peu importantes au début de la maladie, et elles s'apparentent à des troubles de la vision classiques (astigmatisme, myopie).

L'objectif de ce chapitre est de proposer un modèle d'aide au diagnostic médical en ophtalmologie capable de détecter automatiquement la pathologie du Kératocône pour aider les experts du domaine à travers le topographe Orbscan II. C'est dans ce contexte, que se constitue notre travail. La collecte de données locales a été établie dans le cadre d'un projet de fin d'études de Master au sein de la clinique Lazouni.

A partir de cette base de données, nous avons pu mettre au point un modèle qui permet de prendre en charge les fonctionnalités de classification automatique du Kératocône tout en identifiant les paramètres les plus pertinents. Une étude comparative a été réalisée à travers plusieurs techniques de sélection dans l'étape de sélection pour confirmer la puissance et l'avantage d'application des forets aléatoires pour la mesure d'importance des variables.

4 Base de données

Aujourd'hui, il existe dans la littérature un grand nombre de travaux sur l'aide au diagnostic médical dans le domaine ophtalmologique, basé sur la collecte et la sélection des attributs à partir de l'appareil topographique pour renforcer le dépistage du Kératocône. De ce fait, nous avons constitué une nouvelle base de 62 variables, enrichie par de nouveaux paramètres : I-S OSI axial, Tangentiel et OSI Tangentiel (Table 28) ; ce sont des indices topographiques calculés à partir de l'analyse des données recueillies par le topographe et qui permettent d'exploiter de manière plus quantitative les informations topographiques, et servir de variables explicatives.

4.1 Lecture des cartes de topographie cornéenne

La topographie cornéenne permet de recueillir des informations relatives à la courbure ou au relief (élévation) de la cornée, grâce à la projection et l'analyse du reflet d'un motif lumineux éclairant ou balayant la cornée. Les images recueillies sont analysées de façon automatisée par un logiciel, et par la suite des cartes en couleur sont fournies au médecin pour l'interprétation [224].

Les figures suivantes (Figures 39 et 40) montrent ces différentes cartes dans les différents cas :

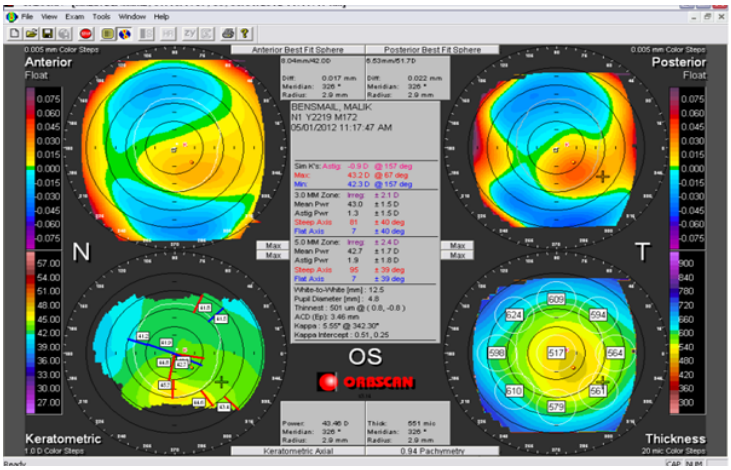


Figure 39 – Les cartes de topographie cornéenne de l'œil gauche d'un patient sain

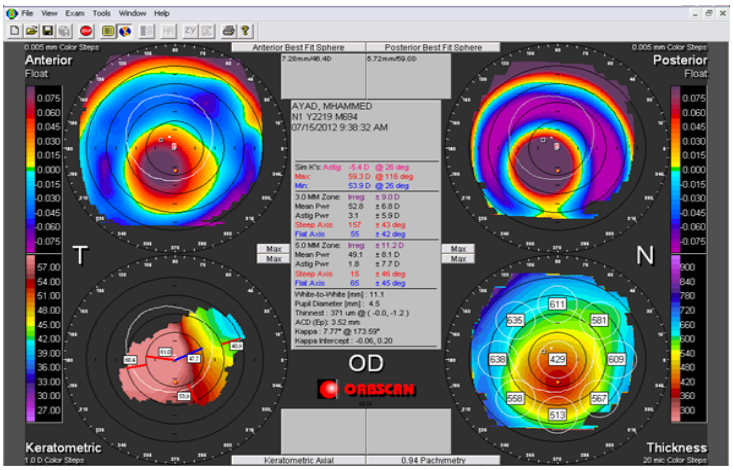


Figure 40 – Les cartes de topographie cornéenne de l'œil droit d'un patient atteint de Kératocône

- Les couleurs chaudes (jaune, orange, rouge, violet) représentent ce qui est bombé (définit les cas Kératocône).
- Les couleurs froides (bleu ciel jusqu'au vert) représentent ce qui est plat.
- Le vert est utilisé pour définir les valeurs moyennes.

La topographie standard est une analyse informatique de la surface cornéenne. Elle permet la réalisation de courbes de niveau de la surface antérieure de la cornée. Ces analyses sont indiquées dans le dépistage et le suivi de certaines affections de la cornée (Kératocône, astigmatisme) dans la chirurgie réfractive ou dans la contactologie (placement de lentilles de contact) [225].

4.2 Présentation de la base Kératocône

La collecte a été réalisée durant l'année 2014 dans le cadre de projet de fin d'études de Master [226] au niveau de la clinique LAZOUNI Tlemcen. Les données sont constituées de 492 échantillons, obtenus à partir de 246 sujets étudiés entre 2010 et 2014 qui se répartissent entre 130 femmes, 111 hommes et 05 enfants (Figure 41) dont l'âge varie de 14 ans à 81 ans. Pour chaque patient nous prenons les paramètres de l'œil gauche et l'œil droit c'est-à-dire 492 cas à traiter dont 292 sont atteints de Kératocône : 90 femmes, 53 hommes et 03 enfants (Figure 41) et 200 non Kératocône.

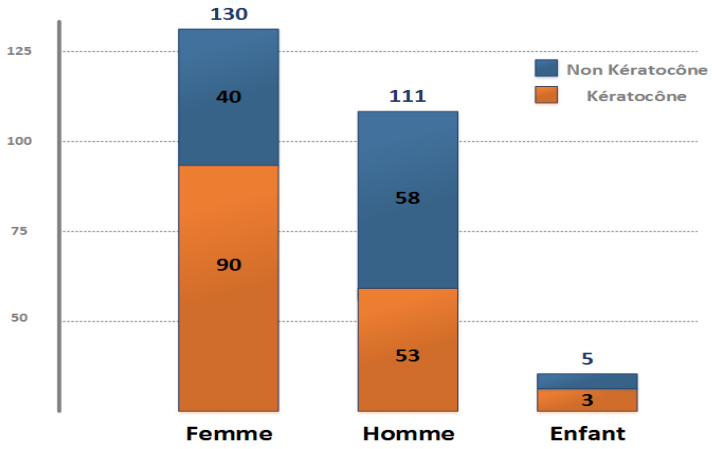


Figure 41 – Répartitions des patients dans la base de données

Chaque cas est formé de 61 attributs, nous donnons les détails des 24 majeurs attributs. Le tableau (Table 28) présente une description de ces attributs suivant la description faite par D. Gatinel [227] :

Sexe	130 femmes, 111 hommes et 05 enfants
Age	Entre 14 ans et 81 ans
Og	Œil gauche
Od	Œil droit
Kératométrie simulée SimK2	Méridien de la cornée le plus plat (la plus faible de Kératometrie) dans la zone du 3 mm
Kératométrie simulée SimK1	Méridien de la cornée le plus cambré (la plus forte de Kératometrie) dans la zone du 3 mm
K moy	La moyenne du rayon de courbure
Couleur	La couleur correspond à un rayon de courbure c'est-à-dire elle correspond à une puissance. Quand il y a beaucoup de couleurs dans une petite surface dans une image topographique (3 jusqu'à 5 mm) cela veut dire qu'il existe des irrégularités cornéennes.
Symétrie	Du fait que la forme de la cornée étant asphérique et régulière cela implique qu'on doit avoir une image presque en miroir. Dans le cas du Kératocône il n y a pas de symétrie
Axe rouge Axe bleu	C'est la même chose que la symétrie (on appelle ça les hémi méridien). Dans le cas normal les axes sont alignés c'est-à-dire symétrique par contre dans un cas de Kératocône les axes ne sont pas alignés.
Point le plus fin	Dans le Kératocône, il existe un amincissement cornéen, par contre dans le cas normal la cornée est fine au centre et s'épaissit vers la périphérie. Dans le Kératocône le point le plus fin est situé au sommet du cône qui est en général en temporal inférieur.
La pachymétrie	La pachymétrie permet la mesure précise de l'épaisseur de la cornée. Elle se mesure en micron. En Algérie la moyenne de la pachymétrie est au tour de 400 microns alors que la moyenne européenne est de 550 microns. Lorsque la pachymétrie d'un patient est inférieur à 470 microns un Kératocône est suspect
Rayon sphère antérieur/ postérieur	Ce rapport est utilisé pour voir la meilleure sphère possible on l'appelle aussi (best fit sphere). Quand ce rapport est : $<$ 1.25 cas normal, Entre 1.25 et 1.27 cas suspect, $>$ 1.27 cas très suspect
I Axial I Tangentiel	I est égal à la moyenne de 5 mesures de kératométrie sur l'anneau des 3 mm et 5 mm centraux de l'hémicornée inférieure

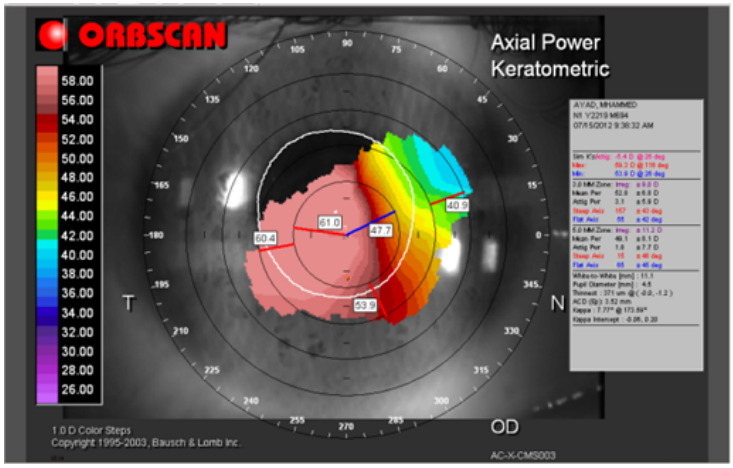
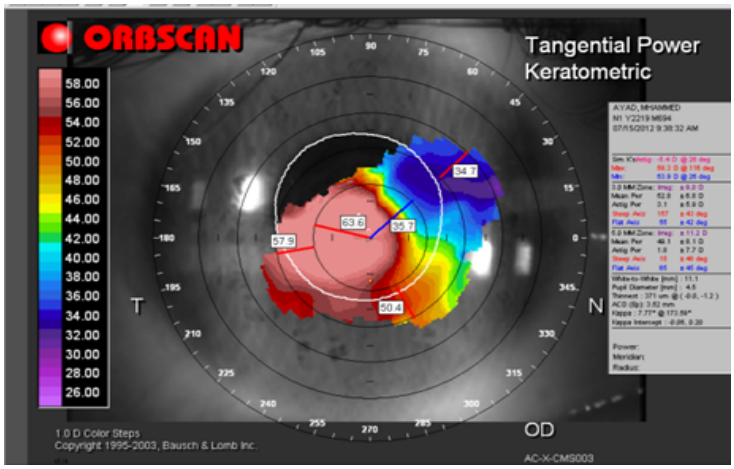
S Axial S Tangentiel	S est égal à la moyenne de 5 mesures de kératométrie sur l'anneau des 3 mm et 5 mm centraux de l'hémicornée supérieure
CYL	Cylindre Kératométrique Simulé Différence entre les valeurs de SimK 1 et SimK 2
OSI : Index de secteur opposé	La surface cornéenne est divisée en 8 secteurs égaux, dont l'angle interne est égal à 45°. La puissance moyenne de chaque secteur est calculée, et l'OSI est égal à la différence maximale mesurée entre deux secteurs opposés. Sa valeur augmente en cas d'astigmatisme irrégulier caractérisé par la présence d'un degré important d'asymétrie.
I-S Axial	Il consiste à moyenner les valeurs kératométriques obtenues en différents points de l'hémicornée inférieure (I) et leur soustraire la moyenne correspondante pour les points correspondants de l'hémicornée supérieure (S).  Fig. La topographie cornéenne (axiale)
I-S Tangentiel	 Fig. La topographie cornéenne (tangentielle)

Table 28 – Les paramètres de la base de données

Les indices calculés sont des données quantitatives qui peuvent être utilisées pour construire des modèles statistiques destinés à séparer en groupes diagnostics les cornées testées. Le logiciel du topographe au niveau de la clinique dispose des critères de Rabinowitz [228] et de Klyce et Maeda [229–231].

5 Étapes de sélection

Les techniques de sélection étudiées reposent sur l'estimation de poids (scores) correspondante à chaque caractéristique. Ces poids sont utilisés pour ordonner puis sélectionner les K parmi D descripteurs les plus pertinents.

5.1 minimum Redondance Maximum Relevance (mRMR)

" Min-Redundancy, Max-relevance" (mRMR) est une méthode de filtrage pour la sélection de caractéristiques proposée par Peng et al. en 2005 [232]. Cette méthode est basée sur des mesures statistiques classiques comme l'information mutuelle, la corrélation etc,....
L'idée de base est de profiter de ces mesures pour essayer de minimiser la redondance (mR) entre les variables et de maximiser la pertinence (MR).

Peng et al. utilisent l'information mutuelle pour calculer les deux facteurs mR et MR. Le calcul de la redondance et de la pertinence d'une variable est donnée par l'équation suivante :

$$Redondance(i) = \frac{1}{|F|^2} \sum_{i,j \in F} I(i,j)$$

$$Pertinence(i) = \frac{1}{|F|^2} \sum_{i,j \in F} I(i,Y)$$

- $|F|$: représente la taille de l'ensemble de variables.
- $I(i,j)$: est l'information mutuelle entre la i^{eme} et la j^{eme} variable.
- $I(i,Y)$: est l'information mutuelle entre la i^{eme} variable et l'ensemble des étiquettes de la classe Y. Le score d'une variable est la combinaison de ces deux facteurs tel que :

$$Score\ (i) = Pertinence\ (i) - Redondance\ (i)$$

5.2 ReliefF

Cet algorithme, introduit sous le nom de Relief dans (Kira et Rendel [233]) puis amélioré et adapté au cas multi-classes par I. Kononenko [80] sous le nom de ReliefF, ne se contente pas d'éliminer la redondance mais définit un critère de pertinence.

Ce critère mesure la capacité de chaque caractéristique à regrouper les données de même étiquette et discriminer celles ayant des étiquettes différentes (voir algorithme 11). L'analyse approfondie de ReliefF est effectuée dans (Robnik-Sikonja et Kononenko, [234]).
Max(d) (resp. min(d)) désigne la valeur maximale (resp. minimale) que peut prendre la caractéristique désignée par l'indice d, sur l'ensemble des données. $x_i d$ est la valeur de la $d$ième caractéristique de la donnée x_i.

Le poids d'une caractéristique est d'autant plus grand que les données issues de la même classe ont des valeurs proches et que les données issues de classes différentes sont bien séparées.

Sa technique aléatoire ne peut garantir la cohérence des résultats lorsque nous appliquons la méthode sur les mêmes données plusieurs fois. De ce fait, pour un modèle d'aide au diagnostic nous ne pouvons pas laisser ces paramètres instables, pour cela nous fixons dans ce projet les paramètres aléatoires par les valeurs suivantes :

- L'exemple choisi est de sortie 0, pour pouvoir extraire les valeurs des patients normaux vu par l'inéquivalance du partitionnement des échantillons de la base de données.
- La variable K pour le calcul des plus proches voisins des hits et misses est fixée à 5.

Algorithm 11 ReliefF Algorithme

1: Initialiser les poids
2: Tirer aléatoirement une donnée Xi
3: Trouver les K plus proches voisins de Xi ayant les mêmes étiquettes (hits),
4: Trouver les K plus proches voisins de Xi ayant une étiquette différente de la classe de Xi (misses)
5: Pour chaque caractéristique mettre à jour les poids

$$W_d = w_d - \sum_{j=1}^{K} \frac{diff(x_i, d_i, hits_j)}{m * k}$$

$$+ \sum_{c \neq class(x_i)} \left(\frac{p(c)}{1 - p(class(x_i))}\right) \sum_{j=1}^{K} \frac{diff(x_i, d_i, misses_j)}{m * k}$$

6: La distance utilisée est définie par :

$$diff(x_i, d_i, x_j) = \frac{|x_i d \urcorner x_j d|}{max(d) \urcorner min(d)}$$

5.3 Las Vegas Wrapper LVW

LVW est une méthode de sélection de caractéristiques proposée en 1996 par Liu et Setiono [235]. Cette méthode consiste à générer aléatoirement et à chaque itération, un sous-ensemble de caractéristiques et à l'évaluer avec un classifieur.

Algorithm 12 pseudo code *Las Vegas Wrapper (LVW)*

1: **Entrée :** Une base d'apprentissage A, Une base de caractéristiques S, Nombre d'itérations T
2: **Sorties :** S : Ensemble sélectionné, $Err = Classifieur(A; S)$, $k = 0$, $N = |S|$
3: **Répéter**
- S1 = Générer Aléatoirement $N1 = |S1|$
- Err_1 = Classifieur(A; $S1$)
- **Si** ($Err_1 < Err$) ou ($Err = Err_1$ et $N1 < N$) **alors**
- $k = 0$; $N = N1$; $S = S1$; $Err = Err_1$
- **Fin Si**
- $k = k + 1$

4: **Jusqu'à** $k = T$
5: **Retourner** S

Après avoir évalué si sa performance est meilleure que celle réalisée auparavant (au départ, l'ensemble de base est supposé comme le meilleur sous-ensemble), ce sous-ensemble devient le meilleur sous-ensemble courant. Ce processus est répété jusqu'à ce que T essais consécutifs soient infructueux pour l'amélioration. L'algorithme 12 résume le pseudo-code de cette méthode [236]. Cette méthode présente l'inconvénient de ne pas garantir l'optimalité de la solution finale ainsi qu'un temps de calcul très élevé.

5.4 La méthode de suppression récursive des paramètres par RFE-SVM

L'algorithme RFE-SVM (Recursive Feature Elimination–Support Vector Machine) proposé par Guyon [237] renvoie un classement des caractéristiques d'un problème de classification par l'apprentissage d'un SVM avec un noyau linéaire, et permet la suppression des paramètres avec le plus petit critère de classement.

RFE-SVM est une méthode basée sur l'élimination backward et utilisant les SVM pour sélectionner un sous-ensemble d'attributs optimaux non redondants. Elle est fondée sur l'estimation de poids relatifs à l'optimisation d'un problème de discrimination linéaire, ce problème étant résolu à l'aide d'une machine à vecteurs de support (SVM). La procédure de sélection est décrémentale et élimine progressivement les attributs de faible poids. L'algorithme est décomposé en trois étapes :

Tant qu'il reste des attributs,
– Apprentissage du classifieur SVM
– Calcul des w_j^2
– Supprimer le (les) attribut(s) correspondant au(x) poids le(s) plus faible(s).

5.5 Mesure d'importance par permutation des Forêts Aléatoires

En plus de construire un ensemble de prédicteurs, l'algorithme des Random Forests-RI calcule une estimation de son erreur de généralisation : l'erreur Out-Of-Bag (OOB). "Out-Of-Bag" signifie ici "en dehors du bootstrap". Cette erreur était déjà calculée par l'algorithme du Bagging, d'où la présence du mot "Bag". Le procédé de calcul de cette erreur est le suivant :

À partir d'une base d'apprentissage A de m exemples, on tire des sous-bases bootstrap de m exemples avec remise. De ce fait, pour chaque échantillon bootstrap 63,2% des exemples sont uniques de A, le reste étant des doublons. Donc pour chaque sous base 1/3 des exemples de A ne sont pas sélectionnés et sont considérés comme OOB [51]. Ils serviront à :

1. L'évaluation interne de la forêt (estimation de l'erreur de classification en généralisation de la forêt).
2. Estimer l'importance des variables pour la sélection de variables.

L'algorithme des forêts aléatoires propose également des critères permettant d'évaluer l'importance des covariables sur la prédiction de Y. Nous considérons ici la mesure d'importance par permutation due à Breiman [51]. Une variable X_j est considérée comme importante pour la prédiction de Y si en brisant le lien entre X_j et Y, l'erreur de prédiction augmente. Pour briser le lien entre X_j et Y, Breiman propose de permuter aléatoirement des valeurs de X_j.

Plus formellement, considérons une collection d'ensembles Out-Of-Bag (OOB) $\bar{D}_n^t = D_n/D_n^t, t = 1, \ldots, Ntrees$ contenant les observations non retenues dans les échantillons bootstrap. Ces ensembles seront utilisés pour calculer l'erreur de chaque arbre h. A partir de ces ensembles, définissons les ensembles out-of-bag permutés $\bar{D}_n^{tj}, t = 1, \ldots, Ntrees$ en permutant les valeurs de la j-ème variable des échantillons out of bag. La mesure d'importance par permutation est alors définie par :

$$I(X_j) = \frac{1}{Ntrees} \sum_{t=1}^{Ntrees} [R(h_t, \bar{D}_n^{tj}) - R(h_t, \bar{D}_n^t)] \tag{1.1}$$

Cette quantité est l'équivalent empirique de la mesure d'importance $I(X_j)$ comme l'ont formulé Zeng et al. [238]

$$I(X_j) = E[(Y - h(X_j)^2] - E[(Y - h(X)^2]$$

Avec : $X(j) = (X_1; \ldots; X_j^{'}; \ldots; X_p)$ est un vecteur aléatoire tel que $X_j^{'}$ est une réplication indépendante de X_j . La permutation de X_j dans la définition de $I(X_j)$ revient donc à remplacer X_j par une variable indépendante et de même loi dans eq 1.1.

6 Étapes de classification

Pour tester la pertinence des variables sélectionnées par les différentes méthodes de sélection qui ont été exposées précédemment, nous utilisons le classifieur SVM (les machines à vecteurs de supports) de la boite à outils LIBSVM [239] comme méthode d'apprentissage sur les sous-ensemble de variables sélectionnées afin de juger de leurs efficacités et valider leurs performances.

7 Expérimentations et résultats

Dans la pratique, il est très utile d'extraire des informations à partir des variables des données utilisées. Nous devons choisir les variables nécessaires pour expliquer les résultats de sortie.

Ces informations extraites peuvent être d'une grande aide dans l'interprétation des données. Elles peuvent également être utilisées pour construire de meilleurs classifieurs : un classifieur construit en utilisant uniquement les variables utiles peut être plus puissant qu'un classifieur construit avec des variables supplémentaires bruyantes.

