République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique
Université Mohamed el-Bachir el-Ibrahimi Bordj Bou Arréridj
Faculté des Mathématiques et Informatique

MEMOIRE

Présente en vue de l'obtention du diplôme

Master en informatique

Spécialité : Technologie de l'information et de la communication

THEME :
La Sélection Des Gènes Et La Classification Des Données Bio-puces

Présenté Par :

- ZERROUGUI SALMA
- MOHAMADI HADJIRA

Soutenue le : 2021

Devant le jury composé de :

Président Mr :

Examinateur Mr :

Examinateur Mr :

Encadreur Mr : Zouache Djaâfar

Année Universitaire 2020-2021

Remerciements

Tout d'abord, nous remercions ALLAH tout-puissant qui nous a apporté la volonté et la force de faire ce travail.

Nous remercions M. Zouache Djâafar, notre directeur et M. Allo Lotfi, de nous avoir accueillis pour la réalisation de ce mémoire.

Nous nous reconnaissons leurs soutiens et leurs conseils au cours de la période de recherche.

Nous remercions sincèrement tous ceux qui nous ont apporté leurs avis éclairés.

Merci à notre collège, le Collège de Mathématiques et Informatique, et à tous les professeurs et les cadres administratifs.

Nous remercions tous les étudiants avec lesquels nous avons étudié et passé des années de dur labeur, dans de bonnes et de mauvaises conditions. Merci pour les moments heureux. Nous n'oublierons jamais ces beaux souvenirs.

Dédicaces

Je dédie ce modeste travail aux défunts de mes chers parents ;

A mon mari et mes enfants ;

A mes sœurs ;

A tous mes enseignants et tous mes collègues et à toute personne qui me connait.

MOHAMADI HADJIRA

Dédicaces

Je dédie ce travail :

A mon père Zerrougui Djamal,

Pour son grand amour, ses encouragements, son sens du devoir et ses sacrifices

pour que je puisse réussir mes études.

À ma chère mère Belagoun Ghania,

Pour sa grandeur pour son amour, sa patience et ses encouragements face à l'adversité, ainsi que ses prières qui m'apportent le bonheur et la réussite.

A mes frères,

Mohamed, Amal, Houda, qui sont toujours à mes côtés, prêts à m'aider.

A ma deuxième famille,

Mon mari Benarroudj Walid et ses parents (Salem et Karima)

Un merci spécial aux parents du fond du cœur pour leur soutien moral et matériel.

Ma belle famille qui a toujours pris soin de ce que nous faisons. Que Je trouve dans cette œuvre le témoignage de notre amour.

A mes chers amis

Pour tous ceux qui m'aiment.

A tous ceux que je connais de près ou de loin.

ZERROUGUI SALMA

LISTE DES FIGURES

LISTE DES TABLEAUX

SOMMAIRE

Introduction générale

La grande taille de données présente un défi réel pour la maitrise et l'exploitation des informations en termes de précision et de vitesse.

La sélection d'attributs est un domaine très actif dans plusieurs domaines comme la bio-informatique et la fouille de données, à cause de sa capacité de minimiser la taille des bases de données en choisissant des sous-ensembles avec un minimum d'attributs qui assurent une bonne performance et une exactitude élevée.

Dans notre projet de fin d'étude, nous donnons un petit détail sur les méthodes de classification supervisées utilisées dans la fouille de donnés et les méthodes de sélections d'attributs.

Nous examinons quelques algorithmes des différents types des méthodes de sélections d'attributs avec une étude comparative entre les résultats obtenus sur trois base de données extraites du site web UCI, afin d'évaluer les sous-ensembles d'attributs sélectionnés.

Organisation du mémoire

Ce mémoire est composé de trois chapitres dont nous allons donner une brève description dans les paragraphes suivants :

– **Le Chapitre 1 :** Nous abordons une petite introduction sur le domaine de la bio-informatique, puis les concepts et les principes de base de la technologie de puces à ADN. En particulier, on traite les différentes étapes de l'analyse de puces ADN.

– **Le Chapitre 2 :** Dans ce chapitre, nous parlons de la classification supervisée et ces méthodes les plus utilisées. Nous discutons aussi les concepts de base du problème de sélection d'attributs et les principales approches de sélection avec une brève description des méthodes Wrapper, Filter et la méthode hybride.

– **Le Chapitre3 :** Dans cette section nous faisons une brève comparaison entre les trois méthodes de sélection comme nous citons les différents algorithmes pour chaque méthode. On a donné aussi une petite vue sur l'environnement de travail Weka, ainsi qu'une explication de jeux des données utilisées dans ce mémoire est détaillée.

Nous détaillons notre travail concernant les données et les algorithmes avec des tests pour choisir la meilleure méthode et la plus performante.

Enfin une conclusion générale et quelques perspectives pour des futures travaux.

CHAPITRE I :

Introduction à la bio-informatique

I .1. Introduction

Le domaine de la bio-informatique est un domaine récent, il a été émergé pour comprendre les phénomènes biologiques, basés sur l'ADN. Ces phénomènes sont complexes et d'une variété énorme ont besoin des interactions entre plusieurs disciplines pour arriver à des études et des résolutions efficaces et raisonnables dans un temps critique.

Nous avons donc décidé dans ce chapitre d'étudier les informations biologiques et d'apprendre le mécanisme d'extraction des séquences d'ADN**.**

I .2. La bio-informatique

I .2.1 Définition de la bio-informatique la bioinformatique est l'utilisation des technologies de l'information en biologie pour stocker des données et analyser des séquences d'ADN, pour comprendre les phénomènes biologiques, les problèmes liés à la gestion de l'anxiété par le séquençage du génome, la modélisation de la structure des protéines, ou la reconstruction d'arbres phylogénétiques. Ces problèmes nécessitent une coopération entre biologistes et informaticiens, car les problèmes à résoudre impliquent souvent des difficultés de calcul importantes.

La bio-informatique est une science interdisciplinaire qui comprend la biologie, l'informatique, les mathématiques et les statistiques dans le but d'analyser les séquences biologiques et de prédire la structure et la fonction des macromolécules. [1.1]

I .2.2 Les tâches de la bio-informatique

Il existe plusieurs activités dont quatre sont principales :

- Acquérir et organiser des données biologiques.
- Faire une conception de logiciel d'analyse, de comparaison et de modélisation des données.
- Analyser, comparer et modéliser les données en réalisant un logiciel approprié.
- Analyser les résultats produits en utilisant le logiciel précédent.
- Prédire la structure des macromolécules ainsi que leur fonction en analysant les séquences biologiques.

I .3. Données manipulées

I .3.1. Biologie moléculaire

La biologie moléculaire est une discipline scientifique au croisement de la génétique, de la biochimie et de la physique, dont l'objet est la compréhension des mécanismes du fonctionnement de la cellule au niveau moléculaire. Le terme ≪ biologie moléculaire ≫est utilisé la première fois en 1938 par Warren Weaver, désigne également l'ensemble des techniques de manipulation d'acides nucléiques (ADN, ARN), appelées aussi techniques de génie génétique.

La biologie moléculaire est apparue au XXe siècle, à la suite de l'élaboration des lois de la génétique, la découverte des chromosomes et l'identification de l'ADN comme support chimique de l'information génétique.

Après la découverte de la structure en double hélice de l'ADN en1953 par James Watson (1928agé de 93ans), Francis Crick (1916-2004), Maurice Wilkins (1916-2004) et Rosalinda Franklin (1920-1958) la biologie moléculaire a connu d'importants développements pour devenir un outil incontournable de la biologie moderne à partir des années 1970.

- **La cellule :**

Une cellule est l'unité de base de tout organisme (excepté les virus). Sa membrane permet de créer une entité, séparée du milieu extérieur dans le cas des organismes unicellulaires, ou des autres cellules dans le cas des organismes pluricellulaires.

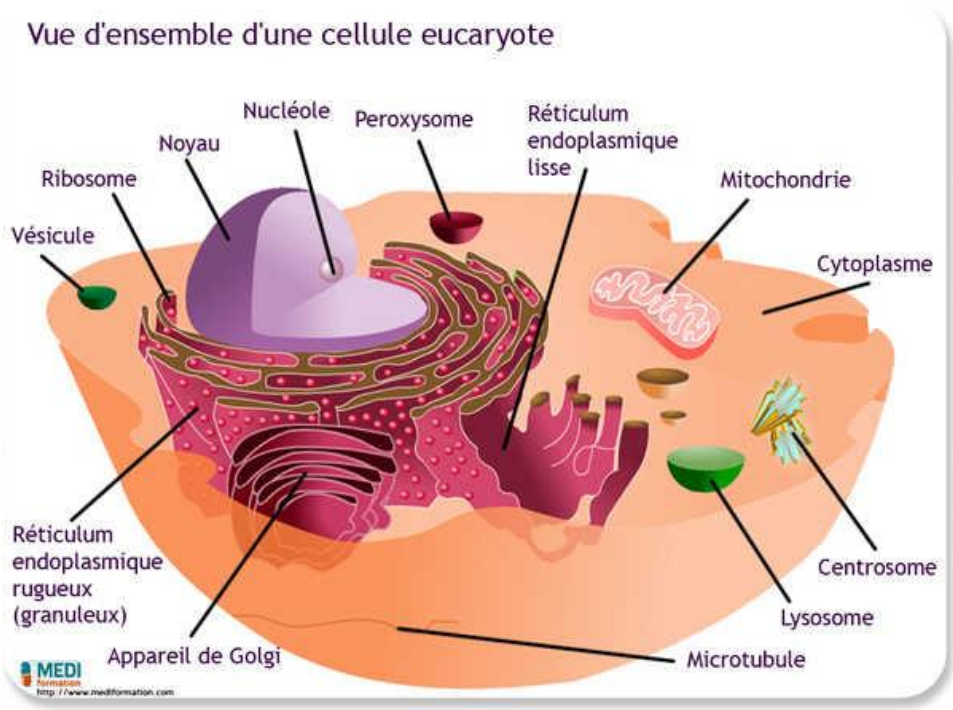


Figure I .1 : La cellule

- **Les chromosomes**

Dans les cellules eucaryotes, les chromosomes se trouvent dans le noyau ; leur nombre varie en fonction des espèces. En dehors des moments où la cellule se divise, les chromosomes ne peuvent pas être visualisés individuellement. Lorsque la division cellulaire se prépare, ils se condensent et deviennent distincts. Au moment de la division cellulaire (mitose), les chromosomes, formés de deux chromatides identiques reliées au niveau du centromère, se

coupe en deux, chaque chromatide part dans une cellule-fille. Une phase de synthèse d'ADN permettra ensuite de doubler l'ADN de la cellule-fille.
Une cellule humaine contient 23 paires de chromosomes (2n = 46 chromosomes), dont 22 sont communes aux deux sexes : les paires d'autosomes. Les deux chromosomes restants sont les chromosomes sexuels. Chez la femme, ce sont deux chromosomes X. Chez l'homme, ils sont différents, l'un est un chromosome X et l'autre est appelé chromosome Y. Dans les gamètes comme le spermatozoïde, seuls 23 chromosomes sont présents.
Certaines maladies génétiques sont liées à des aberrations chromosomiques. Par exemple, la trisomie 21 est due à la présence d'un chromosome 21 surnuméraire. De même, le syndrome de Klinefelter touche des garçons qui possèdent trois chromosomes sexuels : XXY. Les femmes atteintes du syndrome de Turner n'ont qu'un chromosome X.

