MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITÉ ABOU BEKR BELKAID
FACULTÉ DE TECHNOLOGIE
DÉPARTEMENT DE GÉNIE BIOMÉDICAL

MÉMOIRE DE FIN D'ÉTUDES

pour obtenir le grade de

MASTER EN GÉNIE BIOMÉDICAL

Spécialité : **Informatique Biomédicale**

présenté et soutenu publiquement
par

Lakhdari Salsabil & Saidi Amaria

le 11 septembre 2017

Titre:

Étude des techniques d'apprentissage semi- supervisé par regroupement

Jury

Président du jury.	Pr. CHIKH Mohamed Amine,	UABB Tlemcen
Examinateur.	Dr. EL HABIB DAHO Mostafa,	MCB UABB Tlemcen
Encadreur.	Dr. SETTOUTI Nesma,	MCB UABB Tlemcen

Je dédie ce travail :

A mes chers parents qui m'ont toujours épaulé et soutenu, sans eux je n'y serai jamais arrivé, Tous les mots du monde ne sauraient exprimer l'immense amour que je vous porte, ni la profonde gratitude que je vous témoigne pour tous les efforts et les sacrifices que vous navez jamais cessé de consentir pour mon instruction et mon bien-être.

A mes frères et sœurs avec qui j'ai grandi et partagé tant de moment, qui font partie de ce travail et a c'est à eux que je le dédie.

A ma grand-mère qui prie toujours pour moi et à laquelle je tiens tant.

A tout les membres de ma grande famille..

A ma chère binôme pour tout ce qu'elle a fait pour la réussite de cette thèse.

A tous mes amis.

LAKHDARI SALSABIL

Du plus profond de mon cœur, je dédie ce travail à tous ceux qui me sont chers :

A la mémoire de ma mère, aucune dédicace ne saurait exprimer mon respect, mon amour éternel et ma considération pour les sacrifices que vous avez consenti pour mon instruction et mon bien être. Je vous remercie pour tout le soutien et l'amour que vous me portez depuis mon enfance et j'espère que votre bénédiction m'accompagne toujours.

A mon cher père, aucune dédicace ne saurait exprimer mes sentiments, que dieu te préserve et te procure santé et langue vie.

A mes chères sœurs Fatiha, Rachida, Amel, Amina et mon frère Djeloul, vous m'avez toujours soutenu durant toutes mes études, je vous souhaite une vie pleine de joie, de bonheur et de réussite.

A ma chère binôme, pour son entente et sa sympathie.

A tous mes cher(e)s ami(e)s, pour leurs aides et supports.

A toute ma chère famille,

A mes professeurs,

A tous ceux que j'aime,

A tous ceux qui m'aiment.

SAIDI AMARIA

Remerciements

Nous tenons à remercier d'abord le tout puissant ALLAH, d'avoir guidé nos pas vers le chemin du savoir, nos familles et nos amis pour leur soutien et leur encouragement tout au long de la réalisation de ce mémoire.

En guise de reconnaissance, nous tenons à témoigner nos sincères remerciements à toutes les personnes qui ont contribué de près ou de loin au bon déroulement de notre thèse de fin d'étude et à l'élaboration de ce modeste travail.

Nos sincères gratitudes à Mme Settouti Nesma pour la qualité de son enseignement, ses conseils et son intérêt incontestable qu'elle porte à tous les étudiants.

Nous remercions, Monsieur le Doyen de la Faculté de Technologie à l'Université de Tlemcen, Prof. Chikh Mohammed Amine, de nous faire l'honneur d'accepter avec une très grande amabilité de siéger parmi notre jury de ce projet de fin d'études. Veuillez accepter ce travail maître, en gage de notre grand respect et notre profonde reconnaissance.

Nos remerciements vont aussi à Dr. El Habib Daho Mostafa (Maître de conférences à l'Université de Tlemcen) de nous honorés d'accepter avec grande sympathie de siéger parmi notre jury de ce projet de fin d'études. Veuillez trouvez ici l'expression de notre grand respect et nos vifs remerciements.

Dans l'impossibilité de citer tous les noms, nos sincères remerciements vont à tous ceux et celles, qui de près ou de loin, ont permis par leurs conseils et leurs compétences la réalisation de ce mémoire.

Enfin, nous n'oserons oublier de remercier tout le corps professoral du département de génie biomédical, particulièrement l'équipe de spécialité informatique biomédicale, pour le travail énorme qu'il effectue pour nous créer les conditions les plus favorables pour le déroulement de nos études.

Résumé

Dans un monde guidé par les données, l'apprentissage automatique est un outil essentiel pour aider les utilisateurs à appréhender la structure de ces données. Dans ce domaine il existe de nombreuses techniques d'apprentissage, l'apprentissage semi-supervisé reste le plus utilisé dans le cadre applicatif et réel, et cela en raison de son principe qui trouve ses racines dans les problèmes d'apprentissage en présence d'un petit nombre de données labellisées. Dans ce projet de fin d'étude nous nous intéressons à la catégorie d'approche d'apprentissage semi-supervisé par contraintes. Pour ce faire, nous réalisons une étude comparative de trois techniques d'apprentissage semi- supervisé par regroupement qui sont : cop-kmeans, Semi-supervised kernel Mean Shift clustering et le regroupement semi supervisé avec contraintes de distances relatives. Nous discutons et analysons en outre l'influence des contraintes par paires (must-link et cannot-link) sur les performances de clustering en effectuant des expérimentations avec différents pourcentages d'exemples marqués. Nous menons une étude sur 6 ensembles de données médicales. Les résultats expérimentaux indiquent que la méthode Semi-supervised kernel Mean Shift clustering peut généralement surpasser d'autres méthodes semi-supervisées. L'étude expérimentale montre que l'utilisation des contraintes peut améliorer les performances en particulier lorsque le nombre d'exemples étiquetés disponibles est insuffisant pour former un modèle de clustering. Des travaux futurs pourront concerner des systèmes d'aide au diagnostique ou segmentation ciblée et une annotation automatique de structures dans les images biomédicales

Mots clés

Apprentissage automatique. Apprentissage semi-supervisé. Apprentissage semi-supervisé par regroupement. Cop-kmeans. SKMS Semi-supervised kernel Mean-Shift clustering . SKLR le regroupement semi supervisé avec contraintes de distances relatives.Contraintes. Must-link. Cannot-link

Abstract

In a data-driven world, automatic learning was an essential tool to help users understood the structure of this data. In this field there were many techniques of semi-supervised learning that it finds its roots in problems in presence of missing data, in this pursuit that processes the category of learning semi-supervised clustering ,we have proposed a comparative study of three techniques of semi-supervised-learning by cop-kmeans Semi-supervised kernel MeanShift clustering and semi-supervisedclustering with relative distance comparisons We have discuss and analyzed, the influence of paired-constraints (must-link and can not-link) on clustering performance by performing experiments with different percentages of labeled examples.We conducted experiments on six sets of medical data. Experimental results indicate that the Semi-supervised kernel MeanShift clustering method may generally outperform other semi-supervised methods. The discovery shows that the use of constraints can improve performance especially when the number of labeled examples available is insufficient to form a clustering model.Future work may involve targeted diagnostic or segmentation systems and automatic annotation of structures in biomedical images.

Keywords

Automated learning. Semi-supervised learning. SSC semi-supervised clustering. Cop-kmeans. SKMS Semi-supervised kernel MeanShift clustering. SKLR semi-supervised clustering with relative distance. constraints. Must-link .Cannot-link

Table des matières

Table des figures

Liste des tableaux

Glossaire

- ANLS : Alternating Non negativity Least Squares
- ARI : Adjusted Rand Index
- CL : Cannot-Link
- COP-KMEANS : constraints pairwise K-means
- EM : Expectation Maximisation
- KDE : Kernel Density Estimate
- ML : Must-Link
- NMF : Nonnegative Matrix Factorization
- ORM : Overall Risks Minimization
- PDM : Pairwise Distance Matrix
- PID : Pima Indian Diabetes
- RI : Rand Index
- S3VM : Semi-supervised Support Vector Machines
- SKLR : Semi-supervised kernel clustering with relative distance comparisons
- SKMS : Semi-supervised kernel MeanShift clustering
- SSC : Semi-Supervised Clusterin
- SVM :Support Vector Machines
- T-SVM : Transductive Support Vector Machines
- UCI : Machine Learning Database

Introduction

L'apprentissage automatique fait référence à la capacité d'un système à acquérir et intégrer de façon autonome des connaissances. Cette notion contient toute méthode permettant de construire un modèle de la réalité à partir de données, soit en améliorant un modèle partiel ou moins général, soit en créant complètement le modèle [1].

Les algorithmes d'apprentissage automatique ont été appliqués à divers domaines, notamment le traitement du langage naturel et de la parole, la reconnaissance de l'écriture manuscrite, la vision robotisée, la fouille de données, les moteurs de recherche sur Internet, le diagnostic médical, la bio-informatique, etc...
Les techniques d'apprentissage ont ainsi joué un rôle crucial dans des applications qui vont de la mise au point de médicaments à l'analyse de grands réseaux de télécommunication. L'apprentissage automatique est composé de plusieurs types entre autre nous citons :

- **L'apprentissage supervisé :** Lorsque le système apprend à classer selon un modèle de classement prédéterminé ainsi que des exemples connus.

- **L'apprentissage non supervisé :** C'est quand le système ne dispose que d'exemples et que le nombre de classes et leur nature n'ont pas été prédéterminés. On parle d'apprentissage non supervisé ou clustering. Aucune annotation n'est requise.

- **L'apprentissage semi-supervisé :** Il utilise un ensemble de données étiquetées et non-étiquetées. Il se situe ainsi entre l'apprentissage supervisé qui n'utilise que des données étiquetées et l'apprentissage non-supervisé qui n'utilise que des données non-étiquetées. Il a été démontré que l'utilisation de données non-étiquetées, en combinaison avec des données étiquetées, permet d'améliorer significativement la qualité de l'apprentissage non supervisé. Un autre intérêt provient du fait que l'étiquetage de données nécessite l'intervention d'un expert humain. Lorsque les jeux de données deviennent très grands, cette opération peut s'avérer fastidieuse. Dans ce cas, l'apprentissage semi-supervisé, qui ne nécessite que quelques étiquettes, revêt un intérêt pratique évident.

Cette approche semi-supervisée permet d'améliorer les performances de plusieurs jeux de données. En outre, durant ces dix dernières années, beaucoup de travaux sur l'intégration de contraintes dans les méthodes de classification

non supervisée ce sont vu porté un grand intérêt. Elles représentent un nouveau champ d'étude dans l'apprentissage semi-supervisé reconnu sous le nom d'apprentissage semi-supervisé sous contraintes.

Dans ce projet de fin d'études, nous nous intéressons à l'étude de cette approche. Pour ce faire notre mémoire sera structuré en quatre chapitres :

- Dans le premier chapitre, nous présentons brièvement les principaux types d'apprentissage : l'apprentissage supervisé et non-supervisé, et nous indiquons quels sont leurs avantages et inconvénients. Par la suite nous présentons les généralités de l'approche semi-supervisée.

- Dans le deuxième chapitre, nous nous intéressons d'abord aux techniques d'apprentissage semi- supervisé par regroupement, ce chapitre regroupera un état de l'art des différents types de contraintes et méthodes utilisés dans cette approche.

- Le troisième chapitre sera consacré à la présentation des principes des trois techniques de l'apprentissage semi- supervisé par regroupement choisit pour l'étude comparative. Nous détaillons le fonctionnement des algorithmes :

    1. L'ALGORITHME COP-KMEANS : la méthode K-means modifié qui prend en considération des contraintes de comparaison entre paires d'objets est appelé COP-KMEANS.
    2. L'ALGORITHME SKMS : L'algorithme Kernel Mean Shift Clustering semi-supervisé, qui utilise des contraintes paires.
    3. L'ALGORITHME SKLR : regroupement semi supervisé par des contraintes de distances relatives.

- Le quatrième et le dernier chapitre permettra d'évaluer les performances des méthodes étudiées. Nous discutons les résultats expérimentaux obtenus par l'étude comparative des trois méthodes d'apprentissage sous contraintes sur différentes bases de données médicales sélectionnée du dépôt d'UCI [2].

- Enfin, une conclusion générale et des perspectives de travail viennent clôturer ce mémoire.

Chapitre 1

Généralité sur l'apprentissage semi-supervisé

Introduction

Dans la dernière décennie, le développement époustouflant des technologies de l'information et de communication, produit une panoplie d'information riche et accessible. Du fait que cette quantité d'information prend des valeurs dans l'espace dont les dimensions, sont continuellement en progression considérable, les utilisateurs sont confrontés actuellement à une nouvelle difficulté qu'est le classement et le traitement des données afin d'en extraire des informations pertinentes et ciblées.

Ce défi de classification crée le concept de Machine Learning qu'est une discipline consacrée à l'analyse des données. Le but de cette discipline est d'engendrer automatiquement une connaissance à partir de données brutes et à les répartir en ensembles d'objets appelés individus et en sous-groupes nommés classes. Cette répartition a lieu sur la base de ressemblances et de rapports entre les individus. En effet cette connaissance (ou modèle) est énormément utile pour la prise des décisions.
Comme le modèle est construit à partir des échantillons qui sont souvent notés sous forme de vecteur : $x = (x_1, x_2.....x_p)$

Où p est le nombre de coordonnée aussi appelé nombre d'attributs (variables, dimensions, caractéristiques).

Les données sont réparties généralement dans deux grandes familles :

- **Les données labellisées :** ceux sont des données accompagnées d'un label (étiquette) qui identifie la décision à prendre pour chaque échantillon. Il est souvent très coûteux (en temps et en argent) d'avoir de grands volumes de données labellisés.

- **Les données non-labellisées :** Les données ne sont pas accompagnées de labels. Même si les données non labellisées sont plus difficiles à exploiter, elles sont beaucoup plus accessibles. (Un exemple intéressant est la récupé-

ration de millions d'images sur le web pour faire de la reconnaissance de visages) [3]. Il est évident que plus on dispose de données, plus le modèle construit est précis et permettra ainsi de prendre de bonne décisions.

Dans ce chapitre, nous présentons les différentes types d'apprentissage selon la fonction de la nature des échantillons (labellisés ou pas), nous distinguons trois grande famille : l'apprentissage supervisé, non-supervisé, et celui sur lequel nous portons tout notre intérêt dans cette étude l'apprentissage semi- supervisé.

1 Les types d'apprentissage

1.1 L'apprentissage supervisé

Par définition si les échantillons d'apprentissage sont labellisés ou étiquetés, nous sommes dans un contexte supervisé. Son principe est de construire un modèle de classification capable non seulement de décrire la classe des individus classifiés a priori, mais aussi de prédire la classe de nouveaux individus non classifiés a priori.

L'apprentissage supervisé consiste à établir des règles de d'apprentissage à partir d'une base de données contenant des exemples de cas déjà étiquetés. Plus précisément, cette base de données est un ensemble de couples entrées- sorties $(X_i, Y_i)\{1 \leq i \leq n\}$ aléatoires. L'objectif est alors d'apprendre à classifier/prédire, pour toute nouvelle entrée X, la sortie Y. On parle de régression dans le cas où les sorties sont à valeurs continues [4] et de classification dans le cas où elles sont à valeurs discrètes [5].