Dans ce travail, nous nous intéressons aux forêts aléatoires pour la sélection de variables. Très simple à mettre en œuvre où la sélection d'un sous-ensemble de variables explicatives (d'importances) parmi un grand nombre, permet généralement de :

- Réduire beaucoup les temps de calcul : très fructueuse en grande dimension.
- Obtenir une plus grande variété de modèles.
- L'aggrégation des valeurs ou classes prédites (vote majoritaire) par tous les modèles générés devrait alors donner un classifieur plus robuste et plus précis.

L'application de Random Forest requiert deux paramètres principaux : Le paramètre le plus important est le nombre m de variables sélectionnées aléatoirement à chaque nœud de l'arbre. Ce paramètre nommé $mtry$ peut varier de 1 à "p" (observations). Ici, nous l'avons fixé à $\sqrt{(p-1)}$ suivant le paramétrage de Breiman [51]. Nous pouvons également ajuster le nombre d'arbres de la forêt. Ce paramètre est appelé *Ntrees* et sa valeur finale est fixée par la construction incrémentale d'arbres jusqu'à ce que l'erreur soit minimale.

Les résultats, comme on peut le voir dans la Figure 42, montrent qu'à 120 arbres l'erreur converge à 0,065 ce qui donne un taux de classification de 93.5%. Et au-delà de cette valeur, l'erreur reste à peu près constante.

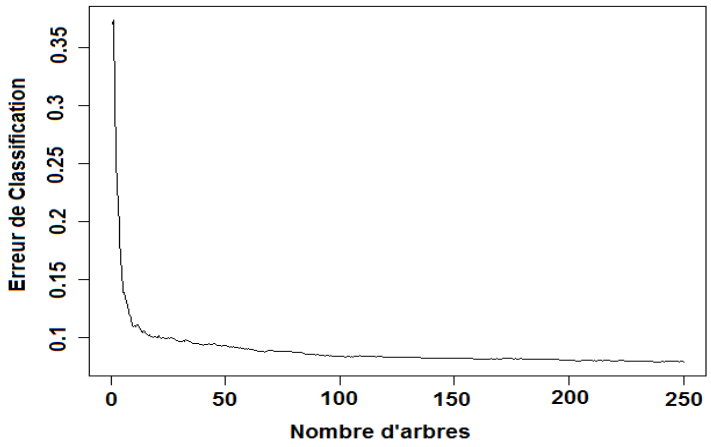


Figure 42 – Performances de classification de la base de données par la forêt aléatoire en fonction du nombre d'arbres.

7.1 Résultats

Les performances de la mesure d'importance des forêts aléatoires sont mises en évidence à travers une comparaison avec d'autres méthodes de sélection des 3 catégories : Filtre, enveloppe et embarqué de la littérature du domaine.

Dans le but d'évaluer la qualité du sous-ensemble de variables obtenues, un classifieur est formé de manière itérative sur les vingts caractéristiques les plus importantes. Nous évaluons ensuite la précision de l'ensemble de données de test avec un classifieur *SVM* (en utilisant le package *LIBSVM* [239]). Les paramètres du classifieur SVM que nous avons mis en exécution sont le noyau polynomiale du premier degré et un paramètre de régularisation.

La Figure 43 montre l'évolution de la précision en fonction des nombres de caractéristiques sélectionnées. Les résultats indiquent une amélioration importante générale de notre approche par rapport aux quatre autres approches. Comme on peut le constater dans la Figure en 43, La mesure d'importance des forêts aléatoires "RF : Random Forest" dépasse nettement les quatre autres méthodes par une marge remarquable. Sur la majeure partie des variables sélectionnées mRMR et LVW effectuent les pires performances.

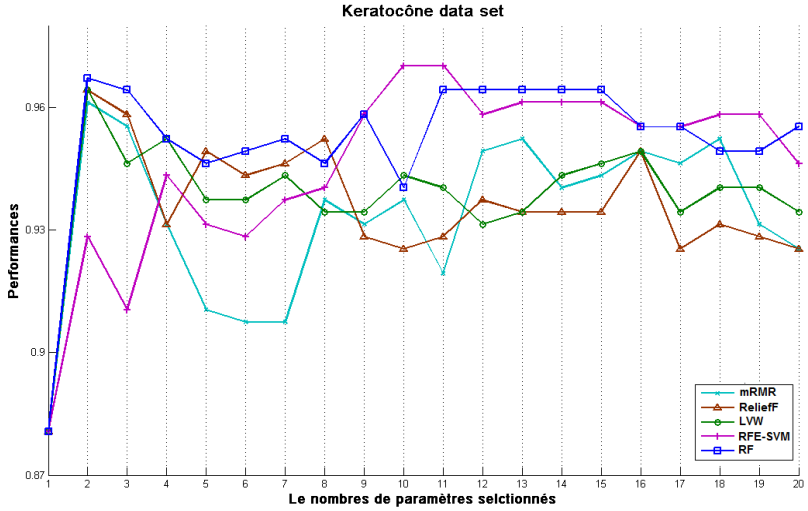


Figure 43 – Courbes de performances en fonction du nombre de variables sélectionnées

Pour plus d'exhaustivité, nous avons également calculé la précision moyenne pour les différents nombres de caractéristiques sélectionnées. Les précisions moyennes entre *RF* et les autres méthodes ainsi que les 20 premières caractéristiques sélectionnées par chacune sont décrites dans le tableau (Table 35).

Techniques	*mRMR*	*ReliefF*	*LVW*	*SVM-RFE*	*Random Forest*
Performances	0.9334	0.9384	0.9354	0.9457	0.9521

Table 29 – Le taux de performance moyen sur les 20 variables les plus pertinentes

7.2 Discussion

Les performances moyennées sur les 20 variables les plus pertinentes après le processus de sélection montrent clairement que la méthode mRMR donne en comparaison des autres approches un faible taux de classification (Table 29), cela est dû à son principe qui se base sur le calcul probabiliste de l'information mutuelle, sachant que la probabilité est un rapport de nombre de cas favorables sur le nombre de cas possibles ; mais dans

le domaine médical la présence d'une caractéristique ne change pas pour chaque classe mais plutôt sa présence qui est évidente en portant sa propre valeur, alors que MI mesure seulement l'indépendance d'attributs par rapport à la classe.

La deuxième approche de type filtre, la méthode ReliefF, réalise une légère amélioration du taux de classification (Table 29), nous remarquons dans la Figue 42 qu'elle atteint des performances très intéressantes avec un nombre de variables inférieur à 10. Son point fort est son principe qui réside dans les différents traitements, comme le calcul des hits (les exemples de la même classe) et les misses (les exemples qui ont une étiquette différente) et la normalisation des valeurs de chaque caractéristique pour éviter la dominance entre les variables et par conséquent éviter d'éliminer les attributs à faibles valeurs.

L'approche enveloppe (LVW : Las Vegas Wrapper) utilisée dans notre cas, donne des résultats d'ordre similaire aux approches filtres, malgré sa capacité d'évaluation d'un sous-ensemble de caractéristiques par la performance de classification de l'algorithme d'apprentissage. Il reste néanmoins sujet au sur-apprentissage même par l'application d'une validation croisée et son principe de recherche itératif ne permet pas de parcourir totalement l'espace de recherche.

L'algorithme de type embarqué RFE-SVM, donne des résultats très intéressants (Table 29), celui-ci exploite essentiellement la richesse du bagage théorique sur lequel sont basées les machines à vecteurs de supports (SVM). Son principe de recherche locale permet d'améliorer de façon itérative, la solution existante en explorant le voisinage de celle-ci. Cependant, cette technique permet la section par un choix localement optimal dans l'espoir que ce choix mènera à une solution globalement optimale, ce qui rend l'algorithme complètement glouton.

Si nous comparons les deux approches embarquées, RFE-SVM et la mesure d'importance de la forêt aléatoire, nous remarquons des performances voisines dans la moyenne partie des variables sélectionnées (Figure 42), ce qui confirme la supériorité des méthodes de type embarqué sur les deux autres catégories.

La démarche de sélection des forêts aléatoires, concède d'une part de sélectionner une batterie de caractéristiques discriminantes et explicatives et d'autre part d'élaborer un modèle prédictif d'une situation donnée en ne faisant aucun a priori sur les variables en entrée. En effet la méthode Random Forest a l'avantage de donner à toutes les variables le même statut.

De ce fait, le principe de permutation aléatoire d'une variable étudiée, permet de hiérarchiser les variables si la modification de sa valeur pour un individu donné entraine sa mauvaise classification. Cette méthode intégrée dans une approche d'ensemble de classifieurs se distingue par ses performances ainsi que la qualité des variables sélectionnées (Table 30).

En effet, les variables sélectionnées (Table 30) par la mesure d'importance par permutation des forêts aléatoires regroupe les 8 variables utilisées par les experts ophtalmologiques dans leur routine clinique, à savoir : le cylindre kératométrique, la pachymétrie, le point le plus fin, le nombre de couleurs postérieur et antérieur, le BFS, OSI et la symétrie en rouge et celle en Bleu.

mRMR	***ReliefF***	***LVW***	***SVM-RFE***	***Random Forest***
Nbre de couleurs post 5mm	Nbre de couleurs post 5mm	Nbre de couleurs post 5mm	Nbre de couleurs post 5mm	Nbre de couleurs post 5mm
Pts le plus fin	Pachy	Pachy	Symétrie en bleu	OSI Tangentiel Sup
Pachy	Pts le plus fin	Age	Nbre de couleurs post 5mm	Œil G/D
Degré K1	Degré K1	Tangentiel 5mm D5 inf	Axial 3mm inf	Sexe
Degré K2	OSI Tangentiel Sup	Tangentiel 5mm D5 sup	Tangentiel 3mm D3 inf	Symétrie en bleu
Sexe	Sexe	Tangentiel 5mm D5 inf	Symétrie de l'image	pts le plus fin centré
Age	Tangentiel 3mm D3 inf	Tangentiel 3mm D3 sup	Axial 5mm D5 inf	Symétrie en Rouge
Axial 3 mm D3 inf	Axial 3 mm D3 inf	Tangentiel 5mm D5 sup	OSI Tangentiel Sup	BFS Post
Axial 3 mm D3 inf	Degré K2	Tangentiel 5mm D5 inf	Axial 3 mm D3 inf	Symétrie de l'image
K1	Axial 5mm inf	Tangentiel 5mm D5 inf	Sexe	BFS Ant
Tangentiel D3 3mm inf	K1	Axial 5mm inf	K1	Cylindre kératométrique
Tangentiel D3 3mm inf	Tangentiel 3mm D3 sup	Tangentiel D3 3mm sup	Axial 5 mm inf	Nbre de couleurs ant 5mm
OSI Tangentiel Sup	Tangentiel D3 3mm inf	Axial 5 mm D5 sup	Tangentiel D3 3mm inf	Tangentiel D3 3mm sup
Axial 5 mm D5 inf	Axial 3 mm D3 inf	Axial 3 mm D3 sup	Axial 3 mm D3 inf	Axial 3 mm D3 sup
Axial 3 mm D3 inf	Age	Tangentiel 5 mm D5 inf	Axial 3 mm D3 sup	Axial 5 mm D5 inf

Table 30 – Les 20 variables les plus pertinentes sélectionnées par les différentes techniques

7.3 Synthèse sur les techniques de sélection

L'analyse de l'existant nous a permis de noter et résumer dans la Table 31 les différentes caractéristiques de chaque méthode lors de son fonctionnement et son traitement durant le processus de sélection.

En examinant bien ces différentes techniques (Table 31), nous remarquons clairement la complémentarité des méthodes. Plusieurs idées peuvent émerger comme l'hybridation entre les méthodes qui sont très performantes au calcul du poids significatifs et celles qui peuvent diminuer les redondances.

Techniques	Types	Avantages	Inconvénients
mRMR	*Filtre*	– Élimine la redondance. – Prend en compte les interactions avec les variables.	Calcul probabiliste ne reflète pas le poids significatif des données biologiques.
ReliefF	*Filtre*	– Précision sur des données bruitées. – Mesure la pertinence globale pour les caractéristiques.	Sa stratégie de recherche est aléatoire.
LVW	*Wrapper*	Les sous-ensembles de caractéristiques sélectionnées sont bien adaptés à l'algorithme de classification utilisé	– Est sujet au sur-apprentissage. – La dépendance des caractéristiques pertinentes sélectionnées par rapport au classificateur utilisé. – La complexité et le temps de calcul nécessaire pour la sélection.
SVM-RFE	*Embedded*	– Utilise les opérateurs de recherche locale qui peuvent être à la fois intuitifs et très efficaces pour chercher à améliorer, de façon itérative, la solution existante en explorant le voisinage de celle-ci. – Moins coûteux en temps de calcul et moins enclins au sur-apprentissage. – La phase de recherche est guidée par le processus d'apprentissage.	Algorithme glouton (i.e. ne fait pas de retours en arrière)
Random Forest	*Embedded*	– Combine le processus d'exploration avec l'algorithme d'apprentissage sans étape de validation, pour maximiser la qualité de l'ajustement et minimiser le nombre d'attributs – Très simple à mettre en œuvre.	Le coût numérique est important.

Table 31 – Caractéristiques des différentes techniques de sélection.

8 Conclusion et Perspectives

La littérature abondante s'intéresse depuis plusieurs décennies sur le problème de sélection de variables (features selection) témoignant ainsi de son importance mais aussi sur ses difficultés ; choisir a priori les caractéristiques pertinentes pour une application donnée n'est pas facile et plus particulièrement dans le domaine médical.

Notre démarche de sélection des identificateurs de la maladie du Kératocône consiste à comparer l'efficacité de plusieurs méthodes de sélection qui peuvent être intégrées dans un processus d'une approche filtre, enveloppe ou embarquée ceci afin de mettre en évidence la transparence du système, avec comme objectif d'extraire les plus pertinents et les plus informatifs. Les expérimentations réalisées ont permis d'évaluer les performances des résultats avec le classifieur SVM.

Nous avons constaté que les forêts aléatoires pouvaient s'avérer très utiles pour faire de la sélection de variables. L'indice d'importance distingue bien les bonnes variables des variables bruitées. Notre procédure automatique permet alors de proposer des sous-ensembles de variables très pertinents, de façon encore relativement rapide.

Bien que les résultats obtenus soient intéressants et encourageants, beaucoup de points sont susceptibles d'être étudiés dans le cadre de travaux futurs, tels que :

- L'utilisation d'autres mesures de sélection de variables pour mettre en valeur les différentes relations entre variables.
- D'après l'étude des avantages et des inconvénients des méthodes de sélection utilisées dans ce travail, une hybridation entre les techniques est envisageable ou la fusion entre les points forts de ces méthodes.
- Faire appel aux méthodes de boosting pour améliorer encore plus le taux de classification.

Ce domaine de recherche restera toujours actif tant qu'il est motivé d'une part par l'évolution des systèmes de collecte et de stockage des données et d'autre part par les exigences de performances. Pour bien juger cette sélection l'approche adéquate serai de collaborer avec des experts (médecins spécialistes) pour une meilleure interprétation des résultats.

Cette collaboration avec les experts permet de nous orienter vers la manière d'utiliser ces données fondamentales en pratique clinique et leurs influences sur la prise en charge des patients, puisque ce domaine de recherche est majeur dans le : dépistage, traitement et prédiction de l'évolution clinique de ces patients.

CHAPITRE 2

Application des Forêts à Inférence Conditionnelle pour la mesure d'importance des variables

1 Objectifs

La mesure d'importance des variables par les forêts aléatoires (RF) a reçu une attention accrue comme moyen de sélection de variables dans les tâches de classification. Le calcul d'importance d'une variable dans les forêts aléatoires est une façon intelligente de sélectionner les variables dans de nombreuses applications de manière embarquée. Cependant, leur efficacité n'est pas fiable dans les situations où les variables prédictives potentielles varient dans leur échelle de mesure ou de catégories.

Dans ce chapitre, nous avons mis en place la Forêt Aléatoire construite par les arbres à inférence conditionnelle (*Conditional Inference Tree* "CIT") qui est appelée Forêt à Inférence Conditionnelle (*Conditional Inference Forest* "CIF"). Sa spécificité réside dans le fait, que dans chaque arbre de la Forêt à Inférence Conditionnelle, la division des nœuds est basée sur la façon d'avoir une bonne associativité par inférence conditionnelle. La méthode statistique $p-value$ est utilisée pour mesurer l'association. En plus d'identifier les variables qui améliorent la précision de classification, la méthodologie identifie aussi clairement les variables qui sont neutres à l'exactitude, et aussi également qui interfèrent dans la bonne classification.

2 État de l'art du domaine

Les arbres classiques ont toujours été utilisés pour la sélection de variables d'importance. Le critère de découpage le plus commun dans l'arborescence de type CART est l'indice de Gini qui vérifie la pureté des nœuds résultants dans l'arbre afin de trouver une division favorable. Plusieurs travaux se sont interessés à la recherche d'une scission favorable, les chances de trouver une bonne division augmente si la variable est continue ou a plusieurs catégories. Par conséquent, même si la variable est non informative, elle pourrait être le nœud père dans la structure hiérarchique de l'arbre.

Le critère de sélection est appliqué sur les variables de type continu et avec une ou plusieurs catégories. L'origine de cette sélection est l'indice 'Gini' pour la division au niveau du noeud (tout en construisant l'arbre) et pour la sélection de variables (en général sur la base de la fréquence ou de la variable choisie pour la division).

Des travaux récents par Abdel-Aty et al. [240] et Harb et al. [241] ont utilisé l'algorithme des forêts aléatoires pour déterminer les variables d'importance. Cependant, Strobl et al. [242] ont montré que la méthode d'échantillonnage (bootstrapping) et l'utilisation de l'indice de Gini dans la sélection tend a biaiser le processus d'extraction de variables d'importance.

Dans cette étude, nous proposons l'application des arbres à inférence conditionnelle, développés par Hothorn et al. [243], et leurs forêts pour la sélection variable, où la procédure de division et la séparabilité des variables sont essentielles pour le développement des arbres.

L'implémentation de cette nouvelle méthodologie permettra d'améliorer les performances de classification et de sélection de variables grâce au principe de division Khi-2 adopté par les arbres à inférence conditionnelle.

Dans la littérature, peu d'études mettent en œuvre cette approche, nous les citons comme suit :

Das et al. Ont proposé dans leur papier [244] une méthodologie basée sur la Forêt à Inférence Conditionnelle pour identifier les variables d'intérêt sur les données du trafic routier (Roadway Characteristics and Inventory database (RCI)). Les auteurs ont réalisé de bons résultats comparés à la forêt aléatoire classique qui leur ont permis d'identifier les endroits où les accidents routiers graves ont tendance à se produire.

Auret et Aldrich présentent dans [245] des mesures d'importance de variables associées aux forêts aléatoires, les forêts à inférence conditionnelle et les arbres stimulés, sur un ensemble de bases de données afin de comparer ces méthodes. Les résultats montrent que les mesures d'importance de variables basées sur les forêts à inférence conditionnelle semblent trouver un bon équilibre entre l'identification des variables significatives et écartent la signalisation inutile des variables corrélées.

Strobl et al. dans [246] mettent en place une Toolbox officielle sous langage *R* nommée *cforest* (Forêt à Inférence Conditionnelle) comme une mise en œuvre alternative de la forêt aléatoire, qui fournit la sélection des variables impartiales dans l'arborescence du classement individuel.

Nagy et al. [247] ont appliqué les techniques : de régression logique, CART et arbres à inférence conditionnelle pour prévoir les facteurs de risque des comportements stéréotypés chez les chevaux. Les deux méthodes CART et CIT atteignent la même précision de prédiction, mais elles aboutissent à une meilleure extraction des facteurs de risque que ceux de la méthode de régression logique.

Nicodemus et Malley [248] ont réalisé une comparaison de trois algorithmes : forêt aléatoire (RF), la Forêt à Inférence Conditionnelle (CIF) et la régression logique de Monte Carlo (MCLR) en combinaison avec la fonction de permutation (VIM) pour la mesure d'importance de variable. Cette fonction est reconnue pour être une méthode puissante pour les ensembles de données à haut débit. Ils ont montré que les distributions de RF VIM étaient sensibles à la structure de corrélation tandis que CIF et MCLR VIM ont été observées à la fois impartiales et moins influencées par la corrélation.

Dans [249] Strempel et al. Implémentent les arbres à inférence conditionnelle et la méthode des forêts aléatoires pour trouver des relations entre les descripteurs physico-chimiques et le facteur de bioconcentration (valeurs BCF), et réalisent des résultats très intéressants.

Récemment, Guelman et al. [250] ont proposé une nouvelle méthode nommée arbres d'inférence conditionnelle causale et son prolongement naturel aux forêts à inférence

conditionnelle causale, pour un traitement optimal personnalisé des règles appliquées dans les interventions de marketing. La méthode de partitionnement récursif impartial proposée par Hothorn et al. [243], à motivée l'idée majeure de cette méthode. Les arbres à inférence causale se caractérisent alors par la séparation entre la sélection de variable et la procédure de séparation, couplée avec un critère d'arrêt statistique efficace basé sur la théorie des tests de permutation développée par Strasser et Weber [251].

Cette approche statistique empêche le sur-apprentissage, sans exiger aucune forme d'élagage ou de validation croisée. Il permet également d'éviter les biais de sélection covariante avec de nombreuses divisions possibles. Les résultats de performances mesurées sur des données synthétiques montrent que la méthode des forêts à inférence conditionnelle causale surpasse souvent les solutions alternatives décrites dans cet article.

Un autre travail de Bessonov et al. [252] sur une application bio-informatique, a permis d'évaluer les performances des méthodes CIT et CIF pour la prédiction du réseau de régulation (GRN) à partir des données d'expression de gènes. Les performances de chaque méthode ont été évaluées par l'aire sous la courbe (AUROC) et (AUPR). Les résultats préliminaires montrent que CIT et CIF prédisent avec succès GRNs à des taux de rendements acceptables mais pas optimaux. Étonnamment, les auteurs ont démontré qu'en utilisant le schéma d'agrégation actuel de mesure d'importance sur les données à grande dimension, un arbre seul CIT donne de meilleures performances que la forêt CIF dans les 5 réseaux.

Dans cette étude [65], nous exploitons l'idée majeure d'inférence conditionnelle dans la Forêt Aléatoire pour la sélection de variables et la procédure de construction des arbres. Les procédures de partitionnement et de choix de variables de séparation sont essentielles pour le développement d'arbres sans tendance à avoir de nombreuses divisions. Notre but est de mettre en évidence la transparence de notre système et d'identifier les variables les plus pertinentes et les plus informatives.

Ce chapitre est réparti comme suit. Dans la section 2, nous donnons les détails des méthodes utilisées et notre approche proposée. Les résultats et la discussion de ces résultats sont présentés dans la section 3. Enfin, dans la section 4, nous concluons ce travail en mettant l'accent sur l'importance de cette méthode dans le domaine de la sélection de variables.

