REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed Khider – BISKRA
Faculté des Sciences Exactes, des Sciences de la Nature et de la Vie
Département d'informatique

N° d'ordre : SIOD15/M2/2018

Mémoire

Présenté pour obtenir le diplôme de master académique en

Informatique

Parcours : **Système Information Optimisation Décisionnelle**

La Classification SVM des données bio-puces

Par:

SAOULI MOHAMED ABDELAZIZ

Soutenu le 25/06/2018, devant le jury composé de :

Dr.DJEFFAL Abdelhamid	MCA	Président
Mme.BELLOUNAR Saliha	MAB	Rapporteur
Mme.BOUGTITICHE Amina	MAA	Examinateur

Remerciement

Je remercie tous d’abord notre seigneur Dieu tout puissant pour m’avoir donné le courage et la santé pour accomplir ce modeste travaille

Ensuite je tiens à présenter mon sincère gratitude à SALIHA BELLOUNAR pour son dévouement personnel son encadrement, son aide précieuse, et sa patience porté à l’égard de mon projet

Je tiens à remercier les jurys pour l’intérêt qu’ils ont bien voulu porter à ce travail en l’honorant par leurs présences et leurs évaluations

Et pour finir je tiens a exprimé mon reconnaissance à ma famille, amis et collègues respective pour leur soutien inconditionnel.

Dédicace

A mes chers parents, pour tous les sacrifices consentis, leur tendresse, leur soutien et leurs prières tout au long de mes études pour me permettre d'atteindre cette étape de ma vie.

A mes chers frères, pour ses encouragements permanents et ces soutiens morals.

A toute ma famille pour leur soutien tout au long de mon parcours universitaire, que ce travail soit l'accomplissement de vos vœux tant allégués.

A tous mes amis Titaouine Aboubaker, Barnaoui Taha, Cherif Mohamed, Amine Khelifa,Saouli Lazhar Amine, Hfayed Fayz, Okba Khelil, Badran Khelil, Ahriz Mohamed Bachir,Youcef Mokhtari,Salim Nouar merci d'être toujours là pour moi.

Je dédie ce travail.

La table des matières

La table des figures

Introduction

Introduction générale

- **Contexte de travail**

Le domaine de la bioinformatique suscite depuis déjà plusieurs années un intérêt très Grand dans la communauté scientifique car il ouvre des perspectives très riches pour la compréhension des phénomènes biologiques. Les problèmes abordés concernent par exemple le séquençage du génome, la modélisation de la structure des protéines, ou la reconstruction d'arbres phylogénétiques (phylogénie). Ces problèmes nécessitent la collaboration entre biologistes et informaticiens car les problèmes à traiter posent souvent de grandes difficultés algorithmiques.

Dans ce mémoire nous abordons un problème de bioinformatique qui est celui de la classification de données de biopuces. La technologie des puces à ADN permet de différencier des tissus tumoraux et des tissus sains à partir de la mesure simultanée d'un grand nombre de gènes au sein d'un échantillon biologique. Pour cette tâche de classification, on dispose d'un faible nombre d'échantillons alors que chaque échantillon est décrit par un très grand nombre de gènes.

Le traitement de ces données nécessite donc de réduire le nombre de gènes pour proposer un sous-ensemble de gènes pertinents et de construire un classiffieur prédisant le type de tumeur qui caractérise un échantillon cellulaire. Il s'agit d'un problème de sélection d'attributs.

La sélection d'attributs est un problème complexe qui a déjà été largement étudié, mais les dimensions des données des biopuces nécessitent des approches spécifiques (plusieurs milliers de gènes).

Dans ce mémoire nous intéressons au développement d'une méthode de la classification des données d'expression qui consiste à regrouper les gènes sur la base de leur profil. Les méthodes que nous proposons utilisent des informations spécifiques au problème de la classification pour proposer des solutions efficaces.

Plan du mémoire

Ce mémoire est composé de cinq chapitres dont nous présentons une brève description.

