Université Abou Bekr Belkaid
Tlemcen - Algérie
جامعة أبي بكر بلقايد
تلمسان الجزائر

République Algérienne Démocratique et Populaire
Université Abou Bakr Belkaid– Tlemcen
Faculté des Sciences
Département d'Informatique

Mémoire de fin d'études

pour l'obtention du diplôme de Master en Informatique

Option: *Modèles Intelligents et décision (MID)*

Thème

Classification des données biologiques par la méthode de sous espace aléatoire.

Réalisé par :

- **KHATER FATIMA**
- **LOUAHAD SAHIMA**

Présenté le 03 Juin 2013 devant le jury composé de MM.

- *Mr .Benazzouz Mourtada* *(Président)*
- *Mr .Chikh Mohamed Amine* *(Encadreur)*
- *Melle.Settouti Nessma* *(co_Encadreur)*
- *Mr .Benmouna Youcef* *(Examinateur)*
- Mr.Chouiti Sidi Mohamed *(Examinateur)*

Année universitaire : 2012-2013

type/leukemia/overview/?region=bc, 2013.

[7] Newsmedical, Leukemia, www.news-medical.net/health/Leukemia-Causes-(French).aspx, 25 juin 2013.

[8] Newsmedical, Leukemia, http://www.news-medical.net/health/What-is- Leukemia-%28French%29.aspx, 25 juin 2013.

[9] E-sant,http://www.e-sante.fr/leucemies-adulte-symptomes-diagnostic/guide/1054, 2012.

[10] FondationARC , Leukemia, http://www.arc-cancer.net/Les-leucemies-de-ladulte/vivre-avec-et-apres-une-leucemie-chez-un-adulte.html, 22-11-2012.

[11] PubMed,http://www.ncbi.nlm.nih.gov/pubmed/10521349.15-10-1999.

[12] FondationARC , Leukemia, http://www.arc-cancer.net/Les-leucemies-de-ladulte/la-recherche-sur-les-leucemies.html, 22-11-2012.

Remerciements

Nous remercions tout d'abord Dieu pour l'accomplissement de ce mémoire.

Ensuit nous adressons nos sincères remerciements à Monsieur Chikh Mohammed Amine, pour sa direction, sa disponibilité, ses orientations, et sa compréhension.

Nous remercions également Melle Settouti Nesma pour sa disponibilité, ses conseils, ses suggestions, et son encadrement.et remercie aussi pour avoir accepté d'examiner ce travail.

Nous remercions bien évidemment nos famille, nos parents, nos frère et nos amis de nous avoir toujours soutenu....

En suite nous désir adresser nos sincères sentiments à *Mr.Benazzouz Mourtada* d'avoir accepté de présider ce jury

nos remerciements à *Mr.Benmouna Youcef* et *Mr.Chouiti Sidi Mohamed* qui ont participé à examiner ce travail.

Enfin nous adressons nos plus sincères remerciements à tous nos professeurs du département informatique pour leurs enseignements

fatima et Sahima

Dédicace

Je dédie ce modeste travail à mes très chers parents que dieu les récompense et les garde, qui ont toujours été la pour moi, et me ont donne un magnifique modèle de labeur et de persévérance.

A mon père Tayeb qui m'a indique la bonne voie et qui ma tout donné. En témoignage de ma profonde gratitude pour les efforts qu'il a consenti a ma formation et m'a aidé à surmonter les moments difficiles, qu'il voit en ce travail la réalisation de ses vœux

A plus merveilleuse mère Abbes Khadidja qui a sacrifie sa vie pour faire épanouir sa famille, sa présence et petits soins qu'elle m'a donné m'ont procuré beaucoup de réconfort et maitrise pour attendre mon but, qu'elle trouve ici le témoignage de ma profond affection

Je vous remercie infiniment de vos sacrifices, que DIEU les protège et garde pour nous.

A ma chère grande mère Yacoubi Fatma

A mes chers frères Mohamed, Ilias, Yasser que je leurs souhaite la bonne réussit Pour leurs soutien morale et leurs sacrifices le long de ma formation

A mes chères sœurs Salima et son époux Ghouti, Souad et son époux Mousa

A Imen l'épouse de mon frère Mohamed

A mes nièces Ismahan, Habiba, Mohamed, Meriem et Fatima Zohra

A ma binome Sahima

A mes meilleurs amies: Karima,Sahima,Fatima.

A toute la famille KHATER

A tous ceux et toutes celles qui m'ont accompagné et soutenu de prés ou de loin durant cette année.

Je dédie ce projet de fin d'études en espérant la réussite et le succès.

Fatima

Dédicace

Je dédie ce modeste travail à mes très chers parents que dieu les récompense et les garde, qui ont toujours été la pour nous, et nous ont donne un magnifique modèle de labeur et de persévérance.

A mon pére qui m'a indique la bonne voie et qui ma tout donné. En témoignage de ma profonde gratitude pour les efforts qu' il a consenti a ma formation et m'a aidé à surmonter *les moments difficiles, qu'il voit en ce travail la réalisation de son vœux*

A plus merveilleuse mère Qui a sacrifie sa vie pour faire épanouir sa famille, sa présence et petits soins qu'elle m'a donné m'ont procuré beaucoup de réconfort et maitrise pour attendre mon but, qu'elle trouve ici le témoignage de ma profond affection

Je vous remercie infiniment de vos sacrifices, que DIEU les protège et garde pour nous.

A ma chére grande mère sadare rahma

A mes chèrs fréres mohamed, abd latif que je leurs souhaite la bonne réussit Pour leur soutien morale et leurs sacrifices le long de ma formation

A ma binome fatima

A mes tantes et à mes oncles.

A chaque cousins et cousines spécialement pour Hanane,Naima,Najat,Souad,Khira,Kotbia et Manal.

A mes meilleurs amies Nouria , Nabila, Khadidja ,Ahlam,Souad,fatima,kawter.

A toute la famille LOUAHAD

A tous ceux et toutes celles qui m'ont accompagné et soutenu de prés ou de loin durant cette année.

Je dédie ce projet de fin d'études *en espérant la réussite et le succès.*

Sahima

Table des matières

Liste des tableaux

Liste des figures

Glossaire

ADN Acide Désoxyribo Nuclieique.

ALL (Aculte Lynphplastic Leukemia),

ARR (Arrhythmia)

BAGGING Boostrap AGGregatING

LAL leucémie aiguë lymphoblastique

LMA leucémie myéloïde aiguë

LMC La leucémie myéloïde Cronique

MRMR minimum Redondance Maximum Relevance.

NFS numération formule sanguine

RSM Random Subspace method

RFE Recursive Feature Elimination

SE SEnsibilité

SP spécificité

SVM Support Vector Machine

TC taux de classification

Introduction générale

Introduction générale

La leucémie est considéré actuellement comme la maladie du siècle vu le nombre de Patient qui ne cesse d'augmenter. L'augmentation du nombre des patients est tellement rapide selon l'Agence de la biomédecine. [5] l'a identifié comme étant une épidémie.

La recherche scientifique facilite l'acquisition et le recueil de nombreuses données, notamment dans le domaine médical lors d'examen des patients. Ces données peuvent être utilisées comme support de décision médicale,

conduisant aux développements d'outils capables de les analyser et de les traiter

Pour la classification des maladies. L'utilisation des méthodes dites intelligentes pour effectuer cette classification sont de plus en plus fréquentes. Même si la décision du médecin est le facteur le plus déterminant dans le diagnostic.

mais les problèmes les plus intéressants sont souvent basés sur des données de grand dimensionnalité, Ces problèmes désignent les situations où nous disposons peu d'observations alors que le nombre de variables explicatives est très grand.