3 Méthodes

3.1 La forêt Aléatoire

Les Forêts Aléatoires (RF), est un algorithme d'ensemble des arbres de type CART (Classification And Regression Tree) indépendants et non élagués, développés par Breiman [51] pour remédier au problème d'instabilité des arbres de décision. L'aspect aléatoire des forêts est accompli en sélectionnant un échantillon bootstrap de l'ensemble d'apprentissage et cela en divisant chaque nœud par la meilleure répartition entre sous-ensemble de variables choisies au hasard.

Les arbres CART [74] appliqués utilisent l'indice de Gini (eq. 2.1) pour mesurer l'impureté des nœuds, elle est maximale lorsque le nœud est divisé en parts égales entre toutes les classes et zéro dans les cas où le nœud appartient à la même classe. Par conséquent, la meilleure partition est avec un indice de Gini minimal. La procédure de partitionnement est arrêtée lorsque tous les nœuds sont purs.

$$i(t) = \sum_{k=1} p(k|t)p(l|t) \tag{2.1}$$

Où : $k, l = 1, \ldots, K$ - indice de classe ;

$p(k|t)$ - la probabilité de la classe k réalisée pour le nœud t.

Cependant, comme vu dans les parties précédentes, cette approche présente certains inconvénients :

- La modification de l'échantillon d'apprentissage pourrait conduire à des changements radicaux dans l'arbre de décision : augmentation ou diminution de la complexité de l'arbre, les changements dans les variables et les valeurs de partitionnement.
- Le sur-apprentissage
- La sélection de variables de séparation est biaisée en faveur de variables avec un nombre élevé de catégories. (Hothorn et al, [253])

3.2 La Forêt et l'Arbre à Inférence Conditionnelle

Traditionnellement, les arbres de classification Breiman et al. [74] sont utilisés pour déterminer la variable d'importance dans une variété de travaux et études. Les arbres de décision sont des structures en forme d'arbres représentant des ensembles de décision qui sont auto-générés pour la classification d'un ensemble de données (par opposition à un échantillon), dans un ordre hiérarchique, en utilisant des algorithmes tels que ID3 et ses améliorations C4.5 et C5.0, ainsi que CART et CHAID (Quinlan [82] ; [78]).

Tandis que la forêt aléatoire classique est une mise en œuvre d'un ensemble d'arbres de type CART comme base d'apprenants ; la Forêt à Inférence Conditionnelle (CIF) utilise les arbres à inférence conditionnelle (CIT). Dans CIF, le système d'agrégation fonctionne en appliquant la moyenne des poids d'observation extraits de chaque arbres et non pas en appliquant la moyenne des prévisions directement comme dans la Forêt aléatoire.

L'arbre à inférence conditionnelle est une nouvelle méthode, qui estime une relation de régression par partitionnement récursif binaire pour les variables continues, manquantes, ordonnées, nominales et multivariées dans un cadre d'inférence à conditionnelle. Ces dernières sont considérées à un certain égard, supérieures aux méthodes de régression linéaire traditionnelle ou de régression par étapes, car elles ne reposent pas sur des hypothèses sous-jacentes de la linéarité. En outre, ce modèle de prévision répond à une préoccupation rencontrée par les arbres classiques qui est le sur-apprentissage des échantillons. Le «sur-apprentissage» fait référence à la question d'un modèle statistique approprié qui approche l'erreur aléatoire ou celles des fluctuations mineures des données, fonctionnant ainsi comme un mauvais outil de prédiction.

Un autre problème fréquemment rencontré avec les différents types d'arbres de décision est le biais dans la sélection des variables de partition pour avoir des seuils plus naturels (Shih et Tsai [254]).

Généralement, les partitions obtenues à partir d'arbres à inférence conditionnelles ont été signalées à être plus près de la vraie partition de données par rapport à celles réalisées à partir d'une procédure de recherche exhaustive avec l'élagage (Hothorn et al, [243]).

CIT utilise une procédure de mesure d'association pour sélectionner les variables de séparation au niveau des nœuds à la place de la sélection de la variable par optimisation d'une mesure d'information utilisée par CART.

Les deux algorithme CART et CIT effectuent de manière récursive des divisions uni-variées de la variable dépendante fondée sur des valeurs d'un ensemble de variables. CART emploie l'indice de Gini comme mesure d'information pour la sélection de la covariable de séparation, tandis que CIT utilise un schéma de sélection des covariables qui est basé sur la théorie statistique (c.à.d sélectionné par les tests d'association en fonction de permutation). Par conséquent, la partition des nœuds est choisie en fonction de la qualité de l'association, et non pas par sa pureté comme dans l'arbre CART. Le nœud résultant

doit avoir une association ultérieure avec la valeur observée de la variable dépendante.

L'arbre à inférence conditionnelle utilise le test $p - value$ pour mesurer l'association. Cette approche élimine les biais dus aux catégories mais permet aussi de choisir les variables qui sont informatives.

L'algorithme de l'arbre à inférence conditionnelle CIT fonctionne comme suit (Algorithme 13) :

Algorithm 13 Les étapes de construction d'un arbre à inférence conditionnelle

Étape 1 Tester l'hypothèse nulle globale d'indépendance entre l'une des variables d'entrée et la réponse (qui peut être multivariée).
Étape 2 *Si* cette hypothèse ne peut être rejetée
Alors Arrêt.
Étape3 *Sinon* sélectionner la variable d'entrée avec l'association la plus forte à la réponse. Cette association est mesurée par une p-valeur correspondante à un test de l'hypothèse nulle partielle d'une variable d'entrée unique et de la réponse.
Étape 4 Mettre en œuvre une scission binaire dans la variable d'entrée sélectionnée.
Étape 5 Répéter les étapes 1) et 4).

Les CITs utilisent un cadre unifié développé par Strasser et Weber [251], pour l'inférence conditionnelle, ou les tests de permutation. Le critère d'arrêt à l'étape 1) est soit basé sur la multiplicité ajustée des p-valeurs ou sur les valeurs p-univariées. Dans les deux cas, le critère est maximisé, à savoir, $1 - p - value$ est utilisé. Une partition est mise en œuvre lorsque le critère dépasse la valeur donnée par min-critère comme spécifié dans CIT. Par exemple, quand min-critère = 0,95, la p-valeur doit être inférieure à 0,05, afin de diviser ce nœud. Cette approche statistique garantit que l'arbre est cultivé à la bonne taille et aucune forme de validation croisée n'est nécessaire. Le choix de la grandeur d'entrée à diviser se base sur les p-valeurs de manière à éviter une sélection de variables biaisée vers des variables d'entrées avec plusieurs seuils possibles.

Les arbres à inférence conditionnelle présentent plusieurs avantages (Hothorn et al. [243]) :

- Ils ne sont pas biaisés,
- Ils ne souffrent pas de sur-apprentissage,
- La précision de la prédiction des arbres à inférence conditionnelle équivaut à la précision de la prédiction des arbres élagués optimaux.

3.3 Forêt à Inférence Conditionnelle (CIT) vs. Forêt Aléatoire (CART)

Il convient de noter, qu'il y a des choses réalisables avec CIF qui ne peuvent pas être faites avec la forêt aléatoire, par exemple la construction d'une forêt en mode non supervisé (voir Hothorn et al., [253]) ou par des données multivariées ou multi-étiquettes.

En outre, lorsque les prédicteurs varient dans leur échelle de mesure et le nombre de catégories, la sélection de variables et la mesure d'importance d'une variable sont biaisées dans les forêts aléatoires en faveur des variables avec de nombreux seuils potentiels. Tandis que dans les Forêts à Inférence Conditionnelle CIF, les arbres sont impartiaux et un régime adéquat de ré-échantillonnage est utilisé par défaut. Voir Hothorn et al. [243] et Strobl et al. [242].

4 Résultats et Interprétations

Dans la pratique, il est très utile d'avoir des informations sur les données employées. Comme par exemple d'identifier « *Quelles sont les variables qui sont vraiment nécessaires pour expliquer les résultats de prédiction ?* » ; « *Quelles sont les variables qui peuvent être rejetées ?* ».

Ces informations peuvent être d'une grande aide dans l'interprétation des données. Elles peuvent également être utilisées pour construire de meilleurs prédicteurs : un prédicteur construit à l'aide de variables pertinentes peut devenir plus performant qu'un prédicteur construit avec des variables supplémentaires et bruyantes.

Dans ce chapitre, les résultats obtenus avec chaque procédé proposé seront examinés et comparés à des algorithmes de la littérature. Les performances de la forêt à inférence conditionnelle ont été évaluées sur six bases de données biologiques. Les bases de données de nos expériences sont décrites dans la Table 32.

Bases	**# Inst.**	**# Variables**	**# Classes**	**Ref**
Colon	62	2000	2	[127]
Data_C	60	7129	8	[132]
Leukemia	72	7129	2	[130]
Lung	197	233	4	[84]
Lymphoma	96	4026	10	[84]
Prostate	102	12533	2	[128]

Table 32 – Les paramètres des bases de données biologiques

Dans ce travail, nous nous concentrons sur la capacité de mesure d'importance de la Forêt à Inférence Conditionnelle. Très simple à mettre en œuvre pour la sélection d'un sous-ensemble de variables explicatives (informatives) à partir d'un grand ensemble de variables, elle permet généralement de :

- Réduire considérablement le temps de calcul spécialement pour les bases de données à grande dimension.
- Obtenir une plus grande variété de modèles.
- L'agrégation des valeurs prédites ou classes par calcul de la moyenne de tous les modèles devrait alors donner un classifieur plus robuste et précis.

La Forêt à Inférence Conditionnelle à pour principe d'ajuster les corrélations entre les variables prédictives. Si la condition est vraie, l'importance de chaque variable est calculée en permutant dans une grille définie par les variables qui sont associées (avec (1 - $p - value$) supérieure à un seuil) à la variable d'intérêt. Le degré d'importance de la variable représente l'effet d'une variable dans les deux effets : prédiction et d'interaction. L'étude de Strobl et al. [246] le démontre clairement dans le cadre de régression et de classification.

Pour prouver l'efficacité de cette méthode, les performances de la Forêt à Inférence Conditionnelle ont été évaluées et comparées avec : l'arbre de décision (CART), l'arbre à inférence conditionnelle (CIT) et la Forêt Aléatoire (RF). Nous avons fixé le nombre de taille de l'ensemble qui sera utilisé à 100 arbres. Les résultats sont résumés dans le Tableau (Table 33).

Bases	CART	Temps	CIT	Temps	RF	Temps	CIF	Temps
Colon	60.00 ± 0.0784	0.7990801 secs	68.75 ± 0.0832	1.532253 mins	65.63 ± 0.0853	2.473097 mins	75.00 ± 0.0778	2.436517 mins
Data_C	62.86 ± 0.0849	6.675668 secs	65.55 ± 0.0304	3.570362 mins	75.00 ± 0.0784	5.530348 mins	80.67 ± 0.0849	8.230207 mins
Leukemia	52.50 ± 0.0775	8.658111 secs	58.54 ± 0.0139	4.042948 mins	65.38 ± 0.0699	7.734867 mins	69.23 ± 0.0785	10.359085 mins
Lung	86.67 ± 0.0333	0.188019 secs	92.19 ± 0.0418	0.4240429 secs	90.48 ± 0.0288	0.2300229 secs	94.29 ± 0.0228	0.8880481 secs
Lymphoma	55.00 ± 0.0139	2.070523 secs	57.31 ± 0.0699	2.053651 mins	61.77 ± 0.0421	3.115863 mins	62.08 ± 0.0771	5.136490 mins
Prostate	81.14 ± 0.0484	22.47125 secs	82.69 ± 0.0530	4.154202 mins	85.71 ± 0.0797	4.428232 mins	86.29 ± 0.0750	20.192710 mins

Table 33 – Performances de classification par la Forêt à Inférence Conditionnelle en comparaison avec d'autres méthodes.

La Forêt à Inférence Conditionnelle a amélioré la performance de la Forêt Aléatoire, celle ci peut être observée avec les différents ensemble de données biologiques. En effet, le tableau (Table 33) montre que pour les deux ensembles de données *Colon* et *Lung*, l'Arbre à Inférence Conditionnelle (CIT) seule obtient de meilleurs résultats sur la forêt aléatoire. Cela revient au processus de division CIT au niveau des nœuds qui est basé sur la qualité de l'association et non pas sur sa pureté comme dans l'arbre CART. Le nœud qui en résulte doit avoir une association ultérieure avec la valeur observée de la variable dépendante. Les résultats obtenus démontrent clairement la supériorité de CIT en comparaison avec RF dans le cas de bases de données de moyenne dimension.

Nous indiquons également dans le tableau (Table. 33), le temps d'exécution avec le même matériel d'exécution (i7-4720HQ CPU @ 2.60GHz, 16Go RAM MATLAB R2014a) des quatre algorithmes pour chaque ensemble de données. Il convient de noter que, bien que CIF obtient des améliorations sur l'ensemble de données en générale, le temps d'exécution est beaucoup plus important dans les méthodes d'ensemble (RF et CIF) par rapport aux classifieurs simple CIT et CART. Enfin, pour chaque ensemble de données, l'algorithme CIF montre un bon compromis en termes de performance et de temps d'exécution et prouve ainsi sa robustesse lorsque le nombre de variables est important.

Afin d'évaluer la qualité du sous-ensemble de variables obtenu par le modèle CIF en comparaison avec ceux de la Forêt Aléatoire RF, nous suivons le même protocole d'évaluation utilisée par [66, 255, 256] afin de mieux évaluer la qualité du sous-ensemble de variables choisies par chaque algorithme (Figure 44). Pour ce faire, nous lançons l'apprentissage à travers *SVM* (en utilisant le package *LIBSVM* [239]) sur l'ensemble de données biologiques et évaluons la précision par l'intégration au fur et à mesure des variables pertinentes générées par chaque approche.

Les nomogrammes dans la figure 44 montrent l'évolution de la précision en fonction du nombre de variables sélectionnées. Les résultats indiquent une amélioration significative de l'ensemble de l'approche Forêt à Inférence Conditionnelle par rapport à la méthode Random Forest. Comme nous pouvons le voir dans la figure 44, CIF dépasse largement la mesure d'importance de la méthode RF par une marge remarquable, il est plus intéressant dans les grands ensembles de données comme : *Prostate, Data_C, Leukemia*.

Un examen plus approfondi révèle que, les différentes courbes de la précision des variables importance sélectionnée par CIF augmente généralement rapidement au début (le nombre de variables sélectionnées est petit) et ralentit par la suite. Ceci suggère que CIF classe les variables les plus pertinentes dès le début. Ainsi, un classifieur peut obtenir un bon classement avec seulement les 5 meilleurs caractéristiques tandis que la forêt Aléatoire nécessite plus de variables pour obtenir des résultats comparables.

Figure 44 – La courbe de performances des Forêts Classique et à Inférence Conditionnelle en fonction du nombre de variables sélectionnées pour chaque base de données biologiques

4.1 Discussion

Contrairement aux conclusions données par Nagy et al. [247] et Berzal et al. [257], nos résultats vont avec les conclusions de Hothorn et al [243]. En effet, les différents critères de fractionnement ont un impact sur la précision de la classification. Nous avons constaté une réelle différence en ce qui concerne la structure et la précision entre la forêt aléatoire classique et la Forêt à Inférence Conditionnelle.

Plusieurs études et expérimentations ont prouvé que lorsque la mesure d'importance au niveau de la forêt aléatoire classique est appliquée sur différents types de données, les résultats sont trompeurs, car les variables prédictives sous-optimales peuvent être artificiellement préférées dans la sélection des variables. En effet, deux mécanismes sous-jacents à cette mesure sont biaisés ce qui se reflète dans la sélection de variables. D'une part dans les arbres de classification individuels utilisés pour construire la forêt aléatoire, et d'autre part, des effets induits sur l'échantillonnage bootstrap avec remplacement. En considérant l'inférence conditionnelle de la forêt CIF et sa mesure d'association comme critère de fractionnement, cette approche permet d'éliminer le biais dans la sélection des variables induites par l'indice de Gini.

En outre, nous remarquons que lorsque la Forêt CIF est appliquée à l'aide de sous-échantillonnage avec remplacement, résultent des mesures d'importance variable très

fiable pour la sélection de variables, même dans des situations où les variables prédictives potentielles varient dans leur échelle de mesure ou le nombre de catégories.

Le taux d'erreur obtenu lors de l'utilisation de la sélection de variables correspond à une diminution significative par rapport aux résultats appliqué à l'ensemble de variables entier. En revanche, la mesure d'association mise en œuvre dans ce procédé tend vers un mécanisme de sélection de variables non biaisé. Nous pouvons conclure en disant que l'approche proposée fournit une sélection de variables impartiale pour une plus grande précision.

5 Conclusion

La Forêt à Inférence Conditionnelle CIF, diffère de la Forêt Aléatoire RF classique sur deux façons. Les apprenants de base utilisés sont des arbres à inférence conditionnelle (Hothorn et al. [243]). En outre, le schéma d'agrégation fonctionne en faisant la moyenne des poids d'observation extraits de chacun des arbres construits à la place du vote majoritaire appliqué dans la version originale.

Recourir aux arbres à Inférence Conditionnelle CITs au lieu des arbres de Classification et de régression CART pour identifier les variables d'importance a permis d'aboutir à une sélection précise et concise malgré les natures différentes des données.

Cette étude expérimente la Forêt à Inférence Conditionnelle comme un modèle de sélection de variables. Nous avons mis en œuvre la Forêt à Inférence Conditionnelle, par un ensemble d'arbres d'inférence conditionnelle tout en exploitant leur capacité de sélection de variables. Les résultats obtenus à l'aide de cette méthode sont très compétitifs par rapport à ceux de l'état de l'art.

D'autres travaux sont actuellement en cours, nous essayons d'exploiter et d'adapter ce principe d'inférence conditionnelle en apprentissage semi-supervisé. Aussi, nous étudions l'efficacité de la mesure d'importance dans les ensembles de données à grande dimension.

CHAPITRE 3

Sélection de variables en classification semi-supervisée

1 Objectifs

La sélection de variables est un domaine de recherche qui donne lieu à de nombreuses études et à de nouvelles approches. Dans ce chapitre, nous apportons une contribution concernant la sélection de variables pour des problèmes de classification semi-supervisée.

Les algorithmes de sélection supervisée nécessitent de définir des labels de toutes les données. Par conséquent, la procédure de labellisation réalisée par un expert humain peut s'avérer fastidieuse et couteuse en temps de travail. C'est pour cette raison qu'on est généralement confronté à des bases de données réelles formées de nombreuses données non labellisées et peu de données labellisées. Ce contexte d'apprentissage est appelé semi-supervisé car l'analyste exploite à la fois les données non labellisées et les quelques données labellisées. De ce fait, nous proposons alors une méthode d'évaluation semi-supervisée de façon à ce que la sélection d'attributs et la classification de données soient réalisées dans le même contexte.

Pour améliorer l'efficacité de la sélection des ensembles de données en grande dimension, nous mettons en œuvre une nouvelle approche d'évaluation de sélection de variables d'importances en apprentissage semi-supervisé (appelée *Optim Co-forest*). Cette méthode combine à la fois les idées de l'algorithme *Co-forest* et le principe de sélection des *forêts aléatoires* basé sur la permutation des *OOB* (Out Of Bag).

Afin de comparer les performances de différentes méthodes proposées dans la littérature, des expérimentations sont menées avec des bases de données à grandes dimensions de référence.

2 Contribution

La tendance actuelle d'un fort accroissement de la taille des bases de données pose un défi sans précédent pour la fouille de données. Non seulement les bases de données s'agrandissent, mais de nouveaux types de données deviennent très répandus, tels que les flux de données sur le web, les données de puces à ADN génomique et les données relatives aux réseaux sociaux.

Les chercheurs se sont rendus compte que la phase de labellisation des données qui a priori nécessite l'intervention d'un expert humain est devenue une opération difficile voire fastidieuse lorsque le nombre de données est important. Dans des applications concrètes,

il est souvent impossible que l'expert puisse assigner toutes les données d'apprentissage aux classes en présence.

Le contexte semi-supervisé qui se situe à l'intersection entre le contexte supervisé et non supervisé, est alors une solution envisagée [4]. L'apprentissage semi-supervisé cherche à extraire une règle de décision ou de régression d'un ensemble d'apprentissage, avec une particularité : cet ensemble contient à la fois des objets supervisés, c.à.d. étiquetés, et d'autres qui ne le sont pas.

Un autre problème s'ajoute dans le cadre de données à grande dimension c'est le nombre très important d'attributs. Un nombre élevé de variables peut en effet s'avérer pénalisant pour un traitement pertinent et efficace des données. D'un côté par les problèmes algorithmiques que cela peut entraîner (liés au coût calculatoire et à la capacité de stockage nécessaire), et d'autre part, du fait que parmi les variables certaines peuvent être non-pertinentes, inutiles et/ou redondantes perturbant ainsi le bon traitement des données. Or, il est très souvent difficile voire impossible de distinguer les variables pertinentes des variables non-pertinentes. L'intérêt d'application de techniques de sélection de variables est devenu essentiel pour que la fouille de données atteigne ses objectifs (Han et al. [10]) ; (Guyon et Elisseeff [11]) ; (Liu et Motoda, [12, 13]).

Le problème de dimensionalité des données peut être résumé par l'aphorisme de Liu et Motoda "Less is more" [12] qui met en exergue la nécessité de supprimer l'ensemble des portions non pertinentes des données de manière préalable à tout traitement si on désire en extraire des informations utiles et compréhensibles.

La sélection de variables en apprentissage semi-supervisé constitue une solution à ces problèmes. Ce processus vise en effet à la détermination d'un sous-ensemble optimal (au sens d'un critère donné) de variables en tenant compte à la fois des données labellisées et non labellisées.

Notre contribution est de proposer une méthode d'apprentissage semi-supervisée en présence d'un grand nombre de variables tout en fournissant une sélection de variables les plus pertinentes de manière intégrée. L'idée est de mettre en exécution les méthodes ensemblistes qui combinent plusieurs modèles pour produire une solution plus performante et plus robuste.

L'application de méthodes ensemblistes est nécessaire dès lors qu'on veut passer un cap de réalisation de meilleurs résultats de prédiction. En effet, au lieu d'essayer d'optimiser une méthode "en une seule fois", les méthodes d'ensemble génèrent plusieurs règles de prédiction et mettent ensuite en commun leurs différentes réponses. L'heuristique de ces méthodes est qu'en générant beaucoup de prédicteurs, on explore massivement l'espace des solutions, et qu'en agrégeant toutes les prédictions, on récupère un prédicteur qui rend compte de toute cette exploration.

Le contenu suivant de ce chapitre est réparti comme suit : premièrement, un état de l'art des techniques de sélection en apprentissage semi-supervisé est effectué. Nous exposons ensuite dans la section 3, le principe de notre approche appelée *optim co-forest* en décrivant comment la mesure de variable d'importance utilisée dans les forêts aléatoires peut être étendue dans le contexte semi-supervisé grâce aux données étiquetées et non étiquetées. Des expériences utilisant des bases à grande dimension et des données réelles sont présentées dans la section 4. Nous terminons par une conclusion qui résume la contribution apportée par ce travail.