Dans les paragraphes suivants :

- **Chapitre 1 :** aborde les concepts et les principes de base de la biologie, bioinformatique et de la technologie de puces à ADN.
- **Chapitre 2 :** est consacré à la classification, dont nous avons décrit la classification, ensuite les différentes méthodes de classification avec une étude comparative entre ces méthodes.
- **Chapitre 3 :** concerne la méthode de classification SVM, et les différentes versions de cet algorithme.
- **Chapitre 4 :** est consacré à la conception de notre système, nous avons donc présenté la conception global ensuite nous avons détaillé les étapes de conception
- **Chapitre 5 :** Pour finir on va montrer l'implémentation de notre application qui contient des captures d'écran, et guide d'utilisation d'application.

Chapitre I :

La bioinformatique et les bio-puces

I.1.Introduction

Au cours de ces trente dernières années, la récolte de données en biologie a connu un boom quantitatif grâce notamment au développement de nouveaux moyens techniques servant à comprendre l'ADN et d'autres composants d'organismes vivants. Pour analyser ces données, plus nombreuses et plus complexes aussi, les scientifiques se sont tournés vers les nouvelles technologies de l'information. L'immense capacité de stockage et d'analyse des données qu'offre l'informatique leur a permis de gagner en puissance pour leurs recherches. Et la rencontre entre la biologie et l'informatique, c'est ce qu'on appelle la bioinformatique. Celle-ci couvre des disciplines des sciences de la vie telles que la génomique et la biologie des systèmes.

I.2.Bioinformatique

I.2.1.Définition de la Bioinformatique

Le domaine de la Bioinformatique est défini de diverses manières.

Le centre américain pour les informations biotechnologiques définit la Bioinformatique comme étant l'union de la biologie, des mathématiques et des technologies de l'information [1].

" Georgia Institute of Technology" (USA) voit la Bioinformatique comme l'intégration des méthodes mathématiques, statistiques et informatiques pour analyser les données biologiques, biochimiques et biophysiques [2].

Le National Institute of Health (NIH), définit la Bioinformatique comme étant la recherche, le développement ou l'application d'outils et d'approches informatiques pour améliorer le traitement des données biologique et médicales [2].

Selon Stanford University (USA), la Bioinformatique est la création et le développement d'informations avancées et de technologies informatiques pour étudier les problèmes biologiques. Elle inclut également les méthodes de stockage, de récupération et d'analyse des données biologiques, telles que les séquences d'acides nucléiques (ADN / ARN), les protéines, les structures, les fonctions, les voies et les interactions génétiques [3].

Jean-Michel Claverie décrit la Bioinformatique comme étant l'ensemble des concepts et des techniques nécessaires à l'interprétation de l'information génétique et structurale. C'est le décryptage de la "bioinformation". Son but est d'élaborer la synthèse des données disponibles, d'énoncer des hypothèses généralisatrices (comment les protéines se replient ou comment les espèces évoluent…etc.), et de formuler des prédictions (localiser ou prédire la fonction d'un gène) [4].

La définition de synthèse indique que la Bioinformatique est une discipline scientifique multidisciplinaire (Figure I-1). Elle consiste à l'analyse et l'interprétation des données biologiques en utilisant des approches, des concepts et des formalismes issus de la biologie, de l'informatique, des mathématiques, des statistiques et de la physique pour le développement d'infrastructures et d'outils, tels que des algorithmes, des modèles statistiques ou des logiciels de bases de données. Ces outils aident à produire de nouvelles connaissances et servent à résoudre des problèmes en biologie.