Ce travail de fin d'étude se situe dans le contexte général de l'Aide au Diagnostic médical, qui a pour but de réaliser un système capable d'aider le médecin pour le diagnostic du leucémie et de sélectionner les variables les plus pertinentes pour la reconnaissance de cette maladie, il est composé de trois chapitres :

_ Le premier chapitre concerne la leucémie : il présente le contexte lié au leucémie et le diagnostique de cette maladie la problématique confrontée dans ce domaine des données a grande dimension.

Le deuxième chapitre concerne l'état de l'art : il est partitionné sur trois parties :

_ la première comporte, l'état de l'art des travaux appliqués sur la leucémie d'une manière générale et plus particulièrement sur les travaux liés au sélection des variables

_ La seconde comporte, l'état de l'art des travaux réalisé par RSM

_ La troisième partie décrit notre contribution dans ce domaine.

_ le troisième chapitre concerne Résultats et discussion : il présente notre élaboration des différentes méthode nécessaires pour la classification de leucémie , puis nous décrivons les implémentations et leurs résultats obtenus et enfin Nous terminons ce chapitre par une comparaison entre nos résultats et ceux de la littérature.

_ En dernier lieu, une conclusion générale et des perspectives de ce travail de Master seront présentées.

Chapitre I :

Leucémie

CHAPITRE 1

Leucémie

1.1. INTRODUCTION

Chaque année, la leucémie touche environ 350.000 personnes dans le monde avec environ 257.000 décès par année [1] Ce qui représente 3% de tous les cancers dans le monde [2], il définit un cancer du sang, ou en termes plus précis des affections hématologiques malignes caractérisées par la prolifération incontrôlée, dans la moelle osseuse, de cellules qui sont à l'origine des globules blancs du sang.

Le nombre de nouveaux cas annuels dépasse les 6 000, et les personnes touchées sont le plus souvent des enfants ou des personnes âgées [3]. En Algérie, 250 nouveaux cas de LMC sont enregistrés chaque année [4], dans ce chapitre, nous présentons les différents types du la leucémie, les causes ainsi que les étapes d'évolution du cette maladie, ensuite nous abordons l'aide au diagnostic de la leucémie et nous finirons par l'aspect biologique de la maladie à travers les gènes.

1.2. DEFINITION DE LA LEUCEMIE

La leucémie est un cancer qui prend naissance dans les cellules souches du sang (cellules sanguines immatures) qui se trouvent dans la moelle osseuse. La moelle osseuse est la matière molle et spongieuse qui remplit le centre de la plupart des os. C'est là que sont fabriquées les cellules sanguines. Les cellules souches du sang peuvent devenir soit des cellules souches myéloïdes, soit des cellules souches lymphoïdes.
Les cellules souches myéloïdes se développent en globules rouges, en globules blancs ou en plaquettes.

- Les globules rouges transportent l'oxygène vers tous les tissus de l'organisme.

- Il existe plusieurs types différents de globules blancs. Les cellules souches myéloïdes peuvent se développer en granulocytes et en monocytes, lesquels détruisent les bactéries et luttent contre les infections.
- Les plaquettes forment des caillots dans les vaisseaux sanguins endommagés afin de prévenir les hémorragies.

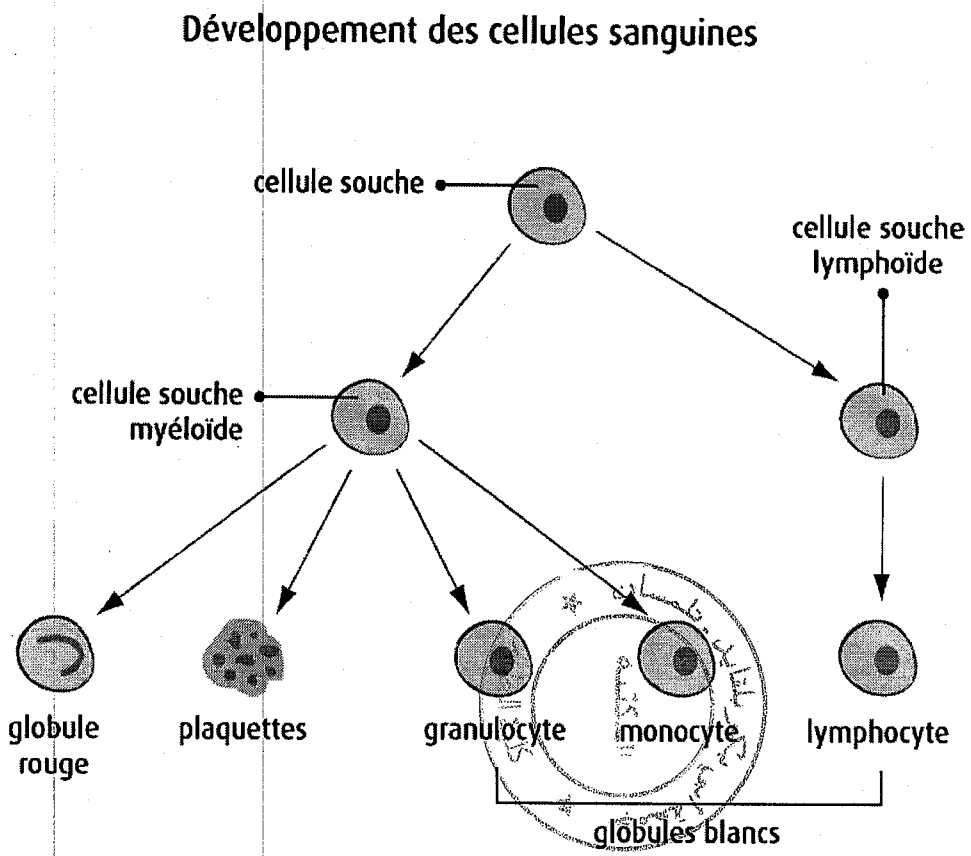


FIGURE 1.1.Représentation de développement des cellules sanguines.

- Les cellules souches lymphoïdes se transforment en lymphocytes, un autre type de globule blanc. Les lymphocytes se trouvent habituellement dans le sang et différentes

parties du système lymphatique, notamment dans les ganglions lymphatiques et la rate. Les lymphocytes fabriquent les anticorps qui aident à combattre les infections.

- La leucémie se développe lorsque des cellules souches sanguines présentes dans la moelle osseuse fabriquent des cellules sanguines anormales. Ces cellules anormales, appelées cellules leucémiques, se multiplient peu à peu et finissent par dépasser en nombre les cellules normales. Il devient alors difficile pour les globules blancs, les globules rouges et les plaquettes d'accomplir adéquatement leurs tâches respectives.

1.3. TYPES DE LEUCEMIE

Il existe plusieurs types de leucémie. Les différents types de leucémie sont d'abord classés selon le type de cellule souche du sang à partir duquel ils se développent :

- La leucémie myéloïde prend naissance dans les cellules myéloïdes anormales.
- La leucémie lymphoïde (aussi appelée leucémie lymphoblastique) se développe à partir de cellules lymphoïdes anormales.

Chaque type de leucémie se subdivise ensuite en sous-catégories, en fonction de la rapidité avec laquelle la maladie se développe et évolue :

- La leucémie aiguë débute de manière soudaine et se développe souvent en quelques jours ou quelques semaines. La quantité de cellules leucémiques dans le sang peut grimper en flèche et les cellules sanguines n'arrivent plus à jouer leur rôle.
- La leucémie chronique se développe lentement au fil des mois ou même des années. Elle ne cause parfois aucun symptôme au début. Les symptômes commencent à se manifester au fur et à mesure que la quantité de cellules leucémiques dans le sang ou la moelle osseuse augmente [6].