Certes l'approche supervisée est très utilisée pour ces raison et ces avantages, néanmoins il reste qu'il y a un manque de stratégies pour les exemples d'auto-apprentissage (c'est-à-dire d'apprendre à partir d'une base sans aucune connaissance préalable) que les méthodes supervisée ne peuvent pas traiter , dans ce cadre vient la deuxième approche des méthodes de classification, qui est : l'approche non-supervisée (ou spécifiquement « la classification automatique ») .

D'autre part la procédure de labellisation réalisée par un expert humain peut s'avérer lassante 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. On parle donc de l'apprentissage non supervisé.

1.2 L'apprentissage non supervisé (clustering, segmentation)

Dans ce type d'apprentissage les données (échantillons) sont non labellisées, le regroupement des individus dans les classes se produisent sans aucune connaissance préalable de ces groupes ou échantillon qui le composent. Le clustering (ou l'analyse en clusters) vise à organiser une collection d'éléments de données en clusters, de sorte que les éléments d'un cluster sont plus «similaires» les uns aux

autres que les éléments des autres clusters.

Cette notion de similarité peut être exprimée de manière très différente, selon le but de l'étude (des hypothèses propres au domaine et sur la connaissance préalable du problème). Il existe deux approches différentes dans le clustering [6] :

- Le clustering partionnel (non-hiérarchique),
- Le clustering hiérarchique.

1.2.1 Le clustering partionnel

vise à obtenir directement une seule partition de la collection d'éléments en clusters (la division d'un ensemble de données en K groupes) jusqu'à obtenir une similarité satisfaisante, qui possèdent les propriétés suivantes :

- Homogénéité dans les groupes (intra-classe) : les données appartenant à un même cluster doivent être les plus similaires possibles.
- Hétérogénéité entre groupe (inter-classe) : les données appartenant à différents clusters doivent être les plus dissemblables possibles.

L'inertie inter-cluster mesure "l'éloignement" des centres des clusters entre eux. Plus cette inertie est grande, plus les clusters sont bien séparés. Il faut minimiser l'inertie intra-cluster et maximiser l'inertie inter-cluster. Bien qu'elles permettent de classifier un ensemble volumineux de données, mais il reste qu'il faut fixer au départ le nombre de classes : choix du nombre de k clusters.

1.2.2 Le clustering hiérarchique

vise à obtenir une hiérarchie de clusters, appelée dendrogramme, qui montre comment les clusters sont liées les uns aux autres. Ces méthodes procèdent soit par :

- **Les méthodes ascendante :** (Agglomérations successives) Un algorithme de clustering hiérarchique est fondé sur l'union entre les deux plus proches clusters c'est-à-dire : consiste à trouver des clusters successifs utilisant des clusters précédemment établis. La première condition est de mettre, au début, chaque objet dans un cluster distinct et les fusionner en clusters successivement plus grand.

- **Les méthodes descendante :** Divisions successives en deux groupes différents. Au départ tous les individus sont dans le même groupe. A chaque étape, un groupe est séparé en deux. Il faut un critère de séparation. Après quelques itérations on atteint le Cluster final voulu qui regroupe tous les sous-clusters (sous-partitions).

Une partition des éléments de données peut être obtenue en coupant le dendrogramme à un niveau souhaité. L'inconvénient majeur de la plupart des fonctions

de clustering, c'est qu'elles sont coûteuses en temps de calcul et sont de plus sensibles à la dimension des données.

Les techniques de clustering actuelles ne traitent pas tous les besoins de façon adéquate (et simultanément), comme le fait que si nous n'avons pas des variables continuées (la longueur), mais les catégories nominales, comme les jours de la semaine. Dans ces cas encore, la connaissance du domaine doit être faite pour formuler le clustering appropriée.

Beaucoup d'algorithmes de clustering exigent la spécification du nombre de clusters à produire en entrée de l'ensemble de données, avant l'exécution de l'algorithme. i.e. : connaissance de la valeur correcte à l'avance, la valeur appropriée doit être déterminée, un problème pour lequel un certain nombre de techniques ont été développées.

1.3 L'apprentissage semi-supervisé

Il existe d'autres types de classification qui s'appuient sur d'autres types de méthodes d'apprentissages comme « l'apprentissage semi-supervisé » En effet, l'apprentissage semi-supervisé est un bon compromis entre les deux types d'apprentissage « supervisé » et « non-supervisé », car il permet de traiter un grand nombre de données sans avoir besoin de toutes les étiqueter, et il profite des avantages des deux types mentionnés.

D'autre part la labellisation a priori de toutes les données nécessite l'intervention d'un expert humain.C'est 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 le contexte non supervisé, est alors une solution alternative [7]. Il se caractérise par la présence de quelques informations disponibles sur l'ensemble des données. Ces informations sont représentées soit sous la forme de quelques données labellisées, soit sous la forme de ressemblance ou dissemblance au sein de couples de données.

Le contexte semi-supervisé utilise des connaissances partielles qui sont soit incomplètes (par exemple, le cas où des relations entre certains individus sont connues) ou tout simplement les exemples étiquetés ne sont pas en quantité suffisante pour que l'on puisse appliquer des algorithmes supervisés.

Parmi les nombreuses méthodes d'apprentissage récemment apparues, de nouvelles méthodes de classification connaissent des succès importants sont les méthodes semi-supervisées, où l'on dispose à la fois d'un (petit) ensemble de données étiquetées, et d'un (grand) ensemble de données non-étiquetées [8]. Les

approches d'apprentissage semi-supervisé peuvent être globalement divisés en deux catégories : l'apprentissage semi- supervisé classique et celui par Regroupement.

1.3.1 L'apprentissage semi-supervisé classique

L'ensemble d'apprentissage est constitué de données libellées et non libellées mais la tâche d'apprentissage qui est essentiellement supervisée. La classification est effectuée premièrement par un apprentissage avec des données libellées, en second lieu, un renforcement de l'apprentissage est réalisé avec des données non-libellées. L'objectif est le même que pour l'apprentissage supervisé mais en tirant profit des observations non labellisée.

Dans le contexte d'apprentissage semi-supervisé avec labels, le nombre de données labellisées est trop faible pour apporter suffisamment d'informations nécessaires à la sélection supervisée d'attributs. Un algorithme de sélection non supervisée peut être alors envisagé mais il ignore l'information fournie par les labels disponibles. Il est donc préférable que la pertinence des attributs soit évaluée en tenant compte à la fois des données labellisées et non labellisées.C'est pour cette on s'intéresse à l'apprentissage semi-supervisé par regroupement. Dans ce contexte il existe deux types d'apprentissage le Transductif et l'inductif :

- Transductif : Fournit le label uniquement pour les données disponibles non labellisées.
- Inductif : Produit non seulement des labels pour données non labellisées, mais aussi produit un classifieur.

L'Apprentissage transductif : L'objectif de cet apprentissage est de faire des prédictions sur les observations de la base de test et de minimiser leur erreur moyenne à partir d'un apprentissage sur la base de données d'apprentissage [9].

L'Apprentissage inductif : Il existe plusieurs technique d'apprentissage semi-supervisé développées de façon générale pour tous les types de classifieurs mais qui sont adaptées au mieux à une classe spécifique de classifieurs. Parmi ces méthodes : L'Auto-apprentissage (Self-Training), le Co-apprentissage, les S3VM : séparateur semi-supervisé à vaste marge et T-SVM : Transductive Support Vector Machines.

1. **Auto-apprentissage (Self-Training) :** C'est une technique très répandue pour faire de l'apprentissage semi-supervisé. Le classifieur est d'abord entraîné par les données étiquetées et on utilise le résultat pour classer les données non étiquetées. Les données non étiquetées qui sont classées avec un haut degré de confiance sont ajoutées aux données d'apprentissage. Le classifieur est ré-entrainé sur les données d'apprentissage et la procédure est répétée jusqu'à satisfaire un critère d'arrêt.

Nous notons que cette technique utilise ses propres prédictions pour s'améliorer à chaque itération. On parle alors de "self-teaching" [10]. Ce processus d'auto-étiquetage peut constituer une faiblesse pour cette technique, surtout dans le cas où la frontière de décision temporaire utilisée est très loin de la frontière réelle. [11]

2. **Co-apprentissage (co-training) :** L'idée du co-apprentissage est que l'espace de caractéristiques peut être divisé en deux sous-espaces procurant chacun un bon cadre d'apprentissage. Ainsi, initialement deux classifieurs sont entraînés avec les données étiquetées sur deux sous-espaces différents. Puis chaque classifieur obtenu pour chaque sous-espace, est utilisé pour déterminer la classe probable des données non étiquetées qui seront utilisées pour ré-entraîner l'autre classifieur. [11]
   Pour utiliser la méthode co-training, il faut avoir deux différentes vues des données à classer, et ces deux différentes vues doivent être compatibles et indépendantes, et Chaque vue désigne la façon utilisée pour extraire les caractéristiques, donc chaque vue donne lieu à une caractérisation différente des formes à reconnaître. La compatibilité permet d'avoir la même étiquette pour un exemple donné selon chaque vue considérée indépendamment. En ce qui concerne l'indépendance, on veut pour un exemple donné, qu'il n'y ait aucune corrélation entre les caractéristiques issues des deux différentes vues.

3. **S3VM : séparateur semi-supervisé à vaste marge :** Nous pouvons citer la transduction proposée par Vapnik (Transductive SVM), qui est une technique spécifique pour entraîner de façon semi-supervisée les machines à vecteurs de support. S3VM sont construits à l'aide d'un mélange de données marquées (l'ensemble de formation) et de données non marquées (le jeu de travail).
   S3VM construit une machine vectorielle de support en utilisant les ensembles de formation et de travail. L'objectif est d'attribuer des étiquettes de classe à l'ensemble de travail de sorte que la "meilleure" machine de vecteur de support (SVM) soit construite.
   Nous utilisons S3VM pour résoudre le problème de transduction en utilisant la minimisation globale des risques (ORM) Posé par Yapnik. Le problème de la transduction est d'estimer la valeur d'une fonction de classification aux points donnés dans l'ensemble de travail.
   Si l'ensemble de travail est vide, la méthode devient la première approche SVM de la classification [12]. Si l'ensemble de formation est vide, la méthode devient une forme d'apprentissage non supervisé. L'apprentissage semi-supervisé se produit lorsque les ensembles de formation et de travail sont non vifs (nonempty).
   Pour formuler le S3VM, nous commençons par la formulation SVM , puis ajoute deux contraintes pour chaque point dans l'ensemble de travail. Une contrainte calcule l'erreur de classification incorrecte comme si le point était dans la classe 1 et l'autre contrainte calcule l'erreur de classification comme si le point était en classe - 1. L'objectif de la fonction calcule le minimum

des deux erreurs possibles de mauvaise classification. La classe finale des points correspond à celle qui aboutit à la plus petite erreur. Plus précisément, nous définissons le problème de machine de vecteur de support semi-supervisé. [13]

4. **T-SVM : Transductive Support Vector Machines :** L'idée du T-SVM est d'induire une fonction globale à l'aide des données annotées et des données de test. La Transductive SVM (TSVM) est une variante intéressante des SVMs utilisant l'approche transductive. La TSVM maximise la marge en utilisant des données étiquetées et non étiquetées en plus de la minimisation de l'erreur sur tout l'ensemble des données [14].
   D'abord, les paramètres du modèle représentant la fonction sont estimés pour tout l'espace d'entrée considéré en utilisant des exemples d'apprentissage. Puis, pour chaque exemple de test donné, on calcule la valeur de la fonction avec les paramètres estimés préalablement. Cependant, il est possible de trouver la valeur de la fonction en un point de test en une étape. Et puisque au cours de l'apprentissage, le but final est d'estimer les valeurs de la fonction en des points spécifiques et non de façon générale pour tout l'espace, cette approche s'est révélée plus exacte que l'inférence inductive et moins difficile. La formulation de la TSVM va au-delà de l'idée première de Vapnik qui est la transduction [15].

1.3.2 L'apprentissage semi- supervisé par regroupement

L'apprentissage semi- supervisée par regroupement ou *Semi-Supervised Clustering (SSC)* en anglais, est une extension de la classification non-supervisée. Le Clustering semi-supervisé introduire des connaissances expertes partielles dans un algorithme du clustering relève du domaine du clustering semi-supervisé.

Contrairement à l'apprentissage semi-supervisé, où l'accent est mis sur le traitement des données manquantes ou insuffisantes dans les algorithmes supervisés, le clustering semi-supervisé est utilisé lorsque que la quantité de supervision est tellement faible ou partielle qu'il est impossible d'appliquer des techniques supervisées.

Les connaissances a priori se présentent soit sous la forme d'étiquettes de classe, soit sous la forme de contraintes sur des paires de données si elles sont similaires et doivent alors être regroupées ensemble, ce sont des contraintes "must-link", ou si elles sont dissimilaires et donc ne doivent pas être regroupées ensemble, ce sont des contraintes "cannot-link". Elles sont ensuite utilisées pour guider le processus de clustering dans l'espace des solutions. Nous utilisons des techniques issues du clustering semi-supervisé pour modéliser et utiliser à la fois les connaissances additionnelles attachées aux données complexes et la dimension temporelle des données.
Selon [16] deux sources d'information sont habituellement disponibles pour une méthode de clustering semi-supervisé : la mesure de similarité (must-link ou cannot-link) et la recherche de clusters appropriés.

Les méthodes d'adaptation de similarité : Un algorithme de clustering existant utilise une certaine mesure de similarité, des points proches représentent des données d'un même groupe et que des points lointains représentent des données qui appartiennent à des groupes différents [17]. La mesure de similarité est adaptée pour que les contraintes disponibles puissent être plus facilement satisfaites. Plusieurs mesures de similarité ont été utilisées pour l'adaptation semi-supervisée de la similarité :
La divergence de Jensen-Shannon entraînée avec la descente en gradient.
La distance euclidienne modifiée par un chemin le plus court algorithme.
La distance Mahalanobis ajustées par l'optimisation convexe.

Les mesures de similarité les plus utilisées sont les mesures de distance :

- **La distance Euclidienne :** La distance Euclidienne, qui est la distance la plus utilisée, est définie comme suit :

$$d_2(x_i, x_j) = (\sum_{k=1}^{d}(x_{i,k}, x_{j,k})^2)^{1/2} = ||(x_i, x_j)||^2$$

  avec $x_i = (x_{i,1}, x_{i,2}.....x_{i,d})$ et $x_j = (x_{j,1}, x_{j,2}.....x_{j,d})$ deux vecteurs de dimension d

- **La distance de Mahalanobis :** La distance de Mahalanobis est une mesure de distance introduite par Prasanta Chandra Mahalanobis en 1936. Elle est basée sur la corrélation entre des variables par lesquelles différents modèles peuvent être identifiés et analysés. C'est une manière utile de déterminer la similarité entre une série de données connues et inconnue.