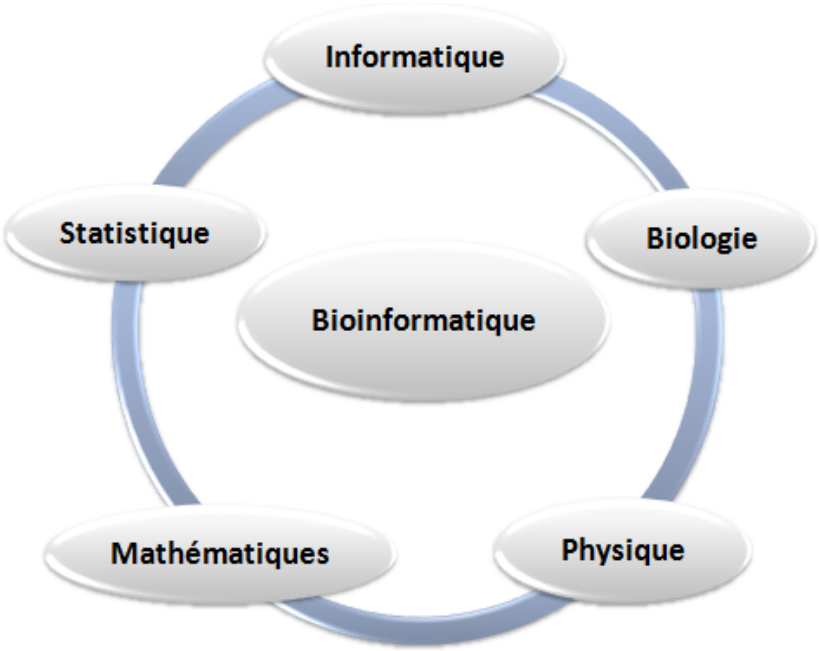


Figure I-1 : L'interaction des disciplines construant en bioinformatique.

I.3.La génomique

La génomique est une discipline récente de la biologie qui étudie le fonctionnement d'un organisme à l'échelle du génome.

Dans cette partie nous présentons le contexte biologique et la technologie des puces à ADN afin de mieux comprendre la nature des données sur lesquelles nous travaillons. Après quelques rappels de génétique, nous montrons les enjeux de l'étude du transcriptome qui ont motivés notre travail de recherche, en mettant en évidence l'importance de l'information pour l'analyse des données issues de cette technologie.

I.3.1.La cellule

C'est la petite unité structurale et fonctionnelle de tous les êtres vivants, il existe des milliers de type de cellules différents par leur forme, leur taille, leur fonction et leur comportement.

I.3.2.Acide désoxyribonucléique (ADN)

L'ADN est la forme de stockage de l'information génétique de tous les êtres vivants. C'est une molécule gigantesque situé dans le noyau de chacune des cellules, elle se présente sous la forme d'un double brin enroulé en double hélice composé d'un enchainement linéaire de millions d'unité de quartes bases azotées (nucléotide) : adénine(A), cytosine(C), guanine(G) et thymine(T) [5] (Figure I-2).

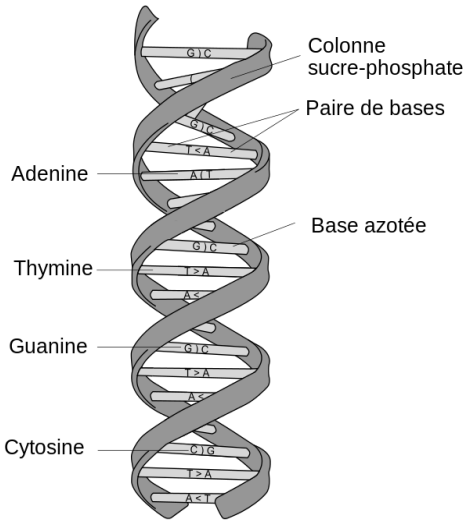


Figure I-2:Structure d'une molécule d'ADN.

I.3.3.Structure primaire de l'ADN

L'acide nucléique ADN est une macromolécule formée de nucléotides, il est constitué de deux chaines (polymère). Ces deux chaines sont appelées « brins d'ADN ».

Les bases azotées des deux brins sont reliées par des liaisons faibles, ils sont associées par paire, dans chaque pair il y a toujours une purine associée à pyrimidine :

- Les bases A sont associées aux bases T.
- Les bases G sont associées aux bases C.

Les deux brins sont complémentaires l'un de l'autres mais ne sont identiques. En connaissant la séquence de l'un on peut déduire la séquence de l'autre. Chaque brin d'ADN ressemble à une hélice, les deux brins forment alors une structure en « double hélice ».