1.4. CAUSE DE LEUCEMIE

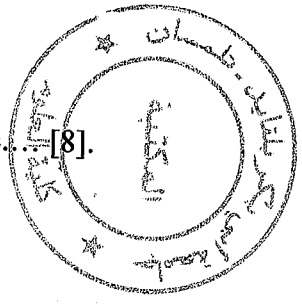
La cause exacte de la leucémie est inconnue mais plusieurs facteurs de risque peuvent soulever la possibilité d'obtenir des leucémies. Ceux-ci comprennent familial et des risques génétiques ainsi que des facteurs environnementaux et de mode de vie [7].

1.5. LES ETAPES D'EVOLUTION DE LEUCEMIE

1.5.1. Signes et symptômes de leucémie

Il y a plusieurs signes et symptômes de leucémie qui sont communs à tous les types. Les prises de sang et les tests Particuliers de moelle osseuse sont nécessaires pour effectuer un diagnostic. Dans les leucémies aiguës les symptômes communs comprennent :

- manque d'énergie
- anémie
- Leucopénie
- Thrombopénie
- Trouble de l'hémostase
- fatigue
- fièvre prolongée
- la nuit sue
- propension d'attraper les infections etc.…[8].



1.5.2. Le diagnostic

Une leucémie peut être suspectée suite à une simple prise de sang, lorsque la numération formule sanguine (NFS) est anormale : l'analyse sanguine montre alors une baisse du nombre de globules rouges, de plaquettes et de polynucléaires. Elle peut aussi mettre en évidence la présence de cellules leucémiques au travers d'une quantité de globules blancs anormalement élevée.

Cependant, le diagnostic de leucémie aiguë ne peut se fonder uniquement sur cette analyse sanguine. Si les résultats de la numération formule sanguine laissent suspecter une leucémie

aiguë, le patient doit être adressé à un centre d'hématologie spécialisé pour confirmer le diagnostic grâce à un myélogramme.

Le myélogramme est l'examen clé permettant de poser un diagnostic de leucémie aiguë. Il consiste à analyser les cellules de la moelle osseuse au microscope. Le prélèvement de moelle osseuse est effectué sous anesthésie locale, par ponction dans le sternum ou dans l'os du bassin (épine iliaque). Ce geste ne dure que quelques secondes mais peut être douloureux. Des antalgiques sont donc souvent prescrits en plus de l'anesthésie.

Le diagnostic est confirmé si l'analyse montre que la moelle contient plus de 20 % de cellules immatures. L'analyse morphologique des cellules permet alors de définir la sous-catégorie de leucémie aiguë. D'autres examens biologiques permettent d'obtenir des données complémentaires afin de mieux caractériser la maladie. C'est le cas notamment de l'étude des chromosomes des cellules anormales qui permet d'affiner le diagnostic et le pronostic, afin de choisir le meilleur traitement pour le patient [9].

1.5.3. Après le diagnostic

Le moment du diagnostic est toujours un instant brutal et déstabilisant. Pour être acteur de sa prise en charge, il peut être utile de compiler les questions à poser à son médecin ou à l'équipe médicale : de quel type de leucémie suis-je atteint ? Quel est mon schéma de traitement et mon planning de soins ? Quelles sont les modifications éventuelles qui pourraient y être apportées, les complications possibles ? Si le traitement peut comporter une greffe de moelle osseuse, il peut être pertinent de commencer à en parler avec sa famille (père, mère, frère, sœur). Il leur sera demandé d'effectuer une prise de sang pour déterminer s'ils peuvent être donneurs.

Il peut être rassurant d'organiser à l'avance son séjour à l'hôpital, en questionnant les infirmières sur son déroulement, les visites possibles, les objets qu'il est possible d'apporter…

Le diagnostic de leucémie, tout particulièrement de leucémie aiguë, est souvent extrêmement rapide. Le patient se trouve brutalement confronté à une maladie grave : diagnostic, isolement dans une chambre stérile et rupture du contact physique avec la famille,

examens contraignants (myélogramme, pose du cathéter, ponctions lombaires…). Le parcours de soins peut créer une angoisse qui peut être particulièrement difficile à vivre.

Le soutien de son entourage est donc primordial. Parallèlement, un psycho-oncologue (psychiatre ou psychologue spécialisé en cancérologie) peut être sollicité pour écouter, voire suivre le patient

1.5.4. Pendant la maladie

La leucémie est liée à la multiplication de cellules anormales. Cette prolifération se fait au détriment des cellules sanguines normales et entraîne de ce fait divers symptômes : fatigue en cas d'anémie, risques de saignements en cas de thrombopénie et d'infections en cas de leucopénie.

Parallèlement, le traitement, notamment la chimiothérapie, peut engendrer des effets indésirables : chute des cheveux, nausées, vomissements, diarrhées, fièvre, mucite (inflammation de la muqueuse buccale)… Ces troubles ne sont pas systématiques et peuvent être pris en charge par des approches adaptées (casque réfrigérant contre la chute de cheveux, anti-nauséeux, antalgiques…).

Si le patient se trouve en aplasie médullaire, il est placé pendant quelques semaines dans une chambre isolée, avec des précautions anti-infectieuses : des antibiotiques permettent de lutter contre les infections bactériennes et de nouveaux médicaments sont utilisés pour prévenir les infections par des champignons (infections fongiques), en particulier au niveau des poumons. Des transfusions de globules rouges et de plaquettes sont parfois réalisées.

Enfin, les leucémies peuvent entraîner des douleurs au niveau de la rate, du foie ou des os. Les traitements de la maladie peuvent réduire ces manifestations. Mais si des douleurs persistent et ne sont pas tolérables, il ne faut pas hésiter à en parler avec l'équipe médicale. Il est possible de soulager ces douleurs par des traitements médicaux ou dans certains cas par une radiothérapie [10].

1.6. L'ASPECT BIOLOGIQUE DE LEUCEMIE

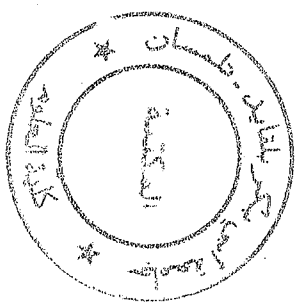
Bien que la classification du cancer se soit amélioré au cours des 30 dernières années, il n'y a pas eu d'approche générale pour l'identification de nouvelles classes de découverte cancer (classe) ou l'attribution de tumeurs à des classes connues (prédiction de classe). Ici, une approche générique de classification du cancer basé sur le suivi de l'expression des gènes par puces à ADN est décrite et appliquée aux leucémies aiguës de l'homme comme un cas de test. Une procédure de découverte de la classe découvert automatiquement la distinction entre la leucémie myéloïde aiguë (LMA) et la leucémie aiguë lymphoblastique(LAL), sans connaissance préalable de ces classes. Un prédicteur de la classe dérivée automatiquement été en mesure de déterminer la classe des nouveaux cas de leucémie. Les résultats démontrent la faisabilité de la classification du cancer basé uniquement sur la surveillance de l'expression des gènes et de proposer une stratégie générale pour la découverte et la prévision des classes de cancer pour les autres types de cancer [11].