Les méthodes de recherche : L'algorithme de clustering lui-même est modifié afin que les contraintes ou les libellés fournis par l'utilisateur puissent être utilisés pour polariser la recherche d'un cluster approprié. Cela peut se faire de plusieurs façons, par exemple en effectuant une fermeture transitive des contraintes et en les utilisant pour initialiser les clusters, en incluant dans la fonction de coût une pénalité pour défaut de respect des contraintes spécifiées, ou par des contraintes à satisfaire lors de l'affectation des grappes dans le processus de regroupement.

2 Conclusion

Dans ce chapitre nous exposons les notions fondamentales de l'apprentissage automatique tout en passant en revue les différents types d'apprentissage. Nous avons mis en avant les situations où les algorithmes de classification classiques peuvent présenter des limites et où l'apprentissage semi-supervisé peut être une solution intéressante et peu coûteuse. Nous décrivons par la suite de manière générale la théorie et les fondements des méthodes de l'apprentissage semi- su-

pervisé avec ces deux familles.

Dans le chapitre suivant nous nous intéresserons à l'apprentissage semi-supervisé par regroupement et plus précisément les méthodes de similarité. Le choix de l'étude de cette approche pour ce projet de fin d'études a été motivé par sa simplicité d'application en plus d'être populaire et très performante.

Chapitre 2

L'apprentissage semi- supervisé par regroupement

Introduction

Dans le contexte d'apprentissage semi- supervisé, les techniques sont regroupées prenant en compte l'entièreté d'un échantillon partiellement étiqueté. Nous notons donc l'échantillon d'apprentissage S composé d'un échantillon superviséS_{sup} et d'un non supervisé S_{nsup} :

$$S = S_{sup} U S_{nsup}.$$

Lors de ces dernières années, le thème de l'apprentissage par regroupement a été abordé par divers chercheurs ; la plupart de ces travaux s'intéressentà l'intégration de contraintes dans ces méthodes. Ces contraintes peuvent être générées [18] à partir des connaissances préalables sur les données, ou à partir d'un sous-ensemble de données étiquetées [19].

La prise en compte de ces connaissances dans un processus de classification, si elles existent, représente un nouveau champ d'étude dans l'apprentissage automatique qui est la classification sous contraintes [20]. Cette approche semi-supervisée permet d'améliorer les performances de plusieurs jeux de données. De plus, l'intérêt portésur l'incorporation de connaissances a priori dans des processus de classification, a pris de l'ampleur dans un nombre important d'applications issus du monde réel telles que l'identification de personnes via des caméras de surveillance [21], le raffinement des cartes GPS [22] et la détection de paysage dans les données hyper-spectrales [23] [24].

1 L'apprentissage par contraintes

Le regroupement sous contraintes est une tâche importante dans le processus de fouille de données. Il permet de modéliser plus finement une tâche de regroupement en intégrant des contraintes d'utilisateurs. Voir figure 2.1.

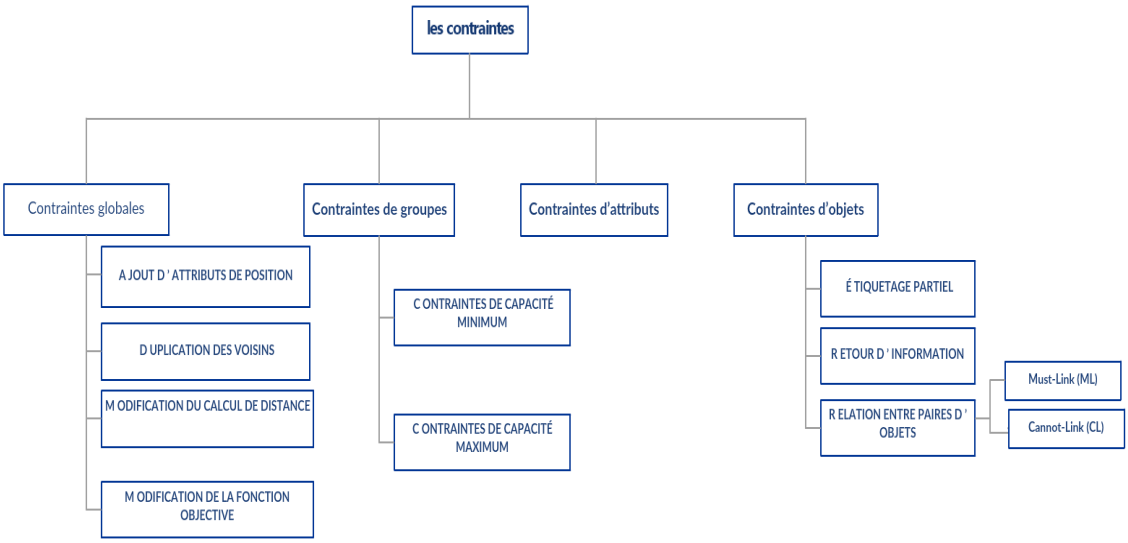

FIGURE 2.1 – Les différents types de contraintes

Plusieurs types de contraintes peuvent être considérés ; elles peuvent porter sur les clusters, comme par exemple sur leur diamètre ou sur leur taille, ou porter sur des paires d'objets qui doivent être ou pas dans une même classe. Une contrainte peut aussi dépendre du type d'information qui est manipulée, cette dernière peut prendre plusieurs formes.

1.1 Contraintes globales

Si l'information est de forme structurelle sur les données, dans ce type il existe 4 façons de procéder :

- **Ajout d'attributs de position :** Cette méthode permet d'ajouter une notion de localisation spatiale dans un ensemble de données, ceci est réalisé d'une façon à adjoindre des attributs de position pour chacun des objets. L'avantage de cette méthode réside dans le fait qu'elle ne requiert pas de modification de l'algorithme de classification utilisé.

- **Duplication des voisins :** Cette technique de duplication est basée sur la notion de voisinage en termes de distance d_{ij} dans l'espace en D dimensions (avec $i \neq j$). Elle permet d'augmenter la taille du vecteur attribut d'un objet avec l'ajout d'un ou plusieurs ensembles d'attributs selon le nombre de voisins considéré. Ceci consiste en l'augmentation de la taille du vecteur attribut d'un objet en ajoutant un ou plusieurs ensembles d'attributs selon le nombre de voisins considéré. Cette approche est souvent impraticable à cause de l'augmentation importante de la dimension des données et donc le cout devient non négligeable tant en termes de temps de calcul qu'en termes d'espace mémoire nécessaire à l'exécution des algorithmes de classification utilisés. Pour cause, Il est affecté à chaque objet, un duplicata des caractéristiques de chacun de ses voisins (au sens de la distance calculée) [25].

  - **Modification du calcul de distance :** Au contraire des méthodes précédentes qui consistent à modifier l'ensemble des données (et plus particulièrement, le vecteur attribut de chaque objet par l'ajout de nouveaux attributs), d'une manière moins directe certaines méthodes permettent d'intégrer des contraintes.En conséquence, il est possible d'incorporer des informations spatiales en changeant la manière de calculer la distance entre deux objets. Ceci consiste par le biais selon une fonction non-linéaire (dans la littérature c'est la fonction exponentielle) à modifier la distance originale séparant deux objets par $\mathrm{d_{ij}}^* = \mathrm{d_{ij}}[1 - \exp(\frac{-\mathrm{u_{ij}}}{\mathrm{w}})]$ avec d^*_{ij} représentant la distance modifiée incorporant l'information spatiale (grâce au voisinage de x_i et x_j), d_{ij} étant la distance originale entre les deux objets, u_{ij} s'exprimant comme une mesure de distance égale au ratio des distances moyennes des voisins de chacun des objets x_i et x_j, et enfin W étant un poids (de valeur arbitraire ou fixée par l'utilisateur).
    Afin d'obtenir une unique source d'information, cette méthode permet d'associer les caractéristiques des objets aux connaissances a priori. De ce fait, le but est de trouver des groupes d'objets de compacité équitable et homogènes.

- **Modification de la fonction objective :** Cette dernière catégorie de méthodes des contraintes globales, et plus particulièrement des informations de voisinage, est celle modifiant un critère à optimiser par la fonction objective d'un algorithme quelconque, utilisant une procédure d'optimisation. Afin d'obtenir les contraintes définies, il est nécessaire d'échanger cette fonction par une somme pondérée de la fonction originale avec l'information de voisinage à disposition. De façon générale, l'optimisation de la variance et de la contigüité des groupes d'objets, avec un paramètre de pondération est réalisée par la spécification de l'importance relative des deux objectifs $\mathrm{f_{original}}^* = \mathrm{f_{original}} + \lambda * f_{contraintes}$ ou encore $\mathrm{f_{original}}^* = (1 - \lambda) * \mathrm{f_{original}} + \lambda * f_{contraintes}$;
avec :
$f^*_{original}$: la nouvelle fonction objective (données originales et contraintes).
$f_{original}$: la fonction originale (données originales).
$f_{contraintes}$: la fonction à optimiser pour l'information a priori (contraintes).
λ : $unparamtredepondration.$

1.2 Contraintes de groupes

Dans l'apprentissage automatique, la connaissance disponible peut également être fournie sous la forme d'information sur les groupes d'objets. Il peut être définit des exigences sur la forme globale, l'orientation ou d'autres caractéristiques des groupes. La capacité minimum ou maximum de ces derniers semble être la caractéristique la plus utilisée dans la littérature.

  - **Contraintes de capacité minimum :** Ceux sont des méthodes utilisant des contraintes sur des groupes d'objets ont été développées récemment, pour éviter d'obtenir des solutions contenant des groupes vides.Cette approche

impose des contraintes sur la structure des groupes. En conséquence, il est alors possible de spécifier un nombre minimum d'objets pour chaque groupe.

– **Contraintes de capacité maximum :** L'utilisation de ce genre de contraintes, est appliquée la plus part du temps par les algorithmes de classification non-supervisée du type regroupement hiérarchique. En conséquence, à partir de la hiérarchie créée, on a la possibilité de sélectionner des groupes d'objets adaptés permettant de faire respecter la contrainte définie.

1.3 Contraintes d'attributs :

La connaissance a priori peut être interprétée comme de l'information dépendante des caractéristiques des objets. Ces contraintes permettent d'orienter la classification des objets selon leurs valeurs pour un attribut donné.

1.4 Contraintes d'objets :

En 1984 Bejar et Cortes [26] ont développé un algorithme de classification hiérarchique contraint et incorporant des contraintes d'attributs. Les contraintes d'objets définissent des délimitations sur des paires individuelles d'objets. Ce type de connaissance a priori sur les données est généralement fournie sous trois formes :

1. **Étiquetage partiel :** La labellisation des objets d'un ensemble d'une dimension plus élevée est représenté par une fonction complexe et coûteuse en temps de calcul ce qui la rend souvent impossible. Comme une solution, il est possible d'étiqueter un sous-ensemble ne contenant que quelques objets. L'utilisation de cette nouvelle source d'information peut être faite de deux façons différentes :

   Identification des groupes obtenus à partir de l'application d'un algorithme de classification non-supervisée,et à l'aide de sous-ensemble d'objets étiquetés au préalable, les groupes d'objets obtenus peuvent être identifiés . Pour cela, il est nécessaire d'utiliser des règles simples. Parmi ces dernières, le vote majoritaire semble être une bonne alternative car il permet d'affecter une étiquette à chaque groupe obtenu.

   Initialisation intelligente des algorithmes de partitionnement L'initialisation est une étape essentiel dans certains algorithmes de classification non-supervisée et peut se révéler cruciale dans la validation et l'interprétation des résultats de partitionnement obtenus6. De ce fait,pour l'orientation de choix des paramètres initiaux, il semble intéressant d'utiliser l'information contenue dans ces étiquettes de classes.

2. **Retour d'information :** Une approche itérative a été adoptée par les systèmes de classification interactifs, ce système produit une partition de données puis l'évalue et la valide par un expert, si l'ensemble des données est de dimensions importantes on peut trouver une difficulté lors de la validation des résultats, ainsi cet expert peut faire clairement l'indication des erreurs de partitionnement (induits par le système) après on peut utiliser cette information dans l'itération suivante de l'algorithme.

3. **Relation entre paires d'objets :** Les contraintes produisent des indications sur la partition souhaitée et mettent en œuvre ces indications dans des algorithmes de classication an d'augmenter leur performance [18]. Soit $X = x_1, ..., x_n$le jeu des observations qui doit être regroupées dans K classes, qu'on note par $u_1, ..., u_K$. Pour chaque paire d'observations x_i, x_j dans X,on note la distance entre eux par$d(x_i, x_j)$. Ce type de contraintes atteste simplement que deux objets soit :

   Must-Link (ML) : qui force deux observations x_i et x_ja être dans la même classe.
   Définition (Must-Link) : Pour deux instances de données x_i etx_j dans l'ensemble de données,$x_i, x_j \in X(1 \leq i, j \leq n)$, si$x_ietx_j$ satisfait la contrainte Must-Link, puis après avoir terminé le clustering, $x_i$et x_j satisfait$x_i \in Cm \wedge x_j \in Cm, Cm \in \prod, 1 \leq m \leq k$, sinon le cluster échoue. La contrainte peut être décrite comme x_i ML x_j.

   Cannot-Link (CL) : si deux observations x_i et $x_j$sont dans deux classes différentes on peut définir cesdeux objets par Cannot-Link (CL)
   Definition (Cannot-Link) :Pour deux instances de données xi et $x_j$dans l'ensemble de données,$x_i, x_j \in X(1, j \leq n)$, si x_i et x_j satisfait la contrainte Cannot-Link, après avoir terminé le clustering, x_i etx_j satisfait$x_i \in Cm \wedge x_j \in Cn, Cm, Cn \in \prod, 1 \leq m, n \leq k, m \neq n$, sinon le cluster échoue. La contrainte peut être décrite comme x_iCL x_j.

Une remarque peut être faite, c'est que les contraintes globales ainsi que les contraintes d'attributs peuvent facilement se mettre sous la forme de contraintes de paires d'objets.

2 État de l'art des méthodes d'apprentissage semi-supervisé par regroupement (SSC)

L'apparition des techniques d'apprentissage semi-supervisé par regroupement (SSC) ont vu le jour par des travaux portant sur l'extension des méthodes non supervisées (clustering) en apprentissage semi-supervisé.
L'adaptation a été réalisée sur les deux groupes de méthodes d'apprentissage non supervisées (clustering) : celles par approche basée sur les caractéristiques et

celles à base de graphe.

1. **Approche basée sur les caractéristiques :** où chaque point de données a une représentation en termes de caractéristique d'un vecteur ou une représentation structurée comme une séquence, des séries chronologiques, ou graphiques, tel que les méthodes : k-means et mixture of Gaussian.

2. **Approche à base de graphes :** où une similitude graphique entre des points de données est accordée. Exemple : les méthodes de clustering spectral.

En effet, la littérature démontre qu'un grand nombre de travaux ont portés sur l'application de deux types de semi-supervision :

- Par point (pointwise) : [27] où l'étiquette de cluster d'un petit nombre de points sont disponibles pour guider la mise en cluster, et deux par deux.

- Par paire (pairwise) : où "must-link" et "cannot-link" les contraintes entre certaines paires de points sont disponibles.