3 Les techniques de sélection semi-supervisée

L'identification de sous-ensembles de variables pertinents parmi des milliers de variables potentiellement inutiles et superflues est un sujet de recherche ardu dans le domaine de reconnaissance de forme, cela qui a attiré énormément d'attention au cours des dernières années. En apprentissage supervisé, les algorithmes de sélection de variables se basent uniquement sur des informations à partir de données étiquetées pour trouver les sous-ensembles de variables pertinents, à savoir, ceux qui s'avèrent utiles pour la construction d'un classifieur efficace. Un modèle de classification qui permet d'atteindre de bonnes solutions avec un sous-ensemble restreint de caractéristiques [11, 41, 42].

D'un autre côté, la sélection de caractéristiques dans l'apprentissage non supervisé vise à trouver des sous-ensembles de variables pertinents qui produisent des regroupements «naturels» en assemblant les objets "similaires" en ensemble basé sur une mesure de similarité. Plusieurs algorithmes de sélection d'attributs opérant dans un contexte non-supervisé ont été proposés dans la littérature. Ces algorithmes utilisent les différents critères d'évaluation cités ci-dessus (mesure de corrélation [258], mesure de consistance [259] . . .etc.).

De toute évidence, la combinaison des deux paradigmes (supervisés ou non) a permis l'émergence d'approches semi-supervisées sophistiquées pour la sélection de variables qui peuvent gérer à la fois les données étiquetées et non étiquetées. Le problème de la sélection de variables semi-supervisées a suscité récemment beaucoup d'intérêt et son efficacité a déjà été démontrée dans de nombreuses applications [47–50].

Récemment, plusieurs études ont porté leurs fruits sur la sélection de variables en semi-supervisé. Pareil à la sélection de variable en supervisé et non supervisé, ces méthodes peuvent être divisées en trois catégories [11], en fonction de la façon dont ils interagissent avec l'algorithme d'apprentissage : filtres, enveloppes (wrapper) et les approches intégrées (embbeded).

Approche filtre (filter) Le filtrage est un processus de pré-traitement des données par filtrage des variables non pertinentes avant que n'intervienne la phase de classification. Il utilise les caractéristiques générales de l'ensemble de variables pour sélectionner certaines variables et en exclure d'autres. La plupart des approches filtres classent les variables selon leur pouvoir individuel de prédiction de la classe. Il peut être estimé de divers moyens tels que le score de Fisher (Furey, et al. [260]), le test de Kolomogorov-Smirnov [261], le coefficient de corrélation de Pearson (Miyahara et Pazzani, [262]) ou encore l'information mutuelle ([263, 264]).

Le principal avantage des méthodes filtre est leur efficacité calculatoire et leur robustesse face au sur-apprentissage. Malheureusement, ces méthodes ne tiennent pas en considération les interactions entre les variables et tendent à sélectionner les variables comportant de l'information redondante plutôt que complémentaire [11].

Dans les approches de type filtre, les méthodes de sélection en semi-supervisé existantes sont basées sur le principe de maximisation de la marge. Elles font également l'hypothèse que toutes les données étiquetées et non étiquetées sont indépendantes et identiquement distribuées. La majorité des travaux réalisés dans cette catégorie se basent sur l'analyse spectrale, mais ils diffèrent dans la manière de l'appliquer [265–267].

Nous citerons en premier lieu les travaux de Zhao et al. [49] qui ont proposé un algorithme de classement des variables en semi-supervisé, dénommé sSELECT, basé sur la théorie de la courbe spectrale. Leur méthode établie d'abord un graphe de voisinage en utilisant les données d'origine, puis évalue chaque vecteur de caractéristiques en le transformant en un indicateur de cluster et vérifie si elle est conforme aux informations de l'étiquette. sSELECT a démontré des résultats prometteurs sur des jeux de données de références.

Une autre approche similaire, par l'utilisation de la théorie de la courbe spectrale est la méthode de sélection de variables en semi-supervisé à sensible Localité (*Locality Sensitive Semi-supervised Feature Selection method (LSDF)*) [50]. C'est une extension de la méthode d'analyse discriminante sensible Localité [268] dans le cas semi-supervisé. Son objectif principal est de découvrir la structure intrinsèque des données, compte tenu de toutes les variables. *LSDF* tente de classer les caractéristiques en fonction de leur contribution à conserver la structure retrouvée à la plus basse dimension. Les données labellisées sont utilisées pour maximiser la marge entre les données de différentes classes. Quant aux données non marquées elles sont utilisées pour détecter la structure géométrique de l'espace de données. L'idée est assez simple, son principal inconvénient est que c'est une méthode unidimensionnelle.

Récemment, d'autres méthodes de type filtre ont vu le jour, elles se basent sur la logistique *I-RELIEF* [47] leur but est de mesurer le pouvoir discriminant de chaque variable.

L'approche enveloppe (Wrapper) Ces approches ont été introduites par Kohavi et John [193, 269]. Pour ces auteurs, les algorithmes de filtrage ne sont pas toujours efficaces car ils ignorent totalement l'influence de l'ensemble de variables sélectionnées sur les performances de l'algorithme de classification. Pour résoudre ce problème, ils proposent une approche différente utilisant le résultat de l'algorithme de classification comme fonction d'évaluation. L'algorithme de classification appliqué aux données prétraitées est utilisé comme un sous-programme et aussi considéré comme une boite noire par cet ensemble de méthodes.

Le risque de sur-apprentissage est important si le nombre d'observations est insuffisant et si le nombre de variables à sélectionner doit être choisi par l'utilisateur. Enfin, le plus grand désavantage de ces méthodes est le temps de calcul qui devient vite important surtout dans le cas d'un nombre élevé de variables.

John, et al. [193] et Aha et Bankert [270] furent les premiers à démontrer (de façon empirique) que la stratégie enveloppe était supérieure à la stratégie filtre en terme de performance de classification.

Conformément à la stratégie enveloppe, la méthode de sélection de variable en semi-supervisé avec une *recherche en avant* est proposée [48]. Ren et al. utilisent le mécanisme de la sélection aléatoire des données non étiquetées pour former de nouveaux ensembles d'apprentissage. La variable la plus fréquemment retenue est obtenue en appliquant la stratégie de *recherche séquentielle en avant* dans le mode supervisé. Ensuite elle est ajoutée au sous-ensemble des variables pertinentes à chaque itération. Dans cette méthode, le sous-ensemble de variables dérivées des ensembles d'apprentissage aléatoires utilisés peuvent être insuffisants, mais une fois que la variable est sélectionnée, elle ne sera jamais éliminée.

Tout récemment, Benabdeslem et al. [271] ont proposé un algorithme à base de pondération de variables dans un paradigme de classification automatique sous contraintes (*CLS*). A cet effet, une vision globale est développée se basant sur la satisfaction relaxée des contraintes intégrées directement dans la fonction objective du modèle proposé. L'approche évalue la pertinence des variables par des poids appris en cours de la construction du modèle de classification. En parallèle à cette tâche principale de sélection de variables, les auteurs s'intéressent au traitement de la redondance. Pour résoudre ce problème, ils proposent une méthode originale combinant l'information mutuelle et un algorithme de recherche d'arbre construit à partir de variables pertinentes en vue de l'optimisation au final de leur nombre.

D'un autre côté, un inconvénient majeur à ces approches resurgit lors de leurs applications, où les performances de score de contrainte sont sévèrement influencées par le

choix de l'ensemble de contraintes. Pour surmonter ce problème, les auteurs de [272] ont proposé une approche d'échantillonnage (*BS*) à la partition de contrainte afin d'améliorer la précision globale de classification. Le principal désavantage de cette méthode est l'ignorance de la partie non marquée de données qui est généralement beaucoup plus importante à celle qui est marquée. Afin de tirer profit des parties à la fois marquées et non marquées de données, un score appelé *C*4 proposé dans [273], introduit une simple multiplication des scores Laplaciens avec les contraintes pour un compromis entre les deux scores. Cependant, la méthode est biaisée vers les variables avec un bon score Laplacien mais avec un score de contrainte très faible et vice-versa.

L'approche intégrée (Embedded) Les approches intégrées incorporent la sélection de variables lors du processus d'apprentissage, sans étape de validation, pour maximiser la qualité de l'ajustement et minimiser le nombre de variables. Un exemple très connu est celui des arbres de décision, où les variables sélectionnées sont celles qui sont présentes au niveau de la division de chaque noeud.

Selon Guyon et al. [11], ces approches seraient bien plus avantageuses en terme de temps de calcul que les méthodes de type wrapper et resteraient robustes face au problème de sur-apprentissage.

Nous retrouvons dans cette catégorie la méthode dite de FS-Manifold [274]. Elle sélectionne un sous-ensemble optimal de caractéristiques en maximisant la marge de classification entre les différentes classes, tout en exploitant la géométrie de la distribution de probabilité que FS- Manifold génère à partir des données (marquées et non marquées). La régularisation de distribution dans la méthode de sélection de variables proposée assuarant une classification est satisfaisante sur la distribution réalisée par les caractéristiques sélectionnées des données non labellisées.

Dans un autre papier de sélection par approche intégrée suivant le paradigme semi-supervisé, les auteurs Hindawi et al. [275] proposent une sélection de variables globale en semi-supervisé appelée *L2GFS : Local to Global Semi-Supervised Feature Selection*. Ils présentent un modèle métrique par le calcul de pondération locale des caractéristiques, basé sur le regroupement des contraintes dans le but d'effectuer une sélection globale de caractéristiques semi-supervisées.

Dans le cadre de sélection de variable par approche ensembliste en apprentissage semi-supervisé, Bellal et al. [256] sont parmi les premiers à avoir réalisé la combinaison à la fois du ré-échantillonnage de données (bagging [37]) et celle des sous-espaces aléatoires [39] par l'application de l'algorithme *co-Training* [5]. Ils ont mis en œuvre une technique appelée SEFR qui de son principe original conduit à l'exploration et l'extraction des variables les plus pertinentes.

4 Approche proposée

Dans le cadre d'une application sur des ensembles de données à grande dimension, les caractéristiques peuvent être corrélées, redondantes ou peuvent-être même bruitées et donc non pertinentes. Dans ce cas de figure *Co-forest* peut sélectionner ces variables et de ce fait apprendre sur des données erronées résultant des performances de classification individuelles médiocres. Cet inconvénient peut être évité par la sélection de variables les plus pertinentes, ce qui implique que l'utilisation d'algorithme de sélection intelligente à l'instar de la sélection aléatoire est nécessaire.

Le procédé de mesure des variables d'importance dans le paradigme des forêts aléatoires (RF) [51] a eu une grande influence sur notre approche proposée. Dans cette étude, nous montrons que ces idées sont également applicables à la sélection de variables en

semi-supervisé. De ce fait, nous proposons une méthode d'évaluation des variables d'importance en semi-supervisé basée sur le principe de *Co-forest* appelée *Optim Co-forest* (présenté dans la partie 3 chapitre 3).

L'algorithme classe les caractéristiques à travers un framework composé d'une méthode d'ensemble, dans lequel la pertinence d'un élément est évaluée par son exactitude prédictive en faisant appel aux données étiquetées et non étiquetées.

Dans *Co-forest*, la mesure de variable d'importance ne peut être estimée qu'à partir des échantillons OOB puisque l'échantillon bootstrap utilisé pour l'apprentissage de chaque arbre aléatoire est modifié après première itération. Les données OOB sont toutes étiquetées. Toutefois, étant donné la quantité très réduite de données marquées, la diversité des données OOB n'est pas suffisante. Les estimations OOB sont biaisées car elles dépendent de trop peu de données.

Optim Co-forest combine à la fois le ré-échantillonnage de données (Bagging [37]) et deux stratégies de sélection. La première implique la sélection intelligente d'un sous-ensemble de paramètres aléatoires inspirée de l'approche *Rel-RASCO* [6] afin de générer l'ensemble des classifieurs suivant le principe de *Co-forest*. Cela permettra de conserver la diversité des classifieurs ainsi que leur capacité à produire la meilleure discrimination pour chaque classe.

Une fois que chaque membre de l'ensemble est obtenu, la seconde stratégie appliquée consiste à une extension de la mesure d'importance de RF [51]. Elle fait appel à un assortiment d'ensemble de données marquées et non marquées, pour mesurer la pertinence des variables. Un classement de toutes les variables est finalement réalisé par rapport à leurs pertinences dans tous les classifieurs semi-supervisés.

La combinaison de ces deux stratégies dans la construction de l'ensemble des classifieurs en semi-supervisé mène à l'exploration d'un plus grand espace de solutions et delà récupérer un prédicteur qui rend compte de toute cette exploration.

Dans ce chapitre, nous proposons d'établir des sous-espaces aléatoires de caractéristiques pertinentes pour *Co-forest*. L'algorithme proposé, *Optim Co-forest* produit des sous-espaces aléatoires pertinents à l'aide du score de pertinence des caractéristiques, obtenu en calculant l'information mutuelle entre les caractéristiques et les étiquettes de classe. Afin de maintenir aussi la diversité (l'aspect aléatoire), chaque variable d'un sous-espace est choisie en fonction des probabilités proportionnelles à la pertinence des scores de caractéristiques.

4.1 La procédure de mesure d'importance des variables

En premier lieu, nous construisons les ensembles *Out Of Bag* « en dehors du sac » pour chaque classifieur, nous sélectionnons les cas bien prédits (classés) des exemples étiquetés OOB_l et nouvellement étiquetés dans un sous-ensemble qu'on appellera U^2 (Algorithme 14).

En second lieu, nous reprenons la confiance de chaque exemple sélectionné. Notons que la confiance des exemples étiquetés, si l'étiquette de classe donnée par h_i correspond à l'étiquette réelle, est mise à 1. Pour les exemples non étiquetés, leurs confiances seront calculées sur le degré d'accord sur l'étiquette parmi les membres de concomitance H^*.

En dernier lieu, les valeurs de chaque variable V_i sont permutées au hasard, et h_i est utilisé pour prédire la classe de ce nouveau modèle *Out Of Bag Perm*. La procédure est répétée pour chaque variable $V \in \{V_1, \ldots, V_i\}$ (Algorithme 14).

Algorithm 14 *Mesure d'importance*

Entrée : Comité H, ensemble de données OOB_l, ensemble de données nouvellement labellisées U^2, nombre de classe K
Sortie : Importance V
Pour chaque $x \in U^2$ **faire**
- $H_x = \{hi \in H \mid x \in Ui_{oob}\}$
- [label(x), conf(x)]= Mesure Confidence (x, H_x, K)
- **fin**

Classer les exemples de U^2 par ordre décroissant de confiance et sélectionner les n_k exemples de confiances pour chaque classe K
Pour chaque $V \in V1 \ldots Vi$ **faire**
- $OOB = [OOB_l \bigcup U^2]$
- *Out Of Bag Perm*= $permute\ (OOB, V)$
  - **Si** (V est selectionné comme de confiance) **alors**
  - Importance = Importance $\bigcup[V; label(V)]$
  - **fin si**
- **fin**

En fin de procédure, la somme des confiances des exemples pour lesquels l'étiquette prédit dans *Out Of Bag Perm* diffère de la première étiquette initiale dans *Out Of Bag*, est calculée. Cette dernière valeur est moyennée sur N (la taille du comité). La valeur ainsi obtenue est prise comme l'importance de la variable V (Algorithme 14). L'idée essentielle de notre approche est l'utilisation de la confiance de l'étiquette dans l'évaluation de mesure d'importance des variables. Ainsi, les exemples non étiquetés jouent un rôle important dans l'évaluation de variable d'importance.

5 Expérimentations et résultats

Dans cette section, nous présentons des résultats empiriques sur plusieurs données de références et des données réelles à grande dimension. Aussi nous évaluons *Optim co-forest* avec les algorithmes de l'état de l'art en semi-supervisé. A cet effet, *Optim co-forest* est comparé à quatre autres méthodes de sélection *BS : Bagging Score* [272], *CLS : Constrained Laplacian Score* [271], *sSELECT :sélection de variable par analyse spectrale* [49] et *SEFR : Semi-supervised Ensemble learning guided Feature Ranking method* [256].

Bases	*#instances*	*#variables*	*#classe*
Arcene	200	10000	2
BaseHock	1993	4862	2
CNAE-9	1080	856	9
Leukemia	73	7129	2
Pancreatic	119	6771	2
PC MAC	1943	3289	2
Promoters	106	57	2
Relathe	1427	4322	2
Robot	88	90	4
Semion	1593	265	2
SMK-CAN	187	19993	2
Spect	267	22	2
SRBCT	83	2308	4
Vehicle	846	18	4

Table 34 – Description des bases d'expérimentation

Douze ensembles de données sont principalement choisis du référentiel Machine Learning UCI [84], et de l'ASU feature selection Repository [184]. Ces derniers sont utilisés pour

évaluer la performance de *Optim co-forest* et leurs caractéristiques sont décrites dans le tableau (Table 34). Nous avons choisi ces ensembles de données, car ils contiennent de nombreuses caractéristiques et relativement peu d'échantillons (comme exemple les bases BaseHock et Prostate) sont de bons candidats pour la sélection de variables. La plupart de ces ensembles de données ont déjà été appliqués par d'autres auteurs et ce pour tester les performances de leurs algorithmes de sélection [44, 45, 48, 49, 256].

5.1 Phase d'évaluation

Dans le but d'une comparaison équitable, les mêmes paramètres expérimentaux dans [271] ont été adoptés ici pour *CLS*, à savoir, le graphe de voisinage avec un degré approprié à 10, et la valeur est définie sur 0,1. Pour *BS*, nous avons fixé la taille de l'ensemble à 100, car il a été démontré dans [272] qu'autour de cette valeur l'algorithme *BS* est moins sensible aux variations de la taille de l'ensemble. *Optim Co-forest* et *SEFR* sont paramétrés de façon similaire. Le nombre de variables par sous espace est égal à p, où p est la taille de l'espace d'entrée. La taille du comité N est équivalente à 100 arbres.

Pareil à l'algorithme *co-forest* [18], le seuil de confiance θ dans *Optim Co-forest* est fixé à 0,75, à savoir, un exemple nouvellement labellisé est considéré comme étant de confiance si plus de 3/4 des arbres sont d'accords sur son étiquette assignée.

Cependant, comme le suggèrent les auteurs dans [256], le nombre d'itérations $maxiter$ et la taille de l'échantillon n dans *SEFR* sont fixés à 10, et 1, respectivement. Pour chaque jeu de données, les résultats expérimentaux sont moyennés sur plus de 10 itérations. A chaque itération, l'ensemble des données est découpé (de façon stratifiée) dans une partition d'apprentissage avec les 2/3 des observations et une partition de test avec le tiers restant. L'ensemble d'apprentissage est également réparti en ensembles de données étiquetées et non étiquetées.

De même que [49], l'ensemble des échantillons étiquetés L se compose de 3 exemples choisis par classe au hasard, et les exemples restants sont utilisés comme ensemble de données non marquées U.

5.2 Résultats

Afin d'apprécier la qualité d'un sous-ensemble de variables obtenues avec les procédés semi-supervisés mentionnés ci-dessus et suivant le protocol de test présenté par [255], [256], nous lançons un apprentissage avec *SVM* (en utilisant le package *LIBSVM* [239]) sur l'ensemble de données d'apprentissage étiquetées et nous évaluons l'exactitude à partir des données de test. La précision réalisée est moyennée et utilisée pour l'évaluation de la qualité du sous-ensemble de variables sélectionnées en fonction de chaque algorithme (Table 35).

Comme nous pouvons le constater dans la Figure 45, *Optim co-forest* dépasse nettement les quatre autres méthodes par une marge remarquable. Sur la plupart des bases de données sSELECT ainsi que CLS, ils effectuent les pires performances. *Optim co-forest* semble combiner plus efficacement les données étiquetées et non étiquetées pour l'évaluation des caractéristiques. Il est plus intéressant sur les données à grande dimension dans le mode semi-supervisé vue sa bonne performance sur les bases *CNAE-9, Relathe, PCMAC, Leukemia*.

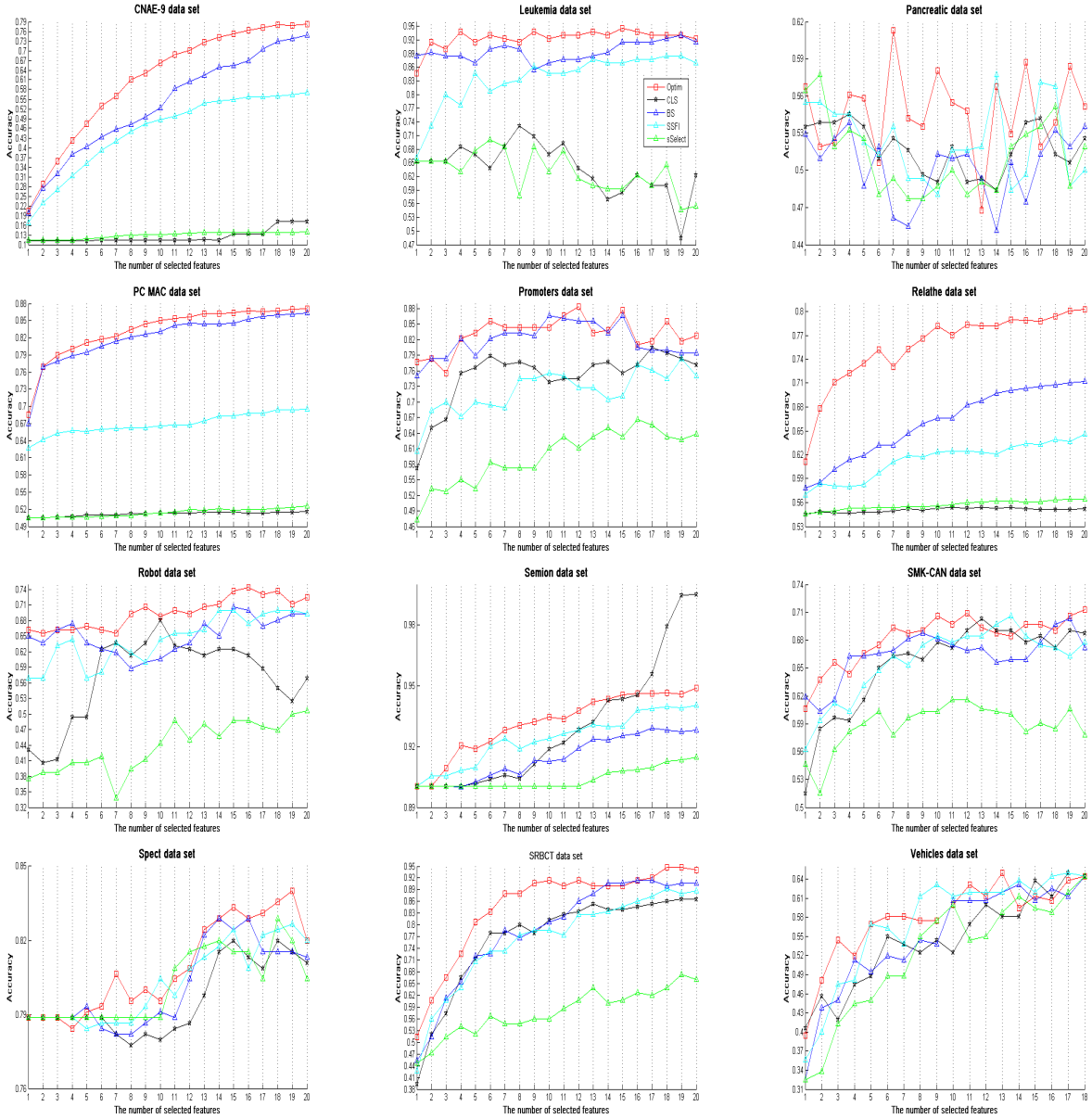


Figure 45 – Courbes de performances des différentes bases de données en fonction du nombre de variables sélectionnées

Les graphes dans la Figure 45 montrent l'évolution du taux de classification en fonction des nombres de caractéristiques sélectionnées. Les résultats indiquent une amélioration importante générale de notre approche par rapport aux quatre autres approches.