1.7. CONCLUSION

La classification de leucémie est devenue l'objet qui attire l'attention de nombreux Chercheurs durant ces dernières années, Le problème spécifique de la classification de variables nécessite une approche particulière des méthodes d'ensembles c'est le RSM (Random Subspace method) qui est plus adéquate a les bases de données biologiques qui se caractérise par des centaines ou des milliers de variables.

ChapitreII :

L'état de l'art



CHAPITRE 2

Etat de l'art

2.1. INTRODUCTION

La recherche sur les leucémies s'oriente dans plusieurs grandes directions. Il s'agit aujourd'hui de mieux comprendre les anomalies génétiques impliquées dans les leucémies et d'apporter des réponses plus adaptées, plusieurs travaux sont dirigés vers la détection de leucémie [12]. Dans ce chapitre nous présentons plusieurs méthodes de la sélection des gènes pour la leucémie, ainsi que la méthode de sous espaces aléatoires tout en passant en revue l'état de l'art dans ce domaine ; et enfin nous exposons notre contribution dans la reconnaissance de cette maladie.

2.2. TRAVAUX REALISE SUR LA MALADIE DE LEUCEMIE

Dans cette partie nous citons des différents travaux réalisés dans la classification de la Leucémie par les techniques de sélection de variables :

Auteurs	Articles	Approches	Expériences	Résultats
YiZhang, ChrisDing, Tao Li, 2008. [ZDL08]	Gene selection algorithmby combining ReliefF andMRMR	Ce papier combine deux méthodes de sélection ReliefF etMRMR où la première consiste à trouver un ensemble de gèneset la seconde est appliquée explicitement pour réduire la redondance ; afin d'avoir un ensemble de gènes	Les bases sont : ALL (Aculte Lynphplastic Leukimia), ARR (Arrhythmia) , GCM, HBC, MLL	Les taux de classification sont évalués après la sélection de 30 gènes pour chaque base

		compacte et efficace. La classification a été réalisé avec SVM et Naivebayes.	(leukemia)	
Yuhang Wing,Fillia Makedon, 2004[WM04]	Applicationof ReliefFto selecting informative genesfor cancer classification usingmicroar ray data	Ce papier implémentela méthode de sélection ReliefF pour sélectionner les gènesles plus pertinents des différentes bases de données avec les classifieurs SVM et K-NN.	Les bases de données sont: ALL leukemia, MLL leukemia	Apres la sélection de150 gènes pour chaque base, les taux de classification sont: SVM : -ALL : 99% -MLL :97% K-NN: -ALL:100% -MLL: 98%
G. Baskar, P.Ponmuthur amalingam 2012 [BP12]	Analysis of Gene Expression Microarray Dataset for Feature Selection	Dans cet article les auteurs proposent 3 méthodes : SVM-RFE ,WeightingK-Means et Sw SVM-RFE.Le processus principal de Svm-RFEélimine périodiquement les variables de bas poids, en utilisant SVM pour déterminer les poids. À partir de l'ensemble complet des variables, à chaque itération	Les bases de donnée est: Lekemia, colon	La meillieur resultat obtenu est avec la methode Sw SVM-RFE

		l'algorithme forme un classifieur de SVM basé surl'ensemble restant des variables élimine un ou plusieurs variables avec le plus bas poids. Ce processus récursif d'élimination des variables s'arrête jusqu'à ce quetoutes les variables sont enlevés ou un nombre désiré de des variables atteintes.		
M.Rosskopf, U.Feldkamp, W.Banzhaf [RPB]	Classification of Leukemia Classes by GP-based DNA-chip Analysis	Dans ce papier il est tester trois methodes PG-based, The mean difference method , Maximum signal-to-noise statistics pour l'obtention de la meilleure procedure qui classifier 100%	La base de Données est: leukemia ALL / MLL	Apres la sélection de 63 gènes pour la base ALL/AML, les taux de classification sont: PG : 91,18% The mean difference method : 82,35%Maximum signal-to-noise statistics: 82,35%

A. Sharma, C.H. Koh, S. Imoto and S. Miyano [SKIM]	Strategy of finding optimal number of features on gene expression data	Dans cet article on implémente les méthodes : SVM, SVM random, InfoGain,naiveBayes,Onedimensional SVM,SVM exhaustive,nearest centroid classifier,KNN ; pour obtenir le meillieur taux de classification	Appliqué sur la base : MLL leucémie	SVM+SVM random: 150 gènes 100%. -InfoGain+ naive Bayes 150 gènes 54%. -One dimensional SVM+SVM Random 150 gènes 100%. -One dimensional SVM+SVM Exhaustive 150 gènes 100%. -InfoGain+ nearest centroid classifier 46 gènes 93.3%. -InfoGain+ nearest neighbour classifier 46 gènes 86.7%. - SVM+nearest

				centroid classifier 37 gènes 93.3% - SVM+Knn 37 gènes 100%

TABLE 2.1.Quelques travaux sur la leucémie

2.3. Méthodes d'ensemble

Le principe général des méthodes d'ensemble est de construire une collection de prédicteurs, pour ensuite agréger l'ensemble de leurs prédictions pour avoir une prédiction finale.

Dans un cadre de la classification, l'agrégation revient par exemple à faire un vote majoritaire parmi les classes fournies par les prédicteurs on choisit un prédicteur final qui soit meilleur que chacun des prédicteurs individuels.

Le succès de ces méthodes d'ensemble se résume ainsi :

- Chaque prédicteur individuel doit être relativement bon : agréger des prédicteurs tous mauvais ne pourra vraisemblablement pas donner un bon prédicteur

- Les prédicteurs individuels doivent être différents les uns des autres : car agréger des prédicteurs tous quasiment pareils donnera encore un prédicteur semblable et n'améliorera pas les prédictions [R12].

Plusieurs techniques peuvent être appliquées, parmi les plus utilisées, on présente les techniques: bagging, boosting, les méthodes de sous espaces aléatoires...

2.3.1.Bagging(BoostrapAGGregatING)

Pour construire chacun des classificateurs de base, la diversité des ensembles d'apprentissage est obtenue par un choix aléatoire par remplacement, à partir d'un ensemble d'apprentissage d'origine. L'échantillon choisi par une technique statistique appelée «bootstrap» est de même taille que l'ensemble d'origine. Avec la réplication de certains exemples, l'échantillon contient en moyenne 2/3 de l'ensemble de données de départ. Les prédictions des modèles de base sont ensuite combinées par la méthode de vote à la majorité [DH] .

2.3.2.Boosting

L'idée de base est de construire un nouveau classificateur, selon la performance d'une série de classificateurs précédents, dans un processus séquentiel. L'ensemble d'apprentissage d'origine est renforcé par des poids qui seront ajustés à chaque étape, dans l'objectif d'amplifier (boost) les exemples mal classés. Les poids des exemples bien classés par le dernier modèle construit seront alors décrémentés, et les poids des exemples mal classés seront incrémentés, en permettant ainsi au système de prêter plus d'attention aux exemples mal classés. Les modèles sont combinés par vote à majorité pondéré où la pondération est déterminée par la précision de prédiction de chaque classifieur.