Au cours de la dernière décennie [20], les méthodes basée sur les caractéristiques ont été largement étudiées pour une généralisation convenable, les algorithmes tels que k-means et ses variantes ont été une des plus réussite adaptation qui avec précision a su tiré parti de la semi-supervision. Ces méthodes ont été généralisées pour incorporer la métrique d'apprentissage en contexte de SSC et ainsi que l'estimation de paramètres et d'inférence dans un model graphique [19] [28] [29]].

D'un autre côté, il y a plusieurs approches de la SSC fondée sur la représentation graphique et les méthodes de clustering basées sur les graphes. La littérature sur le SSC a également exploitée la piste des graphes centrée principalement sur les points pour l'explorer dans le contexte semi-supervisé. Les méthodes de clustering spectral sont largement les plus utilisées pour la classification non supervisée avec la présentation graphique [28] [20] [30] et en même temps, peuvent être considérées comme la solution d'un espace de problèmes de graph-cut. Dans le cadre semi-supervisé, on peut aborder le problème comme l'une des coupes du semi-supervised-graphique, où les points avec la même étiquette de cluster sont censés être dans la même coupe.

2.1 L'apprentissage semi-supervisé par regroupement basé sur les approches Pointwise et Pairwise

L'apprentissage semi-supervisé par regroupement avec par point (pointwise) et par paire (pairwise) a été largement étudié par l'adaptation des méthodes de regroupement en apprentissage semi-supervisé. Dans cette partie, nous passons en revue certaines des approches de la littérature [20].

2.1.1 Semi-supervisé Clustering basé sur Seeding :

L'une des premières idées sur SSC a porté sur la supervision de l'étiquette pointwise, où les identifiants de cluster d'un petit nombre de points sont disponibles. Le SSC basé sur l'ensemencement (seeding) se concentre sur les algorithmes de clustering basés sur le centroïde, tels que k-means [27] [31].

L'idée principale en SSC basée sur seeding est d'améliorer l'initialisation en fonction des identifiants de cluster disponibles. Dans cette famille se démarque en particulier, deux versions appliquée à k-means [27] : seeded k-mean et constraint k-means.

- Seeded k-mean : Dans seeded-kmeans, le clustering de graine (seed) sert à initialiser l'algorithme des k-means, avec l'initialisation des centroïdes de cluster disponible à l'aide des identifiants et exécute ensuite les mises à jour itératives pour k-means. En seeded k-means, l'identifiant de cluster avec l'ensemble de l'étiquette peut changer au cours de la mise à jour itérative si la fonction d'objective est améliorée en conséquence.

- Constraint k-means : initialise également les centroïdes de cluster avec l'identifiant de cluster disponible, mais les affectations de ces identifiants de clusters de points marqués ne sont pas autorisés à changé au cours de la mise à jour itérative. Constraint k-mean est donc plus approprié quand l'étiquetage initial est libre de bruit.

Les méthodes d'apprentissage semi-supervisée par regroupement SSC basée sur seeding peuvent être généralisée à d'autres méthodes de regroupement au-delà de k-means, comme la méthode de spherical-k-mean. L'approche seeding peut être consultée dans le cadre de clusters probabiliste où on utilise l'algorithme EM pour apprendre les modèles de mélange [32], et chaque composant du mélange correspond à un cluster. En particulier, la semi-supervision est utilisé pour définir la probabilité a posteriori $p(z_i|x_i,)$ des points à 1 pour le vrai cluster, et 0 autrement.

Les résultats dans [27] illustrent les avantages de l'approche seeding. La méthode a également été appliquée à un scénario dans lequel le libellé data ne couvre qu'une fraction de l'amas, et les centroïdes des autres clusters doivent être initialisés au hasard.

2.1.2 Semi-supervisé Clustering basé sur les paires de contraintes (Pairwise constraint) :

Elle est l'une des méthodes les plus populaires utilisée dans cette famille, elle utilise les contraintes par paires. Étant donné un jeu de données X, les paires de contraintes sont généralement données sous la forme de deux types de contraintes : must-link (ML) et cannot-link(CL).

– Si $(x_i, x_j) \in M$, l'ensemble de contraintes must -link, puis l'algorithme de clustering est encouragé à garder i (x_i, x_j) dans le même groupe ;

– Si $(x_i, x_j) \in C$, l'ensemble de contraintes cannot-link,l'algorithme de clustering est encouragé à garder garder (x_i, x_j) dans clusters différentes.

L'une des premières approches de la SSC avec contraintes par paires a portée sur la modification de l'algorithme k-means [22] d'intégrer must-link et cannot-link des contraintes. À chaque itération de l'algorithme, un cluster correspondant est considérable pour chaque point classé. L'algorithme converge jusqu'à ce qu'une affectation trouve le cluster qui ne viole aucune des contraintes. En l'absence d'un tel groupe, l'algorithme se termine.

2.1.3 Le regroupement semi-supervisé par l'apprentissage Actif :

L'apprentissage semi-supervisé par regroupement (SSC) travaille avec quelques pairwise (ou pointwise) les étiquettes, l'apprentissage actif est devenu un cadre naturel à envisager dans ce contexte. Étant donné un nombre fixe de requêtes autorisées sur les étiquettes par paire, l'objectif est de déterminer laquelle des relations pairwise pour effectuer une requête afin d'obtenir les informations must-link et cannot-link. Dans [28], deux étapes qui décrit l'apprentissage avec des étiquettes par paires :

1- Exploration : Dans la première étape, appelée exploration, l'accent est mis sur l'obtention d'au moins un point de chaque groupe avec un petit nombre de requêtes. Le premier point est choisi au hasard et affectés à un cluster. Tous les points subséquents sont choisis avec le plus de traversée d'abord, c'est-à-dire, en choisissant le point qui est le plus éloigné de tous les points existants, (le point le plus éloigné d'un ensemble est mesuré par la distance jusqu'au point le plus proche de l'ensemble). Une fois qu'un point est sélectionné, deux par deux, les requêtes sont faites avec un point quelconque de chacune des clusters existants.

– Si une contrainte must-link est trouvée avec l'un des clusters existants, le point est attribué à ce cluster, et la méthode choisit le prochain point le plus éloigné.

– Si une contrainte cannot-link est trouvée avec l'un des clusters existants, donc un nouveau cluster est initialisé avec ce point comme étant membre. L'exploration du processus continue jusqu'à au moins un point de chaque cluster est trouvé ou le budget de requêtes est épuisé.

2- Consolidation : Dans la deuxième étape,appelé consolidation des données supplémentaires, les points sont choisis au hasard et affectés au bon cluster en interrogeant par paires.

Compte tenu d'un point de données, tous les clusters sont d'abord triés en ordre croissant de distances aux centres de concentration (centroids) de cluster correspondants. Les requêtes en paire sont effectuées avec n'importe quel point de chacun des clusters en ordre trié jusqu'à ce qu'un point/cluster avec une

contrainte de liaison obligatoire soit trouvé. Dans ce cas, le nouveau point est attribué à ce cluster, et le processus se poursuit en choisissant un autre point au hasard.

L'interrogation devrait être efficace, c.-à-d., pour un k -clustering, ou moins il y a un nombre de requêtes k qui sera habituellement nécessaire. Empiriquement, dans [28], la stratégie d'apprentissage actif est montré pour être efficace dans la pratique, en vue d'améliorer la performance de l'ensemble des tests avec moins de requêtes avec la comparaison des requêtes qui choisir au hasard.

2.1.4 Semi-supervisé Clustering basé sur l'utilisation (User Feedback)

Un ensemble d'idées pour la coopération SSC a été pour suivi par [33], où la semi- supervision est faite avec l'utilisateur dans la boucle. Le développement reconnaît le fait qu'il peut y avoir plusieurs façons semi-supervision peut être fourni et préconise la mise à jour de la mise en cluster en se fondant sur les commentaires reçus de l'utilisateur. En particulier, [33] envisage un scénario dans lequel un utilisateur est itérativement les commentaires sur la qualité des clusters. En général, une mesure de divergence est définie à l'avance et utilisées pour produire un clustering. Sur la base des commentaires de l'utilisateur sur le premier cluster, la méthode tente d'ajuster ce que signifie être similaires. En d'autres termes la mesure de divergence dans ce cas n'est pas fixe, mais plutôt ce que l'utilisateur fournit des commentaires.

2.1.5 Semi-supervisé Clustering basé sur la factorisation de la matrice non négative (NMF)

Il y a eu un développement important dans l'utilisation des méthodes la factorisation de la matrice non négative (NMF) pour diverses tâches non supervisées. Bien que NMF peut être utilisés pour l'analyse de données classiques, le récent intérêt considérable dans NMF est due à la capacité nouvellement découverte du NMF à l'exploration de données et de résoudre des problèmes d'apprentissage automatique.

En particulier, NMF par sa fonction de coût (somme des erreurs au carré)est équivalente à une approche K-means clustering, qui est la plus utilisée de l'algorithme d'apprentissage non supervisé. Outre les approches classiques comme K-means, les approches récentes sont développées grâce à la formulation en factorisation matricielle non négative (NMF). Optimisé par une mise à jour de l'algorithme de multiplication règle ou en alternant l'algorithme des moindres carrés, NMF peut apporter beaucoup plus de précision que K-means dans le cas haute dimension (des centaines de millions de données) [34], surtout quand non négativité est imposée à chaque itération de la procédure des moindres carrés alternés (ANLS) [35], [36]. Cette méthode d'apprentissage est populairement utilisée dans l'analyse de texte. De nombreuses autres explorations de données et de problèmes d'apprentissage peuvent être reformulées comme un problème NMF.

2.2 Semi-supervisé Clustering basé sur les graphes

Dans cette catégorie, nous considérons avec SSC, la représentation basée sur les graphes, avec l'accent mis sur les semi-supervisé graph-cuts. La principale différence entre les méthodes basée sur les caractéristiques et celles considérées ici est que la base de données est sous la forme d'une similarité graphique G entre les points de données, plutôt que des vecteurs de caractéristiques correspondant à chaque point de données.

Nous définissons un concept pertinent pour des coupes sans surveillance graphique. Soit $G = (V, E)$ un graphe non orienté, pondéré avec matrice de poids W. Si V_1, V_2 est une partition de V, c.-à-d., $V_1 \bigcap V_2 = /\emptyset, V_1 \bigcup V_2 = V,$

alors la valeur de la coupe par le cloisonnement implicite (V_1, V_2) est donnée par la

$coupe(V_1, V_2) = \sum_{v_i \in V_1, v_j \in V_2} w_{ij}$.

Le problème de coupe minimale est de trouver une coupe (V_1, V_2) telles que la coupe (V_1, V_2) est réduit au minimum. En raisons pratiques, l'une travaille souvent avec un objectif coupe normalisé, tel que le ratio-cut [37] ou normaliséescut [38], qui encouragent la partition V_1, V_2 à être plus équilibré. Dans cette approche, nous trouvons quatre sous-approches : [39]

- Semi-supervisé non normalisée graph.
- Semi-supervisé ratio-cut.
- Semi-supervisé normalisée graph.
- Semi-supervised embedding.

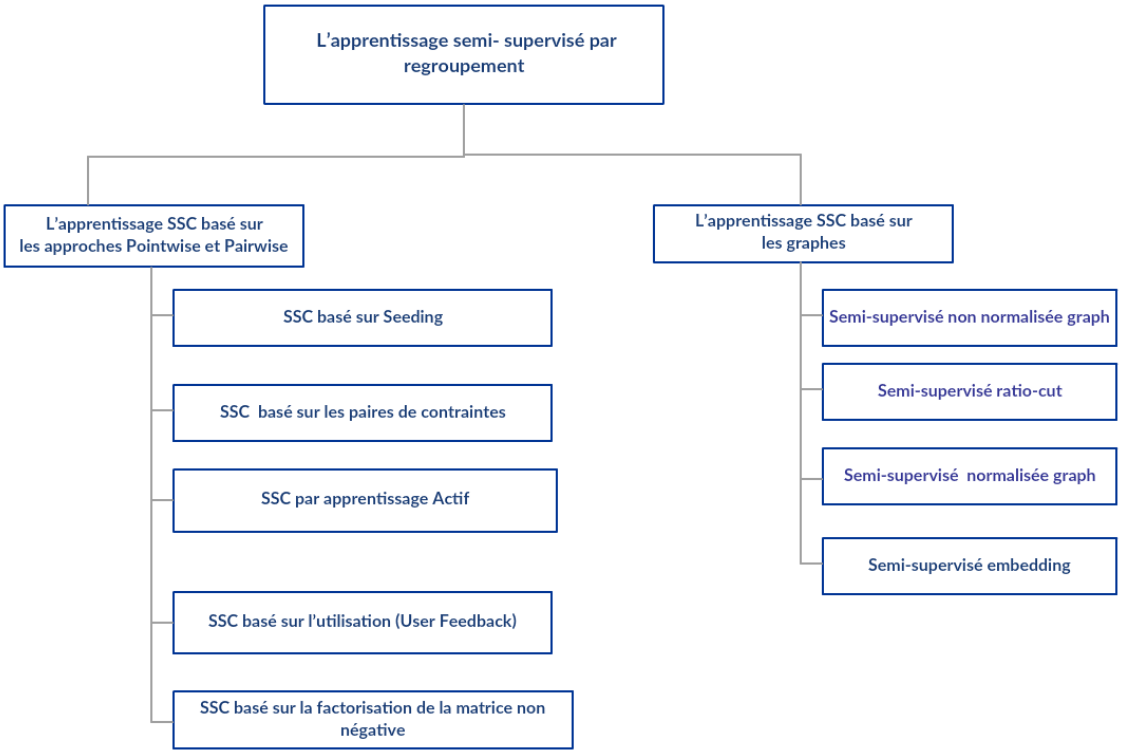

FIGURE 2.2 – L'organigrame Semi-supervisé Clustering

3 Proposition

Dans ce projet de fin d'études de Master, nous avons présenté les différentes approches d'apprentissage semi-supervisé par regroupement. Plus particulièrement, nous nous sommes intéressés au problème de l'apprentissage semi-supervisé sous contraintes. Dans le cadre de cet apprentissage, nous proposons dans ce travail une étude comparative entre les méthodes suivantes :

- L'algorithme des K-means modifié qui applique des contraintes de comparaison entre paires d'objets, appelé COP-KMEANS.
- L'algorithme Kernel Mean Shift Clustering semi-supervisé.
- L'algorithme SKLR : regroupement semi supervisé par des contraintes de distances relatives.

Ces algorithmes d'apprentissage semi-supervisé par regroupement sont basés sur les paires de contraintes (Pairwise constraint) qui sont fournies sous la forme de deux types de contraintes : must-link (ML) et cannot-link(CL).

L'étude comparative concernera à évaluer leur performances sur des bases de données médicales. Ces performances sont déterminées à l'aide de méthodes de comptage de pair, et plus particulièrement, de la méthode de L'index de Rand (RI) et L'indice de Rand Ajusté (ARI) qui permettent de mesurer l'influence de l'intégration des contraintes dans ces algorithme afin de guider et améliorer leur

performance.

4 Conclusion

L'apprentissage semi-supervisé par regroupement est une tâche d'apprentissage essentiellement non-supervisée,qui cherche notamment à guider le processus de clustering à l'aide de contraintes, il regroupe plusieurs méthodes ayant chacune des approches différentes.