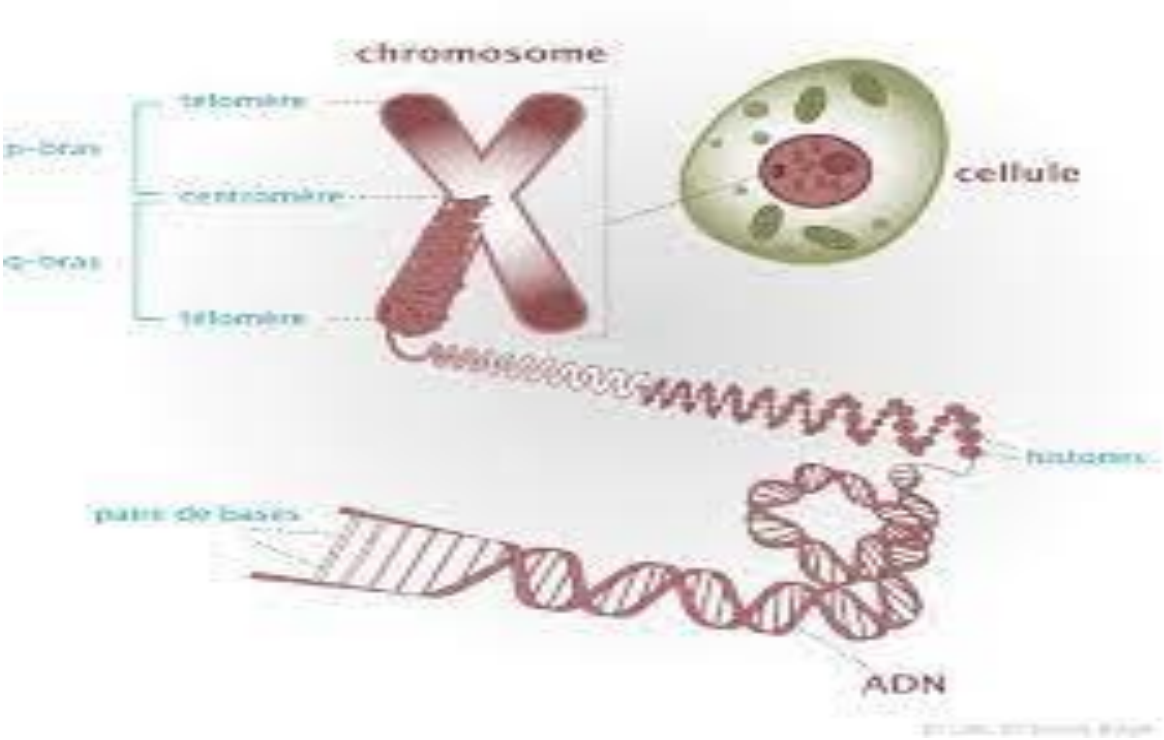


Figure I .2 : Les chromosomes

I .3.2 Définition de l'ADN

L'acide désoxyribonucléique ou acide désoxyribonucléique (ADN) est une énorme molécule qui réside à l'intérieur des cellules de tous les êtres vivants et de nombreux virus et contient l'information génétique qui permet le travail, la reproduction et le développement de ces organismes, l'ADN et l'ARN proviennent d'acides nucléiques qui, avec les protéines, les lipides et les polysaccharides, constituent les quatre macromolécules essentielles à la vie, la plupart des molécules d'ADN sont constituées de deux chaînes de polymères biologiques enroulées l'une autour de l'autre en une double hélice.

Une chaîne d'ADN est appelée poly-nucléotides est composée d'unités plus simples appelées nucléotides. Chaque nucléotide est constitué d'un sucre appelé désoxyribose, d'un groupe phosphate et d'une base azotée parmi les quatre bases azotées (adénine [A], thymine [T], guanine [G], cytosine [C]). Les nucléotides sont liés les uns aux autres dans une chaîne via des liaisons covalentes entre le sucre d'un nucléotide et le phosphate nucléotidique suivant, formant le squelette de l'ADN, les bases des deux chaînes poly-nucléotidiques azotées sont liées entre elles selon les règles de la liaison par paires (A avec T et G avec C) par des liaisons hydrogène complémentaires pour former une molécule d'ADN à deux chaînes par contre parallèles.

Chez les eucaryotes, l'ADN est présent dans le noyau sous forme de chromatine pour faciliter le processus d'expression génique, et il ne se présente pas sous forme de chromosomes sauf au stade de l'iso-division où se déroule le processus de réplication et d'empaquetage, alors qu'il est présent chez les procaryotes (bactéries) dans le cytoplasme. Au cours de l'expression génique, les gènes sont transcrits en molécules d'ARN messager et traduits en protéines par les ribosomes. [1.2]

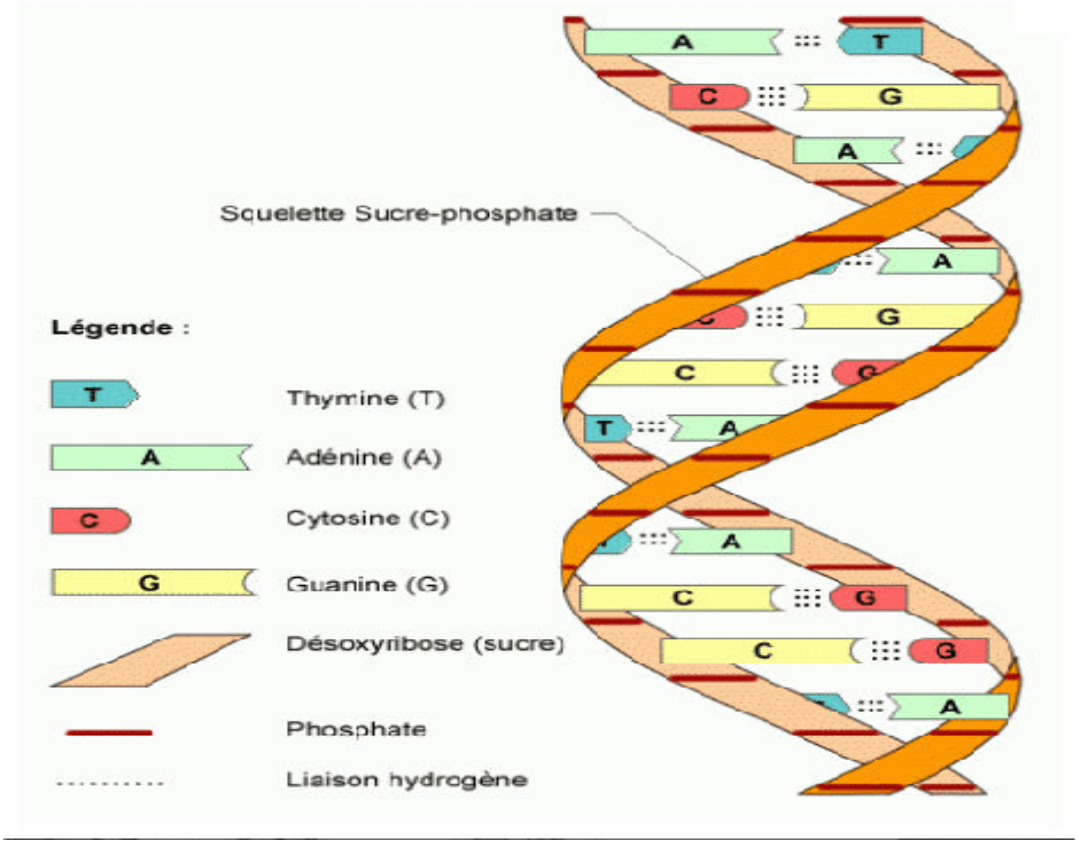


Figure I .3 : L'ADN

I .3.3. Les gènes

Le génome est l'ensemble du matériel génétique d'un organisme. Il contient à la fois les séquences codantes, c'est-à-dire celles qui codent pour des protéines, et les séquences non

codantes. Chez la majorité des organismes, le génome correspond à l'ADN présent dans les cellules.

Les gènes sont responsables des différentes tâches génétiques telles que la fabrication de la protéine et la formation du brin d'ADN, leurs complexité se différent entre les organismes vivants ainsi que leurs taille qui varie de quelques centaines à deux millions de bases. Le nombre de chromosomes est aussi différent d'un être vivant à un autre ; il y a 23 paires, soit 46 chromosomes chez les êtres humains.

La suite des acides aminés ou bien des bases nucléiques (A C G T) déterminent les caractéristiques physiques et psychologiques des personnes et même la susceptibilité à une maladie.

L'unité de base de l'hérédité dans l'organisme est le gène qui vient de nos pères (la mère et le père).Le sperme et l'ovule porte chacun un jeu de 23 chromosomes, dont 22 sont des autosomes et le 23èmeest celui du sexe (**x** ou **y**).La femelles hérite du chromosome **x** de chacun des parents, le mâle hérite du chromosome **x** de la mère et du chromosome **x** du père. Les gènes contiennent toutes les informations nécessaires pour la construction et la maintenance de la cellule ainsi que les informations génétiques transmissent de génération en génération. [1.1]

I .3.4. Les protéines

Les protéines sont de longues chaînes d'acides aminées reliées par des liaisons peptidiques, ils sont nécessaires à la vie, ils constituent 20% du corps humain parce qu'ils sont utilisées par les différentes cellules pour accomplir leurs fonctions. La disposition et les types des acides aminés déterminent la fonction de la protéine.

La synthèse ou bien la fabrication de la protéine se fait à l'intérieur de la cellule, tout instruction de cette opération est contenue dans la molécule d'ADN et passe par deux étapes principales qui sont :

- **La transcription** : elle consiste à faire une copie d'ADN appelée ARN.
- **La traduction** : l'ARN est convertit en une chaîne d'acides aminées au niveau d'une machine complexe dans la cellule appelé ribosome.