2.2.3. Méthode de sous espace aléatoire (Random SubspaceMethod)

Cette technique consiste à choisir aléatoirement un certain nombre d'attributs par des méthodes spécifiques, les sous ensembles d'apprentissage obtenus seront alors utilisés pour construire les classifieurs de base. C'est une approche qui est très bénéfique pour les problèmes ayant un grand nombre d'attributs, avec de multiples redondances. [ZLC08]

Travaux réalisés avec la méthode de sous espaces aléatoires

Dans ce tableau nous présentons l'état de l'art des travaux concernant la résolution du problème de grande dimension, avec les méthodes de sous espaces aléatoireset leurs applications dans les différents domaines

Auteurs	Articles	Approches	Expériences	Résultats
Tin Kam Ho 1998 [H98]	The Random Subspace Method for Constructing Decision Forests	Dans cet article il est proposé d'utiliser 3 méthodes : bootstrapping, boosting, random subspace.	A chaque fois l'exactitude le taux de calssification est mesuré	l'expérience donne les exactitudes suivantes : -random subspace : [95%,99%] -Boosting : [92%,99%] -Bootstrapping : [90%,99%] -C.4.5 : [85%,99%]
Carmen Lai, Marcel J.T. Reinders, LodewykWes sels [LJW]	Random Subspace Method for multivariate feature Selection	Ce papier combine deux Algorithmes de RSM avec deux différentes méthodesde	Les variables réduites par les différentes approches sont	Les résultats de RSM s'améliorent avec l'utilisation de Liknon et RFE

		linéaires(FLD,FLD, NMC, NNC ...)		espace aléatoire RSM peut être utile dans LDA

TABLE 2.2.Quelques travaux sur RSM.

2.4. CONTRIBUTION

La réduction de la dimension de grande base de données (Big data) est un problème complexe qui a été largement étudié dans plusieurs domaines.

Notre contribution porte sur la sélection de variables, La sélection de variables est un processus très important en apprentissage supervisé. Nous disposons d'une série de variables candidates, nous cherchons les variables les plus pertinentes pour expliquer et prédire les valeurs prises par la variable à prédire. Les objectifs sont bien souvent multiples : nous réduisons le nombre de variables à recueillir pour le déploiement du système ; nous améliorons notre connaissance du phénomène de causalité entre les descripteurs et la variable à prédire, ce qui est fondamental si nous voulons interpréter les résultats pour en assurer la reproductibilité ; enfin, mais pas toujours, nous améliorons la qualité de la prédiction, le ratio nombre d'observations et dimension de représentation étant plus favorable.

La meilleure approche pour sélectionner les variables pertinentes est certainement la sélection experte. Seule la connaissance du domaine permet de bien comprendre les causalités sous-jacentes, discerner les vrais liens des simples artefacts, mettre en évidence les interactions, etc.

Malheureusement, elle n'est pas toujours possible, notamment parce que le nombre de variables candidates est élevé, une sélection manuelle devient vite inextricable. De toute

		sélection : RSM-RFE et RSM-Liknon	appliquées avec le classifieur pour évaluer les performances de chaque méthode.	
PancePanov and Saso Dzeroski [OD10]	Combining Bagging and Random Subspaces to Create Better Ensembles	Ce papier se focalise sur la comparaison entre les quatre algorithmes suivant : bagging,RSM, RF et SubBag performant.	Les expériences ont été testé sur 19 base, achaque fois ils sont utilisé les principes suivants J48 ,JRip, IBK .	le meilleure taux de classification obtenus pour les trois expérience est celles de la méthode SubBag .
Marina Skurichina and Robert P. W. Duin [SP02]	Bagging, Boosting and the Random Subspace Method for LinearClassifiers	Ce papier propose l'application des méthodes suivantes : Bagging, Boosting et La méthode de sous-espace aléatoire (RSM)avec des classifieurs	Les expériences de cet article ont été réalisées avec des classifieurs linéaires	-L'instabilité des classifieurs linéaires dépend de la taille de l'échantillon de formation et de leur complexité -Bagging, Boosting et la méthode de sous-

manière, dans une démarche exploratoire, il paraîtrait bien étrange finalement que tout soit connu à l'avance, on se demande alors à quoi servirait la fouille de données dans ce contexte.

La littérature concernant la sélection de variables étant très vaste, nous nous intéressons dans cette partie uniquement aux méthodes de sélection des gènes du cancer du sang(la leucémie), laissant de côté d'autres méthodes couramment utilisées pour réduire la dimension telles que l'Analyse en Composantes Principales.

Pour la validation des gènes sélectionnés dans notre base de cancer du sang, nous testons leurs capacités et leurs taux de classification avec le classfieur RSM qui est caractérisé par son principe de double randomisation et sans remise.

2.5. CONCLUSION

Ce chapitre nous a amené a proposé pour la résolution de notre problématique une approche basé sur un L'analyse de données acquises du domaine médical ou biologique qui a pour but d'extraire de la connaissance, ou bien de créer des modèles permettant de structurer les informations qu'elles contiennent. Dans la littérature un grand nombre d'algorithmes de sélection de variables sont disponibles, mais rares sont les méthodes capables de relever le défi sur lequel nous nous focalisons.

CharpitreIII :

Résultats

et

discussion

CHAPITRE 3

Résultat et discussion

3.1. INTRODUCTION

Les méthodes de prétraitement des données offrent aujourd'hui une technologie mature pour résoudre les problèmes ou le premier enjeu consiste à dépasser le cadre actuel de l'apprentissage pour s'attaquer à cette nouvelle gamme de problème de la réduction de dimension, Les méthodes de sélection de variables peuvent rendre quelques systèmes plus robustes. Cette robustesse est exprimée par la capacité de la méthode de prétraitement à produire des descripteurs fiables permettant une meilleure classification à chaque fois que les données sont perturbées.

Dans ce chapitre nous élaborons notre contribution représentée dans la sélection de Variables par les forets aléatoires, Pour cela nous divisons ce chapitre en2 parties où chacune d'elles traite des résultats d'une méthode appliquée. Les méthodes de classification de la leucémie sont représentées dans le schéma suivant :

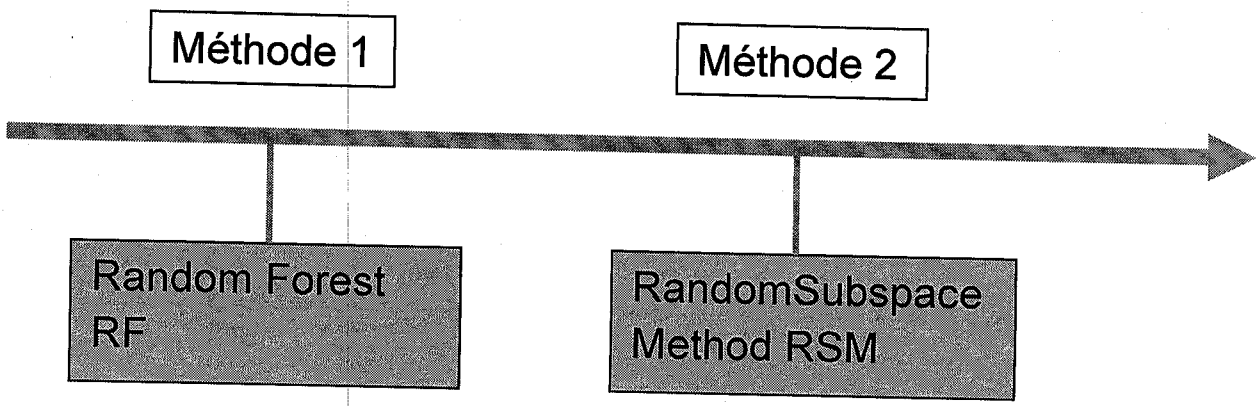


FIGURE 3.1.Schéma représente les méthodes de classification des malades.