Une inspection plus minutieuse des différentes courbes révèle que la précision sur les variables sélectionnées par *Optim Co-forest* augmente rapidement au départ (le nombre de variables sélectionnées est petit) et se ralentit par la suite. Ceci suggère que *Optim Co-forest* classe dès le début les caractéristiques les plus pertinentes. Donc un classifieur peut réaliser une très bonne classification avec le top 5 des caractéristiques tandis que les autres méthodes nécessitent plus de variables pour obtenir des résultats comparables.

Par souci d'exhaustivité, nous avons également établi la précision moyenne pour les différents nombres de caractéristiques sélectionnés. Les précisions moyennes entre *Optim co-forest* et les autres méthodes sur les 20 premières caractéristiques sont décrites dans le tableau (Table 35). Encore une fois, comme nous pouvons le constater, *Optim co-*

forest surpasse nettement *BS, sSELECT, CLS* et *SEFR* par une marge notable, sur tous les ensembles de données. Pour finir, ces expériences confirment la capacité de la mesure d'importance en exploitant efficacement les informations à partir des données non étiquetées pour la classification des caractéristiques pertinentes.

Bases de données	*BS*	*CLS*	*SEFR*	*sSelect*	*Optim Co-forest*
Arcene	0,6318	0,6506	0,6971	0,5537	0,7000
BaseHock	0,8145	0,5114	0,6756	0,5117	0,8865
CNAE-9	0,5333	0,1252	0,4505	0,1293	0,6113
Leukemia	0.8946	0.6388	0.8346	0.6296	0.9238
Pancreatic	0,5032	0,5179	0,5237	0,5116	0,5477
PC MAC	0,8208	0,5118	0,6693	0,5143	0,8332
Promoters	0,8189	0,7489	0,7211	0,5956	0,8317
Relathe	0,6605	0,5507	0,6138	0,5569	0,7562
Robot	0,6516	0,5697	0,6450	0,4384	0,6959
Semion	0,9148	0,9293	0,9240	0,9041	0,9317
SMK-CAN	0,6644	0,6541	0,6573	0,5881	0,6819
Spect	0,8017	0,7964	0,8031	0,8020	0,8094
SRBCT	0,7870	0,7653	0,7627	0,5767	0,8457
Vehicle	0,5493	0,5444	0,5715	0,5226	0,5781

Table 35 – Le taux de performance moyen sur les 20 variables les plus pertinentes

6 Conclusion

Depuis la prolifération des bases de données partiellement étiquetées, la sélection de variables a connu un développement important dans le mode semi-supervisé. Nous avons proposé dans ce chapitre d'aborder cette problématique avec une méthode d'ensemble à base de mesure d'importance de variables. Ce modèle est basé sur la classification semi-supervisée pour sélectionner les variables les plus pertinentes.

Dans ce travail, l'algorithme *Optim Co-forest* est proposé. Il a la faculté de tirer profit des échantillons non marqués pour améliorer les performances du système formé à partir des échantillons marqués. En étendant le paradigme *Co-forest*, *Optim Co-forest* exploite la puissance de *Random Forest*, une méthode d'ensemble bien connue, pour s'attaquer au problème de la sélection des échantillons non labellisés les plus confiants pour le ré-apprentissage de la forêt.

Optim Co-forest exploite également deux stratégies de sélection : la première concerne la sélection de sous-ensemble de paramètres aléatoires qui nous permettra de garder la diversité des classifieurs. La seconde méthode est la mesure d'importance de variables pour évaluer la pertinence de ces variables. La combinaison de ces deux stratégies dans la construction de l'ensemble des classifieurs en semi-supervisé conduit à l'exploration d'un plus grand espace de solutions et delà avoir un prédicteur plus compétitif et adéquat aux espaces à grande dimension.

Des expérimentations réalisées sur plusieurs bases de données de très grande dimension UCI et ASU prouvent l'efficacité de *Optim Co-forest* et confirment ainsi sa capacité de sélection et de mesure d'importance tout en améliorant la performance de l'hypothèse apprise sur une petite quantité d'échantillons labellisés et cela en exploitant les échantillons non labellisés.

D'autres travaux sont actuellement en cours, avec de nouvelles expérimentations sur des bases de données biologiques possédant plusieurs milliers de variables et pourront évaluer la stabilité de la méthode de sélection de caractéristiques [276] dès que de petites modifications seront apportées aux données.

Conclusion

Dans de nombreuses applications d'apprentissage statistique, le nombre de variables est grand devant le nombre de données. Du point de vue de la classification, il est préférable de sélectionner un sous échantillon de variables pertinentes. Le problème de la sélection de variables peut être vu comme un problème du choix de modèle (Raftery et Dean [277] ; Maugis et al. [278] ; Murphy et al. [279]). Le fait de sélectionner les variables à partir de la vraisemblance conditionnelle permet aussi de sélectionner les variables en supposant que les variables non prises en compte sont indépendantes de la classe. Cette approche a l'avantage de ne pas émettre d'hypothèses sur la distribution des variables non prises en compte.

La littérature abondante depuis plusieurs décennies sur le problème de sélection de variables (Features Selection) témoigne non seulement sur son importance mais aussi sur ses difficultés qui résident dans le choix des caractéristiques pertinentes pour une application donnée.

Le problème de la sélection de variables en classification, se pose généralement, lorsque le nombre de variables utilisées pour expliquer la classe d'un individu, est très élevé. Les besoins ont beaucoup évolué ces dernières années avec la manipulation d'un grand nombre de variables dans des domaines tels que les données génétiques ou le traitement d'image. Néanmoins si l'on doit traiter des données décrites par un grand nombre de variables, les méthodes classiques d'analyse, d'apprentissage ou de fouille de données peuvent se révéler inefficaces ou peuvent aboutir à des résultats peu précis.

Dans cette thèse, nous avons proposé des méthodes innovantes pour réduire la taille initiale des données et pour sélectionner des ensembles de variables pertinents dans un cadre de classification supervisée et semi-supervisée.

Notre démarche de sélection et de mesure d'importance consiste dans un premier temps à comparer l'efficacité de plusieurs méthodes de sélection pouvant être intégrées dans un processus de différentes approches de types : filtre, enveloppe et embarquée ; afin de mettre en évidence la transparence de notre système, avec un objectif d'extraire les plus pertinents et les plus informatifs. Les expérimentations réalisées ont permis d'évaluer les performances des résultats avec différents classifieurs.

Cette quatrième partie de la thèse expose le processus de sélection de variables et l'importance de la sélection de variables pour l'amélioration des performances des algorithmes de classification. Dans le premier chapitre, nous avons illustré nos propos avec des algorithmes de sélection de variables proposées dans la littérature dans le cadre de la classification de la maladie du Kératocône. Cette approche est basée sur la combinaison d'algorithmes de sélection de différents types pour évaluer la pertinence des sous-ensembles candidats. Les différentes expérimentations que nous avons menées montrent

de très bonnes performances pour la mesure d'importance utilisée dans les Forêts Aléatoires, avec des résultats en concordance avec les avis d'experts pour le diagnostic du Kératocône.

Notre deuxième contribution porte sur l'application et l'adaptation des Forêts à Inférence Conditionnelle pour la sélection des caractéristiques. Alors que la forêt aléatoire classique est une mise en œuvre d'un ensemble d'arbres de type CART comme apprenants de base, la Forêt à inférence conditionnelle (CIF) utilise les arbres à inférence conditionnelle (CIT). Dans CIF, le système d'agrégation fonctionne en faisant la moyenne des poids d'observation extraits. Cette combinaison avec la mesure d'importance par permutation connue dans la Forêt Aléatoire a permis une sélection de variables les plus pertinentes par une stratégie simple et efficace avec une plus grande précision.

Dans le chapitre trois de cette partie de la thèse, nous proposons l'approche de classification de données semi-supervisées pour la tâche de sélection à grande échelle. Notre méthode de sélection de variables a permis de construire des prédicteurs efficaces pour un problème de données semi-supervisées. Les performances obtenues sont aussi bonnes, que celles des meilleures méthodes publiées à ce jour pour les mêmes bases de données. Notre principale contribution a été d'obtenir ces performances avec un nombre minimal de variables. Cette caractéristique est importante pour la robustesse de nos prédicteurs avec une condition nécessaire à une possible utilisation en routine clinique.

Dans sa globalité, cette quatrième partie de thèse, nous a permis de voir une multitude de pistes de recherche qui s'offrent pour les algorithmes de sélection de variables. Nos travaux dans les chapitres présentés portent sur la proposition de nouvelles approches de sélection de variables.

Ce domaine de recherche restera toujours actif tant qu'il est motivé par l'évolution des systèmes de collecte et de stockage des données d'une part et par les exigences d'autre part. La meilleure approche pour juger cette sélection est de collaborer avec des experts pour une interprétation des résultats et mettre en évidence les points suivants :

– Les variables qui sont en prédisposition et en cause des maladies.
– Les variables qui contribuent au développement de ces dernières.

Cette collaboration avec les experts permet de nous orienter vers la manière d'utiliser ces données fondamentales en pratique clinique et leurs influences sur la prise en charge des patients, car ce domaine de recherche est majeur dans le : dépistage, traitement et prédiction de l'évolution clinique de ces patients. En parallèle, nous œuvrons à mieux comprendre l'évolution de la réponse immunitaire du patient à tous les stades d'évolution des différentes pathologies étudiées.

Cinquième partie

V

Conclusion et Perspectives

Conclusion

Le sujet de cette thèse concerne la classification semi-supervisée des données médicales et biologiques. Dans ce contexte, nous nous intéressons à la question du choix des modèles qui sont conçus et réalisés en utilisant conjointement des données étiquetées et celles non étiquetées disponibles en nombre plus important.

La classification semi-supervisée trouve ses racines dans les problèmes d'apprentissage en présence de données manquantes (McLachlan, [280]). De nombreux travaux y ont été consacrés dans les années 1970 (Fryer et al. [281] ; Dempster et al. [119]). Ce thème de recherche a connu ensuite un regain d'intérêt à la fin des années 1990 dans la communauté du Machine Learning, avec la disponibilité croissante de grands jeux de données grâce aux nouvelles technologies.

En parallèle, les recherches en apprentissage artificiel initiées dans les années 1990 ont montré qu'il est possible d'atteindre une décision aussi précise que souhaitée par une combinaison judicieuse d'hypothèses imparfaites mais correctement entraînées. De ce fait, les techniques consistant à distribuer la tâche d'apprentissage entre plusieurs apprenants et aussi combiner les solutions partielles pour obtenir la solution générale ont vu un grand succès. Les méthodes de bagging/boosting, très étudiées actuellement, sont l'archétype de cette tendance.

Nous avons choisi les méthodes d'ensemble pour la classification semi-supervisée des données médicales, mais pour ce faire, nous avons réalisés plusieurs éxpérimentations pour justifier nos choix et nos réalisations.

Dans une première partie, nous avons dressé trois expérimentations différentes des méthodes d'ensemble « Forêt Aléatoire » en classification supervisée. Celles-ci nous ont permis de mettre en évidence la multitude de questions posées dans ce cadre, ainsi que les nombreuses méthodes développées pour y répondre. Cela justifie alors notre choix pour les méthodes d'ensemble puisque celles-ci permettent de prendre en considération l'information présente dans les données.

Les modèles de classification concus ont été utilisés dans la seconde partie dans laquelle nous avons détaillé plus les méthodes d'ensemble utilisées dans le cadre semi-supervisé ainsi que l'estimation de leurs paramètres à travers la mise en œuvre de l'algorithme *co-Forest*. Nous avons constaté durant les expérimentations réalisées dans cette partie, que les résultats obtenus pouvaient varier, et de ce fait, être améliorés selon la pertinence des données d'entrée. Si l'algorithme d'apprentissage semi-supervisé dégrade les performances du supervisé, alors le modèle établit n'est pas robuste et ce qui est largement

justifié dans les applications à grande échelle.

Pour mettre davantage en évidence le potentiel applicatif de la méthode d'ensemble *co-Forest*, nous réalisons la segmentation semi-automatique des images fond d'œil de la rétine pour le suivi médical du Glaucome. Dans le but de mesurer automatiquement le rapport cup/disque (C/D) et réaliser une prise en charge toujours plus précoce de la maladie afin de retarder le développement du déficit visuel, nous proposons une approche de segmentation semi-supervisée par la classification de super-pixels des régions cup et disque pour le calcul du rapport (C/D). Pour ce faire, la méthode proposée SP3S (*Super-Pixel for Semi-Supervised Segmentation*) comporte deux étapes principales. Dans la première, la forêt aléatoire en apprentissage semi-supervisé « *co-Forest* » est formée uniquement sur 10% des images super-pixels annotées par un expert ophtalmologiste. Dans la seconde étape, les super-pixels non labellisés sont impliqués pour le renforcement de l'apprentissage du classifieur afin de mieux discriminer les régions cup et disque des images rétiniennes. Pour calculer le rapport VC/D, un modèle de forme géométrique actif est utilisé pour dresser le contour du cup et disque.

Dans cette même partie nous avons proposé une nouvelle approche d'ensemble en mode semi-supervisé, qui exploite deux stratégies de sélection : la première concerne la sélection de sous-ensemble de paramètres aléatoires qui nous permettra de garder la diversité des classifieurs. La seconde méthode concerne la mesure d'importance de variables pour mesurer la pertinence de ces dernières. La combinaison de ces deux stratégies dans la construction de l'ensemble des classifieurs en mode semi-supervisé conduit à l'exploration d'un espace plus large de solutions et delà avoir un prédicteur plus compétitif et adéquat aux espaces à grande dimension. Des tests statistiques et non paramétriques ont été réalisés pour tester de la pertinence du modèle, prouvant ainsi sa supériorité par rapport à d'autres modèles proposés dans le même domaine.

Dans la dernière partie, nous avons abordé la problématique de sélection de variables pertinentes avec les méthodes d'ensemble à base d'une fonction de mesure d'importance de variables. Nous avons mené des expérimentations sur plusieurs bases de données réelles à grandes et moyennes dimensions, elles ont prouvé l'efficacité de la méthode des forêts aléatoires pour la sélection de variables. En se basant sur le principe même des *Forêts Aléatoires*, notre contribution concernait la sélection des variables en mode d'apprentissage semi-supervisé. Notre approche confirme ainsi sa capacité de mesure en améliorant la performance de l'hypothèse apprise sur une petite quantité d'échantillons labellisés tout en exploitant les échantillons non labellisés.

Perspectives

Nous sommes conscients que notre travail ne constitue qu'un début aux réflexions qui devrons se poursuivre dans le futur, et cela en suivant les nombreuses pistes qui nous sont apparues durant ces dernières années. A court terme pour ce travail, nous pensons à quelques issues de recherche :

- Développer d'autres types de méthodes d'évaluation en plus de la combinaison de classifieurs simples utilisée actuellement, comme par exemple l'évaluation en cascade [282].

- Exploiter le principe d'apprentissage en cascade (cascading) au contexte semi-supervisé, afin de remédier à la complexité temporelle. Cette technique a un avantage évident, qui est d'optimiser le temps de décision, en particulier si on ordonne les classifieurs élémentaires par complexité de calcul croissante.

- Toujours dans le but de minimiser le temps de calcul et en terme de complexité algorithmique, d'autres pistes sont à exploiter, par le biais des GPUs, ces unités de calcul

disposent d'une architecture particulière qui permet de paralléliser principalement des opérations matricielles ou vectorielles. Pour des problèmes de moyenne et grande taille, les mesures réelles obtenues par différents travaux démontrent des gains importants lorsqu'elles sont comparées à celles d'une implémentation séquentielle réputée.

Nos travaux offrent également des perspectives à long terme. Par exemple, nous envisageons :

- La mise en place d'une approche de classification non-supervisée dans un environnement semi-supervisé, et cela par l'utilisation des données étiquetées pour la validation des résultats de regroupement des données non étiquetées en apprentissage non supervisé. Dans ce contexte, nous pourrons envisager de laisser les données étiquetées pour "guider" ou "ajuster" le processus de regroupement, à savoir fournir une forme limitée de supervision. L'approche résultante est appelé *le clustering semi-supervisé*. Cette approche peut être considérée dans le cadre où les données labellisées disponibles sont loin d'être représentatives pour aboutir à une classification ciblée des exemples, de sorte que l'apprentissage supervisé est difficile à réaliser, même sous une forme transductive. Contrairement au principe de regroupement traditionnel, l'approche de regroupement semi-supervisé se résume en quelques méthodes publiés jusqu'à présent. La principale distinction entre ces méthodes concerne la façon dont les deux sources d'information sont combinées : soit en adaptant la mesure de similarité [283–287] ou en modifiant la recherche de grappes (clusters) appropriées [288–290].

- Proposer une architecture d'annotation semi-automatique des images médicales pour extraire les informations médicales spécifiques (i.e. modalité médicale (image globale), région anatomique (au niveau des pixels de l'images)) à partir du contenu et du contexte des images. L'apprentissage semi-supervisée par les experts du domaine peut être une piste interessante a exploiter pour l'annotation du contenu et du contexte adaptée aux images médicales.

- Appliquer les algorithmes semi-supervisés existants pour résoudre la problématique de poly-pathologies chez les patients. Cette problématique de classification *multi-labels* est affirmée par les praticiens, vu que, sur le terrain, les patients peuvent être atteints de plusieurs pathologies simultanément. En parcourant l'état de l'art, nous avons compris l'intérêt de deux notions importantes :

  1. L'exploitation de dépendances entre les labels (pathologies),
  2. L'impact de la sélection de variables dans ce domaine.

  Au stade actuel, nous avons pu mettre en place une méthode d'ensemble à classifieurs multi-labels en apprentissage supervisé [118], apportant des résultats très prometteurs pour la suite des travaux, un thème de Doctorat est déjà en cours pour aborder ce sujet de recherche.

- Adaptation de la méthode d'ensemble *Forêt Rotationnelle* dans le contexte d'apprentissage semi-supervisé. Elle est reconnue dans la littérature, comme l'une des techniques de génération d'ensemble d'arbres indépendamment la plus performante. Où la projection des données dans différents nouveaux espaces porte une amélioration simultanée sur l'exactitude individuelle et la diversité au sein de l'ensemble. Notre intérêt se porte plus sur la manière dont la diversité est conservée par l'application d'extraction de caractéristiques et la rotation des axes, et ainsi la précision est favorisée en gardant tous les composants. Une première application a été réalisée par l'optimisation et l'adaptation de la forêt rotationnelle aux données supervisées parkinsoniennes [291].

- Extrapoler le principe de l'approche *Boosting* [292] qui est par sa définition, une technique d'apprentissage qui vise à rendre plus performant un système d'apprentissage «

faible ». Pour ce faire, le système d'apprentissage est entrainé successivement sur des échantillons d'apprentissage sur-pondérant les exemples difficiles à apprendre. Dans le cadre des données semi-supervisées, ce système de pondération pourra être appliqué aux données nouvellement labellisées à faible confiance, permettant ainsi d'enrichir d'avantage l'espace d'apprentissage et d'acquérir plus de précision.

– Dans les applications réelles actuelles, nous sommes confrontés devant une contrainte plus grande que celle du flux de données partiellement labellisées, déséquilibrées ou à multi-étiquettes. Les exigences futures du cadre médical sont des applications et des traitements en "temps réel" avec la plus grande précision. Deux grands axes qui laissent le domaine de la recherche ouvert pour une amélioration continue.

Productions scientifiques

De nombreuses contributions méthodologiques ont résulté de cette thèse, ces productions sont parallèlement ou simultanément integrées à certaines grandes parties ou à certains chapitres de la thèse. Plusieurs travaux en collaboration avec des collègues ont servi à répondre à plusieurs réflexions directes et indirectes à la thèse.

Journaux internationaux

– N. SETTOUTI, MA. CHIKH and V. BARRA. *A New feature selection approach based on ensemble methods in semi-supervised classification*. Journal of Pattern Analysis and Applications. 10/2015 DOI : 10.1007/s10044-015-0524-9
– I. NEDJAR, M. El HABIB DAHO, N. SETTOUTI, S. MAHMOUDI and MA. CHIKH. *Random forest based classification of medical X-ray images by using a genetic algorithm for feature selection*. Journal of Mechanics in Medicine and Biology 01/2015 ; 15(02) :1540025. DOI :10.1142/S0219519415400254
– N. SETTOUTI, MA. BECHAR and MA. CHIKH. *Statistical Comparisons of the Top 10 Algorithms in Data Mining for Classification Task.* International Journal of Artificial Intelligence and Interactive Multimedia. 01/2016 ; vol. 4(1-1) :, pp 46-51. DOI :10.9781/ijimai.2016.419
– N. SETTOUTI, M. SAIDI and MA. CHIKH. *Generating Fuzzy Rules for Constructing Interpretable Classifier of Diabetes Disease*. Australasian Physical & Engineering Sciences in Medicine. Vol. 35, No. 3 (2012), pp. 257-270, DOI : 10.1007/s13246-012-0155-z
– N. SETTOUTI, M. El HABIB DAHO and MA. CHIKH. *Using of Conditional Inference Forest to identify Variable importance*. In press : International Journal of Bioinformatics Research and Applications. 2016.

Conférences Internationales

– **Année 2015-2016**

– N. SETTOUTI, MA. BECHAR and MA. CHIKH. *Statistical comparisons of the Top 10 algorithms in data mining for classification task*. International conference Advanced Information Technology, Services and Systems (AIT2S-15). December 16-17, 2015 Faculty of Sciences & Technologies, Settat, Morocco.
– MA. BECHAR, N. SETTOUTI, M. El HABIB DAHO and MA. CHIKH. *Croissance de région par classification pixellique : Application aux images cytologiques*. Systeme Conjoint de Compression et Indexation des Objets Videos : SCCIBOV'2015, December 02-03, 2015, Djillali Liabes University -Faculty of Technology, Sidi Bel Abbes, Algeria.
– S. BENIKHLEF, E. BENDIMERAD, K. DOUIBI and N. SETTOUTI. *An improved Rotate Forest classifier for the recognition of Parkinson's disease*. Knowledge Discovery and Data Analysis KDDA'2015, November 15- 18, 2015. ESI, Algiers, Algeria.
– K. DOUIBI, N. SETTOUTI and MA. CHIKH. *Bagged MLknn : Bootstrap and Aggre-*

gating K Nearest Neighbors for Multi-label data classification. Knowledge Discovery and Data Analysis KDDA'2015, November 15- 18, 2015. ESI, Algiers, Algeria.