Après la fixation de l'ADN sur le ribosome, et à partir du code « START », le ribosome commence son lecture des messages et des instructions. Chaque trois lettres ou acide aminé construit une chaîne d'acides aminés basées sur les codes (ARN), le processus est arrêté si le code « STOP » est apparut, et voila la protéine est maintenant prête à fonctionner. [1.3]

I .4. Les données de puces à ADN

La technologie par «puces à ADN» utilise le principe de l'hybridation génomique comparatif (CGH) et consiste à Co-hybrider une même quantité d'ADN d'un patient et d'un témoin contrôle, marqués chacun par un fluor chrome de couleur différente, sur un

réseau de séquences d'ADN (puce à ADN) représentant l'entier du génome. La lecture des signaux fluorescents est réalisée grâce à un scanner laser automatisé. Une analyse bio-informatique des données est ensuite effectuée à l'aide d'un logiciel qui enregistre l'intensité des différentes fluorescences. Après normalisation de ces données, un rapport sous forme de représentation graphique est effectué à l'aide d'un logiciel, les rapports d'hybridation étant proches de «0» pour toutes les régions sans anomalie structurelle. Une déviation significative de cette valeur sur plusieurs signaux d'une même région génomique suggère la présence d'anomalies, perte ou gain d'ADN. Cette analyse est répétée à l'ensemble du génome. Il est alors possible de voir le caryotype avec une loupe contenant un grossissement d'environ 1000 fois. [1]

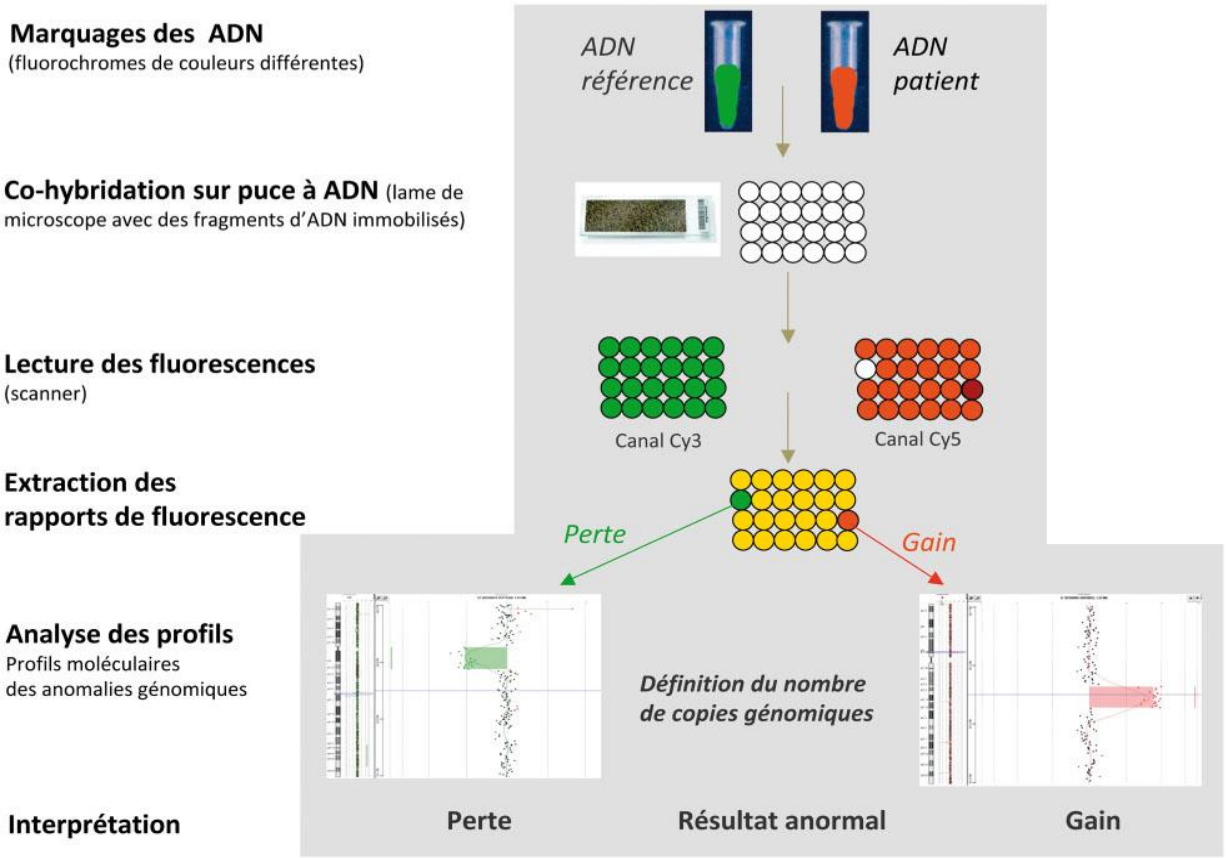


Figure I .4 : Les données de puces à ADN

I .5. **Conclusion**

Nous avons vu une généralité sur la bio-informatique, la nouvelle science multidisciplinaire qui ne cesse de se développer et qui utilise les méthodes mathématiques, statistiques et informatiques pour la maitrise des données biologiques en vision de résoudre les problèmes difficiles qui constituent un obstacle scientifique en biologie.

Nous avons pris aussi les notions générales de la biologie moléculaire et les données manipulées pour permettre la compréhension des différentes techniques de prédictions des séquences d'ADN que nous allons discuter dans le deuxième chapitre.

CHAPITRE II:
La sélection d’attributs pour la classification

II.1.Introduction :

Vue l'explosion de la quantité de données due des différents sources et suite à la complexité, la sensibilité à traiter ,manipuler, stocker et exploiter efficacement ces données , il est devenu indispensable de trouver de nouvelles méthodes puissantes qui marchent en parallèle avec ce progrès et qui permettent une meilleure façon d'analyser et de classer cette volumineuse masse d'informations ,ainsi de purifier celles les plus intéressantes et les plus utiles afin d'avoir la connaissance la plus importante, la plus implicite et la plus nécessaire.

II.2 Fouilles de Données

Ce sont des processus et des moyens automatiques (mathématiques, statistiques et algorithmiques) pour analyser les données (data) qui permettent de trouver les modèles cachés (Paterns). Le rôle principal de ces processus est la recherche des dépendances et des relations entre les données ainsi que les corrélations, les associations, les structures et les classes pour les classer. Cet analyse profonde est presque impossible manuellement et qui demande un très long temps de réalisation avec un taux d'erreur élevé.

Ces besoins en processus et méthodes utiles de formalisme, de simplification des données massives donnent naissance et émergent un nouveau domaine qui est « La fouille de données » ou « Data Mining » en anglais.

La fouille de données est un processus ou un ensemble d'approches statistiques qui permettent d'extraire, de découvrir une information utiles ou bien une connaissance parmi un grand jeu de données dans une perspective d'aide à la décision.

Figure II.1 : Fouille de données comme source de connaissance

La fouille de données est définit aussi comme un ensemble de techniques d'exploration de données permettant d'extraire une base de données des connaissances sous la forme de

modèles de description afin de décrire le comportement actuel des données et de prédire le comportement futur de ces données.[2.1]

II.3 Problème de la classification

Suite aux progrès scientifique, la capacité de mesure a augmenté sensiblement ainsi que les bases de données, mais jusqu'à présent l'opérateur humain reste incapable de traiter cette massive quantité de données avec toute cette dimensionnalité.

La prise de décision est difficile surtout si les données sont de différentes sources (plusieurs variables et attributs), ce qui pose le problème de classification.

La classification est un ensemble d'algorithmes et des méthodes de structurations utilisée pour répartir et organiser des données ou des objets en sous-ensembles ou en groupes homogènes, en prenant compte leurs similarités ou dés-similarités. L'objectif principal est de minimiser la dimensionnalité (la variabilité intra-groupe) tout en maximisant les distances intergroupes.

Ces algorithmes visent à regrouper les données dont les individus sont très similaires (pertinents) mais distants des autres (non redondants) dans une même classe, et permet de prédire si un élément est un membre d'un groupe ou d'une catégorie donnée ou non.

Les méthodes de classifications se répartissent en deux grandes approches, l'approche non supervisée et l'approche supervisée. [2.2]

II.3.1 La classification non supervisée

Elle consiste à regrouper les données sans aucune information à priori, s'appelle aussi clustring ou bien classement automatique (les données ne sont pas étiquetées).La classification supervisée dont les groupes de classement sont définis ou existent à priori, donc les données sont étiquetées à l'avance.

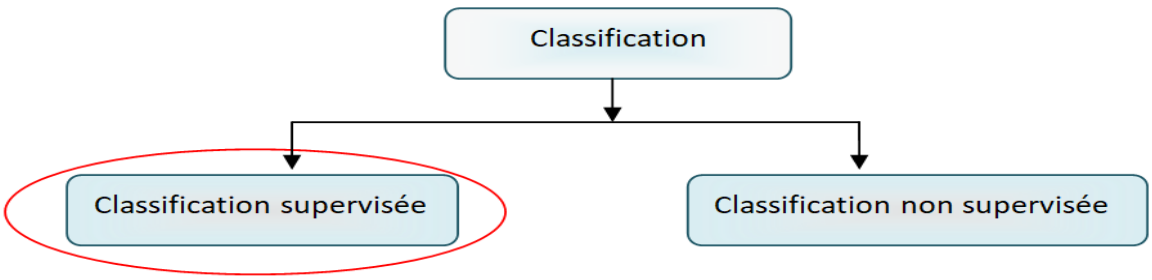


Figure II.2. Les types de classification

II.3.2. Définition de la classification supervisée

Dans l'objectif de classer des individus à partir de leurs caractéristiques ou leurs attributs dont les classes sont connus préalablement. On dispose de **N** exemples de pairs Entrée/ Sorties : **(x_i, y_i)**.

(x_1 ,y_1), (x_2 ,y_2) …….. (x_n ,y_n) tels que :

$x_i \in X$: est la variable descriptif de l'objet à classer.

$y_i \in Y$: est la supervision (la classe) de x_i , ou chaque y_i a été généré par une fonction F(x) = Y inconnu.

Il reste à découvrir la fonction **f** qui se rapproche de **F.**

En s'appuyant sur l'ensemble des exemples pour prédire la classe de toute nouvelle donnée x_i. [2.3]

II.3.2.1 Méthodes de classifications supervisées

Il existe plusieurs méthodes de classifications voici quelques unes :

- La méthode de K-Plus Proche Voisin
- La méthode de noyau et des Machines à Vecteurs de Support (SVM)
- La méthode de l'arbre de décision.
- Réseaux de neurones.

II.3.2.1.1 La méthode K-Plus Proche Voisin

K plus proche voisin ou (K-nearest Neighbors K-NN) est un algorithme de classification parmi les algorithmes les plus simples et les plus utilisés pour effectuer une prédiction. Son principe est de chercher les instances du jeu de données les plus proches (selon la mesure de similarité) par rapport aux autres voisins de notre observation **X,** pour prendre une décision si l'élément **X** appartient t-il à cette classe (dit majoritaire) parmi l'ensemble d'apprentissage des **K** voisins qui lui ressemble.