Nous passerons par la suite aux expérimentations ainsi que les résultats obtenus avec leurs interprétations. Enfin nous finissons par une comparaison avec les travaux de l'état de l'art, Pour l'implémentation de ces résultats nous avons fait ce travaille aux logiciels : Matlab 7.11.0 (R2010b)

3.2. BASE DE DONNEE

Dans ce mémoire de Master nous utilisons la base de données biologique. Cette base de données réalisée par Golub et al. (1999) concerne le cancer du sang(leucémie) [WM04]. Elle est constitué de 38 patients dont 25 sont des tissus tumoraux et 13 des tissus normaux

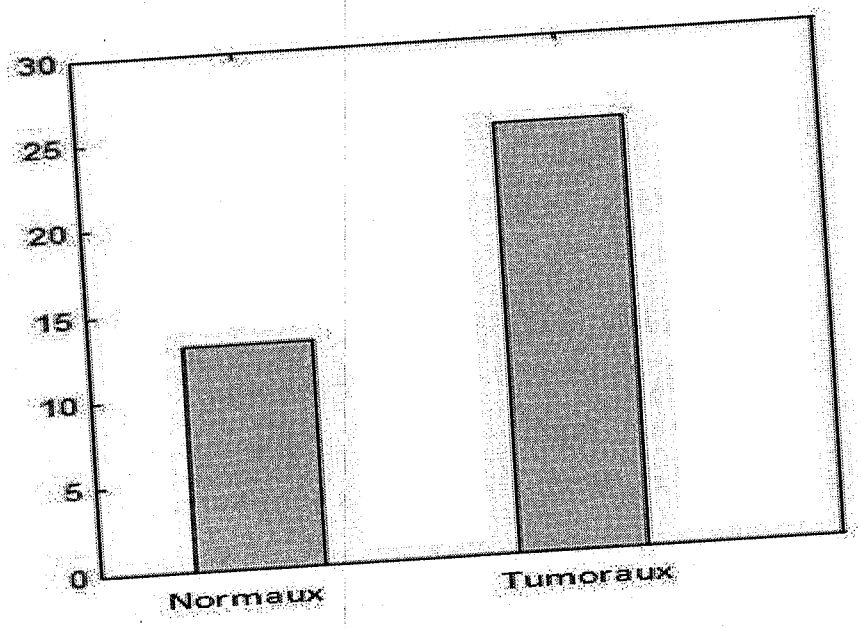


FIGURE 3.2.La base de données de Leucémie.

Le diagnostic est une valeur binaire variable «classe» qui permet de savoir si le patient montre des signes de leucémie selon les critères de l'Organisation Mondiale de la Santé.

D'après le FIGURE 3.1 nous remarquons que le taux de classification n'est pas stable par le changement de nombres des arbres à chaque fois il prendre une valeur différente que la valeur précédente et il est stable dans l'intervalle [300 ;400] par le taux égale à 61,9% finalement on conclu que le meilleur nombre d'arbre égale à 40 pour le quel le taux de classification égale à 79,17%.

Cette méthode donne de bons résultats mais leurs inconvénients c'est la redondance des informations Au niveau de la construction des arbres par le tirage avec remise d'après les baggings pour cela nous proposons une autre méthode plus efficace pour les bases de grande dimension

3.4.METHODE 2: LA METHODE DE SOUS_ESPACE ALEATOIRE

RSM utilise le même principe que bootstrap la seule différence est le tirage sans remise pour la construction des arbres. Cette méthode fonctionne efficacement sur les bases de données de grande dimensionnalité.

Résultats et discussion

Après l'application de la méthode de sous-espaces aléatoire sur notre base de donnée (leucémie) on a obtenu les résultats suivant :

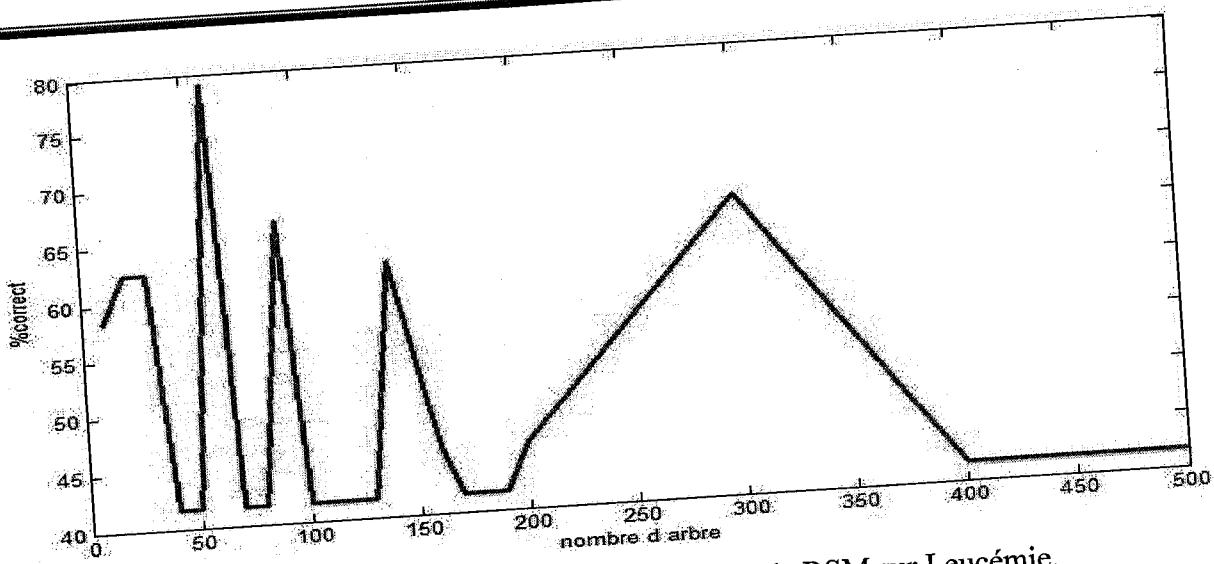


FIGURE 3.4.Graphe représente les résultats de RSM sur Leucémie.

D'après le FIGURE 3.2nous conclu que le meilleur nombre d'arbre égale à 60 pour le quel le taux de classification atteint les 79,17%

3.5. COMPARAISON ENTRE RF ET RSM

L'importance et la difficulté de retrouver le meilleur taux de classification nécessite l'application des différentes méthodes et plusieurs testes

Afin de situer la performance de l'approche proposée nous avons réalisé une étude comparative entre les résultats obtenus par RSM et celles de RF classique avec la base Leukemia. La FIGURE 3.3résume la comparaison

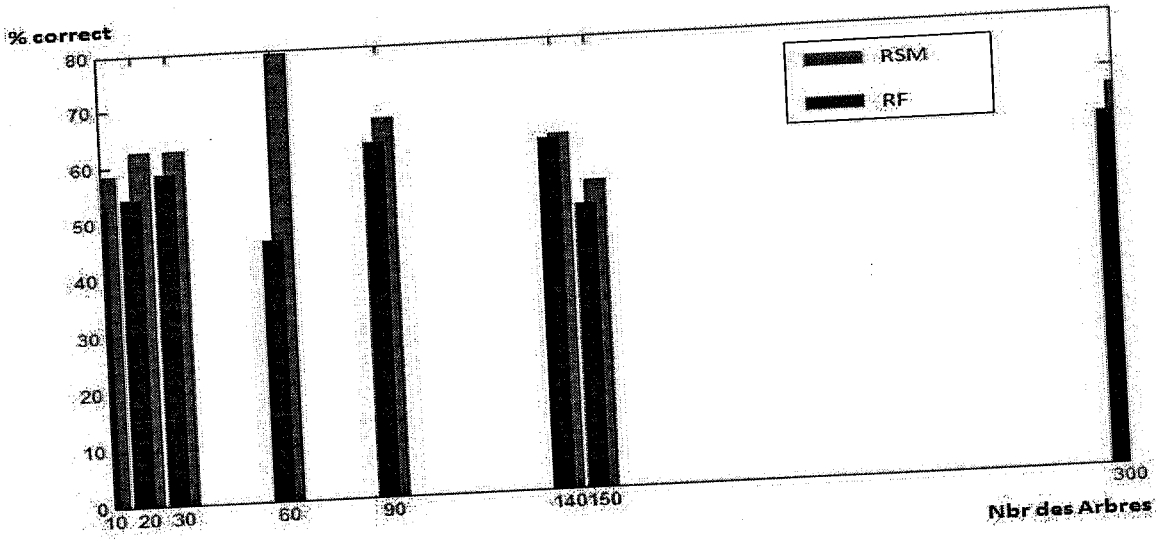


FIGURE 3.5.Histogramme représente la comparaison entre RSM et RF pour Leucémie.