Dans ce chapitre nous avons voulu en premier lieu, présenter les différentes types de contraintes qui sont utilisées dans l'apprentissage semi-supervisé par regroupement. En second lieu, nous avons donnés un aperçu détaillé sur l'apprentissage semi-supervisé qui désigne un vaste ensemble de méthodes tout en explorant les différentes approches le regroupant.

Chapitre 3

Principe des approches de regroupement semi-supervisé par contraintes

Introduction

Nous allons dans ce chapitre décrire brièvement les 3 méthodes choisies parmi les nombreuses méthodes qui ont été proposées pour résoudre le problème du regroupement sous contrainte. De ce fait, nous allons décrire quelques-unes des méthodes les plus couramment utilisées.

La première méthode est une version améliorée de l'algorithme k-means [40] qui met en application des contraintes, appelée COP- kmeans [22].

La seconde méthode est l'algorithme Semi-Supervised Kernel Mean Shift Clustering (SKMS) [41] il intègre la supervision par des paires de contraintes pour guider la procédure de mean shift clustering [41].

La troisième méthode est l'algorithme de regroupement semi supervisé par contraintes de distances relatives (SKLR) [42] . Sa principale contribution est d'étendre l'algorithme SKMS de sorte qu'il gère les comparaisons de distance relative.

1 Le principe de l'algorithme COP K-means

L'objectif des méthodes d'analyse de clustering vise à partager un ensemble de données dans des sous-groupes homogène. Nous remarquons que la plupart des méthodes de classification semi-supervisée de regroupement existants sont des versions modifiés de k-means [40]. L'algorithme COP K-means étant le plus abouti de ces versions, par l'intégration des contraintes, il a été démontré que cette approche permet de guider et améliorer le clustering, et ainsi améliorer les résultats des algorithmes de classication. [22]

Dans le cadre de la classification semi- supervisée,Il existe de nombreuses façons d'effectuer le regroupement des données par contrainte. COP K-means considérer particulièrement deux types de contraintes possibles entre les obser-

vations : si les deux observations sont dans même cluster on a le type "must - link" ; sinon ils sont dans différent cluster donc cette contrainte est "cannot-link".

L'algorithme prend un ensemble de données (D), un ensemble de contraintes must-link $(Con_{=})$ et un ensemble de contraintes cannot-link $(Con_{\neq})$. Il renvoie une partition des instances dans D qui satisfait toutes les contraintes spécifiées.

L'algorithme se décompose selon les étapes suivantes :

1. L'algorithme sélectionne aléatoirement K points comme centres de clusters initiaux.
2. Chaque point des point restants est assigné à son cluster le plus proche tout en assurant qu'aucune contrainte must-link et cannot-link n'est brisée.
3. Chaque centre de cluster est mis à jour pour être le moyen de ses points constitutifs
4. répéter (2) et (3) jusqu'à la convergence (jusqu'à ce que les centres des classes ne varient plus).

Selon Wagstaff et al. [22] le point n'est assigné à aucune classe si :

– Un point $d_{=}$ qui doit être affecté au même cluster que d, mais qui est déjà dans un autre cluster, ou s'il y a un autre point $d_{\neq}$ qui ne peut pas être groupé avec d mais qui est déjà en C.
– Si aucun cluster ne peut être trouvé pour d (Les contraintes ne sont jamais brisées), il est renvoyé à la partition vide

Algorithm 1 :COP K-means

D **: un ensemble de données**
$(Con_{=})$ **: un ensemble de contraintes must-link**
$(Con_{\neq})$ **:un ensemble de contraintes cannot-link**

1: **Sélectionnez aléatoirement :** $Kpoints$ **: centres de clusters initiaux.**
2: **Chaque point** $d_i \in D$ **est assigné à son cluster le plus proche tout en assurant qu'aucune contrainte** $(Con_{=})$ et $(Con_{\neq})$ **n'est brisée.**
3: **Mis à jour de Chaque centre de cluster pour être le moyen de ses points constitutifs**
4: **Répéter**(2)**et**(3) **jusqu'à la convergence**

2 Le principe de l'algorithme SKMS (Semi-supervised Kernel Mean Shift clustering)

L'objectif de cette méthode est d'intégrer la supervision dans la méthode mean shift clustering [41] qui utilise uniquement des contraintes paires pour guider la procédure de clustering.

Le mean shift clustering est une technique (algorithme) non paramétrique de recherche de mode populaire puissante qui ne nécessite pas une connaissance préalable du nombre de clusters et ne limite pas la forme des clusters.

- Mode populaire qui localise itérativement les modes dans les données en maximisant l'estimation de la densité du noyau (kernel density estimate KDE).
- La nature non paramétrique du mean shift fait un outil puissant pour découvrir des clusters de forme arbitraire présents dans les données. En outre, le nombre de clusters est automatiquement déterminé par le nombre de modes découverts.

L'algorithme de SKMS (Semi-supervised kernel Mean Shift clustering) généralise l'opération de projection linéaire à une transformation linéaire de l'espace du noyau qui va permettre d'escalader la distance entre les points de contrainte. À l'aide de cette transformation, les points de must-link sont rapprochés, tandis que les points de cannot-link peuvent être déplacés plus loin, voir Figure 3.1.

Cette transformation est faite comme suite :

Pour chaque cluster dans la base de données, une petite quantité de données labellisées est utilisée pour générer les contraintes paires (must-link et cannot-link).

1. Les données sont d'abord mappées sur un espace de noyau (kernel).
2. Le mean shift est appliqué d'abord sur les données non labélisées (apprentissage non-supervisé).
3. La matrice de distance par paire (PDM) à l'aide des modes découverts par le regroupement de mean shift non supervisé est effectuée dans l'espace du noyau.
4. Après l'ajout des données supervisées par le biais des contraintes paires.
5. Sélectionner le paramètre σ pour la fonction kernel gaussienne initiale avec la minimisation de la métrique log det divergence [41] [43] et calculer la matrice initiale.
6. Après le calcul de la matrice de noyau de faible rang $(r \leq n)$.
7. Sélection de paramètre de largeur de bande (k) avec l'utilisation de la matrice de noyau de faible rang et seulement les contraintes must-link.
8. Application de l'algorithme SKMS pour affecter les données aux les clusters.

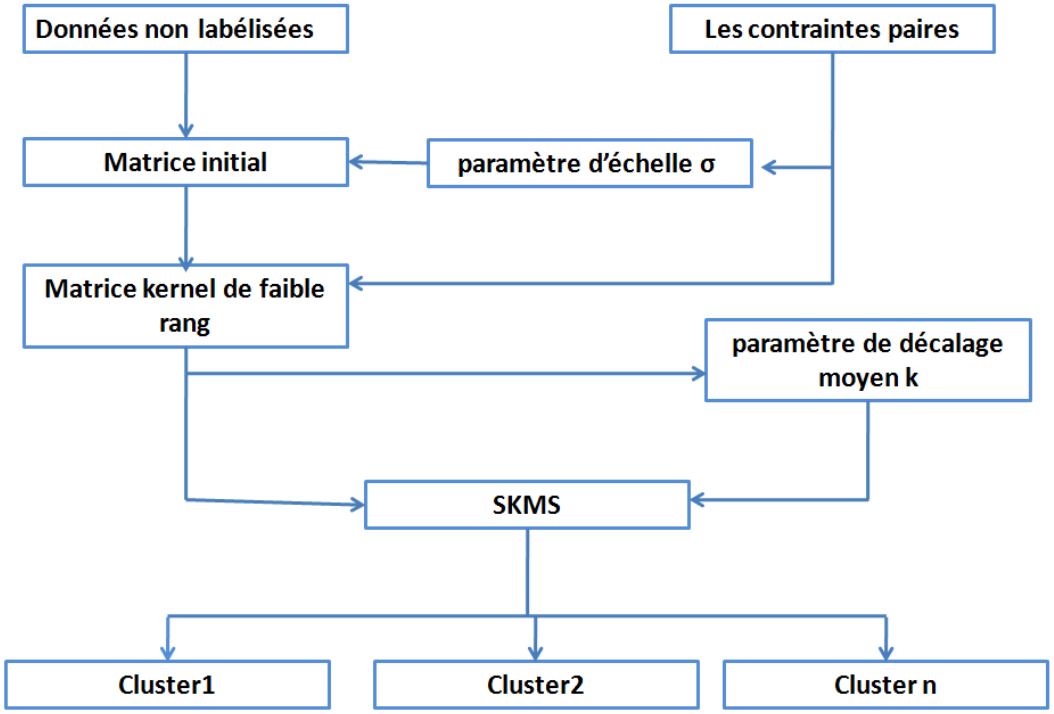

FIGURE 3.1 – L'organigramme de fonctionnement de l'algorithme de SKMS

La sélection de paramètre d'échelle σ : Soient ML et CL les ensembles de must-link et cannot-link respectivement, tels que $m + c = nc$. Laissez dm et dc être les seuils de distance carrés cibles pour must-link et cannot-link.

La sélection de paramètre d'échelle σ pour la fonction kernel gaussienne initiale est mesurée à l'aide des ensembles M et C et les distances d_m, tel que, dans l'espace initial du noyau, les distances entre les points de must-link sont petits, tandis que ceux entre les points de cannot-link sont grands. Il en résulte une bonne régularisation et une convergence plus rapide de l'algorithme d'apprentissage.
Les auteurs [41] ont supposé que les distances par paires entre les points d'échantillonnage dans l'espace caractéristique induite par matrice initiale de noyau (K_σ)se trouvent dans l'intervalle [0,2]. Ceci donne un moyen efficace de fixer les distances $d_m = min(d_1, 0.05)$ et $d_c = max(d_{99}, 1.95)$, où d_1 et d_{99} sont le premier et le 99ème percentile des distances entre toutes les paires de points de l'espace du noyau.

Apprentissage du noyau de faible rang(r) : Le calcul de la matrice de faible rang est tel que, $r \leq n$, et $\frac{||(K)||_F}{||(K)_\sigma||_F} \leq 0.99$. Avec K_σ, la matrice initiale de noyau (initial kernel matrix) et K la matrice de noyau de faible rang (low-rankkernel matrix).Les contraintes must-link sont générées à l'aide des points marqués de chaque classe, tandis qu'un nombre égal de contraintes cannot-link peut être utilisé.

Pour une paire de contraintes donnée, la mise à jour correspondante du noyau est équivalente à minimiser la divergence de log det entre la matrice de noyau mise à (k) et la matrice de noyau initiale(K_σ).Les données non linières sont cartographiées dans un espace de noyau de dimension supérieure où les contraintes sont effectivement imposées en appliquant une transformation linéaire. Cette transformation est apprise en minimisant une divergence de log detBregman entre le noyau initial et le noyau appris. Le procedure de SKMS selon les étapes suivantes :

Algorithm 2 :SKMS

Entrée :
D **: un ensemble de données : D**
$(Con_=)$ **: un ensemble de contraintes must-link**
$(Con_\neq)$ **:un ensemble de contraintes cannot-link**
γ **: Facteur de distance constante**

Mappez les données sur un espace de noyau (kernel).
Appliquez le mean shift sur les données non labélisées
Sélectionner le paramètre σ pour la fonction kernel gaussienne initiale avec la minimisation de la métrique log det divergence
calculer la matrice initiale
Calculez la matrice Kde noyau de $(n * n)$ de low-rank K_0
Sélection de paramètre de largeur de bande (k) avec l'utilisation de la matrice de K et seulement les contraintes$(Con_=)$
•**Répéter** pour $X_i \in D, i = 1...n$
L' application de l'algorithme **SKMS** pour affecter les données aux les clusters.
Jusqu'à ce que toutes les contraintes soient satisfaites
• **Retour :**lables de classe

3 Le regroupement semi supervisé avec contraintes de distances relatives SKLR

Cet algorithme est largement inspiré par l'algorithme SKMS. La principale contribution est d'étendre l'algorithme SKMS de sorte qu'il gère les comparaisons de distance relative. Ceci afin de considérer le problème de la mise en grappe d'un ensemble de données en k groupes soumis à un ensemble supplémentaire de contraintes sur les comparaisons de distance relative entre les éléments de données. [42]
Les contraintes supplémentaires sont destinées à présélectionner des informations latérales qui ne sont pas exprimées directement dans les vecteurs de caractéristiques.
Les comparaisons relatives peuvent exprimer des structures à un niveau de détail plus étroit (finer) que les contraintes de must-link (ML) et de cannot-link (CL) qui

sont couramment utilisées pour le regroupement semi-supervisé.
Les comparaisons relatives sont particulièrement utiles dans les contextes où l'octroi d'une contrainte ML ou CL est difficile parce que la granularité du clustering réel est inconnue. L'algorithme SKLR se décompose selon les étapes suivantes :

Algorithm 3 :SKLR

Entrée :
initiale $(n * n)$ **matrice kernel** K_0
C_{neq} **et** C_{eq} **: ensemble de comparaisons relatives**
γ **: Facteur de distance constante**
Sortie : matrice kernel K

1: **Trouvez une représentation de bas niveau : low-rank**
2: **Calculez la matrice** K**de noyau de** $(n * n)$ **de low-rank** K_0
3: **En utilisant une décomposition incomplète de Cholesky**
4: **Trouvez** $(n * r)$ **colonne de la matrice orthogonale** Q
5: **Appliquer la transformation** $\hat{M} \leftarrow Q^\top MQ$ **Sur toutes les matrices**
6: **Initialiser la matrice du noyau** $\hat{K} \leftarrow \hat{K}_0$
7: **•Répéter**
8: **(1) Sélectionnez une contrainte insatisfaisante** $C \in C_{neq} \bigcap C_{eq}$
9: **(2) Appliquer la projection de Bregman**
10: **Jusqu'à ce que toutes les contraintes soient satisfaites**
11: **• Retour** $K \leftarrow Q$ $K \succ Q$

4 Les métriques d'évaluation de l'apprentissage semi-supervisé par regroupement

La nécessité de comparer les regroupements a été abordée dans plusieurs domaines. Par conséquent, différentes mesures ont été utilisés et il n' y a pas de consensus général sur le choix de la mesure de comparer les regroupements [44] [45]. Une stratégie fréquente est basée sur le dépouillement des paires d'entités sur lesquels deux regroupements sont d'accord ou non. Les indices de cette classe sont souvent connus sous le nom d'accord par paire, un récent examen répertorie 28 paires différentes de mesures d'accord [45]. Cependant, après correction pour l'occasion, plusieurs de ces mesures deviennent équivalentes [45]. Un grand nombre de mesures d'accord ont été proposés dans la littérature, elles peuvent être classées en trois types de mesures :

4.1 Comptage de pair

qui est fondé sur le dépouillement des paires de points et la comparaison de l'accord et le désaccord entre les deux regroupements. L'indice de Jaccard [46] ,l'indice de Rand [47] l'indice de Fowlkes et Mallows [48] et l'indice de Rand ajusté [49] sont des exemples de ce groupe de mesures.

4.2 Mettre en correspondance

qui est basées sur la mesure de l'ensemble commun de la cardinalité entre deux regroupements. F-mesures [50] et les taux de classification [51] sont des exemples de ce groupe de mesures.