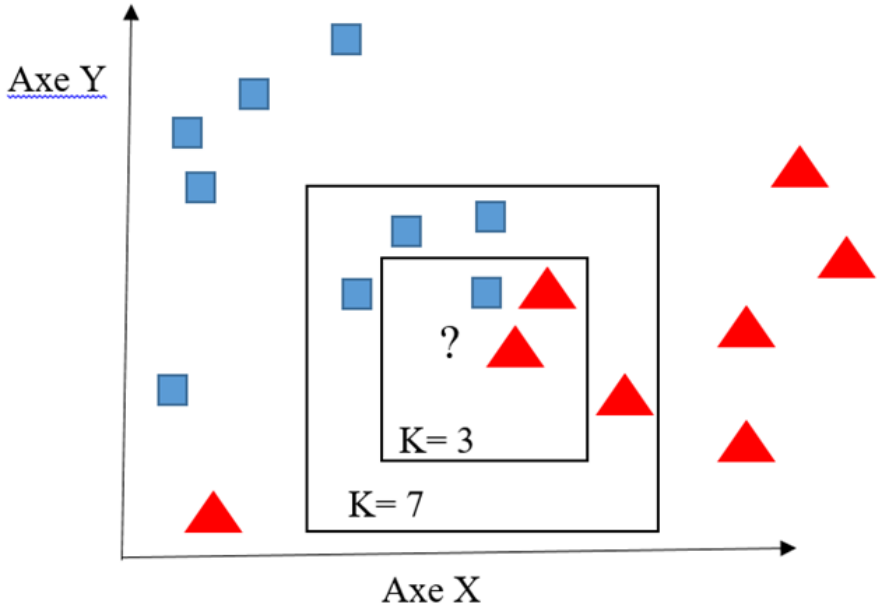


Figure II.3 La méthode du K plus poche voisin

Dans cette figure la notion de K-Plus Proche voisin est bien claire, nous avons deux classes des instances : la classe de triangles rouges et celle des carrés bleus; pour classer une nouvelle instance **(?)** la technique est de visualiser l'ensemble de ses voisins dont la distance est minimale, la classe de l'élément **(?)** est la classe majoritaire de ses voisins.

Dans notre exemple :

- Pour **k = 3**, nous avons deux éléments triangles rouges et un carré bleu donc la classe majoritaire est celle des triangles donc l'élément est un triangle rouge.
- Pour **k = 7**, nous avons trois triangles rouges et quatre carrés bleus donc la classe est donc «carré bleu ».

II.3.2.1.2 Méthodes des Machines à Vecteurs de Support (SVM) :

Les machines à vecteurs de support (SVM ou Support Vector Machine) sont des classificateurs linéaires, leurs principes est simple, dont le but est de séparer les données en classes par une frontière (dit marge) .Il faut dessiner une droite afin de séparer deux classes d'apprentissages à **n** dimensions des données d'expressions. L'objectif est de maximiser la distance des échantillons aux frontières de la marge, les vecteurs de support étant les données les plus proches de la frontière.

Dans notre exemple la droite noir représente la frontière (l'hyperplan); les vecteurs de supports sont les points entourés (les plus proches de la frontière), la marge est la distance entre la frontière et la droite bleue et la droite rouge. [2.4]

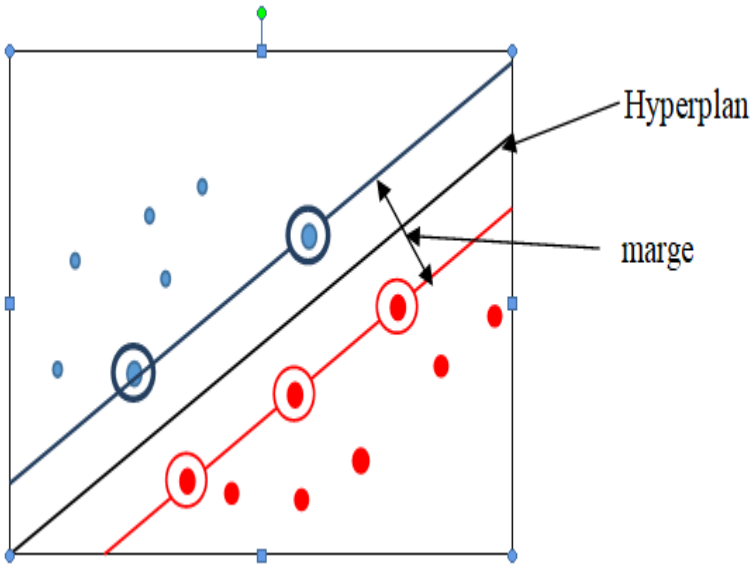


Figure II.4 Méthode des Machines à Vecteurs de Support(SVM)

II.3.2.3 La méthode de l'arbre de décision: Les arbres de décisions sont les méthodes les plus utilisées et connues en classification; ils permettent de répartir une population d'individus en groupes homogènes selon des attributs choisis préalablement.

Le principe est d'effectuer une suite de tests récursifs sur les attributs afin de réduire un problème en un ensemble de sous-problèmes jusqu'à atteindre un niveau de simplicité qui permet de le résoudre facilement. C'est un classificateur graphique hiérarchique structuré comme un arbre réel, il est composé d'un nœud racine, un ensemble de nœuds internes qui représentent les attributs et un ensemble de feuilles représentent les classes d'affectations (résultats) ainsi que des arcs entre les nœuds qui représentent les tests sur ses derniers, une relation de parenté existe entre les différents nœuds de l'arbre.

Une règle de décision est de la forme (si…..alors) est crée pour chaque chemin à partir de la racine et passant par arcs (tests) jusqu'à la feuille (classe).

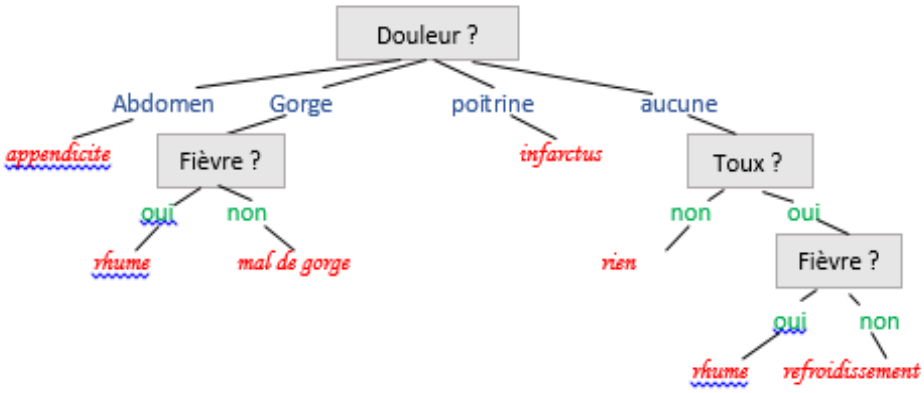


Figure II.5 Arbre de décision

II.3.2.4. Réseaux de neurones : Un réseau de neurones (neural network) et un système informatique inspirant du cerveau humain pour apprendre. Un réseau de neurone repose sur un grand nombre de processus opérants et organisés en tiers. Le premier tiers reçoit les entrées d'informations brutes par la suite chaque tiers reçoit les sorties d'informations du tiers précédent, le retour en arrière n'est pas possible (dans l'apprentissage supervisé) l'algorithme s'entraine sur un ensemble de données étiquetées et se modifie jusqu'à être capable de traiter le dataset pour obtenir le résultat souhaité.

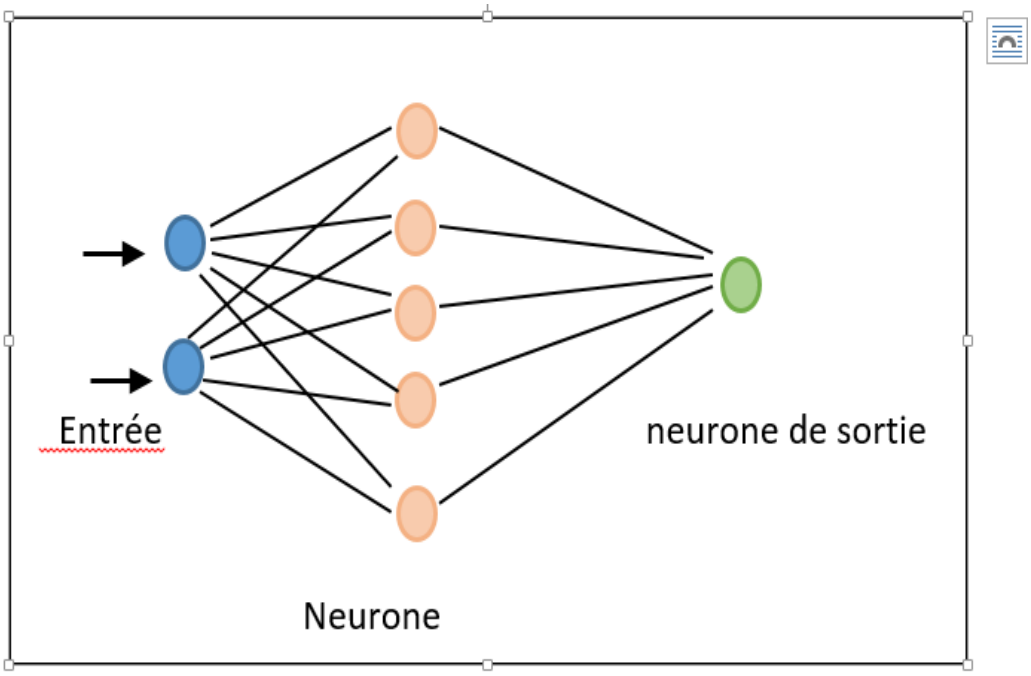


Figure II.6 Vue simplifiée d'un réseau artificiel de neurone

II.4. Sélections d'attributs :

Actuellement les bases de données disponibles sont en constante augmentation, dont beaucoup de données sont non significatifs ou redondantes ; ce qui crées du bruit et rendre les modèles incapables de créer la prédiction sur de nouvelles données.

En effet pour remédier à ce problème, et pour plus de performance, nous utilisons la sélection des attributs comme une phase préliminaire pour réduire la dimensionnalité des données afin de permettre de découvrir les meilleurs modèles. [2.7]

La sélection des données est devenue un sujet de recherche très actif depuis des années dans le domaine de traitement d'image, de fouilles de données et les domaines d'apprentissage artificiels.

La sélection d'attributs (de variables) consiste à choisir parmi un ensemble d'attributs de très grande taille, un sous-ensemble d'attributs plus petit qui sont importants (significatifs) pour le problème étudié. [2.8]

Les principales conséquences de la sélection d'attributs sont:

- Elle permet de déterminer les attributs considérés comme pertinents pour avoir un sous ensemble de données optimal d'attributs.
- Elle permet de supprimer le bruit généré par les variables non pertinentes et / ou redondantes ainsi que leur sur-apprentissage.
- Le petit sous-ensemble choisi permet de réduire l'espace de représentation et de réduire aussi le coût de la phase d'apprentissage et d'exécution des algorithmes de classification.
- Elle permet une meilleure compréhension du problème et améliore la précision et augmente la performance. [2.9]

II.4.1 Processus général de la sélection d'attributs :

Idéalement, les méthodes de sélection de caractères vérifient toute les caractéristiques et essayent de trouver les meilleurs sous-ensembles d'entre eux. Il y a quatre étapes de base à une méthode de sélection typique :

1. **Une procédure de génération :** pour générer le prochain sous-ensemble de caractéristiques.
2. **Une fonction d'évaluation :** pour évaluer le sous-ensemble choisi par la procédure de génération.
3. **Un test d'arrêt :** pour décider quand arrêter.
4. **Une procédure de validation :** pour vérifier si le sous-ensemble est valide ou non.