– **Année 2014-2015**

– MA. BECHAR, N. SETTOUTI and MA. CHIKH. *L'impact de la mesure de similarité en auto-apprentissage*. Colloque sur l'Optimisation et les Systèmes d'Information COSI'2015, 1 au 3 juin 2015, Oran, Algérie Université d'Oran 1, Ahmed Ben Bella.
– N. SETTOUTI, MA. LAZOUNI, M. El HABIB DAHO and MA. CHIKH. *Identification automatique des facteurs importants qui influent sur le contrôle du diabète en Algérie*. Biomedical Engineering International Conference (BIOMEIC'14), 15-16 october 2014, Université de Tlemcen, Algérie. 2014
– N. SETTOUTI, M. A. BECHAR, M. EL HABIB DAHO, M. A. LAZOUNI, M. A. CHIKH. *Bagged Nearest Neighbor Classifiers in Semi Supervised Learning*. The 19 International Conference on Mechanics in Medecine and Biology (ICMMB'14), September 3-5 2014, Bologna, Italy.
– M. El HABIB DAHO, N. SETTOUTI, MA. LAZOUNI and MA. CHIKH. *Dynamic Pruning For Random Forest*. The 19 International Conference on Mechanics in Medecine and Biology (ICMMB'14), September 3-5 2014, Bologna, Italy.

– **Année 2013-2014**

– M. El HABIB DAHO, N. SETTOUTI, MA. LAZOUNI and MA. CHIKH. *Weighted vote for trees aggregation in Random Forest*. The 4th International Conference on Multimedia Computing and Systems (ICMCS'14). April 14-16 2014, Marrakesh, Morocco, Pages : 453 - 458, IEEE Xplore, DOI :10.1109/ ICMCS.2014.6911235
– M. El HABIB DAHO, N. SETTOUTI, MA. LAZOUNI and MA. CHIKH. *Amélioration des Forêts Aléatoires pour une Meilleure Prédiction*. ICA2IT International Conference on Artificial Intelligence and Information Technology. Ouargla, Algeria, March 10-12, 2014.
– N. SETTOUTI, M. El HABIB DAHO, MA. LAZOUNI, A. KOULOUGHLI, S. KALACHE, and MA. CHIKH. *Optimisation des Forêts Aléatoires Floues*. The First International Symposium on Informatics and its Applications ISIA. M'sila, Algeria, February 25-26, 2014.

– **Année 2012-2013**

– N. SETTOUTI, M. El HABIB DAHO, M.A LAZOUNI and MA. CHIKH. *Random Forest in Semi Supervised Learning «Co-forest»*. WoSSPA'13 The 9th International Workshop on Systems, Signal Processing and their Applications 12-15 May 2013, Hotel Safir, Mazafran, Algiers, Algeria. IEEE : 10.1109/WoSSPA.2013.6602385

– **Année 2011-2012**

– N. SETTOUTI, M. El HABIB DAHO, M. SAIDI and MA. CHIKH. *Conditional Inference Forest for Variables Selection of Medical Data*. Biomedical Engineering International Conference (BIOMEIC'12), October 10-11, Tlemcen, Algeria.
– A. HAFA, N. SETTOUTI and MA. CHIKH. *Approche Filtre pour l'identification des gènes pertinents des données biopuces du Cancer du Côlon*. Biomedical Engineering International Conference (BIOMEIC'12), October 10-11, Tlemcen, Algeria.
– MA. BEKHTI N. SETTOUTI and MA. CHIKH. *Sélection de Variables neuronale pour le diagnostic du Diabète*. Biomedical Engineering International Conference (BIOMEIC'12), October 10-11, Tlemcen, Algeria.
– N. SETTOUTI, MA. CHIKH and M. SAIDI. *Applications des forêts Aléatoires pour La sélection de descripteurs de données médicales*. COSI (Colloque sur l'Optimisation et les Systèmes d'Information) COSI'2012, 12-15 Mai 2012, Tlemcen, Algérie.

Bibliographie

[1] Reinhold Haux, "Medical informatics : Past, present, future.," *I. J. Medical Informatics*, vol. 79, no. 9, pp. 599–610, 2010.

[2] Oded Maimon and Lior Rokach, *Data Mining and Knowledge Discovery Handbook*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[3] Michael Goebel and Le Gruenwald, "A survey of data mining and knowledge discovery software tools," *SIGKDD Explor. Newsl.*, vol. 1, no. 1, pp. 20–33, June 1999.

[4] Anil Jain and Douglas Zongker, "Feature selection : Evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 153–158, Feb. 1997.

[5] Avrim Blum and Tom Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*, New York, NY, USA, 1998, COLT' 98, pp. 92–100.

[6] Yusuf Yaslan and Zehra Cataltepe, "Co-training with relevant random subspaces," *Neurocomput.*, vol. 73, no. 10-12, pp. 1652–1661, June 2010.

[7] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*, MIT Press, Cambridge, MA, 2006.

[8] Antoine Cornuéjols and Laurent Miclet, *Apprentissage artificiel : Concepts et algorithmes*, Eyrolles, June 2010.

[9] Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby, "Selective sampling using the query by committee algorithm," *Mach. Learn.*, vol. 28, no. 2-3, pp. 133–168, Sept. 1997.

[10] Jiawei Han, Micheline Kamber, and Jian Pei, *Data Mining : Concepts and Techniques*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[11] Isabelle Guyon and André Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

[12] Huan Liu and Hiroshi Motoda, *Feature Extraction, Construction and Selection : A Data Mining Perspective*, Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[13] Huan Liu and Hiroshi Motoda, *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*, Chapman & Hall/CRC, 2007.

[14] Sally Goldman and Yan Zhou, "Enhancing supervised learning with unlabeled data," in *Proc. 17th International Conf. on Machine Learning*. 2000, pp. 327–334, Morgan Kaufmann, San Francisco, CA.

[15] Yan Zhou and Sally Goldman, ''Democratic co-learning,'' in *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence*, Washington, DC, USA, 2004, ICTAI '04, pp. 594–202, IEEE Computer Society.

[16] Zhi-Hua Zhou and Ming Li, ''Tri-training : Exploiting unlabeled data using three classifiers,'' *IEEE Trans. on Knowl. and Data Eng.*, vol. 17, no. 11, pp. 1529–1541, Nov. 2005.

[17] Xiaojin Zhu, ''Semi-Supervised learning literature survey,'' Tech. Rep., Computer Sciences, University of Wisconsin-Madison, 2005.

[18] Ming Li and Zhi-Hua Zhou, ''Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples,'' *Trans. Sys. Man Cyber. Part A*, vol. 37, no. 6, pp. 1088–1098, Nov. 2007.

[19] Christian Leistner, Amir Saffari, Jakob Santner, and Horst Bischof, ''Semi-supervised random forests,'' in *ICCV IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, 2009, pp. 506–513.

[20] Neil Beagrie, ''E-infrastructure strategy for research : Final report from the osi preservation and curation working group,'' Tech. Rep., Beagrie Publishing. http ://www.nesc.ac.uk/documents/OSI/preservation.pdf, 2007.

[21] Conseil Scientifique, *Rapport du groupe de travail sur la gestion et le partage des données*, INRA, Juin 2012.

[22] Clement Jonquet, Mark A. Musen, and Nigam H. Shah, ''Help will be provided for this task : Ontology-based annotator web service,'' Tech. Rep. BMIR-2008-1317, 2008.

[23] Emily J. Richardson and Mick Watson, ''The automatic annotation of bacterial genomes.,'' *Briefings in Bioinformatics*, vol. 14, no. 1, pp. 1–12, 2013.

[24] Clement Jonquet, Mark A. Musen, and Nigam Shah, ''A system for ontology-based annotation of biomedical data.,'' in *DILS*, Amos Bairoch, Sarah Cohen Boulakia, and Christine Froidevaux, Eds. 2008, vol. 5109 of *Lecture Notes in Computer Science*, pp. 144–152, Springer.

[25] Nigam H. Shah, Clement Jonquet, Annie P. Chiang, Atul J. Butte, Rong Chen, and Mark A. Musen, ''Ontology-driven indexing of public datasets for translational bioinformatics.,'' *BMC bioinformatics*, vol. 10 Suppl 2, no. Suppl 2, pp. S1+, 2009.

[26] Clement Jonquet, Mark A. Musen, and Nigam H. Shah, ''Building a biomedical ontology recommender web service,'' *J. Biomedical Semantics*, vol. 1, no. S-1, pp. S1, 2010.

[27] Olivier Bodenreider and Robert Stevens, ''Bio-ontologies : current trends and future directions.,'' *Brief Bioinform*, vol. 7, no. 3, pp. 256–274, Sept. 2006.

[28] Robert Moskovitch, Susana B. Martins, Eytan Behiri, Aviram Weiss, and Yuval Shahar, ''A comparative evaluation of full-text, concept-based, and context-sensitive search,'' *Journal of the American Medical Informatics Association*, vol. 14, no. 2, pp. 164 – 174, 2007.

[29] Catia Pesquita, Daniel Faria, André O. Falcão, Phillip Lord, and Francisco M. Couto, ''Semantic similarity in biomedical ontologies,'' *PLoS Comput Biol*, vol. 5, no. 7, pp. e1000443, 07 2009.

[30] Thierry Mathieu, Laurent Bermont, Jean-Christophe Boyer, Céline Versuyft, Alexandre Evrard, Isabelle Cuvelier, Remy Couderc, and Katell Peoc'h, ''Champs lexicaux de la médecine prédictive et personnalisée,'' *Ann Biol Clin*, vol. Vol 70, num 6,, pp. 651–8, 2012.

[31] Fabio Roli, ''Semi-supervised multiple classifier systems : Background and research directions.,'' in *Multiple Classifier Systems*, Nikunj C. Oza, Robi Polikar, Josef Kittler, and Fabio Roli, Eds. 2005, vol. 3541 of *Lecture Notes in Computer Science*, pp. 1–11, Springer.

[32] Nigam Shah, ''Biomedical data/content acquisition, curation.,'' in *Encyclopedia of Database Systems*, Ling Liu and M. Tamer Ozsu, Eds., pp. 224–229. Springer US, 2009.

[33] Nicolas Hervé and Nozha Boujemaa, ''Automatic image annotation,'' in *Encyclopedia of Database Systems*, pp. 180–187. 2009.

[34] Nguyen Bach and Sameer Badaskar, ''A Review of Relation Extraction,'' 2007.

[35] David Nadeau and Satoshi Sekine, ''A survey of named entity recognition and classification,'' *Linguisticae Investigationes*, vol. 30, no. 1, pp. 3–26, January 2007, Publisher : John Benjamins Publishing Company.

[36] Asma Ben Abacha and Pierre Zweigenbaum, ''Medical entity recognition : A comparison of semantic and statistical methods,'' in *Proceedings of BioNLP 2011 Workshop*, Stroudsburg, PA, USA, 2011, BioNLP '11, pp. 56–64, Association for Computational Linguistics.

[37] Leo Breiman, ''Bagging predictors,'' *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[38] Yoav Freund and Robert E. Schapire, ''Experiments with a new boosting algorithm,'' in *Proceedings of the Thirteenth International Conference on Machine Learning (ICML 1996)*, Lorenza Saitta, Ed. 1996, pp. 148–156, Morgan Kaufmann.

[39] Tin Kam Ho, ''The random subspace method for constructing decision forests,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.

[40] Thomas G. Dietterich and Ghulum Bakiri, ''Error-correcting output codes : A general method for improving multiclass inductive learning programs,'' in *IN PROCEEDINGS OF AAAI-91*. 1991, pp. 572–577, AAAI Press.

[41] Jianping Hua, Waibhav D. Tembe, and Edward R. Dougherty, ''Performance of feature-selection methods in the classification of high-dimension data,'' *Pattern Recogn.*, vol. 42, no. 3, pp. 409–424, Mar. 2009.

[42] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga, ''A review of feature selection techniques in bioinformatics,'' *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Sept. 2007.

[43] Jennifer G. Dy and Carla E. Brodley, ''Feature selection for unsupervised learning,'' *J. Mach. Learn. Res.*, vol. 5, pp. 845–889, Dec. 2004.

[44] Haytham Elghazel and Alex Aussem, ''Feature selection for unsupervised learning using random cluster ensembles,'' *2013 IEEE 13th International Conference on Data Mining*, vol. 0, pp. 168–175, 2010.

[45] Yi Hong, Sam Kwong, Yuchou Chang, and Qingsheng Ren, ''Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm,'' *Pattern Recogn.*, vol. 41, no. 9, pp. 2742–2756, Sept. 2008.

[46] Pabitra Mitra, C. A. Murthy, and Sankar K. Pal, ''Unsupervised feature selection using feature similarity,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 301–312, Mar. 2002.

[47] Yubo Cheng, Yunpeng Cai, Yijun Sun, and Jian Li, ''Semi-supervised feature selection under logistic i-relief framework.,'' in *ICPR*. 2008, pp. 1–4, IEEE.

[48] Jiangtao Ren, Zhengyuan Qiu, Wei Fan, Hong Cheng, and Philip S. Yu, ''Forward semi-supervised feature selection,'' in *Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, Berlin, Heidelberg, 2008, PAKDD'08, pp. 970–976, Springer-Verlag.

[49] Zheng Zhao and Huan Liu, ''Semi-supervised feature selection via spectral analysis.,'' in *SDM*. 2007, SIAM.

[50] Jidong Zhao, Ke Lu, and Xiaofei He, ''Locality sensitive semi-supervised feature selection,'' *Neurocomput.*, vol. 71, no. 10-12, pp. 1842–1849, June 2008.

[51] L. Breiman, ''Random forests,'' *Machine Learning*, vol. 45, pp. 5–32, 2001.

[52] Yali Amit and Donald Geman, ''Shape quantization and recognition with randomized trees,'' *Neural Computation*, vol. 9, no. 7, pp. 1545–1588, 1997.

[53] Marko Robnik-Sikonja, ''Improving random forests,'' *Machine Learning ECML 2004*, pp. 359–370, 2004.

[54] Joaquan Abellan and Andres R. Masegosa, ''A filter-wrapper method to select variables for the naive bayes classifier based on credal decision trees,'' *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 17*, vol. 6, pp. 833–854, 2009.

[55] Andrés Cano, Andrés R. Masegosa, and Serafín Moral, ''A bayesian random split to build ensembles of classification trees,'' in *Symbolic and Quantitative Approaches to Reasoning with Uncertainty, 10th European Conference, ECSQARU 2009, Verona, Italy, July 1-3, 2009. Proceedings*, 2009, pp. 469–480.

[56] Evanthia E. Tripoliti, Dimitrios I. Fotiadis, Maria Argyropoulou, and George Manis, ''A six stage approach for the diagnosis of the alzheimer's disease based on fmri data,'' *Journal of Biomedical Informatics*, vol. 43, no. 2, pp. 307–320, 2010.

[57] Mostafa El Habib Daho, Nesma Settouti, Mohamed Amine Lazouni, and Mohamed Amine Chikh, ''Weighted vote for trees aggregation in random forest,'' in *The 4th International Conference on Multimedia Computing and Systems (ICMCS'14). Marrakesh, Morocco, Pages : 453 - 458, IEEE Xplore, DOI :10.1109/ ICMCS.2014.6911235*, April 14-16 2014.

[58] Mostafa El Habib Daho, Nesma Settouti, Mohammed Amine Lazouni, and Mohamed Amine Chikh, ''Amélioration des forets aléatoires pour une meilleure prédiction.,'' in *ICA2IT International Conference on Artificial Intelligence and Information Technology. Ouargla, Algeria.*, March 10-12, 2014.

[59] Nesma Settouti, Mohamed Amine Lazouni, and Mohamed Amine Chikh, ''An optimized fuzzy random forest,'' *Submitted to : Special Issue on Advances in "Fuzzy Data Analysis and Classification"*, Mai 2015.

[60] Nesma Settouti, Mostafa El Habib Daho, Mohammed El Amine Lazouni, and Mohammed Amine Chikh, ''Random forest in semi-supervised learning (co-forest),'' in *8th International Workshop on Systems, Signal Processing and their Applications (WoSSPA), 2013*, May 2013, pp. 326–329.

[61] Nesma Settouti, Mohamed Amine Chikh, and Vincent Barra, ''An optimized semi supervised learning approach for large dimensional dataset,'' *Second Lecture in : Advances in Data Analysis and Classification.*, November 2014.

[62] Nesma Settouti, Mohamed Amine Chikh, and Meryem Saidi, ''Applications des forets aléatoires pour la sélection de descripteurs de données médicales,'' in *COSI (Colloque sur l'Optimisation et les Systemes d'Information) COSI'2012, Tlemcen, Algérie*, 12-15 Mai 2012.

[63] Nesma Settouti, Mohamed Amine Lazouni, Mostafa El Habib Daho, and Mohamed Amine Chikh, ''Identification automatique des facteurs importants qui influent sur le controle du diabete en algérie,'' in *Biomedical Engineering International Conference (BIOMEIC'14), Université de Tlemcen, Algérie.*, 15-16 october 2014.

[64] Nesma Settouti, Mohamed Amine Lazouni, and Mohamed Amine Chikh, ''Identification of the most important ocular parameters for the keratoconus diagnosis,'' *Second Lecture in : International Journal of Biomedical Engineering and Technology (IJBET).*, October 2015.

[65] Nesma Settouti, Mostafa El Habib Daho, and Mohamed Amine Chikh, ''Using of conditional inference forest to identify variable importance,'' *Accepted and published soon in the : International Journal of Bioinformatics Research and Applications*, November 2013.

[66] Nesma Settouti, Mohamed Amine Chikh, and Vincent Barra, ''A new variable selection approach based on ensemble methods in semi supervised classification.,'' *Journal of Pattern Analysis and Applications.*, vol. 10/2015 DOI : 10.1007/s10044-015-0524-9, June 2015.

[67] T. Dietterich, ''Ensemble methods in machine learning,'' *Lecture Notes in Computer Science*, vol. 1857, pp. 1–15, 2000.

[68] Saso Dzeroski and Bernard Zenko, ''Is combining classifiers with stacking better than selecting the best one ?,'' *Mach. Learn.*, vol. 54, no. 3, pp. 255–273, Mar. 2004.

[69] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, New York, 1993.

[70] Simon Bernard, Laurent Heutte, and Sébastien Adam, ''Influence of hyperparameters on random forest accuracy,'' in *Multiple Classifier Systems*, JónAtli Benediktsson, Josef Kittler, and Fabio Roli, Eds., vol. 5519 of *Lecture Notes in Computer Science*, pp. 171–180. Springer Berlin Heidelberg, 2009.

[71] Pance Panov and Saso Dzeroski, ''Combining bagging and random subspaces to create better ensembles,'' in *Proceedings of the 7th international conference on Intelligent data analysis*, Berlin, Heidelberg, 2007, IDA'07, pp. 118–129, Springer-Verlag.

[72] Gilles Louppe and Pierre Geurts, ''Ensembles on random patches,'' in *Machine Learning and Knowledge Discovery in Databases*, PeterA. Flach, Tijl Bie, and Nello Cristianini, Eds., vol. 7523 of *Lecture Notes in Computer Science*, pp. 346–361. Springer Berlin Heidelberg, 2012.

[73] Stéphane Caron, ''Une introduction aux arbres de décision,'' *https ://scaron.info/doc/intro-arbres-decision/intro.pdf*, 2011.

[74] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A., 1984.

[75] N. Sirikulviriya and S. Sinthupinyo, ''Integration of rules from a random forest.,'' in *International Conference on Information and Electronics Engineering IPCSIT vol.6 (2011) IACSIT Press, Singapore*, 2011.

[76] Nesma Settouti, Mostafa El Habib Daho, Mohammed El Amine Lazouni, and Mohammed Amine Chikh, ''Conditional inference forest for variables selection of medical data,'' in *Biomedical Engineering International Conference (BIOMEIC'12), Tlemcen, Algérie*, 2012, vol. 1, pp. 84–90.

[77] Mostafa El Habib Daho, *Classification and Recognition of Biomedical Data with Ensemble Methods*, Ph.D. thesis, Computer Sciences Departement, Faculty of Sciences, Tlemcen University, Juin 2015.

[78] J. Ross Quinlan, *C4.5 : Programs for Machine Learning*, Morgan Kaufmann, 1993.

[79] Igor Kononenko, ''On biases in estimating Multi-Valued attributes.,'' in *International Joint Conference on Artificial Intelligence*, 1995, pp. 1034–1040.

[80] Igor Kononenko, ''Estimating attributes : Analysis and extensions of relief,'' in *Proceedings of the European Conference on Machine Learning on Machine Learning*, Secaucus, NJ, USA, 1994, ECML-94, pp. 171–182, Springer-Verlag New York, Inc.

[81] Marko Robnik-Sikonja, David Cukjati, and Igor Kononenko, ''Comprehensible evaluation of prognostic factors and prediction of wound healing,'' *Artificial Intelligence in Medicine*, vol. 29, pp. 25–38, 2003.

[82] J. R. Quinlan, ''Induction of decision trees,'' *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986.

[83] J. Abellan and S. Moral, ''Building classification trees using the total uncertainty criterion,'' *Int. J. Intell. Syst.*, vol. 18, no. 12, pp. 1215–1225, 2003.

[84] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, ''Uci repository of machine learning databases,'' 1998.

[85] Pierre Geurts, Damien Ernst, and Louis Wehenkel, ''Extremely randomized trees,'' *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.

[86] JR Petrella, R Edward Coleman, and P Murali Doraiswamy, ''Neuroimaging and early diagnosis of alzheimer disease,'' *a look to the future. Radiology*, vol. 226, pp. 315–336, 2003.

[87] P Scheltens, ''Early diagnosis of dementia,'' *neuroimaging. J Neuro*, vol. 246, pp. 16–20, 1999.

[88] Padraig Cunningham, ''A taxonomy of similarity mechanisms for case-based reasoning,'' Tech. Rep., Technical Report UCD-CSI-2008-01, 2008.

[89] D. Randall Wilson and Tony R. Martinez, ''Improved heterogeneous distance functions,'' *J. Artif. Intell. Res. (JAIR)*, vol. 6, pp. 1–34, 1997.

[90] Alexey Tsymbal, Mykola Pechenizkiy, and Padraig Cunningham, ''Dynamic integration with random forests,'' in *ECML*, 2006, pp. 801–808.

[91] H Hu, J Li, H Wang, G Daggard, and M Shi, *A maximally diversified multiple decision tree algorithm for microarray data classification*, The 2006 workshop on intelligent systems for bioinformatics (WISB2006), 2006.