4.3 La théorie de l'information

qui est fondée sur les probabilités conditionnelles résultant du nombre de points répartis entre les deux groupes de clustering.

5 Conclusion

Dans ce chapitre nous avons présenté les 3 méthodes utilisées dans notre étude comparative, pour résoudre le problème d'apprentissage semi-supervise sous contraintes. De leurs principe de fonctionnement,nous distinguons que ces méthodes sont des algorithmes d'apprentissage semi-supervisé par regroupement basé sur l'approche par paires de contraintes (Pairwise constraint).

Dans le chapitre suivant, nous allons présenté les Résultats et les expérimentations de notre étude comparative des 3 méthodes pour résoudre le problème d' approches de regroupement semi-supervisé par contraintes.

Chapitre 4

Résultats et expérimentations

Introduction

Dans ce chapitre nous allons exposer l'étude expérimentale entre les trois algorithmes sélectionnés d'apprentissage sous contraintes. Ces trois algorithmes seront testés sur des bases de données médicales choisissaient du dépôt d'UCI. L'interprétation des résultats sera réalisée à l'aide de métriques d'évaluations comme l'indice de Rand et l'indice de Rand Ajusté.

1 Expérimentations

1.1 Les bases de données

Avant de présenter les différents résultats, nous détaillons les différentes bases utilisées.

1.1.1 La base de données Pima

La base Pima Indienne Diabètes (PID) [52] D'Arizona est constituée de 768 femmes dont 268 sont diabétiques et 500 non diabétique. Chaque cas est formé de 9 attributs, dont le 8éme représente des facteurs de risque et le 9éme représente la classe du patient. Classe : permet de savoir si le patient montre des signes de diabètes selon les critères de l'organisation mondiale de la santé.

1.1.2 La base de données Bupa

Bupa est une base de données sur les troubles hépatiques collectée par BUPA Medical Research Ltd [53], elle contient 345 exemples de sexe masculin 8 (200 non malades et 145 malades) définit par 7 attributs dont le dernier représente la classe.

1.1.3 La base de données Pancreatic :

est une base de données sur laquelle contient 181 exemples définit par 6771 attributs [53]

1.1.4 La base de données Heartstatlog :

est une base de données prédisent l'absence ou la présence de maladies cardiaques elle contient exemples270 définit par 13 attributs dont le dernier représente la classe [53]

1.1.5 La base de données New-thyroid :

est une base de données sur laquelle contient 215 exemples définit par 15 attributs. [53]

1.1.6 La base de données Dermatologie :

La base de données Dermatologie dénommée «Derythemato-squamous» .Elle contient les informations médicales de 358 exemples définit par 35 attributs et six classes [53].

La base de données	# d'exemples	# d'attributs	# classes
Pima	768	8	2
Bupa	345	6	2
pancreatic	181	6771	2
heartstatlog	270	13	2
New-thyroid	215	15	3
dermatologie	358	34	6

TABLE 4.1 – Les caractéristiques des bases de données

1.2 Les métriques d'évaluation de l'apprentissage semi-supervisé par regroupement

Dans notre travail, nous allons utiliser l'approche de comptage de pair pour la comparaison des trois méthodes d'apprentissage semi-supervisé par regroupement.Bien que de nombreuses mesures existent qui résument la comparaison par paires, nous avons sélectionnés les deux les plus communément appliqués :

Indice de Rand (RI, Rand (1971)) : est une mesure de similarité permettant de comparer les regroupements. Il exprime la proportion de paires qui sont dans le même cluster ou non.Il est basé sur le dépouillement et la comparaison de paires de points de l'accord et le désaccord entre les deux regroupements ou deux règles de classement. RI peut être utilisé quand il y a un grand nombre de clusters [54].

L'indice de Rand [55] a été motivé par des problèmes classiques de classification dans lesquels le résultat d'une méthode de classification doit être comparé à une classification correcte. La mesure de performance la plus courante pour ce problème, calcule la partie des éléments correctement classés (respectivement mal classés) à tous les éléments. Pour Rand, la comparaison de deux regroupements n'était qu'une extension naturelle de ce problème qui est une extension correspondante de la mesure de performance : au lieu de compter des éléments

uniques, il compte correctement des paires d'éléments classés. Ainsi, l'indice de Rand est défini par :

$$R(C, C') = \frac{2(n_{11} + n_{00})}{n(n-1)}$$

Avec :

- (C,C') : les paires d'échantillons et les paires de comptage qui sont attribuées dans les même ou différents grappes dans les groupements prédits et réels,
- n_{11} : le nombre de couples qui sont dans le même groupe.
- n_{00} : le nombre de couples qui sont dans des groupes différents.

R varie de 0 (aucune paire classifiée de la même manière sous les deux groupes) à 1 (regroupements identiques). La valeur de R dépend de deux nombres : celui des grappes et des éléments.

Morey et Agresti ont montré que l'indice Rand dépend fortement du nombre de grappes [56]. Dans [57] Fowlkes et Mallows montrent que dans le cas (non réaliste) des regroupements indépendants, l'indice Rand converge vers 1 lorsque le nombre de grappes augmente, ce qui n'est pas souhaitable pour une mesure de similarité.

L'indice de Rand Ajusté (ARI) : L'indice de Rand Ajusté (ARI) reste le plus connu et largement utilisé, est souvent appliqué dans la validation du cluster puisque c'est une mesure de l'accord entre deux partitions : une donnée par le processus de regroupement et l'autre défini par des critères externes. Il est basé sur le dépouillement et la comparaison de paires de points de l'accord et le désaccord entre les deux regroupements ou deux règles de classement.

Certaines méthodes offrent une mesure globale de la concordance entre les regroupements, qui tienne également compte des distances inter-clusters, tels que Rand [58] offrant une vue plus fine. D'autres méthodes offrent une vue asymétrique de concordance, dans lequel l'accord de regroupement avec B peut être différente de l'accord de B à A. Un exemple de ce type de mesure est le coefficient de Wallace (W), qui a été appliqué à l'analyse de données de typage microbienne. [59] [60] [61].

2 Résultats

Pour comprendre comment les propriétés de configuration de contraintes affectent divers algorithmes, nous avons effectué différentes expérimentations en variant le taux de labellisation de manière aléatoire et graduelle sur les trois algorithmes sélectionnés d'apprentissage par contraintes.

2.1 Expérimentation 1 : l'algorithme cop-kmeans

L'approche cop-kmeans utilise une distance Euclidienne pour le regroupement des clusters initiaux. Par la suite, un ensemble de contraintes sont ajoutées

au niveau de l'instance sur le processus de clustering (background knowledge). Le choix de k nombre de classe est effectué à partir de données labellisées.

cop- kmeans														
lables %		5	10	15	20	25	30	35	40	45	50	60	70	80
pima	AR	0.02	0.03	0.12	0.04	0.06	0.04	0.06	0.06	0.05	0.05	0.06	0.07	0.10
	RI	0.51	0.52	0.57	0.54	0.55	0.54	0.54	0.54	0.54	0.54	0.54	0.55	0.56
bupa	AR	0.00	-0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	RI	0.56	0.71	0.50	0.49	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
pancreatic	AR	0.02	0.06	0.03	0.05	0.07	0.05	0.17	0.21	0.16	0.16	0.21	0.10	0.01
	RI	0.50	0.53	0.52	0.53	0.53	0.54	0.60	0.61	0.58	0.58	0.61	0.55	0.51
dermatologie	AR	-0.05	0.03	0.05	0.09	0.06	0.12	0.09	0.08	0.05	0.05	0.07	0.03	0.03
	RI	0.69	0.71	0.71	0.73	0.72	0.74	0.73	0.72	0.71	0.71	0.71	0.69	0.69
heartstatlog	AR	0.02	0.02	-0.02	0.02	0.06	0.05	0.07	0.03	0.04	0.04	0.02	0.02	0.02
	RI	0.51	0.49	0.49	0.51	0.53	0.53	0.54	0.52	0.52	0.52	0.51	0.51	0.51
newthyroid	AR	0.21	0.32	0.52	0.41	0.34	0.28	0.38	0.35	0.34	0.40	0.38	0.45	0.51
	RI	0.62	0.64	0.76	0.71	0.67	0.65	0.68	0.68	0.68	0.71	0.70	0.72	0.76

TABLE 4.2 – Résultat de l'algorithme cop-kmeans

Le protocole d'expérimentations consiste à augmenter à chaque fois la quantité de données labellisées.

De manière général, nous remarquons que même si nous augmentons les données labellisées, la valeur de RI tourne au alentour de 0.5 ce qui nous donne des résultats moyens pour chaque base de données, Il est évident qu'au fur et à mesure que la quantité de données labellisées, et le nombre de contraintes varient, la valeur de RI et AR augmentent selon la base de données utilisée.

Du Tableau 4.2. Nous enregistrons les meilleurs résultats pour la base Pima avec un indice RI égale à 0.57 avec 15%de données labellisées. Pour la base Bupa le taux de labellisation égale à 10%est suffisant pour obtenir un taux d'accord élevé avec l'indice RI égale à 0.71, malgré que nous ajoutons par la suite des contraintes, la valeur de RI reste moyenne et stable (RI=0.50). Pour la base pancreatic qui a la caractéristique d'avoir plusieurs variables, le RI= 0.61, à partir de 40%de quantités de donnes de labellisés. En ce qui concerne la base heart statlog , le taux de connaissance est moyen et ne dépasse pas un RI=0.54 , sur les différents taux de labellisation appliqués. La base multi-classes dermatologie atteins un bon taux d'accord à un taux de 30%avec un RI égale à 0.74. L'indice RI le plus élevé a été enregistré par la base Newthyroid mais avec le taux de contraintes atteignant les 80%.

Nous constatons dans le tableau 4.2. Que la deuxième métrique d'évaluation qui est L'indice de Rand Ajusté (ARI), a une valeur qui tourne aux alentours de 0. Nous expliquons cela par le fait que nous utilisons deux classe (binaire avec k=2) dans l'algorithme, il correspondant à un accord très faible ceci est clairement le cas dans les bases Pima- Bupa - Pancreatic - heart statlog. Alors que la valeur d'ARI est moyenne (0.50) pour les données multi-classe comme dermatologie et newthyroid.

Réflexion personnelle ou critique : Nous notons le désavantage de cette méthode qui réside dans le fait qu'un expert peut se retrouver en difficulté aussi

bien : lors de la validation des résultats si l'ensemble des données est de dimensions importantes, que dans l'apprentissage binaire, mais aussi dans les générations des contraintes et le choix de distance.

2.2 Expérimentation 2 : L'algorithme SKMS

SKMS														
lables %		5	10	15	20	25	30	35	40	45	50	60	70	80
pima	AR	0.42	0.92	0.95	0.96	0.95	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.97
	RI	0.70	0.96	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.99
bupa	AR	0.16	0.82	0.94	0.99	0.95	0.99	0.98	0.98	0.97	0.98	1	0.98	1
	RI	0.58	0.91	0.97	0.99	0.98	0.99	0.99	0.99	0.98	0.99	1	0.99	1
pancreatic	AR	0.01	0.04	0.93	1	1	1	1	1	1	1	1	1	1
	RI	0.50	0.52	0.97	1	1	1	1	1	1	1	1	1	1
dermatologie	AR	0.47	0.89	0.91	0.92	0.92	0.93	0.94	0.94	0.94	0.94	0.95	0.94	0.95
	RI	0.87	0.96	0.97	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
heartstatlog	AR	0.02	0.88	0.94	0.99	1	1	1	0.99	0.97	0.99	0.99	0.99	1
	RI	0.51	0.94	0.97	0.99	1	1	1	0.99	0.99	0.99	0.99	0.99	1
newthyroid	AR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	RI	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53

TABLE 4.3 – Résultat de l'algorithme SKMS

Nous reprenons le même protocole d'expérimentations appliqué précédemment sur le même benchmark de données. Nous notons que les valeurs de AR et RI varient de manière croissante pour chaque base de données utilisée en fonction de l'augmentation du taux de contraintes, sauf pour la base new-thyroid qui enregistre a taux d'accord moyen pour les deux partitions avec un indice de Rand ajusté égale à zéro ce qui correspondant à un accord très faible il y a aucun un effet de contraints sur l'apprentissage.

Nous remarquons de manière générale qu'à partir de 10٪à 15٪de taux de labellisation, nous enregistrons une augmentation rapide du taux d'accord pour toute les bases de données (binaire/multi-classes). Pour le cas des bases Pancreatic et Heart statlog elles atteignent un taux d'accord total (RI=1, AR=1) à partir de 20-25٪de labels, ce qui indique que les regroupements sont identiques.

Réflexion personnelle ou critique : L'intégrité de la supervision dans la méthode mean-shift clustering avec l'utilisation unique des contraintes par paires permet de guider la procédure de clustering, cette expérimentation nous a montré que les contraintes améliorent bien la performance de l'algorithme, cet avantage obtenu est prononcé quand le nombre de clusters dans les données est grand, et quand l'ensemble des données est de dimensions importante.

2.3 Expérimentation 3 : L'algorithme SKLR

SKLR														
lables %		5	10	15	20	25	30	35	40	45	50	60	70	80
pima	AR	0.11	0.10	0.14	0.12	0.09	0.10	0.11	0.17	0.15	0.09	0.11	0.13	0.08
	RI	0.56	0.55	0.57	0.56	0.54	0.55	0.55	0.59	0.57	0.55	0.55	0.56	0.54
bupa	AR	0.06	0.04	0.04	0.03	0.04	0.04	0.03	0.09	0.03	0.04	0.05	0.00	0.03
	RI	0.53	0.52	0.52	0.51	0.52	0.52	0.52	0.55	0.52	0.53	0.53	0.50	0.52
pancreatic	AR	0.00	0.00	0.00	0.00	0.02	0.02	0.04	0.03	0.01	0.02	0.02	0.02	0.02
	RI	0.50	0.50	0.50	0.50	0.51	0.51	0.52	0.51	0.51	0.51	0.51	0.51	0.51
dermatologie	AR	0.85	0.92	0.93	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.96
	RI	0.95	0.97	0.98	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
heartstatlog	AR	0.21	0.31	0.43	0.44	0.46	0.50	0.47	0.50	0.44	0.46	0.39	0.44	0.43
	RI	0.61	0.66	0.72	0.72	0.73	0.75	0.74	0.75	0.72	0.73	0.70	0.72	0.72
newthyroid	AR	0.56	0.70	0.76	0.79	0.80	0.85	0.88	0.79	0.85	0.85	0.85	0.85	0.85
	RI	0.78	0.85	0.88	0.90	0.90	0.92	0.94	0.90	0.92	0.92	0.92	0.92	0.92

TABLE 4.4 – Résultat de l'algorithme SKLR

Le tableau 4.4. Représente la variation de la valeur de RI et AR de la méthode SKLR, en augmentant par pas de 5%le pourcentage des données labellisées injectées pour l'apprentissage du modèle.

Nous remarquons que l'algorithme SKLR, enregistre des valeurs moyennes en générale pour les bases binaires comme Pima, Bupa et Pancréatic ; malgré les différents changements de quantités de données labellisées, le taux d'accord RI varie dans un intervalle de 0.52- 0.59.