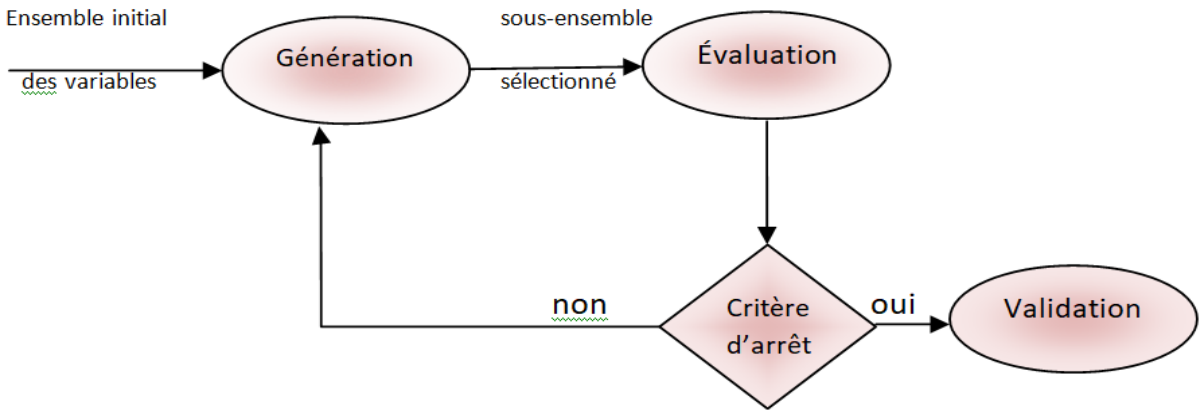


Figure II.7 Processus de sélection des caractéristiques

II.4.1.1 -La procédure de génération :

La procédure de génération est une procédure de recherche dans l'espace des sous-ensembles de cardinal (2^n), ou n est le nombre d'attributs initiaux, elle permet à chaque itération, de générer un sous-ensemble de variables qui seront évalués lors de la deuxième étape de la procédure de sélection selon un critère bien défini. [2.10]

La méthode de génération se poursuit par des ajouts ou suppressions successifs d'attributs. Il peut commencer avec un ensemble vide d'attributs ou un ensemble de tous les attributs, ou avec un sous-ensemble d'attributs sélectionnés au hasard.

Les méthodes de génération peuvent être classées en trois principales approches de génération.

1. Méthode exhaustive (complète) :

On génère et on test tout les sous-ensembles (l'espace des solutions possibles), ce qui est d'ordre (2^n).Ce ci est prohibitive dés que le nombre est supérieur à 10.

Un tel parcours garantira l'optimalité mais le coût est excessif.

2. Méthode heuristique (aléatoire) :

Un seul sous-ensemble est recherché, le nombre maximum d'itérations est définis préalablement. Dans cette catégorie, on peut commencer par un ensemble vide puis ajouter les attributs un par un c'est ce qu'on appelle la méthode ascendante ou en anglais (Forward addition), ou bien en commençant par tout les attributs puis supprimer certains (Backword elimination) c'est la méthode descendante, comme on peut générer les attributs aléatoirement dans chaque itération (Random selection).

3. Méthodes non déterministes (exploratoires) :

Les sous-ensembles sont générés suivant un processus basé sur les algorithmes évolutionnistes (evolutionary Algorithms) dont le principe de l'évolution constitue la clé de leur fonctionnement. [2.10]

II.4.1.2 - Fonction d'évaluation :

Après la procédure de génération chaque nouveau sous-ensemble produit doit être évalué par des critères d'évaluation pour gérer à la fois la fréquence (la redondance) et l'importance (la pertinence).Le sous-ensemble d'attributs candidats est évalué par une échelle d'évaluation et comparé avec le meilleur sous-ensemble ayant les caractéristique obtenu précédemment par rapport à cette échelle de mesure .Si le sous-ensemble actuel est meilleur, il remplace le meilleur sous-ensemble enregistré. Ce processus est répéter jusqu'à obtenir le sous-ensemble optimal. [2.8]

Les fonctions d'évaluation sont divisées en cinq catégories :

1-Mesure de distance:

Soit X et Y des attributs de classes différentes. Si la distance entre X et la classe à laquelle appartient Y est supérieure à la distance entre Y et la classe à laquelle X appartient, alors X est pris, pas Y. Les distances euclidiennes sont les plus couramment utilisées pour ce type de mesure.

2- Gain d'informations :

Il s'agit d'une véritable mesure de l'information ou de la richesse de l'information, qui est tirée de chaque attribut. L'entropie de Shannon est l'une des mesures du gain d'informations les plus utilisées.

3-Mesure de dépendance :

Cette échelle détermine l'association et la corrélation entre deux attributs. C'est à dire pour voir dans quelle mesure la valeur d'un attribut peut être prédite par la valeur d'un autre. Si l'attribut représente la classe bien, mieux il est évalué. La corrélation est la plus couramment utilisée pour ce type de mesure. Une petite différence permet de mesurer la dépendance entre les attributs, donc le degré de fréquence pour chaque attribut.

4- Mesure de cohérence :

Il s'agit d'une mesure relativement nouvelle, elle diffère des autres mesures par le fait qu'elle donne une association forte et profonde avec la sémantique des attributs. Cette échelle trouve plus un petit sous-ensemble qui satisfait le taux de précision spécifié par l'utilisateur.

5-Taux d'erreur du classificateur :

Le but est de construire un classificateur capable d'indiquer correctement chaque cas, tout en réduisant la possibilité d'erreur.

II.4.1.3- Procédures de validation :

Le sous-ensemble choisi doit être validé par différents tests avec des données du monde réel ou non réel ou les deux. La procédure de vérification est une étape pour choisir les attributs qui devrait être pratiquement et comparer les résultats obtenus avec les résultats précédents, ou avec les résultats d'autres méthodes de sélection. La vérification des changements des performances par rapport aux changements des attributs avant et après l'expérience. [2.11]

II.4.1.4 - Test d'arrêt :

Le critère d'arrêt définit une borne et une condition pour arrêter la procédure de sélection d'attribut, il est nécessaire que chaque sous ensemble d'attributs après son évaluation soit comparé au critère d'arrêt définit, celui-ci peut être :

- Accès à un nombre prédéterminé d'attributs sélectionnés.

- Atteindre un temps de calcul fixé auparavant.

- Accès à un nombre prédéfini d'itérations.

- Si l'ajout ou la suppression d'attributs ne fournit pas un sous-ensemble d'attributs meilleur (sous-ensembles homogènes).

- Si nous arrivons à un bon sous-ensemble de thèmes, qui répond à la contrainte de temps et la qualité imposée. [2.9]

II.4.2 Types des méthodes de la sélection d'attributs.

Comme nous avons vus précédemment, après la phase de génération des sous-ensembles, ces derniers doivent être évalués. Il existe trois grandes classes d'algorithmes : les algorithmes basés sur les méthodes enveloppantes (Wrapper) et ceux qui sont basés sur les algorithmes filtrantes et la méthode hybride.

II.4.2.1. Méthodes Filter (filtres) :L'approche filtrante repose sur l'idée d'attribuer un score à chaque sous-ensemble. Aussi, le sous-ensemble avec le plus grand score représente le sous-ensemble d'attributs pertinents.

Pour cela, plusieurs solutions sont possibles, la première solution consiste à donner un score à chaque attribut indépendamment des autres et faire la somme des scores, en cas de problème de classification, on peut retenir le coefficient de corrélation comme le score de l'attribut avec la classe, mais cette solution n'élimine pas les attributs redondants. De plus, il est possible qu'une corrélation existe entre un attribut et la classe ou le contexte des autres attributs. La deuxième solution consiste à évaluer le sous-ensemble dans sa globalité.

L'avantage de cette approche est l'efficacité calculatoire et robustesse face au sur-apprentissage grâce à l'utilisation des mesures statistiques pour le filtrage des caractéristiques informatives qui est réalisé avant d'appliquer tout algorithme de classification.

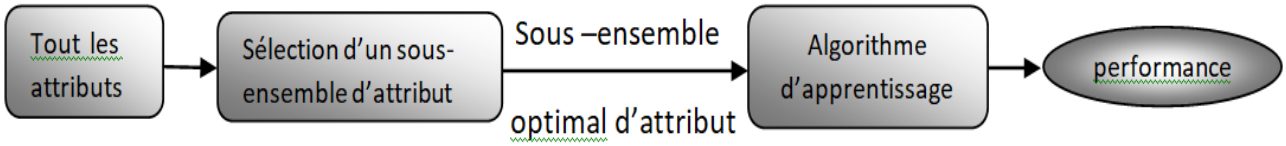


Figure II.8 Schéma de l'approche Filtre

II.4.2.2. Méthodes Wrapper (enveloppantes) : Introduite pour la première fois par John, Kohavi en 1994, leurs principe est de générer des sous-ensembles candidats et les évaluer grâce à un algorithme de classification. Ces méthodes génèrent des sous-ensembles bien adaptés à l'algorithme de classification mais ils ne restent pas valides si le classificateur est changé donc c'est une solution non parfaite.
Leurs avantages est la simplicité conceptuelle et la précision, elles explorent tout l'espace des sous-ensembles des caractéristiques afin de trouver un sous-ensemble optimal.
Cependant l'algorithme d'apprentissage rend ces méthodes plus couteuses en temps de calcul qui devient fastidieux ou irréalisable lorsque le nombre d'attributs croît (appel de l'algorithme de classification à chaque évaluation). Mais ils sont considérés comme étant meilleurs que celles de filtrage pour la sélection des sous-ensembles de petites taille.