Discussion

D'après le FIGURE 3.3 on peut dire que RSM est meilleur que RF pour un certain nombre d'arbres : 10, 20,30, 60,90, 140,150 et 300. Le résultat maximal de RSM est à 79,17%, obtenu pour le nombre d'arbre égal à 60.

Les performances des méthodes RF et RSM implémenté ont été évaluées par le calcul du pourcentage de sensibilité (SE), la spécificité (SP) et taux de classification (TC), les définitions de ces derniers sont respectivement comme suit :

– Sensibilité (Se%) : [Se = 100 * TP / (VP + FN)] on appelle sensibilité (Se) du test sa capacité de donner un résultat positif quand la maladie est présente. Représente ceux qui sont correctement détectés parmi tous les événements réels.

– Spécificité (Sp %) : [Sp = 100 TN * / (VN + FP)] on appelle spécificité du test cette capacité de donner un résultat négatif quand la maladie est absente. Elle est représentée pour détecter les patients non diabétiques.

– Taux de classification (TC %) : [CC = 100 * (TP + TN) / (TN + TP + FN +FP)] est le taux de reconnaissance.

– VP : malade classé malade ;

– FP : non malade classé malade ;

– VN : non malade classé non malade ;

– FN : malade classé non malade[S11].

Les approches	Tc%	Se%	Sp%
RF	58.33%	1%	28.57%
RF-RSM	79.17%	90%	71.43%

TABLE 3.1 – Performances des classifieurs RF et RSM appliqués sur la Leucémie pour nombre des arbres=60

Pour assurer la performance de ces méthodes on propose de tester avec d'autres bases colon et dataset_c .Le tableau suivant comporte les caractéristiques de ces bases

Les bases	# gènes	# d'exemples	# classes	Référence
colon	2000	62	2	Alon et al(1999)
Dataset_c	7129	60	2	Pomeroy et al (2002)

TABLE 3.2- Description des bases biologies.

En comparant les résultats des deux expérimentations avec la base de colon qui constitué de 62 patients dont 21 sont des tissus tumoraux et 41 des tissus normaux

Nous remarquons que la méthode proposée à améliorer le taux de classification pour la forêt aléatoireest représentée dans l'histogramme ci-dessous :

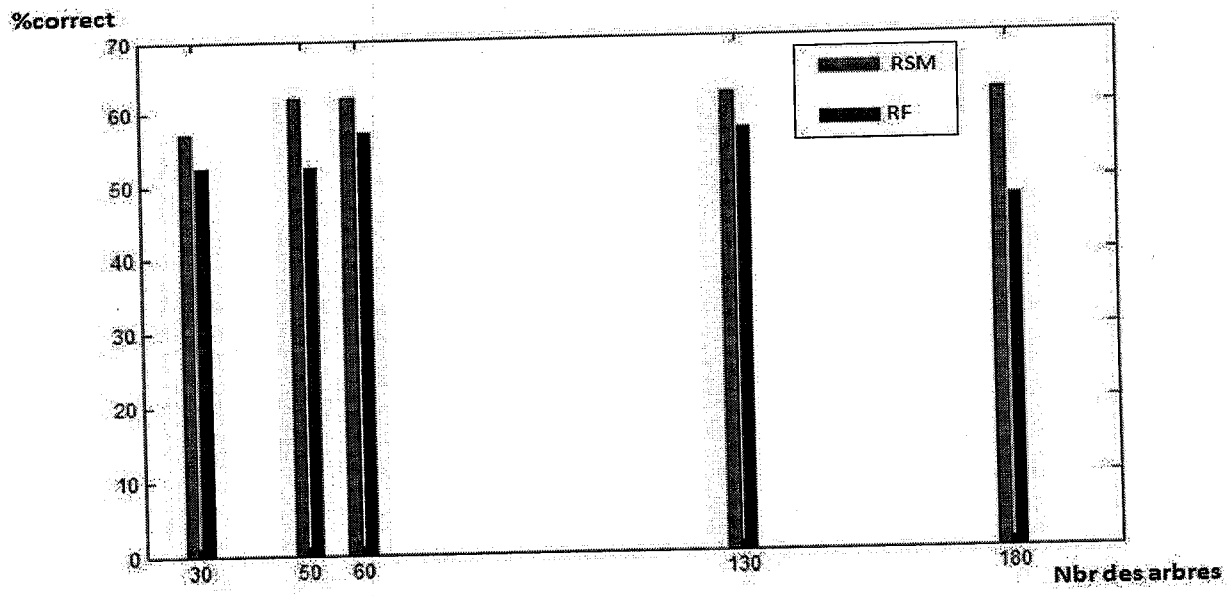


FIGURE 3.6.Histogramme représente la comparaison entre RSM et RF pour colon.

Les performances des méthodes RF et RSM appliqué sur la base du Colon avec un nombre d'arbre égal à 50 ont été évaluées par le calcul du pourcentage du taux de classification(TC), sensibilité (SE) et la spécificité (SP) leurs valeurs respectives sont les suivantes :

Les approches	Tc%	Se%	Sp%
RF	52.38%	84.62%	0%
RF-RSM	61.9%	1%	0%

TABLE 3.2 – Performances des classifieurs RF et RSM appliqués sur colon pour nombre des arbres=50

Un autre test est effectué sur une nouvelle base de donnée « Dataset C » qui constitué de 60 patients dont 20 sont des tissus tumoraux et 40 des tissus normaux
On remarque dans l'histogramme suivant que le taux de classification augmente pour certain nombre d'arbres présenté dans la figure suivante :