[92] S Gunter and H Bunke, ''Optimization of weights in a multiple classifier handwritten word recognition system using a genetic algorithm,'' *Electron Lett Comput Vis Image Anal 2004*, vol. 3, pp. 25–41.

[93] Kevin Woods, W. Philip Kegelmeyer, Jr., and Kevin Bowyer, ''Combination of multiple classifiers using local accuracy estimates,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 4, pp. 405–410, Apr. 1997.

[94] *Hybrid ensembles and coincident-failure diversity*, vol. 4, 2001.

[95] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas, ''Effective voting of heterogeneous classifiers,'' in *In Proceedings of the 15th European Conference on Machine Learning*, 2004, pp. 465–476.

[96] Shun Bian and Wenjia Wang, ''On diversity and accuracy of homogeneous and heterogeneous ensembles,'' *Int. J. Hybrid Intell. Syst.*, vol. 4, no. 2, pp. 103–128, Apr. 2007.

[97] Adele Cutler and Guohua Zhao, ''Pert - perfect random tree ensembles,'' *Computing Science and Statistics*, p. 497, 2001.

[98] Simon Bernard, *Random Forests : From the Study of Behaviors to Dynamic Induction*, Theses, Université de Rouen, Dec. 2009.

[99] Rachid Baroudi, ''Une approche d'estimation de la criticité des agents basée sur la classification floue,'' in *Confèrence Internationale sur l'Informatique et ses Applications CIIA'08, SAIDA, Algérie*, 2008.

[100] Franck Dernoncourt, ''Fuzzy logic : between human reasoning and artificial intelligence,'' M.S. thesis, ENS Ulm, January 2011.

[101] Laurence Cornez, ''Discrimination automatique à base de connaissances expertes d'événements sismiques,'' in *CIIA*, 2007.

[102] Mostafa El Habib Daho, Nesma Settouti, Mohamed Amine Lazouni, and Mohamed Amine Chikh, ''Optimization du systeme nefclass,'' in *WoSSPA'13 The 9th International Workshop on Systems, Signal Processing and their Applications, Hotel Safir, Mazafran, Algiers, Algeria*, Juin 12-15 May 2013.

[103] H. Laanaya, A. Martin, D. Aboutajdine, and A. Khenchaf, ''Régression floue et crédibiliste par svm pour la classification des images sonar,'' in *Extraction et Gestion des Connaissances (EGC)*, Namur, Belgique, 24-26 January 2007, pp. 21–32.

[104] Sabeur Elkosantini, *Introduction a la logique floue : Les concepts fondamentaux et applications*, 2010.

[105] Mohammed Amine Chikh and Pierre Yves Glorennec, ''Application des arbres de décision flous á la reconnaissance des bvps,'' in *Seconde édition des Journées d'Etude algéro-francaises en Imagerie Médicale, USTHB 21-22 Novembre, Alger, Algérie.*, 2006.

[106] Marsala Christophe, ''Application of fuzzy rule induction to data mining,'' in *FQAS*, 1998.

[107] Pascal Garcia, *Utilisation d'arbres de décision flous*, Ph.D. thesis, l'INSA de Rennes, July 2004.

[108] Piero Bonissone, José Manuel Cadenas, M. Carmen Garrido, and R. Andrés Diaz-Valladares, ''A fuzzy random forest : Fundamental for design and construction,'' *Int. J. Approx. Reasoning*, vol. 51, no. 8, pp. 729–747, 2008.

[109] Piero Bonissone, José Manuel Cadenas, Maria del Carmen Garrido, and Ramon A.Diaz-Valladares, ''Combination methods in a fuzzy random forest,'' in *SMC*, 2008, pp. 1794–1799.

[110] Piero Bonissone, José Manuel Cadenas, M.Carmen Garrido, Ramon A.Diaz-Valladares, and Raquel Martinez, ''Weighted decisions in a fuzzy random forest,'' in *IFSA/EUSFLAT Conf.*, 2009, pp. 1553–1558.

[111] Piero Bonissone, José Manuel Cadenas, M. Carmen Garrido, and R. Andrés Diaz-Valladares, ''A fuzzy random forest,'' *Int. J. Approx. Reasoning*, vol. 51, no. 7, pp. 729–747, 2010.

[112] Jose M. Cadenas, M. Carmen Garrido, Raquel Martinez, and Piero P. Bonissone, ''Towards the learning from low quality data in a fuzzy random forest ensemble.,'' in *FUZZ-IEEE*. 2011, pp. 2897–2904, IEEE.

[113] Marsala Christophe, ''Apprentissage artificiel et raisonnement flou,'' in *FUZZ-IEEE*, 30 novembre 2010.

[114] M. Sugeno and G. T. Kang, ''Structure identification of fuzzy model,'' *Fuzzy Sets Syst.*, vol. 28, pp. 15–33, October 1988.

[115] James C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Kluwer Academic Publishers, Norwell, MA, USA, 1981.

[116] J. Dunn, ''A fuzzy relative of the isodata process and its use in detecting compact, well-separated clusters,'' *Journal of Cybernetics*, vol. vol. 3,, pp. 32–57, 1974.

[117] Nesma Settouti, M Amine Chikh, and Meryem Saidi, ''Generating fuzzy rules for constructing interpretable classifier of diabetes disease,'' *Australasian Physical & Engineering Sciences in Medicine*, vol. 35, no. 3, pp. 257–270, 2012.

[118] Khalida Douibi, Nesma Settouti, and Mohamed Amine Chikh, ''Bagged mlknn : Bootstrap and aggregating k nearest neighbors for multi-label data classification,'' in *Knowledge Discovery and Data Analysis KDDA'2015, ESI, Algiers, Algeria*, November 15- 18, 2015.

[119] A. P. Dempster, N. M. Laird, and D. B. Rubin, ''Maximum likelihood from incomplete data via the EM algorithm,'' *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[120] David Yarowsky, ''Unsupervised word sense disambiguation rivaling supervised methods,'' in *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, 1995, ACL '95, pp. 189–196, Association for Computational Linguistics.

[121] Kamal Nigam and Rayid Ghani, ''Analyzing the effectiveness and applicability of co-training,'' in *Proceedings of the Ninth International Conference on Information and Knowledge Management*, New York, NY, USA, 2000, CIKM '00, pp. 86–93, ACM.

[122] Mohamed Farouk Abdel Hady and Friedhelm Schwenker, ''Co-training by committee : A generalized framework for semi-supervised learning with committees.,'' *Int. J. Software and Informatics*, vol. 2, no. 2, pp. 95–124, 2008.

[123] Fabio G. Cozman and Ira Cohen, ''Unlabeled data can degrade classification performance of generative classifiers,'' in *in Fifteenth International Florida Artificial Intelligence Society Conference*, 2002, pp. 327–331.

[124] Geoffrey J. Mclachlan, *Discriminant Analysis and Statistical Pattern Recognition (Wiley Series in Probability and Statistics)*, Wiley-Interscience, Aug. 2004.

[125] L. G. Valiant, ''A theory of the learnable,'' *Commun. ACM*, vol. 27, no. 11, pp. 1134–1142, Nov. 1984.

[126] Boaz Leskes and Leen Torenvliet, ''The value of agreement a new boosting algorithm,'' *J. Comput. Syst. Sci.*, vol. 74, no. 4, pp. 557–586, June 2008.

[127] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, ''Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays,'' in *Proceedings of the National Academy of Sciences of the United States of America*, Department of Molecular Biology, Princeton University, Princeton, NJ 08540, USA., June 1999, vol. 96, pp. 6745–6750.

[128] C. Best, ''Prostate cancer - comparison of androgen-dependent and -independent microdissected primary tumor. nci, nih ; labmolecular therapeutics. usa.,'' http ://www.biolab.si/supp/bi-cancer/projections/datasets/prostateGSE2443.tab, June 2012.

[129] C. Baer, M. Nees, S. Breit, B. Selle, AE. Kulozik, K. Schaefer, Y. Braun, D. Wai, and C. Poremba, ''Pediatric sarcoma : rhabdomyosarcoma & ewing's sarcoma. heidelberg university, children's hospital department pediatric oncology & hematology germany.,'' http ://www.biolab.si/supp/bi-cancer/projections/datasets/EWSGSE967.tab., Mars 2012.

[130] T R Golub, D K Slonim, P Tamayo, C Huard, M Gaasenbeek, J P Mesirov, H Coller, M L Loh, J R Downing, M A Caligiuri, C D Bloomfield, and E S Lander, ''Molecular classification of cancer : class discovery and class prediction by gene expression monitoring,'' *Science*, vol. 286, no. 5439, pp. 531–537, Oct. 1999.

[131] JC. Chang, EC. Wooten, A. Tsimelzon, SG. Hilsenbeck, MC. Gutierrez, R. Elledge, S. Mohsin, CK. Osborne, GC. Chamness, DC. Allred, and P. O'Connell, ''Therapeutic response to docetaxel in patients with breast cancer. baylor college of medicine ; department breast center.,'' http ://www.biolab.si/supp/bi-cancer/projections/datasets/BCGSE349/350.tab, June 2012.

[132] Scott L. Pomeroy, Pablo Tamayo, Michelle Gaasenbeek, Lisa M. Sturla, Michael Angelo, Margaret E. McLaughlin, John Y. H. Kim, Liliana C. Goumnerova, Peter M. Blackand Ching Lau, Jeffrey C. Allen, David Zagzag, James M. Olson, Tom Curran, Cynthia Wetmore, Jaclyn A. Biegel, Tomaso Poggio, Shayan Mukherjee, Ryan Rifkinand Andrea Califano, Gustavo Stolovitzky, David N. Louisand Jill P. Mesirov, Eric S. Lander, and Todd R. Golub, ''Prediction of central nervous system embryonal tumour outcome based on gene expression.,'' *Nature journal*, vol. 415, pp. 436–442, 24 January 2002.

[133] T. Shaarawy, M.B. Sherwood, and J.G. Crowston, *Glaucoma : Medical diagnosis & therapy*, ClinicalKey 2012. Saunders/Elsevier, 2009.

[134] J. G. Crowston, C. R. Hopley, P. R. Healey, A. Lee, and P. Mitchell, ''The effect of optic disc diameter on vertical cup to disc ratio percentiles in a population based cohort : the blue mountains eye study,'' *The British Journal of Ophthalmology,*, vol. 88(6), pp. 766–770, 2004.

[135] David Yarowsky, ''Unsupervised word sense disambiguation rivaling supervised methods,'' in *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, 1995, ACL '95, pp. 189–196, Association for Computational Linguistics.

[136] Beatriz Maeireizo, Diane Litman, and Rebecca Hwa, ''Co-training for predicting emotions with spoken dialogue data,'' in *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL-2004)*, 2004.

[137] Chao Deng and Maozu Guo, ''A new co-training-style random forest for computer aided diagnosis.,'' *J. Intell. Inf. Syst.*, vol. 36, no. 3, pp. 253–281, 2011.

[138] Marc Lalonde, Mario Beaulieu, and Langis Gagnon, ''Fast and robust optic disc detection using pyramidal decomposition and hausdorff-based template matching.,'' *IEEE Trans. Med. Imaging*, vol. 20, no. 11, pp. 1193–1200, 2001.

[139] Huiqi Li and Opas Chutatape, ''Automated feature extraction in color retinal images by a model based approach,'' *IEEE Trans. Biomed. Engineering*, vol. 51, no. 2, pp. 246–254, 2004.

[140] Michael B. Merickel, Jr., Michael D. Abramoff, Milan Sonka, and Xiaodong Wu, ''Segmentation of the optic nerve head combining pixel classification and graph search,'' *Proc. SPIE*, vol. 6512, pp. 651215–651215–10, 2007.

[141] Yanwu Xu, Dong Xu, Stephen Lin, Jiang Liu 0001, Jun Cheng, Carol Yim lui Cheung, Tin Aung, and Tien Yin Wong, ''Sliding window and regression based cup detection in digital fundus images for glaucoma diagnosis.,'' in *MICCAI (3)*, Gabor Fichtinger, Anne L. Martel, and Terry M. Peters, Eds. 2011, vol. 6893 of *Lecture Notes in Computer Science*, pp. 1–8, Springer.

[142] Jun Cheng, Jiang Liu, Yanwu Xu, Fengshou Yin, Damon Wing Kee Wong, Ngan-Meng Tan, Ching-Yu Cheng, Yih Chung Tham, and Tien Yin Wong, ''Superpixel classification based optic disc segmentation,'' in *Proceedings of the 11th Asian Conference on Computer Vision - Volume Part II*, Berlin, Heidelberg, 2013, ACCV'12, pp. 293–304, Springer-Verlag.

[143] Michael D. Abramoff, Wallace L. M. Alward, Emily C. Greenlee, Lesya Shuba, Chan Y. Kim, John H. Fingert, and Young H. Kwon, ''Automated segmentation of the optic disc from stereo color photographs using physiologically plausible features,'' *Investigative Ophthalmology & Visual Science*, vol. 48, no. 4, pp. 1665, 2007.

[144] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, ''Slic superpixels compared to state-of-the-art superpixel methods,'' *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2274–2282, Nov 2012.

[145] P. Neubert and P. Protzel, ''Superpixel benchmark and comparison.,'' in *Proc. of Forum Bildverarbeitung, Regensburg, Germany*, 2012.

[146] Thomas Walter and Jean-Claude Klein, ''Segmentation of color fundus images of the human retina : Detection of the optic disc and the vascular tree using morphological techniques,'' in *Proceedings of the Second International Symposium on Medical Data Analysis*, London, UK, UK, 2001, ISMDA '01, pp. 282–287, Springer-Verlag.

[147] N.A. Mohamed, M.A. Zulkifley, and A. Hussain, ''On analyzing various density functions of local binary patterns for optic disc segmentation,'' in *Computer Applications Industrial Electronics (ISCAIE), 2015 IEEE Symposium on*, April 2015, pp. 37–41.

[148] Pardha Saradhi Mittapalli and Giri Babu Kande, ''Segmentation of optic disk and optic cup from digital fundus images for the assessment of glaucoma,'' *Biomedical Signal Processing and Control*, vol. 24, pp. 34 – 46, 2016.

[149] Gopal Datt Joshi, Jayanthi Sivaswamy, and S. R. Krishnadas, ''Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment,'' *IEEE Trans. Med. Imaging*, vol. 30, no. 6, pp. 1192–1205, 2011.

[150] Muhammad Salman Haleem, Liangxiu Han, Jano van Hemert, and Baihua Li, ''Automatic extraction of retinal features from colour retinal images for glaucoma diagnosis : A review,'' *Computerized Medical Imaging and Graphics*, vol. 37, no. 7(8), pp. 581 – 596, 2013.

[151] D.W.K. Wong, J. Liu, J.H. Lim, X. Jia, F. Yin, H. Li, and T.Y. Wong, ''Level-set based automatic cup-to-disc ratio determination using retinal fundus images in argali,'' in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, Aug 2008, pp. 2266–2269.

[152] J. Lowell, A. Hunter, D. Steel, A. Basu, R. Ryder, E. Fletcher, and L. Kennedy, "Optic nerve head segmentation," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 2, pp. 256–264, Feb 2004.

[153] Chunming Li, Chiu-Yen Kao, John C. Gore, and Zhaohua Ding, "Implicit active contours driven by local binary fitting energy.," in *CVPR*. 2007, IEEE Computer Society.

[154] Giri Babu Kande, P. Venkata Subbaiah, and Satya Savithri Tirumala, "Unsupervised fuzzy based vessel segmentation in pathological digital fundus images.," *J. Medical Systems*, vol. 34, no. 5, pp. 849–858, 2010.

[155] Yanwu Xu, Lixin Duan, Stephen Lin, Xiangyu Chen, Damon Wing Kee Wong, Tien Yin Wong, and Jiang Liu, "Optic cup segmentation for glaucoma detection using low-rank superpixel representation," in *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention*, Boston, Massachusetts, United States, September 2014, pp. 788–795.

[156] Chisako Muramatsu, Toshiaki Nakagawa, Akira Sawada, Yuji Hatanaka, Takeshi Hara, Tetsuya Yamamoto, and Hiroshi Fujita, "Automated segmentation of optic disc region on retinal fundus photographs : Comparison of contour modeling and pixel classification methods.," *Computer Methods and Programs in Biomedicine*, vol. 101, no. 1, pp. 23–32, 2011.

[157] Brian Fulkerson, Andrea Vedaldi, and Stefano Soatto, "Class segmentation and object localization with superpixel neighborhoods.," in *ICCV*. 2009, pp. 670–677, IEEE.

[158] C. H. Wu and H. H. Chang, "Gaussian noise estimation with superpixel classification in digital images," in *International Congress on Image and Signal Processing CISP'12, 16 Oct - 18 Oct 2012, Chongqing University of Posts and Telecommunications Chongqing, Sichuan, China*, 2012, pp. 373–377.

[159] Greg Mori, "Guiding model search using segmentation.," in *ICCV*. 2005, pp. 1417–1423, IEEE Computer Society.

[160] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vision*, vol. 59, no. 2, pp. 167–181, Sept. 2004.

[161] Alex Levinshtein, Adrian Stere, Kiriakos N. Kutulakos, David J. Fleet, Sven J. Dickinson, and Kaleem Siddiqi, "Turbopixels : Fast superpixels using geometric flows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2290–2297, Dec. 2009.

[162] James D. Foley, Andries van Dam, Steven K. Feiner, and John F. Hughes, *Computer Graphics : Principles and Practice (2Nd Ed.)*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1990.

[163] Nikolaos Giannakeas, Petros S. Karvelis, Themis P. Exarchos, Fanis G. Kalatzis, and Dimitrios I. Fotiadis, "Segmentation of microarray images using pixel classification - comparison with clustering-based methods.," *Comp. in Bio. and Med.*, vol. 43, no. 6, pp. 705–716, 2013.

[164] Quan Wang and Kim L. Boyer, "The active geometric shape model : A new robust deformable shape model and its applications," *Comput. Vis. Image Underst.*, vol. 116, no. 12, pp. 1178–1194, Dec. 2012.

[165] M Kass, A Witkin, and D Terzopoulos, "Snakes - Active Contour Models," *International Journal Of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1987.

[166] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models&mdash ;their training and application," *Comput. Vis. Image Underst.*, vol. 61, no. 1, pp. 38–59, Jan. 1995.

[167] Chenyang Xu and Jerry L. Prince, "Snakes, shapes, and gradient vector flow," *IEEE TRANSACTIONS ON IMAGE PROCESSING*, vol. 7, no. 3, pp. 359–369, 1998.

[168] F. Fumero, Silvia Alayen, J. L. Sanchez, Jos. Sigut, and M. Gonzalez-Hernendez, ''Rim-one : An open retinal image database for optic nerve evaluation.,'' in *CBMS*. 2011, pp. 1–6, IEEE Computer Society.

[169] Tom M. Mitchell, ''The role of unlabeled data in supervised learning,'' in *Proceedings of the Sixth International Colloquium on Cognitive Science. San Sebastian, Spain.*, 1999.

[170] Nitesh V. Chawla and Grigoris Karakoulas, ''Learning from labeled and unlabeled data : An empirical study across techniques and domains,'' *J. Artif. Int. Res.*, vol. 23, no. 1, pp. 331–366, Mar. 2005.

[171] Maria F. Balcan, Avrim Blum, and Ke Yang, ''Co-Training and expansion : Towards bridging theory and practice,'' in *Advances in Neural Information Processing Systems 17*, Lawrence K. Saul, Yair Weiss, and Léon Bottou, Eds., pp. 89–96. MIT Press, Cambridge, MA, 2005.

[172] Rie K. Ando and Tong Zhang, ''Two-view feature generation model for semi-supervised learning,'' in *ICML '07 : Proceedings of the 24th international conference on Machine learning*, New York, NY, USA, 2007, pp. 25–32, ACM.

[173] Jiao Wang, Siwei Luo, and Xianhua Zeng, ''A random subspace method for co-training.,'' in *IJCNN*. 2008, pp. 195–200, IEEE.

[174] Yuan Jiang and Zhi hua Zhou, ''Editing training data for knn classifiers with neural network ensemble,'' in *Lecture Notes in Computer Science, Vol.3173*. 2004, pp. 356–361, Springer.

[175] B. Efron, ''Bootstrap methods : Another look at the jackknife,'' *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, 01 1979.

[176] Zhi-Hua Zhou, Jianxin Wu, and Wei Tang, ''Ensembling neural networks : Many could be better than all,'' *Artif. Intell.*, vol. 137, no. 1-2, pp. 239–263, May 2002.

[177] David E. Goldberg and Kalyanmoy Deb, ''A comparative analysis of selection schemes used in genetic algorithms,'' in *Foundations of Genetic Algorithms*. 1991, pp. 69–93, Morgan Kaufmann.

[178] Janez Demsar, ''Statistical comparisons of classifiers over multiple data sets,'' *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006.

[179] Nesma Settouti, Mohammed El Amine Bechar, and Mohammed Amine Chikh, ''Statistical comparisons of the top 10 algorithms in data mining for classification task,'' *International Journal of Interactive Multimedia and Artificial Inteligence*, vol. 4, no. 1, pp. 46–51, 09/2016 2016.

[180] R. A. Fisher, *Statistical Methods for Research Workers*, Cosmo study guides. Cosmo Publications, 1925.

[181] Bogdan Trawinski, Magdalena Smetek, Zbigniew Telec, and Tadeusz Lasota, ''Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms.,'' *Applied Mathematics and Computer Science*, vol. 22, no. 4, pp. 867–881, 2012.

[182] David J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, 4 edition, 2007.

[183] Salvador García, Alberto Fernández, Julián Luengo, and Francisco Herrera, ''Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining : Experimental analysis of power,'' *Inf. Sci.*, vol. 180, no. 10, pp. 2044–2064, May 2010.

[184] Reza Zafarani and Huan Liu, ''Asu repository of social computing databases,'' 1998.

[185] Mohammed Hindawi, Haytham Elghazel, and Khalid Benabdeslem, ''Efficient semi-supervised feature selection by an ensemble approach,'' in *International Workshop on Complex Machine Learning Problems with Ensemble Methods COPEM@ECML/PKDD'13*, Sept. 2013, pp. 41–55.

[186] L. Yu and H. Liu, ''Feature selection for high-dimensional data : a fast correlation-based filter solution,'' in *Machine Learning-International Workshop Then Conference-*, 2003, vol. 20, p. 856.

[187] K. Pearson, ''On lines and planes of closest fit to systems of points in space,'' *Philosophical Magazine*, vol. 2, pp. 559–572, 1901.

[188] Gilbert Saporta, *Probabilités, analyse des données et statistique*, Editions Technip, 1990.

[189] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller, ''Nonlinear component analysis as a kernel eigenvalue problem,'' *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, July 1998.

[190] Sam T. Roweis and Lawrence K. Saul, ''Nonlinear Dimensionality Reduction by Locally Linear Embedding,'' *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[191] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford, ''A global geometric framework for nonlinear dimensionality reduction,'' *Science*, vol. 290, no. 5500, pp. 2319, 2000.