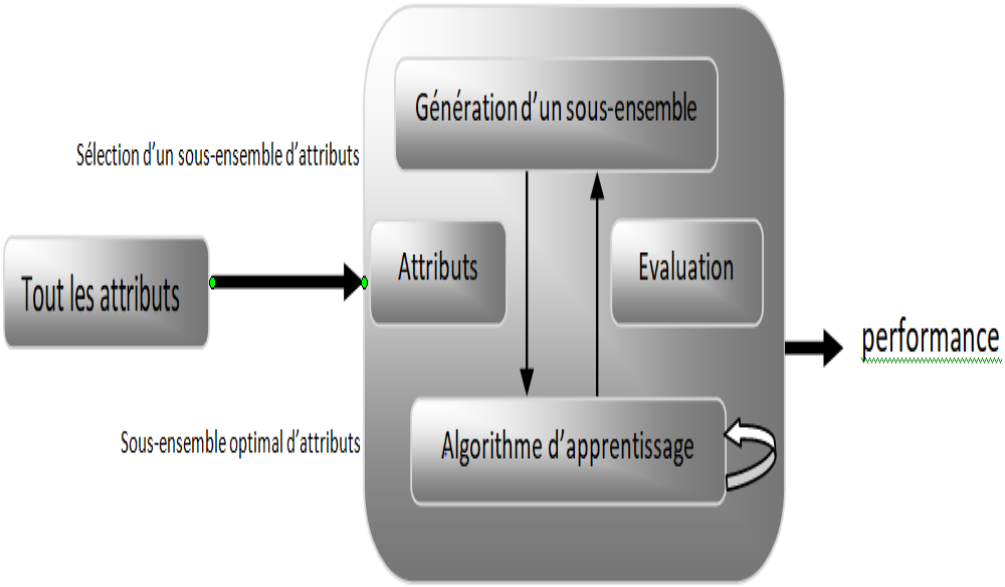


Figure II.9 Schéma de l'approche Wrapper

II.4.2.3. Méthodes hybrides

Ces méthodes sont proches des méthodes Wrapper et il y a une combinaison entre le processus d'exploration et l'algorithme d'apprentissage donc la sélection des attributs et la construction du modèle se fait en même temps. Ils sont rapides par rapport aux méthodes Wrapper parce qu'elles évitent que le classificateur recommence de zéro à chaque sous ensemble de caractères. [2.10]

II.5 Comparaison des approches Wrapper et filtre

	Wrapper	**Filtre**
Performances	- Performante - Corrélation prise en compte	- Non performantes - corrélation entre attributs et interclasse n'est pas prise en compte
Vitesse	Lente	Rapide
Théorie	+/- Simplicité	+/- Explications
Spécificité	Non spécifique	Spécifique

Tableau I.1 Tableau comparatif entre les méthodes de sélection

II.6. Conclusion :

Nous avons introduit l'étape de prétraitement « Choisir les attributs sous supervision » qui joue un rôle important dans l'exploration de données.

Dans ce chapitre, nous avons donné une vue globale sur la fouille de données puis nous avons parlé du processus de la classification supervisée et les méthodes de classification, puis nous avons vu les différents points nécessaires dans le processus de sélection des attributs, ainsi que les différentes méthodes utilisées pour la sélection avec une petite comparaison entre eux.

Chapitre III :

Réalisation et résultats

III. 1. Introduction

Dans ce chapitre nous allons faire une étude comparative entre des algorithmes des trois méthodes de sélection dans le but de choisir la meilleure d'entre eux.

Dans ce chapitre la meilleure méthode est celle qui nous amène au nombre minimum d'attributs, et une exactitude élevée pour trouver une solution au problème de sélection des gènes et la classification des données à puces à ADN.

Afin d'évaluer les algorithmes choisis, nous avons effectué des expérimentations poussées sur trois jeux de données, Langcancer, Lymphography et Tumor (Tumeur) puis comparer les résultats obtenus par rapport aux trois méthodes précités, pour déterminer la plus efficace et la plus précise dans ce choix.

III. 2. Méthodes de sélection d'attributs

Plusieurs catégorisations se trouvent dans la littérature, les développeurs les partagent selon les stratégies de recherche ou selon les critères d'évaluation, suite à cette dernière catégorisation il existe trois approches principales pour la sélection des attributs ; Wrapper, Filter et la méthode hybride, dont chacune ayant des avantages et des inconvénients.

Méthode	Avantages	Inconvénients
Filtre	• Exécution rapide. • Cout de calcul faible	• Aucune interaction avec la classification.
Wrapper	• Interaction avec la classification. • Bonne performance de classification.	• Coute de calcul élevé. • Exécution lente.
Hybride	• Interaction avec la classification. • Bonne performance de classification.	• Côut de calcule élevé mais plus faible que Wrapper • Pas adapté à tous les types du classifieur.

Tableau III.1 : Avantages et inconvénients des méthodes de sélection.

Le principe des Wrappers a besoin d'un classificateur pour effectuer son évaluation ce qui lui rendre très lent malgré leur bonne performance.

L'approche Filtre est réalisée comme un prétraitement, et elle est indépendante de toute sorte de classifieur donc elle est plus rapide que l'approche Wrapper en terme de génération des résultats et mauvaise par rapport à la qualité des résultats et de la performance.

Cependant, cette dernière a l'avantage de fournir généralement des résultats plus pertinents pour la classification, mais elle est couteuse en temps d'exécution et d'espace.

Pour essayer de trouver une solution meilleure et d'éviter les inconvénients de chacune des deux méthodes précédentes, les développeurs ont proposé une méthode intermédiaire qui est l'approche hybride.

Dans ces algorithmes, une mesure basée sur les caractéristiques des données est utilisée pour choisir les meilleurs sous-ensembles pour une cardinalité donnée. Par la suite, la validation croisée est exploitée pour choisir le meilleur sous-ensemble optimal.

Les trois méthodes précédentes ont des algorithmes de sélection qu'on va les cité dans le schéma suivant :

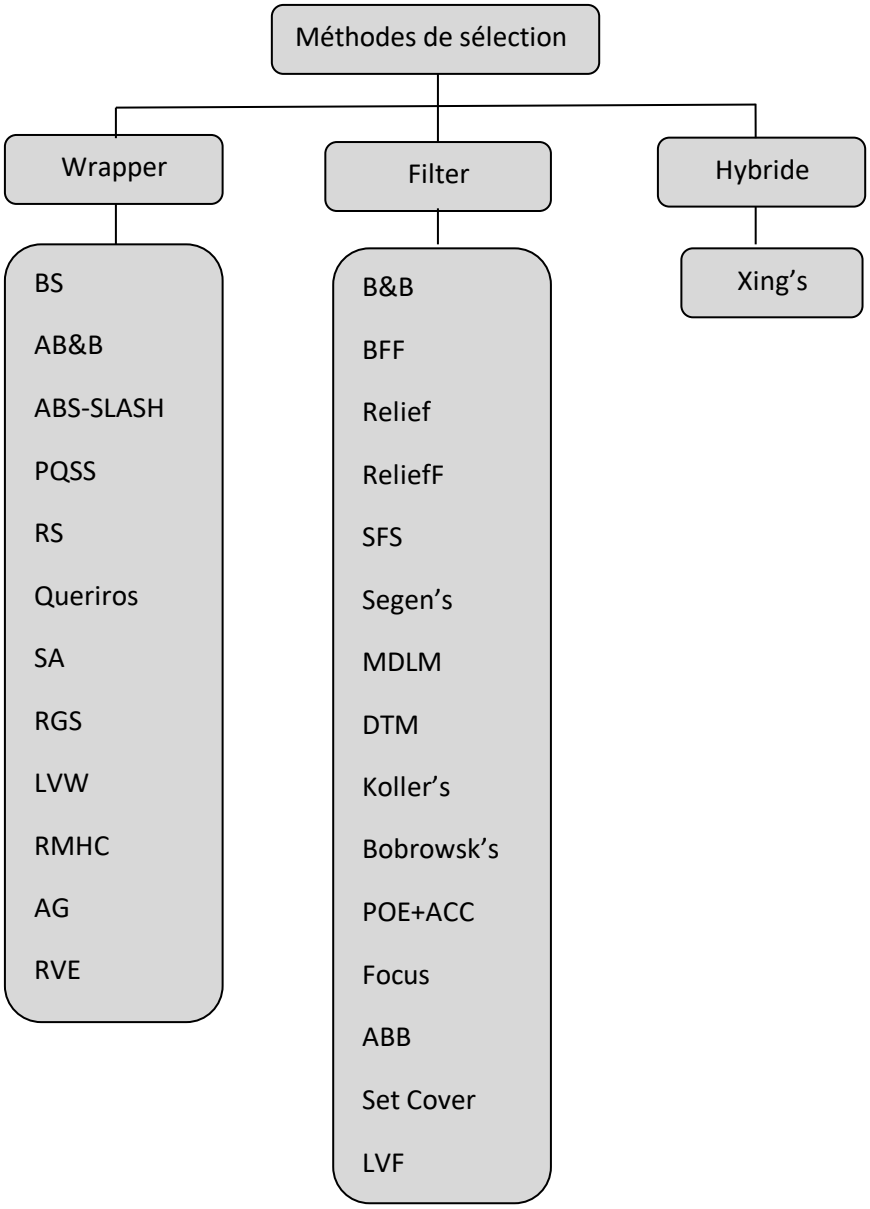


Figure III. 1. Schéma des algorithmes des méthodes de sélection

III. 3. Choix des algorithmes

Nous allons citer quelques algorithmes dans le but de savoir les principes d'évaluation des méthodes

1) CfsSubsetEval (Correlation-basedFeatureSubsetSelection) :

Cet algorithme évalue la valeur d'un sous-ensemble d'attributs en tenant compte de la capacité prédictive individuelle de chaque caractéristique ainsi que du degré de redondance entre elles.

Les sous-ensembles de caractéristiques qui sont fortement corrélés avec la classe (tout en ayant une faible inter-corrélation) sont préférés.

2) ClassifierAttributeEval : Cet algorithme évalue la valeur de l'attribut à l'aide du classificateur spécifié par l'utilisateur.

3) **ClassifierSubsetEval :** Il évalue les sous-ensembles d'attributs sur les données d'apprentissage ou un ensemble de test de commentaire séparé. Le classifieur est utilisé pour estimer la "caractéristique" d'un ensemble d'attributs.

4) CorrelationAttributeEval : Il évalue la valeur d'un attribut en mesurant la corrélation entre celui-ci et la classe. Les attributs nominaux sont considérés valeur par valeur en traitant chaque valeur comme un indicateur. Une corrélation globale pour une caractéristique nominale est obtenue via une moyenne pondérée.

5) PrincipalComponents : Cet algorithme effectue l'analyse des principaux composants et la transformation des données, il s'exécute en conjonction avec la recherche Ranker. La réduction dimensionnelle est obtenue en sélectionnant suffisamment de vecteurs propres pour tenir compte d'un pourcentage de la variance dans les données d'origine par défaut 0,95 (95 %). Le bruit d'attribut peut être filtré en convertissant dans l'espace PC, en éliminant certains des pires auto-transmetteurs, puis en revenant à l'espace d'origine.

6) ReliefFAttributeEval : La valeur d'un attribut en échantillonnant à plusieurs reprises d'une instance et en considérant la valeur de l'attribut donné pour l'instance la plus proche de la même classe et d'une classe différente. Elle peut fonctionner à la fois sur des données de classes discrètes et continues.

7) **WrapperSubsetEval** : Il évalue l'ensemble d'attributs à l'aide d'un schéma d'apprentissage, la validation croisée est utilisée pour estimer la précision du schéma d'apprentissage pour un ensemble d'attributs.

III. 4. Bases de données utilisées

Dans cette section, nous avons parlé des bases de données suivantes (Lungcancer, Lymphography, Tumor).Nous l'avons étudié avec un ensemble d'informations (Propriétés et attributs de l'ensemble des données, nombre d'instances et nombre d'attributs), comme il est présenté dans les tableaux ci-dessous.

III. 4.1. Lungcancer : Le cancer du poumon commence lorsque des cellules anormales se développent dans le poumon, qui peut envahir les tissus voisins et forment des tumeurs dans n'importe quelle partie du système respiratoire. Les cellules cancéreuses se propagent aux ganglions lymphatiques et à d'autres organes du corps, l'une des causes les plus courantes de la maladie est le tabagisme.