I.3.4.Gènes

En génétique, est une unité de base d'hérédité qui en principe prédétermine un trait précis de la forme d'un organisme vivant, tel que défini en 1909 par Wilhelm Johannsen. Au niveau physique, un gène est un fragment ou locus déterminé d'une séquence d'ADN [6].

I.3.5.Génome

Le génome est l'ensemble du matériel génétique d'une espèce codé dans son acide désoxyribonucléique (ADN) à l'exception de certains virus dont le génome est constitué d'ARN. Il contient en particulier tous les gènes codant des protéines ou correspondant à des ARN structurés [7].

I.3.6. Acide ribonucléique (L'ARN messager)

L'ARNm est comme l'ADN une chaine composée de nucléotide, ou l'Uracile(U) remplace la Thymine, c'est une photocopie de l'ADN produite durant la phase de transcription. En effet pour chaque gène, L'ADN qui le compose est transcrit en une molécule complémentaire qui est l'ARN. Le noyau de la cellule est comme une bibliothèque ou sont stockées les informations. Et comme toutes bibliothèque, les originaux ne peuvent pas sortir donc l'ADN est transcrit en un nombre variable d'exemplaire d'ARN messager sans jamais

quitter le noyau de la cellule, c'est l'ARN qui transport l'information génétique à travers le reste de la cellule [5].

I.3.7.Transcriptome

On appelle transcriptome l'ensemble des ARN messagers, c'est-à-dire les milliers de transcriptions d'ADN différente présentes dans une cellule à un moment donné [5].

I.3.8.La méthode PCR

La PCR (Polymérase Chain Réaction) est une technique d'amplification d'ADN. Elle permet d'obtenir un très grand nombre de copies d'une séquence d'ADN choisie [8].

I.4.Les puces à ADN

I.4.1.Introduction

Le transcriptome est la partie du génome transcrite en ARN, et en particulier en ARN messager. Il permet de définir un niveau d'expression pour chacun des gènes dans une cellule donnée et à un temps donné. Des techniques telles que les puces à ADN offrent la possibilité d'étudier les variations d'expression de milliers de gènes simultanément.

Dans la suite de cette section, nous détaillons le fonctionnement de cette biotechnologie [9].

I.4.2.Les puces à ADN

Les puces à ADN, appelées DNA chips ou microarrays (array =rang ordonné), ont été développées au début des années 1990. Depuis leur apparition, les puces à ADN sont devenues des outils majeurs pour la recherche en biologie fondamental. [10][11].

Elle est une technique qui permet d'étudier le transcriptome par l'observation simultanée de l'expression de plusieurs milliers de gènes dans une cellule ou un tissu donné, mesurant ainsi les modifications des différents états cellulaires. La technique des puces à ADN est basée sur le principe d'hybridation qui stipule que deux fragments d'acides nucléiques complémentaires peuvent s'associer et se dissocier de façon réversible sous l'action de la chaleur et de la concentration saline du milieu. Concrètement, une puce à ADN est un support rigide (verre ou nylon) de quelques centimètres carrés, sur lequel de courtes séquences d'ADN ont été déposées.

Ces courtes séquences sont nommées des “sondes” correspondant à des oligonucléotides de synthèse ou à des produits de PCR. Les sondes ont la particularité d’avoir été choisies de manière à être spécifique d’un seul et unique gène. Ce microdispositif est mis au contact des ARNm extraits des échantillons à analyser appelés des “cibles”. Ces cibles sont marquées par incorporation de radioéléments ou de fluorochromes. Après acquisition des images d’hybridation, La quantification des signaux d’hybridation reflète le niveau d’expression, dans l’échantillon initial, de chacun des gènes représentés sur la puce (Figure I-3).