Par contre, pour les bases dermatologie et newthyroid qui sont des bases multi-classe avec respectivement 6 classes et 3 classes, l'indice de Rand est fort c'est à dire un accord élevé pour les deux partitions atteignant les 0.99 et 0.92 respectivement.

Réflexion personnelle ou critique : L'utilisation des comparaisons de distances relatives comme des contraintes améliore la performance de l'algorithme et particulièrement quand l'apprentissage est réalisé sur des données multi-classes.

3 Analyse comparative

Dans cette section, nous proposons une étude comparative des trois algorithmes avec les bases d'expérimentations. Afin de désigner la meilleure performance pour chaque base avec tous les algorithmes, nous nous focalisons sur la métrique d'évaluation RI en fonction de la quantité de données labellisées ainsi que des graphes qui représentent les différents résultats pour chaque base de données.

3.1 Analyse comparative de la base de données PIMA

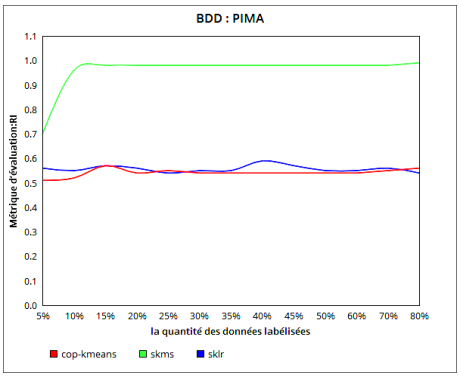

FIGURE 4.1 – La performance des 3 algorithmes pour la BDD Pima.

Algorithme	Ri	AR	Taux de données labellisées
Cop-kmeans	0.57	0.12	15 %
SKMS	0.99	0.97	15%
SKLR	0.59	0.17	40%

TABLE 4.5 – Les meilleures performances des 3 algorithmes pour BDD Pima.

Nous en déduisons des courbes des métriques en fonction des taux de labellisation (figure 4.1)., le tableau (table 4.5) qui représente les valeurs les plus élevées pour chaque algorithme avec la quantité de données labellisées correspondante. Pour la base de données PIMA, la meilleure performance est enregistrée par l'algorithme SKMS avec la quantité de données labellisées égale 15%.

3.2 Analyse comparative de la base de données BUPA

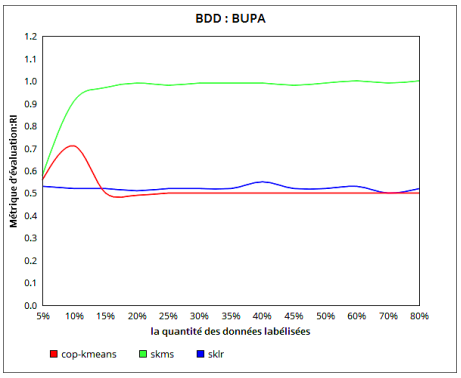

FIGURE 4.2 – La performance des 3 algorithmes pour la BDD Bupa.

Algorithme	Ri	AR	Taux de données labellisées
Cop-kmeans	0.71	0.04	10 %
SKMS	1	1	20%
SKLR	0.53	0.05	60%

TABLE 4.6 – Les meilleures performances des 3 algorithmes pour la BDD Bupa.

La figure 4.2 représente le graphe de performance des trois algorithmes appliqués à la base de données BUPA. Nous remarquons que l'algorithme SKLR enregistre une performance parfaite égale 1 avec une faible quantité de données labellisées égale 15%. Par contre l'algorithme SKLR reste presque stable avec une valeur moyenne qui n'a pas dépassé 0.53. Les performances de l'algorithme cop-kmeans, ont atteint un maximum égale 0.71, Ensuite, sont revenus s'installer aux valeurs moyennes pendant le changement de la quantité de données labellisées pour chaque algorithme.

3.3 Analyse comparative de la base de données Pancreatic

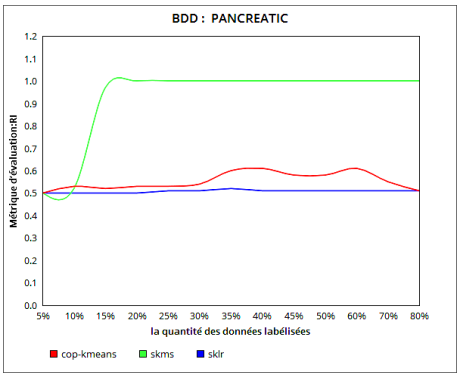

FIGURE 4.3 – La performance des 3 algorithmes pour la BDD Pancreatic.

Algorithme	Ri	AR	Taux de données labellisées
Cop-kmeans	0.61	0.21	40% et 60%
SKMS	1	1	20% jusqu'à 80%
SKLR	0.52	0.04	35%

TABLE 4.7 – Les meilleures performances des 3 algorithmes pour BDD Pancreatic.

Grâce à cette courbe (figure 4.3), nous continuons de noter la bonne performance de l'algorithme SKMS , les regroupements sont identiques le taux d'accord est exactement 1, à partir de 20% les données labellisées. En parallèle les algorithmes SKLR et cop-kmeans gardent leur performance moyenne pendant toute l'expérience. Dans la lecture numérique de la courbe, le tableau (table 4.7) représente l'extraction des meilleures performances pour chaque algorithmes.

3.4 Analyse comparative de la base de données Dermatologie

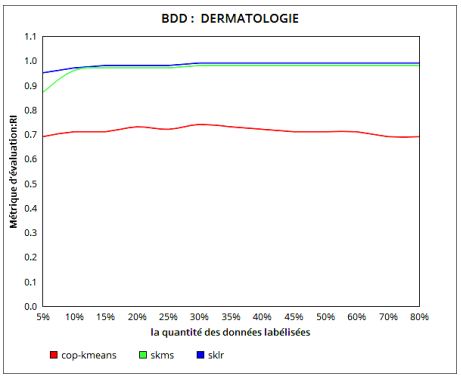

FIGURE 4.4 – La performance des 3 algorithmes pour la BDD Dermatologie.

Algorithme	Ri	AR	Taux de données labellisées
Cop-kmeans	0.74	0.12	30%
SKMS	0.98	0.95	15%
SKLR	0.99	0.96	15%

TABLE 4.8 – Les meilleures performances des 3 algorithmes pour BDD Dermatologie.

Ce qui distingue ce résultat, ils sont bons, de manière générale, particulièrement les deux algorithmes SKLR et cop-kmeans par rapport à leurs résultats enregistrés avec pima, bupa et pencreatic. Nous remarquons que la courbe de SKLR est identique avec le courbe de SKMS, avec des performances élevés au début (10% à15%). Par rapport à la méthode de SKMS, l'algorithme cop-kmeans reste stable avec une valeur ne dépassant pas les 0.74 comme la meilleure performance enregistrée. Ici, nous pouvons dire que l'algorithme de SKLR est le meilleur pour cette base de données multi-classes.

3.5 Analyse comparative de la base de données Heartstatlog

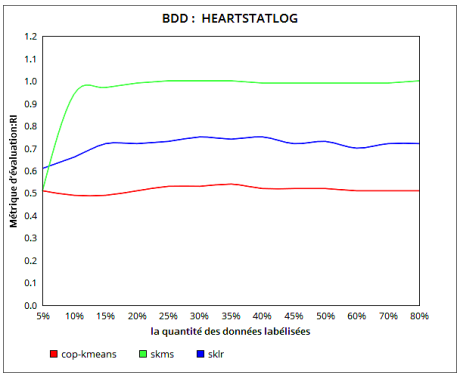

FIGURE 4.5 – La performance des 3 algorithmes pour la BDD Heartstatlog.

Algorithme	Ri	AR	Taux de données labellisées
Cop-kmeans	0.54	0.07	35%
SKMS	1	1	25%
SKLR	0.75	0.50	30%

TABLE 4.9 – les meilleures performances des 3 algorithmes pour BDD Heartstatlog.

De la figure 4.5 nous notons que les performances de l'algorithme SKMS garde toujours les meilleurs résultats, alors qu'une baisse de performances de SKLR est enregistrés. La courbe rouge qui représente la performance de cop-kmeans reste faible pendant l'expérience de celle-ci. Grâce à cette analyse dans le tableau (Table 4.9) et le graphique (Figure 4.5), nous déduisons que le meilleur algorithme est SKMS.

3.6 Analyse comparative de la base de données Newthyroid

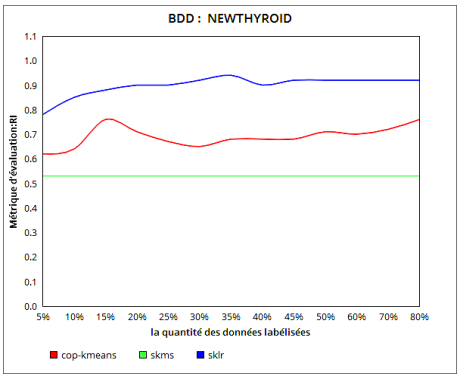

FIGURE 4.6 – La performance des 3 algorithmes pour la BDD Newthyroid.

Algorithme	# Ri	# AR	# Taux de données labellisées
Cop-kmeans	0.76	0.51	80%
SKMS	0.53	00	5% jusqu'à 80%
SKLR	0.92	0.35	30%

TABLE 4.10 – Les meilleures performances des 3 algorithmes pour BDD Newthyroid.

La performance des algorithmes change complètement lors de l'utilisation de la base de données Newthyroid. En effet, par rapport aux expérimentations précédentes, nous trouvons à partir du tableau (Table 4.10) et le graphique (Figure 4.6), que les résultats de SKMS prennent une valeur constante et moyenne ; ce qui indique que l'algorithme n'est pas affecté par les contraintes ajoutées. De son côté, l'algorithme SKLR garde les meilleurs résultats. Enfin pour cop-kmeans, leur meilleur taux d'accord, a dû nécessité un grand taux de données labellisées. Nous déduisons grâce à cette analyse que le meilleur algorithme est SKLR pour

le BDD newthyriod.

En résumé :
Dans cette partie nous comparons les différents algorithmes pour chaque base de données avec la représentation des graphes pour chaque base.

D'après ces différentes figures nous pouvons remarquer l'écart entre l'algorithme SKMS et les autres algorithmes (cop-kmeans et SKLR). Ces résultats de comparaison nous laissant présumer que l'algorithme le plus performant pour toutes les bases est l'algorithme de SKMS.

Base de données	Approches	AR	RI
Pima	Cop-kmeans	0.12	0.57
	SKMS	0.95	0.98
	SKLR	0.14	0.57
Bupa	Cop-kmeans	0.00	0.50
	SKMS	0.94	0.97
	SKLR	0.04	0.52
Pancreatic	Cop-kmeans	0.03	0.52
	SKMS	0.93	0.97
	SKLR	0.00	0.50
Dermatologie	Cop-kmeans	0.05	0.71
	SKMS	0.91	0.97
	SKLR	0.93	0.98
Heartstatlog	Cop-kmeans	0.02	0.49
	SKMS	0.94	0.97
	SKLR	0.43	0.72
Newthyroid	Cop-kmeans	0.52	0.76
	SKMS	0.00	0.53
	SKLR	0.76	0.88

TABLE 4.11 – Résultat des meilleures performances des 3 algorithmes pour chaque base de données à 15% de labels

Dans l'objectif de résoudre le problème de regroupement sous contraintes où le contexte de nombre de données labellisées est trop faible. Nous proposons dans cette partie d'analyse d'enregistrer les performances des trois algorithmes pour toutes les bases de données avec le taux de contraintes atteignant les 15% dans le tableau (Table4.11). Nous remarquons qu'à chaque fois l'algorithme SKMS garde la meilleure performance pour cette quantités de données labellisées.

4 Conclusion

A travers ces expérimentations, nous pouvons dire selon les expérimentations réalisées et l'étude comparative consacrée, que les algorithmes qui sont utilisée

s'adonnent à de bonnes performances en généralisation. Nous pouvons mentionner aussi à travers de ces expérimentations que les contraintes qui sont créé d'une manière aléatoire permettre d'améliorer les performances de ces méthodes et aussi la quantité de données labellisées a un impact direct sur les résultats.

Les résultats présentés dans ce chapitre montrent bien l'intérêt des différente techniques qui sont étudiées dans ce projet de fin d'études, et l'influence de contraintes sur l'apprentissage non supervisé.

Au terme de cette étude, les résultats des comparaisons ont révélé que : comme meilleur compromis est réalisé est par l'algorithme SKMS par rapport SKLR et cop-kmeans qui nécessitent des conditions et améliorations spécifiques.

Conclusion et perspectives

Par le biais de ce projet de fin d'études, nous avons découvert les différentes approches d'apprentissage comme l'apprentissage supervisé, non- supervisé, mais aussi leurs problèmes, qui consistent en pour la première type à un apprentissage à partir d'une base sans aucune connaissance préalable (auto-apprentissage), et pour le second à la spécification du nombre de clusters à produire en entrée de l'ensemble de données, avant de passer à la phase d'apprentissage.

Ces différentes lacunes ont fait naître un nouveau contexte qui est l'apprentissage semi-supervisé, il utilise des connaissances partielles qui sont soit incomplètes ou tout simplement les exemples étiquetés ne sont pas en quantité suffisante pour que l'on puisse appliquer des algorithmes supervisés.

Dans ce travail, notre intérêt a été porté plus par les approches de regroupement semi-supervisé par contraintes, et plus particulièrement celles basées sur les paires de contraintes. Ce projet de fin d'étude établi une étude comparative entre 3 algorithme qui sont :

- L'algorithme des K-means modifié qui applique des contraintes de comparaison entre paires d'objets, appelé COP-KMEANS.
- L'algorithme SKLR : regroupement semi supervisé par des contraintes de distances relatives.
- L'algorithme SKMS : Kernel Mean Shift Clustering semi-supervisé.

Nous pouvons dire à partir des résultats obtenus qu'en augmentant nos exigences sur la similarité, nous sommes tout de même parvenus à extraire de l'information pour enrichir, et obtenir un meilleur taux d'accord.

À la lumière des expériences faites, l'utilisation de ces 3 méthodes se révèle très prometteuse. Elles ouvrent de nouvelles perspectives au sein du « data-mining ». Les différentes pistes explorées pendant ce travail nous ont amenées à envisager de nombreuses perspectives. Nous présentons ici celles qui nous paraissent les plus prometteuses, Mentionner parmi eux des travaux futurs à court et à long terme.

Une des perspectives à court terme consiste à comparer ces approches avec d'autre techniques de clustering semi-supervisé dans le contexte de défis tels que le challenge des création des système d'aide de diagnostic dans domaine médicale, avec des données volumineuses, et spécialement dans le cas où l'on dispose

des bases ayant d'insuffisantes annotation.

Il sera aussi intéressant de réaliser une étude comparative des différents type de contraintes et la meilleure manière de les appliquer le but d'améliorer l'efficacité des méthodes.

Par ailleurs dans une perspective à long terme, l'intérêt sera porté sur la segmentation et l'annotation automatique de structures dans les images biomédicales qui sont des tâches essentielles à une multitude d'applications clés dont le diagnostique assisté, le suivi de pathologies et la recherche clinique. Le processus de segmentation est très complexe, dû notamment au faible contraste, à la superposition des régions d'intérêt et au bruit, typiquement présents dans les images médicales. Avec l'application des techniques regroupements semi-supervisé par contraintes, il sera possible d'étiqueter automatiquement les régions d'intérêt dans l'image ou le volume à segmenter. L'apport des contraintes permettra de guider l'algorithme pour une segmentation efficace et ciblée.