Caractéristiques de l'ensemble de données	Caractéristiques des attributs	Nombre d'instances	Nombre d'attributs
Multi-varié	Entier	32	57

Tableau III.1 : Tableau représentatif de la base de donnés du Langcancer

III. 4.2. Lymphography : Le lymphœdème survient généralement parce que les ganglions lymphatiques ont été retirés ou endommagés dans le cadre du traitement du cancer. Elle est causée par un blocage du système lymphatique, qui fait partie du système immunitaire. Le blocage empêche le liquide lymphatique de bien s'écouler et l'accumulation de liquide provoque un gonflement.

Caractéristiques de l'ensemble de données	Caractéristiques des attributs	Nombre d'instances	Nombre d'attributs
Multi-varié	Catégorique	148	19

Tableau III.2 : Tableau représentatif de la base de donnés du Lymphography

III. 4.3. Tumor : Une tumeur est un type de croissance anormale et excessive des tissus. Il continue de croître anormalement, même si le facteur d'origine qui l'a déclenché est supprimé. Cette croissance anormale forme généralement une masse cancéreuse.

Caractéristiques de l'ensemble de données	Caractéristiques des attributs	Nombre d'instances	Nombre d'attributs
Multi varié	Catégorique	339	18

Tableau III.3 : Tableau représentatif de la base de donnés du Tumor

III. 5. Les outils de simulation

III.5.1. Présentation de Weka :

Weka est un ensemble d'algorithmes d'apprentissage destinés à effectuer des tâches d'exploration de données. Les algorithmes d'IA peuvent être appliqués directement à un ensemble de données ou invoqués via une application Java. Weka contient des outils pour le prétraitement, la classification, la régression, le regroupement, les règles d'association et la visualisation des données.

En fait, Weka permet de prétraiter sur un ensemble de données, d'appliquer un algorithme d'apprentissage, et d'analyser les résultats et les performances du classifieur.

Weka est disponible pour toutes les plateformes : Windows, Mac-os X, Linux...etc,
Le site officiel est : www.cs.waikato.ac.nz/ml/weka.
Le format d'entrée par défaut de Weka est le ARFF (AttributeRelation File Format) D'autres formats peuvent être importés : CSV, binaire, BDD, SQL (avec JDBC), `a partir d'une URL. [3.3]

III.5.2 Composants de l'environnement Weka

Weka possède plusieurs composants à savoir :
- **Explorer** : ce module regroupe tous les packages importants de Weka à savoir le prétraitement, les algorithmes d'apprentissage, le groupement (clustering), les associations, la sélection des attributs et la visualisation.
- **Experimenter** : permet d'exécuter plusieurs algorithmes d'apprentissage en mode lot (batch) et de comparer leurs résultats.
- **KnowledgeFlowenvironment**: fournit les mêmes fonctionnalités que le composant "Explorer". Ces fonctionnalités sont représentées sous forme graphique.[3.5]

Figure III.2 : L'environnement Weka

III.5.3. Matériel utilisé

L'outil matériel utilisé est un micro portable Dell munit d'un processeur intel(R) Core (TM) i5-6440HQ CPU @2 ,62GHz et avec une Ram de 8 .00 Go et un système d'exploitation 64bits sous Windows 10.
La version utilisée dans ce document est : Version 3.9.5

III. 6. Réalisation et résultats

Dans cette partie nous allons suivre des étapes d'exécutions puis comparer les résultats trouvés.

Etape 1 : Ouvrir le fichier contenant la base de données et choisir la de base approprié.

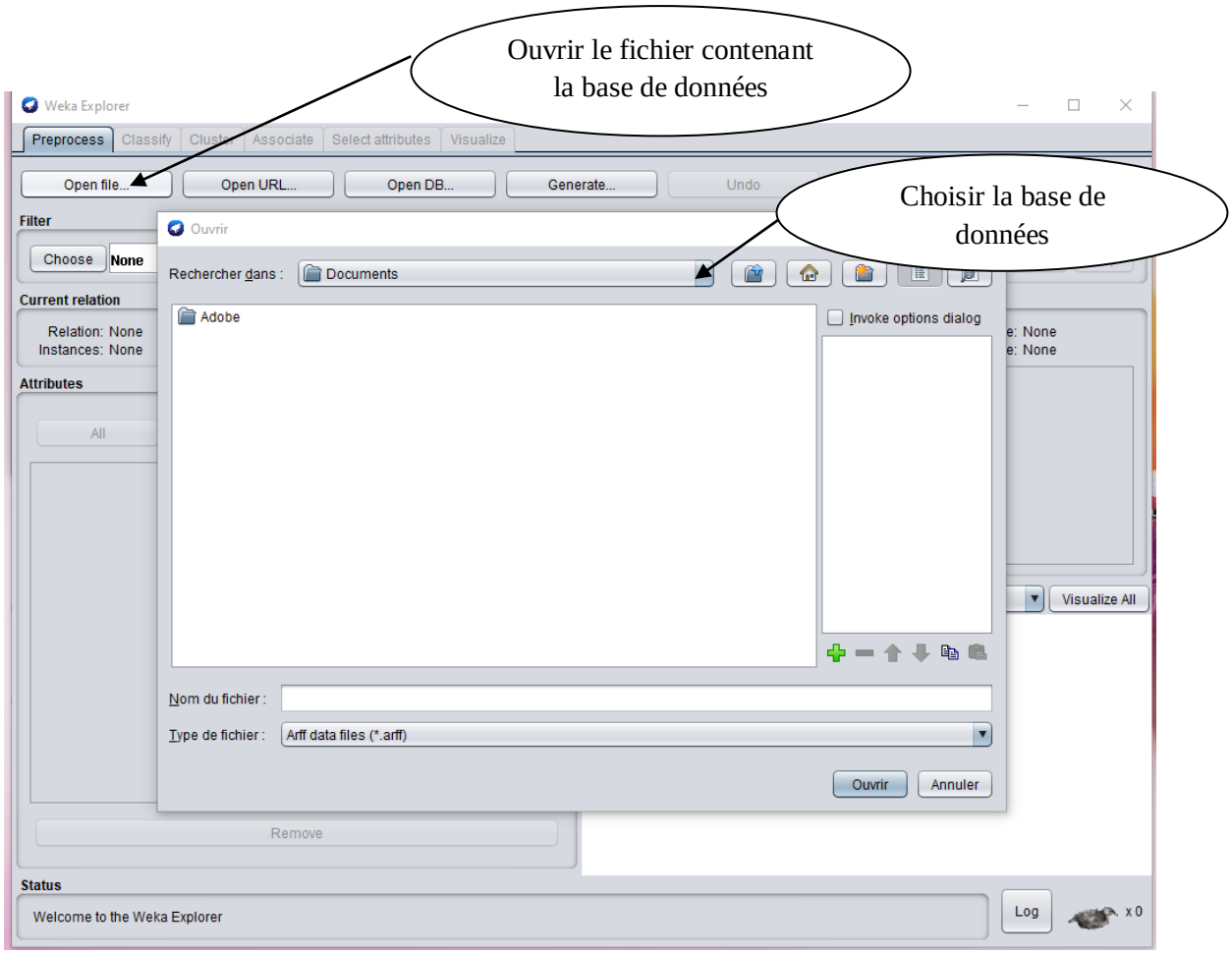
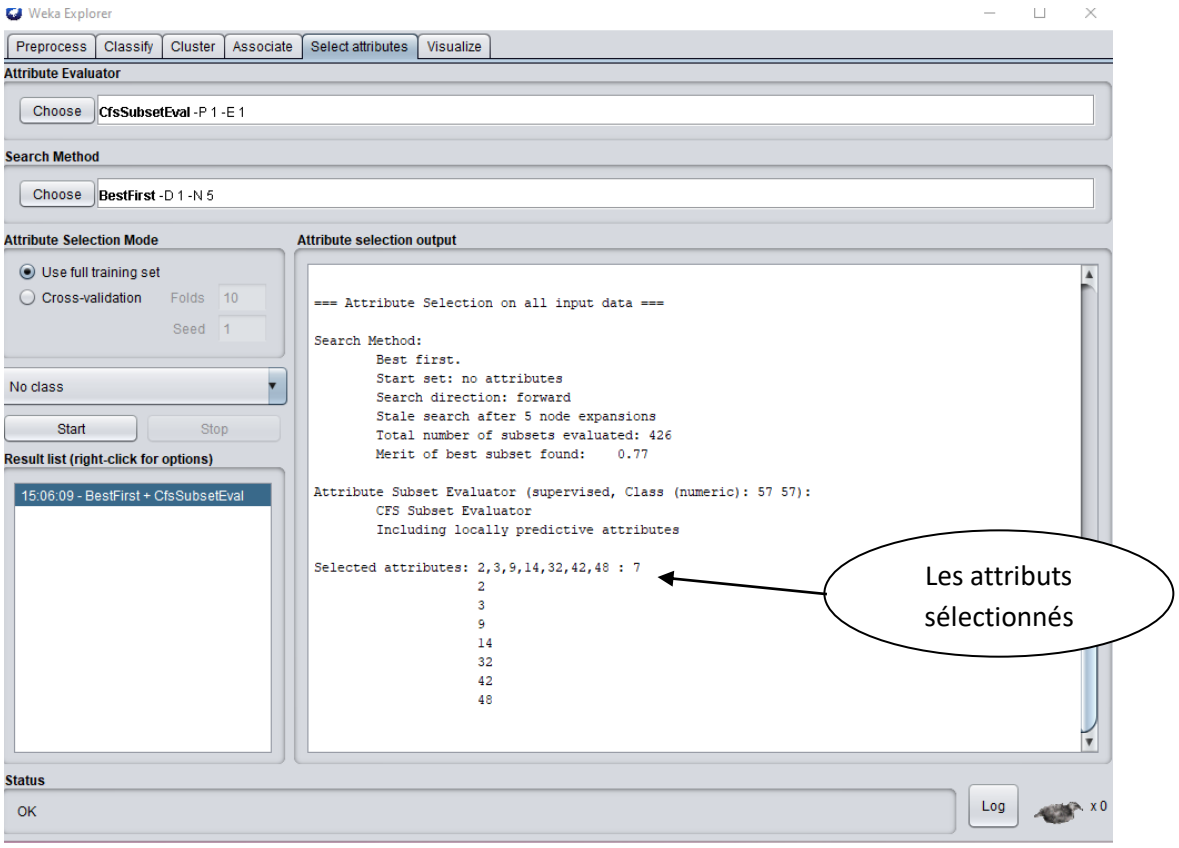


Figure III.3 Ouverture du fichier contenant la base de données

Etape 2 : Sélectionnez la base de données à étudier
Etape 3 : Après avoir sélectionné l'algorithme, les informations qui lui concerne et ses attributs apparaîtront.
Etape 4 : Un groupe d'algorithmes apparaîtra, nous choisissons l'un d'entre eux.
Etape 5 : Exécuter l'algorithme.