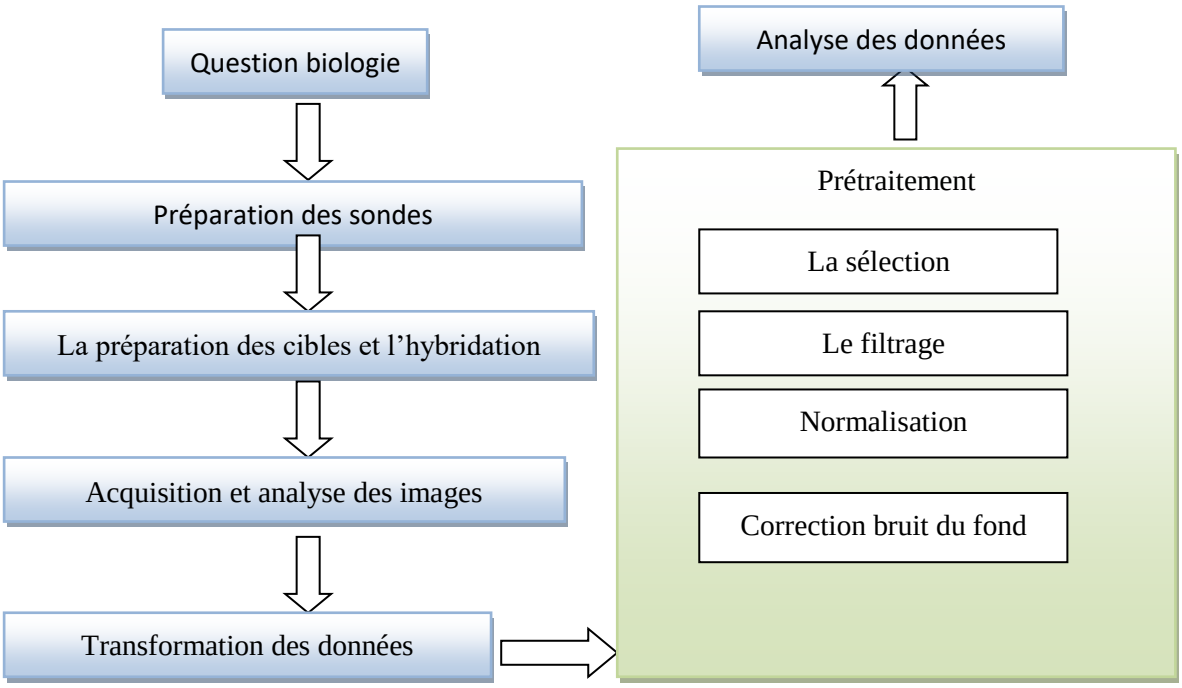


Figure I-3: Les étapes d’une analyse par puces à ADN

I.4.3.Les étapes d’une analyse à puces ADN

I.4.3.1.La préparation des sondes

Selon l’application désirée et le modèle biologique étudié, plusieurs types de biopuces peuvent être utilisés :

- Etude du transcriptome : biopuces ADNc ou oligonucléotides.
- Diagnostic : biopuces oligonucléotides.
- Génomique comparative : biopuces ADN génomique.

Dans le cadre d’une étude transcriptome, les sondes peuvent être les séquences d’ADN double brin d’une taille en moyenne de 400 paire de bases (ADNc produits PCR) ou simple brin [12].

La conception des biopuces de type ADNc nécessite une première étape de l'obtention des sondes grâces à une amplification PCR (polymérase Chain réaction) à partie d'une collection d'ADNc ou directement sur l'ADN génomique.

Un contrôle de ces produits d'amplification doit être effectué (taille, séquence) pour être certain qu'il s'agit de la bonne sonde. L'ADN des sondes est dénaturé pour le dépôt en simple brin pour permettre par la suite une hybridation avec les cibles marquées. Les sondes ainsi préparées sont déposes sur un robot (le spotter) sur le substrat de la puce, recouvert d'une fine couche de polymère permettant la fixation des sondes par des liaisons électrostatique. Le dépôt, ou spot (emplacement sur la puce), est réalisé à l'aide d'aguilles creuses ou par la technique de l'anneau pin and ring. Le diamètre des spots peut varier de 80 à 300µm, et la distance entre deux dépôts consécutifs sur la lame est de l'ordre de 250 µm dans les deux directions [12].