[192] Daphne Koller and Mehran Sahami, ''Toward optimal feature selection.,'' in *ICML*, Lorenza Saitta, Ed. 1996, pp. 284–292, Morgan Kaufmann.

[193] George H. John, Ron Kohavi, and Karl Pfleger, ''Irrelevant features and the subset selection problem,'' in *Machine Learning : Proceedings Of The Eleventh International*. 1994, pp. 121–129, Morgan Kaufmann.

[194] S. Guerif, ''Unsupervised feature selection : when random ranking sound are irrelevancy,'' *In JMCR workshop and conference proceding, New challenges for feature selection in data mining and knowledge discovery*, vol. 4, pp. 161–175, 2008.

[195] M. Hall, ''Correlation-based feature selection for machine learning,'' 1998.

[196] George H . John, Ron Kohavi, and Karl Pfleger, ''Feature and the subset selection problem.,'' *In international conference on Machine learning.*, pp. 121–129, 1994.

[197] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth, ''Warmuth and ocam's razor. information processing letters,'' *In international conference of machine learning*, vol. 24(6), pp. 377–380, 1987.

[198] Y Bennani, ''La sélection de variables,,'' *Numéro spécial de la Revue d'Intelligence Artificielle. Hermes, Paris.*, pp. 351–371, 2001a.

[199] Y. Bennani, ''Systemes d'apprentissage connexionnistes : sélection de variables,'' *Numéro spécial de la Revue d'Intelligence Artificielle. Hermes, Paris.*, , no. g, 2006.

[200] .J. G Dy and C. E. Brodley, ''Feature subset selection and order identification-for unsupervised learning.,'' in *Proceedings of the 17th International Conference onMachine Learning (ICML'2000), Stanford University, CA*, 2000.

[201] A Dean Raftery, ''Variable selection for model-based clustrering,'' *Journal of the American Statistical Association*, vol. 473, pp. 169–178, 2006.

[202] Bennani and S. Guerif, ''Sélection de variable en apprentissage numérique non supervise,'' *In cap 07 : conférence francophone sur l'apprentissage automatique*, 2007.

[203] N. Sorh Madesen, C. Thomsen, and Pena J, ''Unsupervised feature sebset selection,'' *Proceding of the workship on probabilistic graphical models for classification*, pp. 71–82, 2003.

[204] André Elisseeff and Isabelle Guyon, ''An introduction to variable and feature selection.,'' *Journal of Machine learning Research*, 2003.

[205] Daphane Koller and Mehran Sahami, ''To waerd optimal feature selection,'' *In international conference of machine learning.*, pp. 284–292, 1996.

[206] Edwin E. Ghiselli, *Theory of psuchological Meansurement*, Mc Graw-Hill Bokk Company, 1964.

[207] Mark A. Hall, ''Correlation-based feature selection and numeric class machine learning,'' *17th Internatinnal conf. on machine learning*, pp. 359–366, 2000.

[208] Sir Maurice Kendall and Alan Stewart, ''The advanced theory of statistics,'' *4th Edition. Mcmillan Publishing, New York*, vol. 1, 1977.

[209] Huan Liu and Rudy Setiono, ''Feature selection and discretization of numeric attributes,'' *knowledge and data engineering*, vol. 16, pp. 145–153, 1995.

[210] Olivier Gaudoin, ''Méthodes statistiques pour l'ingénieur,'' *Grenoble, France*, 2002.

[211] E. Shannon, ''A mathematical theory of communication,'' *The bell System Techical Journal*, vol. 27, pp. 623–654, 1948.

[212] Nesma Settouti, Mohamed Amine Lazouni, and Mohamed Amine Chikh, ''Filter approach for selecting biochip data relevant genes of colon cancer,'' *Second lecture in : Journal of Medical Imaging and Health Informatics.*, 2013.

[213] M K Smolek and S D Klyce, ''Current keratoconus detection methods compared with a neural network approach.,'' *Investigative Ophthalmology & Visual Science*, vol. 38, no. 11, pp. 2290, 1997.

[214] P.Agostino Accardo and Stefano Pensiero, ''Neural network-based system for early keratoconus detection from corneal topography,'' *Journal of Biomedical Informatics*, vol. 35, no. 3, pp. 151 – 159, 2002.

[215] Luis Alberto Vieira de Carvalho and Marconi Soares Barbosa, ''Neural networks and statistical analysis for classification of corneal videokeratography maps based on Zernike coefficients : a quantitative comparison,'' *Arquivos Brasileiros de Oftalmologia*, vol. 71, pp. 337 – 341, 06 2008.

[216] Ashraf M. Mahmoud, Cynthia Roberts, Richard Lembach, Edward E. Herderick, and Timothy T Mcmahon, ''Simulation of machine-specific topographic indices for use across platforms,'' *Optometry & Vision Science*, vol. Volume 83 - Issue 9, pp. 682–693, September 2006.

[217] Zuzana Schlegel, Thanh Hoang-Xuan, and Damien Gatinel, ''Comparison of and correlation between anterior and posterior corneal elevation maps in normal eyes and keratoconus-suspect eyes,'' *Journal of Cataract & Refractive Surgery*, vol. 34, no. 5, pp. 789 – 795, 2008.

[218] Ugo de Sanctis, Carlotta Loiacono, Lorenzo Richiardi, Davide Turco, Bernardo Mutani, and Federico M. Grignolo, ''Sensitivity and specificity of posterior corneal elevation measured by pentacam in discriminating keratoconus/subclinical keratoconus,'' *Ophthalmology*, vol. 115, no. 9, pp. 1534 – 1539, 2008.

[219] Murilo Barreto Souza, FW Medeiros, Souza DB, R Garcia, and MR. Alves, ''Evaluation of machine learning classifiers in keratoconus detection from orbscan ii examinations,'' *Clinics*, vol. 65.12, pp. 1223 –1228, 2010.

[220] D Gatinel and A Saad, ''Prévention de l ectasie cornéenne par une nouvelle méthode de détection du kératocone frustre,'' *Réalité Ophtalmologiques*, vol. Cahier1, mars 2011.

[221] David Smadja, David Touboul, Ayala Cohen, Etti Doveh, Marcony R. Santhiago, Glauco R. Mello, Ronald R. Krueger, and Joseph Colin, ''Detection of subclinical keratoconus using an automated decision tree classification,'' *American Journal of Ophthalmology*, vol. 156, no. 2, pp. 237 – 246, 2013.

[222] Yaron S. Rabinowitz, ''Keratoconus,'' *Survey of Ophthalmology*, vol. 42, no. 4, pp. 297 – 319, 1998.

[223] JL Arné, ''Emc (elsevier sas, paris),'' *Ophtalmologie*, vol. 21-200-D-40, 2005.

[224] Damien Gatinel, ''Topographie cornéenne,'' http ://www.gatinel.com/chirurgie-refractive/Bilan-préopératoire/Topographie-cornéenne.htlm, Aout 2014.

[225] La clinique de la vision, ''Topographe cornéen orbscan iiz,'' http ://www.cliniquedelavision.be/technologies/topographe-corneen-orbscan/, 2014.

[226] Nawel Fandi, ''Collecte et classification de données pour la reconnaissance du kératocone.,'' M.S. thesis, Master II Génie Biomcal option EBM (Electronique Biomédicale), 2014.

[227] Damien Gatinel, ''Chapitre 3 - principaux axes et angles utiles en topographie cornéenne,'' in *Topographie cornéenne (2e édition)*, Damien Gatinel, Ed., pp. 21 – 29. Elsevier Masson, Paris, 2e édition edition, 2014.

[228] YS Rabinowitz and PJ. McDonnell, ''Computer-assisted corneal topography in keratoconus.,'' *Refract Corneal Surg*, vol. 5(6), pp. 400–8, 1989.

[229] S D Klyce, ''Computer-assisted corneal topography. high-resolution graphic presentation and analysis of keratoscopy.,'' *Investigative Ophthalmology & Visual Science*, vol. 25, no. 12, pp. 1426, 1984.

[230] N Maeda, S D Klyce, M K Smolek, and H W Thompson, ''Automated keratoconus screening with corneal topography analysis.,'' *Investigative Ophthalmology & Visual Science*, vol. 35, no. 6, pp. 2749, 1994.

[231] Sanjay N Rao, Tal Raviv, Parag A Majmudar, and Randy J Epstein, ''Role of orbscan {II} in screening keratoconus suspects before refractive corneal surgery1,'' *Ophthalmology*, vol. 109, no. 9, pp. 1642 – 1646, 2002.

[232] H Peng, F Long, and C Ding, ''Feature selection based on mutual information : criteria of max-dependency, max-relevance, and min-redundancy.,'' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226–1238, 2005.

[233] Kenji Kira and Larry A. Rendell, ''A practical approach to feature selection,'' in *Proceedings of the Ninth International Workshop on Machine Learning*, San Francisco, CA, USA, 1992, ML92, pp. 249–256, Morgan Kaufmann Publishers Inc.

[234] M. Robnik Sikonja and I. Kononenko, ''Theoretical and empirical analysis of relief and relieff,'' *Mach. Learn*, vol. 53, pp. 23–69, 2003.

[235] Huan Liu and Rudy Setiono, ''Feature selection and classification-a probabilistic wrapper approach,'' in *Proceedings of 9th International Conference on Industrial and Engineering Applications of AI and ES*, 1997, pp. 419–424.

[236] H. Chouaib, *Sélection de caractéristiques : Méthodes et applications*, 2011.

[237] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik, ''Gene selection for cancer classification using support vector machines,'' *Mach. Learn.*, vol. 46, no. 1-3, pp. 389–422, Mar. 2002.

[238] Donglin Zeng, Ruoqing Zhu, and Michael R. Kosorok, ''Reinforcement Learning Trees,'' Tech. Rep., The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series. Working Paper 33. http ://biostats.bepress.com/uncbiostat/art33, 04 2012.

[239] Chih-Chung Chang and Chih-Jen Lin, ''Libsvm : A library for support vector machines,'' *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27 :1–27 :27, May 2011.

[240] Mohamed Abdel-Aty, Anurag Pande, Chris Lee, Abhishek Das, Alexis Nevarez, Ali Darwiche, and Premchand Devarasetty, ''Reducing fatalities and severe injuries on florida's high-speed multi-lane arterial corridors,'' Tech. Rep., Center for Advanced Transportation Systems Simulation, University of Central Florida, P.O. Box 162450, Orlando, Fl 32816-2450, April 28, 2009.

[241] Ali Harb, Michel Beigbeder, and Jean-Jacques Girardot, ''Evaluation of question classification systems using differing features,'' in *International Conference for Internet Technology and Secured Transactions, 2009. ICITST 2009.*, Londres, United Kingdom, Nov. 2009, p. ...

[242] Carolin Strobl, Boulesteix Anne-Laure, Zeileis Achim, and Torsten Hothorn, ''Bias in random forest variable importance measures : Illustrations, sources and a solution,'' *BMC Bioinformatics*, pp. 8–25, 2007.

[243] Torsten Hothorn, Hornik Kurt, and Zeileis Achim, ''Unbiased recursive partitioning : A conditional inference framework,'' *Journal of Computational and Graphical Statistics*, vol. 15 (3), pp. 651–674, 2006.

[244] Abhishek Das, Mohamed Abdel-Aty, and Anurag Pande., ''Using conditional inference forests to identify the factors affecting crash severity on arterial corridors,'' *Journal of Safety Research*, vol. 40, pp. 317–327, 2009.

[245] Lidia Auret and Chris Aldrich, ''Empirical comparison of tree ensemble variable importance measures,'' *Chemometrics and Intelligent Laboratory Systems*, vol. 105, pp. 157–170, 2011.

[246] Carolin Strobl, Boulesteix Anne-Laure, Kneib Thomas, Augustin Thomas, and Zeileis Achim, ''Conditional variable importance for random forests,'' *BMC Bioinformatics*, vol. 9, pp. 307, 2008.

[247] Krisztina Nagy, Jeno Reiczigel, Andrea Harnos, Aniko Schrott, and Peter Kabai., ''Tree-based methods as an alternative to logistic regression in revealing risk factors of crib-biting in horses,'' *Journal of Equine Veterinary Science*, vol. Vol 30. No1, 2010.

[248] Kristin K Nicodemu and James D Malley, ''Predictor correlation impacts machine learning algorithms : implications for genomic studies,'' *Bioinformatics Original Paper Genetics and population analysis*, vol. Vol. 25 No. 15, pp. 1884–1890, 2009.

[249] Sebastian Strempel, Monika Nendza, Martin Scheringer, and Konrad Hungerbuhler, ''Using conditional inference trees and random forests to predict the bioaccumulation potential of organic chemicals,'' *Environmental Toxicology and Chemistry*, vol. Vol. 32, No. 5, pp. 1187–1195, 2013.

[250] Leo Guelman, Montserrat Guillen, and Ana M. Pérez-Marín, ''Optimal personalized treatment rules for marketing interventions : A review of methods, a new proposal, and an insurance case study,'' Working Papers 2014-06, Universitat de Barcelona, UB Riskcenter, May 2014.

[251] Helmut Strasser and Christian Weber, ''On the asymptotic theory of permutation statistics,'' *Mathematical Methods of Statistics*, vol. 8, pp. 220–250, 1999.

[252] K. Bessonov, ''Gene regulatory network inference via conditional inference trees and forests,'' in *23rd Annual Conference of International Genetic Epidemiology society (IGES2014)*, 2014.

[253] Torsten Hothorn, Buhlmann Peter, Dudoit Sandrine, Molinaro Annette, and J. van der Laan Mark, ''Survival ensembles,'' *Biostatistics*, vol. 7(3), pp. 355–373, 2006.

[254] Yu-Shan Shih and Hsin-Wen Tsai, ''Variable selection bias in regression trees with constant fits,'' *Computational Statistics & Data Analysis*, vol. 45, no. 3, pp. 595–607, 2004.

[255] Hasna Barkia, Haytham Elghazel, and Alex Aussem, ''Semi-supervised feature importance evaluation with ensemble learning,'' *IEEE 13th International Conference on Data Mining*, vol. 0, pp. 31–40, 2011.

[256] Fazia Bellal, Haytham Elghazel, and Alex Aussem, ''A semi-supervised feature ranking method with ensemble learning,'' *Pattern Recognition Letters*, vol. 33, no. 10, pp. 1426–1432, 2012.

[257] F Berzal, Cubero JC, Cuenca F, and Marty n Bautista MJ., ''On the quest for easy-to-understand splitting rules,'' *Data and knowledge Engineering*, vol. 44, pp. 31–48, 2003.

[258] Mark A. Hall, ''Correlation-based feature selection for discrete and numeric class machine learning,'' in *Proceedings of the Seventeenth International Conference on Machine Learning*, San Francisco, CA, USA, 2000, ICML '00, pp. 359–366, Morgan Kaufmann Publishers Inc.

[259] A. W. Whitney, ''A direct method of nonparametric measurement selection,'' *IEEE Trans. Comput.*, vol. 20, no. 9, pp. 1100–1103, Sept. 1971.

[260] T S Furey, N Cristianini, N Duffy, D W Bednarski, M Schummer, and D Haussler, ''Support vector machine classification and validation of cancer tissue samples using microarray expression data,'' *Bioinformatics*, vol. 16, no. 10, pp. 906–914, Oct. 2000, Evaluation Studies.

[261] Amit Moscovich Eiger, Boaz Nadler, and Clifford Spiegelman, ''The calibrated kolmogorov-smirnov test,'' 2013, cite arxiv :1311.3190.

[262] Koji Miyahara and Michael J. Pazzani, ''Collaborative filtering with the simple bayesian classifier,'' in *Proceedings of the 6th Pacific Rim International Conference on Artificial Intelligence*, Berlin, Heidelberg, 2000, PRICAI'00, pp. 679–689, Springer-Verlag.

[263] Kari Torkkola, ''Feature extraction by non parametric mutual information maximization,'' *J. Mach. Learn. Res.*, vol. 3, pp. 1415–1438, Mar. 2003.

[264] R. Battiti, ''Using mutual information for selecting features in supervised neural net learning,'' *Trans. Neur. Netw.*, vol. 5, no. 4, pp. 537–550, July 1994.

[265] Yoshiyuki Nakatani, Kuangyi Zhu, and Kuniaki Uehara, ''Semisupervised learning using feature selection based on maximum density subgraphs,'' *Systems and Computers in Japan*, vol. 38, no. 9, pp. 32–43, 2007.

[266] Xiangnan Kong and Philip S. Yu, ''Semi-supervised feature selection for graph classification,'' in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2010, KDD '10, pp. 793–802, ACM.

[267] Gauthier Doquire and Michel Verleysen, ''Graph laplacian for semi-supervised feature selection in regression problems.,'' in *IWANN (1)*, Joan Cabestany, Ignacio Rojas, and Gonzalo Joya Caparros, Eds. 2011, vol. 6691 of *Lecture Notes in Computer Science*, pp. 248–255, Springer.

[268] Deng Cai, Xiaofei He, Kun Zhou, Jiawei Han, and Hujun Bao, ''Locality sensitive discriminant analysis,'' in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, San Francisco, CA, USA, 2007, IJCAI'07, pp. 708–713, Morgan Kaufmann Publishers Inc.

[269] Ron Kohavi and George H. John, ''Wrappers for feature subset selection,'' *Artif. Intell.*, vol. 97, no. 1-2, pp. 273–324, Dec. 1997.

[270] David W. Aha and Richard L. Bankert, ''A comparative evaluation of sequential feature selection algorithms,'' in *Learning from Data : Artificial Intelligence and Statistics V*, Douglas H. Fisher and Hans-Joachim Lenz, Eds., Lecture Notes in Statistics, chapter 4, pp. 199–206. Springer-Verlag, 175 Fifth Avenue. New York, New York 10010, USA, 1996.

[271] Khalid Benabdeslem and Mohammed Hindawi, ''Efficient semi-supervised feature selection : Constraint, Relevance and Redundancy.,'' *IEEE Transactions on Knowledge and Data Engineering*, July 2013.

[272] Dan Sun and Daoqiang Zhang, ''Bagging constraint score for feature selection with pairwise constraints,'' *Pattern Recogn.*, vol. 43, no. 6, pp. 2106–2118, June 2010.

[273] Mariam Kallakech, Philippe Biela, Ludovic Macaire, and Denis Hamad, ''Constraint scores for semi-supervised feature selection : A comparative study,'' *Pattern Recognition Letters*, vol. 32, no. 5, pp. 656–665, Apr. 2011.

[274] Zenglin Xu, Irwin King, Michael R. Lyu, and Rong Jin, ''Discriminative semi-supervised feature selection via manifold regularization.,'' *IEEE Transactions on Neural Networks*, vol. 21, no. 7, pp. 1033–1047, 2010.

[275] Mohammed Hindawi and Khalid Benabdeslem, ''Local-to-global semi-supervised feature selection.,'' in *CIKM*, Qi He, Arun Iyengar, Wolfgang Nejdl, Jian Pei, and Rajeev Rastogi, Eds. 2013, pp. 2159–2168, ACM.

[276] Ludmila I. Kuncheva, ''A stability index for feature selection,'' in *Proceedings of the 25th Conference on Proceedings of the 25th IASTED International Multi-Conference : Artificial Intelligence and Applications*, Anaheim, CA, USA, 2007, AIAP'07, pp. 390–395, ACTA Press.

[277] Adrian E. Raftery, Nema Dean, and Nema Dean Is Graduate, ''Variable selection for model-based clustering,'' *Journal of the American Statistical Association*, vol. 101, pp. 168–178, 2006.

[278] Cathy Maugis, Gilles Celeux, and Marie-Laure Martin-Magniette, ''Variable selection for clustering with gaussian mixture models,'' *Biometrics*, vol. 65, no. 3, pp. 701–709, 2009.

[279] Thomas Brendan Murphy, Nema Dean, and Adrian E. Raftery, ''Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications,'' *Ann. Appl. Stat.*, vol. 4, no. 1, pp. 396–421, 03 2010.

[280] G. J. Mclachlan, ''Estimating the linear discriminant function from initial samples containing a small number of unclassified observations,'' *Journal of the American Statistical Association*, vol. 72, no. 358, pp. 403–406, 1977.

[281] J. G. Fryer and C. A. Robertson, ''A comparison of some methods for estimating mixed normal distributions,'' *Biometrika*, vol. 59, pp. 639–648, 1972.

[282] João Gama and Pavel Brazdil, ''Cascade generalization,'' *Mach. Learn.*, vol. 41, no. 3, pp. 315–343, Dec. 2000.

[283] Dan Klein, Sepandar D. Kamvar, and Christopher D. Manning, ''From instance-level constraints to space-level constraints : Making the most of prior knowledge in data clustering,'' Technical Report 2002-10, Stanford InfoLab, February 2002.

[284] Aharon Bar-Hillel, Tomer Hertz, Noam Shental, and Daphna Weinshall, ''Learning distance functions using equivalence relations,'' *ICML*, vol. 3, pp. 11–18, 2003.

[285] Mikhail Bilenko and Raymond J. Mooney, ''Adaptive duplicate detection using learnable string similarity measures,'' in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2003, KDD '03, pp. 39–48, ACM.

[286] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell, ''Distance metric learning, with application to clustering with side-information,'' in *Advances in neural information processing systems*. 2003, pp. 505–512, MIT Press.

[287] Yukihiro Hamasuna, Yasunori Endo, and Sadaaki Miyamoto, ''Semi-supervised agglomerative hierarchical clustering using clusterwise tolerance based pairwise constraints.,'' in *MDAI*, Vicenrra, Yasuo Narukawa, and Marc Daumas, Eds. 2010, vol. 6408 of *Lecture Notes in Computer Science*, pp. 152–162, Springer.

[288] Kristin Bennett and Ayhan Demiriz, ''Semi-supervised support vector machines,'' *Advances in Neural Information processing systems*, pp. 368–374, 1999.

[289] Janne Sinkkonen and Samuel Kaski, ''Clustering by similarity in an auxiliary space,'' in *Intelligent Data Engineering and Automated Learning - IDEAL 2000, Data Mining, Financial Engineering, and Intelligent Agents, Second International Conference, Shatin, N.T. Hong Kong, China, December 13-15, 2000, Proceedings*, 2000, pp. 3–8.

[290] Sugato Basu, Arindam Banerjee, and Raymond J. Mooney, ''Semi-supervised clustering by seeding,'' in *Proceedings of the Nineteenth International Conference on Machine Learning*, San Francisco, CA, USA, 2002, ICML '02, pp. 27–34, Morgan Kaufmann Publishers Inc.

[291] Soumia Benyekhlef, El Batoul Bendimered, and Nesma Settouti, ''An improved rotate forest classifier for the recognition of parkinson's disease,'' in *Knowledge Discovery and Data Analysis KDDA'2015, ESI, Algiers, Algeria.*, November 15- 18, 2015.

[292] Yoav Freund and Robert Schapire, ''A short introduction to boosting,'' *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, pp. 1612, 1999.