Bibliographie

[1] A. Cornuéjols, L. Miclet, and Y.Kodratoff, *Apprentissage Artificiel, Concepts et algorithmes*, 2002.

[2] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, "Uci repository of machine learning databases," 1998.

[3] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, "Uci repository of machine learning databases," 1998.

[4] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk, *A distribution-free theory of nonparametric regression*, Springer Science & Business Media, 2006.

[5] Luc Devroye, László Györfi, and Gábor Lugosi, *A probabilistic theory of pattern recognition*, vol. 31, Springer Science & Business Media, 2013.

[6] Anil K Jain, M Narasimha Murty, and Patrick J Flynn, "Data clustering : a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[7] Anil Jain and Douglas Zongker, "Feature selection : Evaluation, application, and small sample performance," *IEEE transactions on pattern analysis and machine intelligence*, vol. 19, no. 2, pp. 153–158, 1997.

[8] O Chapelle, B Schölkopf, and A Zien, "Semi-supervised learning, ser. adaptive computation and machine learning," 2006.

[9] Apprentissage Transductif and Arnaud Revel, "Apprentissage semi-supervisé," .

[10] Ellen Riloff, Janyce Wiebe, and Theresa Wilson, "Learning subjective nouns using extraction pattern bootstrapping," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics, 2003, pp. 25–32.

[11] Mathias Mahouzonsou Adankon, *Apprentissage semi-supervisé pour les SVMS et leurs variantes*, Ph.D. thesis, École de technologie supérieure, 2009.

[12] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[13] Kristin Bennett, Ayhan Demiriz, et al., "Semi-supervised support vector machines," in *NIPS*, 1998, vol. 11, pp. 368–374.

[14] Mathias Mahouzonsou Adankon, *Apprentissage semi-supervisé pour les SVMS et leurs variantes*, Ph.D. thesis, École de technologie supérieure, 2009.

[15] Olivier Chapelle, Vikas Sindhwani, and Sathiya S Keerthi, "Optimization techniques for semi-supervised support vector machines," *Journal of Machine Learning Research*, vol. 9, no. Feb, pp. 203–233, 2008.

[16] Michel Nizar and Nozha, "Unsupervised and semisupervised clustering : a brief survey," *Review of Machine Learning Techniques for Processing Multimedia Content*, 2005.

[17] Bassem Alsahwa, Basel Solaiman, É Bossé, Shaban Almouahed, and Didier Gueriot, "A method of spatial unmixing based on possibilistic similarity in soft pattern classification," *Fuzzy information and engineering*, vol. 8, no. 3, pp. 295–314, 2016.

[18] Ian Davidson and SS Ravi, "Clustering with constraints : Feasibility issues and the k-means algorithm," in *Proceedings of the 2005 SIAM international conference on data mining*. SIAM, 2005, pp. 138–149.

[19] Mikhail Bilenko, Sugato Basu, and Raymond J Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 11.

[20] Sugato Basu, Ian Davidson, and Kiri Wagstaff, *Constrained clustering : Advances in algorithms, theory, and applications*, CRC Press, 2008.

[21] Aharon Bar-Hillel, Tomer Hertz, Noam Shental, and Daphna Weinshall, "Learning a mahalanobis metric from equivalence constraints," *Journal of Machine Learning Research*, vol. 6, no. Jun, pp. 937–965, 2005.

[22] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, et al., "Constrained k-means clustering with background knowledge," in *ICML*, 2001, vol. 1, pp. 577–584.

[23] Zhengdong Lu and Todd K Leen, "Semi-supervised learning with penalized probabilistic clustering," in *Advances in neural information processing systems*, 2005, pp. 849–856.

[24] Kais Allab and Khalid Benabdeslem, "Sélection de contraintes pour la classification topologique semi-supervisée," in *Conférence Francophone d'Apprentissage CAP'11*, 2011, pp. 39–54.

[25] Sophia Roberts, G Gisler, and JAMES Theiler, "Spatio-spectral image analysis using classical and neural algorithms," *Smart Engineering Systems : Neural Networks, Fuzzy Logic, and Evolutionary Programming*, vol. 6, pp. 425–430, 1996.

[26] ULISES CORTÉS and JAVIER BÉJAR, "Experiments with domain knowledge in unsupervised learning : Using and revising theories," *Computación y Sistemas*, vol. 1, no. 003, 1969.

[27] Sugato Basu, Arindam Banerjee, and Raymond Mooney, "Semi-supervised clustering by seeding," in *In Proceedings of 19th International Conference on Machine Learning (ICML-2002*. Citeseer, 2002.

[28] Sugato Basu, Arindam Banerjee, and Raymond J Mooney, "Active semi-supervision for pairwise constrained clustering," in *Proceedings of the 2004 SIAM international conference on data mining*. SIAM, 2004, pp. 333–344.

[29] Sugato Basu, Mikhail Bilenko, Arindam Banerjee, and Raymond J Mooney, "Probabilistic semi-supervised clustering with constraints," *Semi-supervised learning*, pp. 71–98, 2006.

[30] Fan RK Chung, *Spectral graph theory*, Number 92. American Mathematical Soc., 1997.

[31] Christopher M Bishop, *Pattern recognition and machine learning*, springer, 2006.

[32] Arthur P Dempster, Nan M Laird, and Donald B Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[33] David Cohn, Rich Caruana, and Andrew McCallum, "Semi-supervised clustering with user feedback," *Constrained Clustering : Advances in Algorithms, Theory, and Applications*, vol. 4, no. 1, pp. 17–32, 2003.

[34] Jingu Kim and Haesun Park, "Sparse nonnegative matrix factorization for clustering," Tech. Rep., Georgia Institute of Technology, 2008.

[35] Hyunsoo Kim and Haesun Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM journal on matrix analysis and applications*, vol. 30, no. 2, pp. 713–730, 2008.

[36] Da Kuang, "Matrix factorization for clustering : Nmf and beyond," .

[37] Lars Hagen and Andrew B Kahng, "New spectral methods for ratio cut partitioning and clustering," *IEEE transactions on computer-aided design of integrated circuits and systems*, vol. 11, no. 9, pp. 1074–1085, 1992.

[38] Antonio Criminisi, Jamie Shotton, Ender Konukoglu, et al., "Decision forests : A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Foundations and Trends® in Computer Graphics and Vision*, vol. 7, no. 2–3, pp. 81–227, 2012.

[39] Charu C Aggarwal and Chandan K Reddy, *Data clustering : algorithms and applications*, CRC press, 2013.

[40] James MacQueen et al., "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Oakland, CA, USA., 1967, vol. 1, pp. 281–297.

[41] Saket Anand, Sushil Mittal, Oncel Tuzel, and Peter Meer, "Semi-supervised kernel mean shift clustering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 6, pp. 1201–1215, 2014.

[42] Ehsan Amid, Aristides Gionis, and Antti Ukkonen, "A kernel-learning approach to semi-supervised clustering with relative distance comparisons," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2015, pp. 219–234.

[43] Srujana Merugu and Suvrit Sra, "Learning with bregman divergences," 2007.

[44] Allan D Gordon, "Classification, (chapman & hall/crc monographs on statistics & applied probability)," 1999.

[45] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell, "Text classification from labeled and unlabeled documents using em," *Machine learning*, vol. 39, no. 2, pp. 103–134, 2000.

[46] Paul Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901.

[47] William M Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.

[48] Edward B Fowlkes and Colin L Mallows, "A method for comparing two hierarchical clusterings," *Journal of the American statistical association*, vol. 78, no. 383, pp. 553–569, 1983.

[49] Lawrence Hubert and Phipps Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[50] C. J. Van Rijsbergen, *Information Retrieval*, Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[51] Marina Meil ?, "Comparing clusterings : an axiomatic view," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 577–584.

[52] A Frank Pima and A Asuncion, "Pima indians diabetes dataset," *UCI Machine Learning Repository, University of California, Irvine*, 2010.

[53] Andrew Frank and Arthur Asuncion, "Uci machine learning repository [http ://archive. ics. uci. edu/ml]. irvine, ca : University of california," *School of information and computer science*, vol. 213, 2010.

[54] Celine Vens, Bart Verstrynge, and Hendrik Blockeel, "Semi-supervised clustering with example clusters," in *Proceedings of the 5th International Conference on Knowledge Discovery and Information Retrieval and the 5th International Conference on Knowledge Management and Information Sharing*, 2013, pp. 45–51.

[55] William M Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.

[56] Silke Wagner and Dorothea Wagner, *Comparing clusterings : an overview*, Universität Karlsruhe, Fakultät für Informatik Karlsruhe, 2007.

[57] Edward B Fowlkes and Colin L Mallows, "A method for comparing two hierarchical clusterings," *Journal of the American statistical association*, vol. 78, no. 383, pp. 553–569, 1983.

[58] Francisco R Pinto, João A Carriço, Mário Ramirez, and Jonas S Almeida, "Ranked adjusted rand : integrating distance and partition information in a measure of clustering agreement," *BMC bioinformatics*, vol. 8, no. 1, pp. 44, 2007.

[59] JA Carrico, C Silva-Costa, J Melo-Cristino, FR Pinto, H De Lencastre, JS Almeida, and M Ramirez, "Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant streptococcus pyogenes," *Journal of Clinical Microbiology*, vol. 44, no. 7, pp. 2524–2532, 2006.

[60] Nuno A Faria, João A Carrico, Duarte C Oliveira, Mário Ramirez, and Hermínia de Lencastre, "Analysis of typing methods for epidemiological surveillance of both methicillin-resistant and methicillin-susceptible staphylococcus aureus strains," *Journal of clinical microbiology*, vol. 46, no. 1, pp. 136–144, 2008.

[61] Anna C Shore, Angela S Rossney, Peter M Kinnevey, Orla M Brennan, Eilish Creamer, Orla Sherlock, Anthony Dolan, Robert Cunney, Derek J Sullivan, Richard V Goering, et al., "Enhanced discrimination of highly clonal st22-methicillin-resistant staphylococcus aureus iv isolates achieved by combining spa, dru, and pulsed-field gel electrophoresis typing data," *Journal of clinical microbiology*, vol. 48, no. 5, pp. 1839–1852, 2010.

Résumé

Dans un monde guidé par les données, l'apprentissage automatique est un outil essentiel pour aider les utilisateurs à appréhender la structure de ces données. Dans ce domaine il existe de nombreuses techniques d'apprentissage, l'apprentissage semi-supervisé reste le plus utilisé dans le cadre applicatif et réel, et cela en raison de son principe qui trouve ses racines dans les problèmes d'apprentissage en présence d'un petit nombre de données labellisées. Dans ce projet de fin d'étude nous nous intéressons à la catégorie d'approche d'apprentissage semi-supervisé par contraintes. Pour ce faire, nous réalisons une étude comparative de trois techniques d'apprentissage semi- supervisé par regroupement qui sont : cop-kmeans, Semi-supervised kernel Mean Shift clustering et le regroupement semi supervisé avec contraintes de distances relatives. Nous discutons et analysons en outre l'influence des contraintes par paires (must-link et cannot-link) sur les performances de clustering en effectuant des expérimentations avec différents pourcentages d'exemples marqués. Nous menons une étude sur 6 ensembles de données médicales. Les résultats expérimentaux indiquent que la méthode Semi-supervised kernel Mean Shift clustering peut généralement surpasser d'autres méthodes semi-supervisées. L'étude expérimentale montre que l'utilisation des contraintes peut améliorer les performances en particulier lorsque le nombre d'exemples étiquetés disponibles est insuffisant pour former un modèle de clustering. Des travaux futurs pourront concerner des systèmes d'aide diagnostique ou segmentation ciblée et une annotation automatique de structures dans les images biomédicales

Abstract

In a data-driven world, automatic learning was an essential tool to help users understood the structure of this data. In this field there were many techniques of semi-supervised learning that it finds its roots in problems in presence of missing data, in this pursuit that processes the category of learning semi-supervised clustering ,we have proposed a comparative study of three techniques of semi-supervised-learning by cop-kmeans Semi-supervised kernel MeanShift clustering and semi-supervisedclustering with relative distance comparisons We have discuss and analyzed, the influence of paired-constraints (must-link and can not-link) on clustering performance by performing experiments with different percentages of labeled examples.We conducted experiments on six sets of medical data. Experimental results indicate that the Semi-supervised kernel MeanShift clustering method may generally outperform other semi-supervised methods. The discovery shows that the use of constraints can improve performance especially when the number of labeled examples available is insufficient to form a clustering model.Future work may involve targeted diagnostic or segmentation systems and automatic annotation of structures in biomedical images.

الملخـــــــــــص

في هذا العالم الذي تسيره البيانات يعتبر التصنيف الآلي هو الأداة الأساسية لمساعدة المستعملين على فهم بنية هذه البيانات .في هذا المجال هناك العديد من التقنيات مثل التعلم بالإشراف النصفي الناتج عن مشاكل التعلم في غياب البيانات الموسومة , في هذا العمل الذي يرتكز على دراسة واحدة من الموضوعات المهمة التي تخص الذكاء الاصطناعي في عملية تجميع البيانات ضمن مجموعات متشابهة لهذا تستخدم خوارزميات التجميع على نطاق واسع ليس فقط لتنظيم وتصنيف البيانات ولكن لضغط البيانات وبناء نم اذج لترتيبها. يمكن تلخيص مشكلة البحث في معالجة التجميع بالإشراف النصفي باقتراح دراسة لمقارنة ثلاث تقنيات لهذا النوع من التعلم الآلي وهم خوارزمية التجميع K -العنقودية المعدلة cop-kmeans و خوارزمية تجميع نواة التحول المتوسط SKMS و خوارزمية التجميع مع قيود المسافات النسبية SKLR ، بحيث نناقش ونحلل تأثير القيود الثنائية (قيود يجب الارتباط - قيود لا يمكن الارتباط) في أداء التجميع وذلك عن طريق إجراء التجارب مع استعمال نسب مختلفة من مقدار الأمثلة الموسومة، على ست مجموعات من البيانات الطبية المختلفة. وتشير النتائج التجريبية أن الطريقة SKMS تتفوق عموما على الطرق الأخرى. ن ستخلص أن استخدام القيود يمكن أن تحسن الأداء وخاصة عندما يكون عدد الأمثلة الموسومة في قاعدة البيانات غير كاف لتشكيل نموذج التصنيف .وكرؤية مستقبلية نقترح أنظمة تشخيص طبية و أخرى تدعم التجزئة المستهدفة للشرح التلقائي في الصور الطبية الحيوية.

الكلمات المفتاحية :التصنيف الآلي. التعلم بالإشراف النصفي. الذكاء الاصطناعي. خوارزمية التجميع K العنقودية المعدلة Cop-kmeans . خوارزمية نواة تجميع التحول المتوسط SKMS . خوارزمية التجميع مع قيود المسافات النسبية SKLR . قيود يجب الارتباط. قيود لايكمن الارتباط.