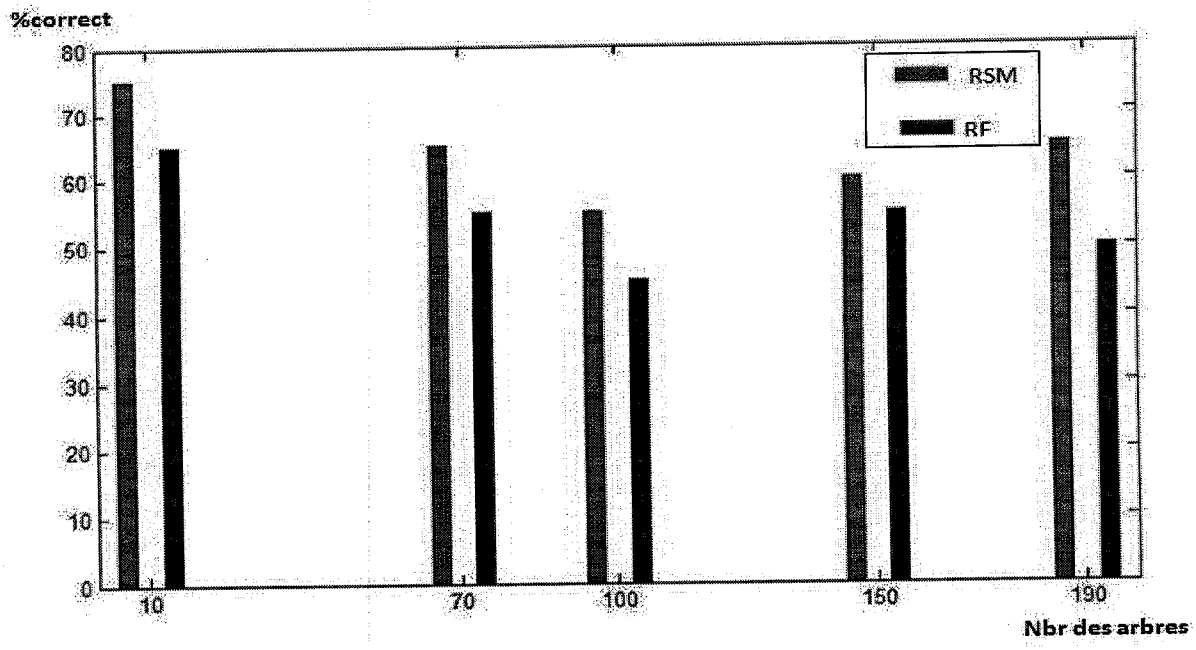


FIGURE 3.7.Histogramme représente la comparaison entre RSM et RF pour Dataset C.

L'évaluation des performances de cette base par les deux méthodes pour un nombre d'arbre égal à 70 est estimée par le taux de classification, la sensibilité et la spécificité.

Les résultats obtenus sont résumés dansle tableau 3.3.

Les approches	Tc%	Se%	Sp%
RF	55%	60%	40%
RF-RSM	65%	80%	2%

TABLE 3.3 – Performances des classifieurs RF et RSM appliqués sur Dataset_C pour nombre des arbres=70

D'après les tests appliqués sur les trois bases de données confirme que la méthode proposée RF-RSM est plus efficaces que la forêt aléatoire classique et donne une amélioration des taux de classification.

3.6. COMPARAISON AVEC DES TRAVAUX DE LITTERATURE

Afin de situer la performance de la méthode proposée nous avons réalisé une étude comparative avec les résultats obtenus avec la base de leucémie et celles des travaux déjà réalisés dans ce domaine (étudiés dans l'état de l'art) avec d'autres bases.

Nous constatons dans l'état de l'art que la méthode RSM est appliquée sur des base médicales et n'est pas testé sur les base biologique de grand dimensionnalité et pour cela on a proposé d'appliquer RSM.

Après plusieurs expérimentations sur cette problématique nous pouvons confirmer que la méthode proposée est une méthode efficace pour les grandes bases biologiques.

3.7. CONCLUSION

Dans ce chapitre, nous avons implémenté deux méthodes pour l'aide au diagnostic du la base de données sur la Leucémie.

La première est une application de la méthode RF classique. La deuxième méthode RSM est une modification de RF au niveau de la sélection des attributs pour la construction des arbres.

Après la comparaison des résultats obtenus par les deux méthodes, nous avons remarqué que les résultats trouvés par RSM sont comparables et même meilleurs par rapport aux résultats des forets aléatoires classiques.

Notre méthode proposée, nous a permis non seulement d'améliorer le taux de classification avec certain nombre des arbres pour les bases des données à grandes dimension.

Conclusion

Conclusion

Afin de créer une application performante utilisée pour la reconnaissance du leucémie, nous avons implémenté une méthode d'ensemble, qui a comme but d'aide au diagnostique pour la reconnaissance des maladies d'une part et d'autre part d'éliminer la redondance des informations pour les bases de grand dimensions. Cette méthode a sélectionné les variables qui effectués une meilleure classification. Pour la validation de ces résultats nous avons fait des testes sur d'autre base biologie

Notre démarche de sélection des gènes de leucémie consiste dans un premier temps de comparer l'efficacité de deux méthodes de sélection RF et RSM. Les expérimentations réalisées ont permis d'évaluer les performances des résultats avec les différentes bases.

Le taux de classification obtenu avec notre méthode est donne des bonnes résultats pour la classification du leucémie, ce taux est aussi donne des bonnes résultats par rapports aux autres base de donnée. on conclu que les résultats obtenus soient intéressants et encourageants

Bibliographie

[ZDL08] Yi Zhang, Chris Ding, and Tao Li. Gene selection algorithm by combining relieff and mrmr. BMC Bioinformatics, 9 :S27, 2008.

[WM04] Yuhang Wing and Fillia Makedon. Application of reliefF feature filtering algorithm to selecting informative genes for cancer classification using microarray data, pages 497_498, 2004.

[SP02] Marina Skurichina and Robert P. W. Duin.Bagging, Boosting and the Random Subspace Method for Linear Classifiers, CTARP , pages 708-715, 2004.

[H98] Tin Kam Ho. The Random Subspace Method for Constructing Decision,computer vision and image understanding, pages 101-715, 1998.

[LJW] Carmen Lai, Marcel J.T. Reinders, LodewykWessels . Random Subspace Method for multivariate feature Selection,TCPR, pages 2977-2980,2012.

[OD10] PancePanov and Saso Dzeroski. Combining Bagging and Random Subspaces to CreateBetter Ensembles, BMCBioinformatics, pages 2,2010.

[BP12] G. Baskar,P.Ponmuthuramalingam, Analysis of Gene Expression Microarray Dataset for Feature Selection, IJDMB, pages 333-345,2009.

[RPB] M.Rosskopf,U.Feldkamp,W.Banzhaf, Classification of Leukemia Classes by GP- based DNA-chip Analysis, CTARR,2012.

[SKIM] A. Sharma, C.H. Koh, S. Imoto and S. Miyano, Strategy of finding optimal number offeatures on gene expression data,CTARP,2009.

[DH] Mokeddem Djamila, Belbachir Hafida, Utilisation des méthodes d'apprentissage Ensembliste dans le Datamining distribué ,2010.

[ZLC08] Yulian Zhu, Jun Liu, Songcan Chen, Semi-random subspace method for face recognition, 2008.

[SH05] Tao Shi and Steve Horvath, Unsupervised Learning with Random Forest Predictors.2005

Mémoire

[S11] Nesma Settouti, Amine Chikh. Renforcement de l'Apprentissage Structurel pour la Reconnaissance du Diabète.2011 Ce rapport de recherche fait parti du projet de MAGISTER EN ÉLECTRONIQUE BIOMÉDICALE.

[R12] Robin Genuer, Forêts aléatoires : aspects théoriques, sélection de variables et applications.2012 Ce rapport de recherche fait parti du projet de DOCTORAT EN SCIENCES.

Site web

[1] AsuraGen, leukemia, http://www.asuragen.com/ Diagnostics/US/ Educational_pages/leukemia.aspx, 2012.

[2] http://www.ilo.org/safework_bookshelf/french?content&nd=857170018.

[3] Le point,Leukemia ,http://hopitaux.lepoint.fr/tout-savoir-95/leucemie.php.25 juin 2013.

[4] La tribune, Leukemia, http://www.latribuneonline.com/ suplements/sante/70039.rss, 25 juin 2013.

[5] Planetoscope, Leukemia, http://www.planetoscope.com/Maladie/584-nombre-De personnes-touchees-par-la-leucemie-en-france.html.2012

[6] Cancer.ca,Leukemia,http://www.cancer.ca/fr-ca/cancer-information/cancer