Etape 6 : Enfin, après avoir étudié toutes les propriétés, les résultats (les propriétés sélectionnées) nous sont présentées.



III. 7. Résultats et discussion

D'après l'exécution des algorithmes sur les trois bases de données nous avons les résultats suivants :

La base de données : Langcancer

Les algorithmes	Nombre des attributs sélectionnés	Attributs sélectionnées	Pourcentage
CfsSubsetEval	7 sur 57	2,3,9,14,32,42,48	77 %
ClassifierAttributeEval	57 sur 57	56,18,19,17,55,16,20,21,22, 23,26,25,24,15,14,13,6,4,3,2 ,5,7,12,8,11,10,9,27,28,29,4 9,47,46,45,48,50,43,51,54,5 3,52,44,42,30,35,33,32,31,3 4,36,41,37,40,39,38,1	0 %
ClassifierSubsetEval	/	/	45 %
CorrelationAttributeEval	57 sur 57	9,14,15,3,24,22,13,7,27,11,1 0,16,19,18,36,5,35,28,25,29, 23,12,8,41,17,43,4,34,53,54, 26,6,44,31,52,30,40,51,39,3 7,32,33,47,55,2,56,38,50,20, 46,49,48,42,45,21,1	0 %
PrincipalComponents	21 sur 56	1,2,3,4,5,6,7,8,9,10,11,12,13 ,14,15,16,17,18,19,20,21	37,5 %
ReliefFAttributeEval	/	2, 3, 4, 6, 9, 10, 14, 15, 16, 18, 19,20, 22, 24, 25,27 30, 31, 32,33, 34, 36, 38,39 41,42, 44, 43, 47, 48,49, 51,52,53,54,55, 26,46,5,45,50,35,37,40,23,1, 7,11,56,12,28,13,8,21,17,29	**1 %**
WrapperSubsetEval	/	/	**47 %**

La base de données : Lymphography

Les algorithmes	Nombre des attributs sélectionnés	Attributs sélectionnées	accuracy
CfsSubsetEval	7 sur 19	1, 5, 6, 8, 11, 16, 18	**71,5 %**
ClassifierAttributeEval	/	/	**/**
ClassifierSubsetEval	/	/	**/**
CorrelationAttributeEval	19 sur 19	1,2,3 ,4,5,6,7,8,9,10,11,12, 13, 14 ,15, 16, 17, 18, 19	**4,88 %**
PrincipalComponents	15 sur 18	1,2,3,4,5,6,7,8,9,10,11,12,13 ,14,15	
ReliefFAttributeEval	17 sur 19	1,10,8,11,12,9,16,17,4, 5,7,14 ,6,2,15,3 , 13	**7%**
WrapperSubsetEval	/	/	**47 %**

La base des données : Tumor

Les algorithmes	Nombre des attributs sélectionnés	Attributs sélectionnées	Pourcentage
CfsSubsetEval	7 sur 18	2, 4, 8, 10, 11, 15, 17	45 %
ClassifierAttributeEval	/	/	/
ClassifierSubsetEval	/	/	/
CorrelationAttributeEval	17 sur 18	11 ,17,8,10,9,7,12,13,1,16,3, 4,5,6,14,15 ,2	Résultat pour chaque attribut
PrincipalComponents	15 sur 18	1,2,3,4,5,6,7,8,9,10,11,12,13 14,15	19,7 **%**
ReliefFAttributeEval	17 sur 18	1, 10, 8, 11, 12, 9, 16, 17, 4, 5, 7,14, 6, 2, 15, 3, 13	1 %
WrapperSubsetEval	/	/	47 ,6 **%**

Pour les résultats des différents algorithmes nous remarquons que les algorithmes CfsSubsetEval, CorrelationAttributeEval, PrincipalComponents et ReliefFAttributeEval affichent tout le détail d'exécution alors que ClassifierAttributeEval, ClassifierSubsetEval ; WrapperSubsetEval n'affichent pas tout les informations demandées dons ces derniers ne répondent pas aux besoins de choix des attributs et ne donnent pas une bonne exactitude cause de la configuration par défaut de Weka.

III. 8. Conclusion

Dans ce chapitre nous avons cité les différents algorithmes des méthodes de sélection des attributs, puis nous les avons testés sur trois bases de données dans l'environnement Weka.

Enfin nous avons fait une étude comparative entre les résultats obtenus selon la métrique des attributs sélectionnés et l'occuracy.

Conclusion générale

Ce mémoire traite le problème de la classification et de la sélection d'attributs afin de réduire la dimensionnalité des données bio-puces traité en bioinformatique en réservant les données les plus pertinents, les plus explicatifs , non redondants et non cohérés.

Dans ce mémoire nous sommes intéressés à la sélection des attributs pour trouver le sous-ensemble optimal parmi un ensemble de données de très grande taille.

Nous avons suivi les points suivants :

- Une présentation du domaine de la bio-informatique et les données bio-puces.
- Une étude détaillée sur la fouille des données et les différents types d'algorithmes de classification supervisée.
- Un détail sur les méthodes de sélection d'attributs.

Nous avons également testé un échantillon des bases de bases de données, en examinant des algorithmes sur le logiciel Weka.

Enfin nous avons fait une comparaison des résultats selon le nombre d'attributs sélectionné et l'occuracy pour chaque algorithme.

Notre travail n'est qu'un pas dans le domaine de la sélection d'attributs et les données bio-puces d'ADN pour prédire les cellules saines des cellules cancéreuses.

Ainsi nous présentons ici quelques perspectives pour améliorer les travaux de ce mémoire.

- Améliorer les méthodes Wrapper, Filter et hybride pour trouver facilement les scores élevés des attributs les plus pertinents.
- Développer l'outil Weka pour maitriser des algorithmes qui aident à la prise des décisions.
- Améliorer l'outil Weka pour définir directement à quel type de méthode (Wrapper, Filter ou hybride) chaque algorithme.
- Enrichir Weka pour l'obtention et la préparation de la base de donnés sans passer par les étapes actuelles.
- Essayer de faire un travail collaboratif entre les biologistes et les informaticiens pour la maitrise des données biologiques en vue informatique (base de données informatiques) afin d'effectuer des prédictions qui aident à restreindre la maladie avant son évolution en cancer mortel.

Annexe

ABB Automatic Branch & Bound

BB Branch & Bound

AG Algorithmes Génétiques

LVF Las Vegas Filter

LVW Las Vegas Wrapper

SFS Sequential Forward Selection

SBS Sequential Backward Selection

RELIEF Relevance In Estimating Features

SA Search Algorithm

PQSS Potential Quadruplex-forming Sequence

RSA Rivest-Shamir-Adlman

Références bibliographiques

[1.1]- HAL Id: tel-00447684,https://tel.archives-ouvertes.fr/tel-00447684

Submitted on 15 Jan 2010

[1.2]-https://www.eshamel.net/vb/t12462.html

[1.3]- Samir Anair; thèse de magister (les phénomènes de repliement et de dépliement des protéines) 05.05.2009.

[2.1]http://www.mytimes.com /2009/08/06 technology/ 06 stats.html?_r =3.

[2.2]Le Meur Nolewen, thèse de doctorat (de l'acquisition des données vers leur interprétation).

[2.3] Nicole Challita. Contributions à la sélection des attributs de signaux non stationnaires pour la classification. Traitement du signal et de l'image [eess.SP]. Université de Technologie deTroyes, 2018.Français. ffNNT : 2018TROY0012

[2.4] Dataanalyticspost.com

[2.5]www.match.univ-toulouse.fr

[2.6]www.lebigdata.fr>dossier

[2.7]Search Lucian Mousin, Lacticia Jourdar Marie sélection d'attribut par learning tab.

[2.8] José Crispin Hernandez Hernandez, Algorithmes Métaheuristiques hybrides pour la sélection de gènes et la classification de données de biopuces, novembre 2008, page 22.

[2.9] Meddah Leila, Benlefki Amina, Optimisation Par Essaim De PSO Pour La Sélection D'attributs.2016 – 2017, page 33.

[2.10] Abdélah Balmane, Sélection D'attributs Par Dimension Fractal, décembre 2007.

[2.11] Mémoire de notification et commande des systèmes non linéaires par LEMMOU Amira- BELLAKHDAR Khaoukha- LEDJEDEL Adila
université de M'Sila Algérie - Ingénieur en électronique 2011

[3.1] Weka Tutoriel - logiciel de machine learning - Kongakura_files, le 25-06-2019.

[3.2]. Weka 2009.pdf .apprentissage à partir d'exemple, janvier 2009.

[3.3] Fichier internet (A1 Weka) crée par Annexe : Weka A. Zemmari, La BRI- Université de Bordeaux ,Jan. 09, 2018.

[3.4]Présentation du logiciel de datamining Weka (-pentaho) v3.8 ,26 octobre 2016, Auteurs : Eric Yabas; Manel Merad ; Marc Holzwarth;
Charlemagne Adechina.

[3.5] Université de Québec À Montréal, Département d'Informatique ; Document d'utilisation : Environnement Weka H. Yazid
Dirigé par : Prof. Hakim Lounis 26/01/2006

[3.6] Etude comparative des principaux outils open source de datamining, mémoire de fin d'études pour l'obtention du diplôme de master en informatique, option : ingénierie des systèmes d'information, réalisé par Chaouch Sadek, Debbab Abbou Bakr, 2014-2015.

ملخص:

تعد تقنية التنقيب عن البيانات أساسية للحصول على المعلومات المفيدة ، وتصنيف السمات واختيارها إلزامي بالإضافة إلى انجاز النماذج المناسبة لاتخاذ قرار جيد بعيدًا عن أي بيانات غير مهمة و متكررة ، ليست ذات صلة ويفضل أن تكون ذات أبعاد صغيرة .

استخدمنا برنامج WEKA المعتمد على JAVA لتنفيذ الخوارزميات من الطرق (Filtre, Wrapper, Hybride) **لعرض النتائج.**

بالنسبة لأنظمة إدارة قواعد البيانات ، استخدمنا القواعد Lungcancer, Lymphography et Tumor

Résumé :

La technique de fouille de données est primordiale pour l'obtention de l'information utile, la classification et la sélection des attributs sont obligatoires ainsi que la réalisation des modèles adéquats pour la prise d'une bonne décision loin de toute donnée non significative, redondante, non pertinente et avec une petite dimensionnalité.

Nous avons utilisé le logiciel WEKA basé sur JAVA comme nous avons expérimenté les algorithmes des méthodes Filtre, Wrapper, Hybride pour diffuser les résultats.

Pour les systèmes de gestion de base de données, nous avons utilisé Lungcancer, Lymphography et Tumor.

Abstract:

The data mining technique is essential for obtaining useful information, the classification and selection of attributes are mandatory as well as the realization of suitable models for making a good decision far from any insignificant, redundant data. , irrelevant and with a small dimensionality.

We used the WEKA software based on JAVA as we experimented with the algorithms of the Filter, Wrapper, Hybrid methods to disseminate the results.

For the database management systems we used Lungcancer, Lymphography and Tumor.