I.4.3.2.La préparation des cibles et l'hybridation

Des milliers de transcrits différente sont présents dans les cellules à un moment donné, et leur abondance relative est révélatrice de l'activité cellulaire a cet instant. La préparation des cibles consiste tout d'abord d'extraire du milieu cellulaire ces transcrits puis leur intégrer un marqueur fluorescent (rouge ou vert) qui permettre d'évaluer et de quantifier l'appariement sonde /cible. Le marquage, soit direct (incorporation d'un nucléotide fluorescent), soit indirect, se fait suite à une transcription inverse des ARNm permettant l'obtention d'un brin d'ADNc fluorescent. Du fait de la complémentarité des nucléotides, le dépôt des cibles marquées sur la puce déclenche l'appariement des séquences sondes/cibles complémentaire. Cette hybridation, qui dure quelques heures en milieu liquide, est suivie d'un lavage du substrat qui permet d'éliminer les cibles non fixées, ou fixées non spécifiquement. Après séchage la puce est passée au scanner pour repérer les hybridations [12] (Figure I-4).

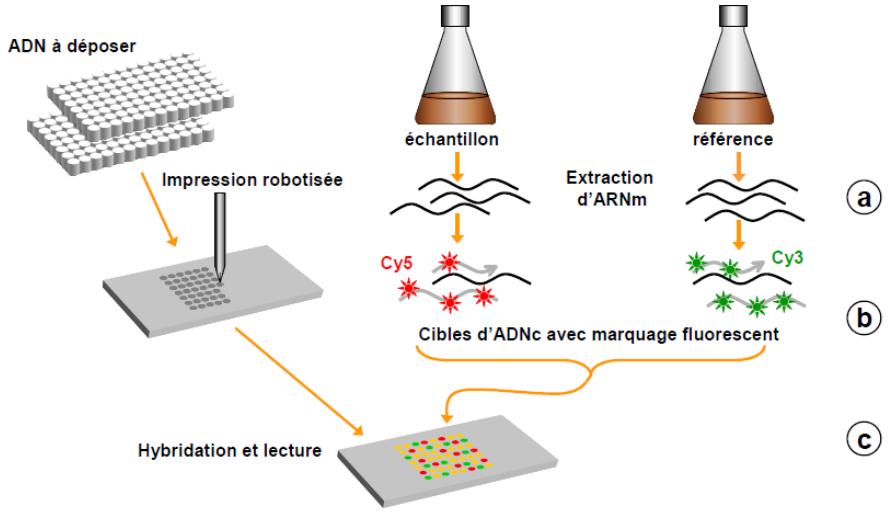


Figure I-4:étape d'hybridation

I.4.3.3.Acquisition et analyse des images

Suite à l'hybridation, une étape de lecture de la puce permet de repérer les sondes ayant réagi avec l'échantillon testé. Cette lecture est une étape clé [13]. En effet, sa qualité conditionne de façon importante la précision des données et donc, la pertinence des interprétations. L'obtention des images est réalisée par lecture des puces sur des scanners de haute précision, adaptés aux marqueurs utilisés. Le procédé de détection combine deux lasers, pour exciter les fluorochromes Cy3 et Cy5. On obtient alors deux images dont le niveau de gris représente l'intensité de la fluorescence lue. Si on remplace les niveaux de gris par des niveaux de vert pour la première image et des niveaux de rouge pour la seconde, on obtient en Les superposant une image en fausses couleurs composée de spots allant du vert au rouge quand un des fluorophores domine, en passant par le jaune (même intensité pour les deux fluorophores). Le noir symbolise l'absence de signal. L'intensité du signal de fluorescence pour chaque couple (gène, spot) est proportionnelle à l'intensité d'hybridation donc à l'expression du gène ciblé. Les images sont traitées par des logiciels d'analyse qui permettent de mesurer la fluorescence de chaque spot sur la lame (estimant les niveaux d'expression pour chacun des gènes présents sur la puce), mais aussi de relier chaque sonde à l'annotation correspondante (nom de gêne, numéro de l'ADNc utilisé, séquence de l'oligonucléotide, etc.). Ainsi, pour chaque spot, l'intensité de chaque marqueur est calculée puis comparée au bruit de fond (Figure I-5).

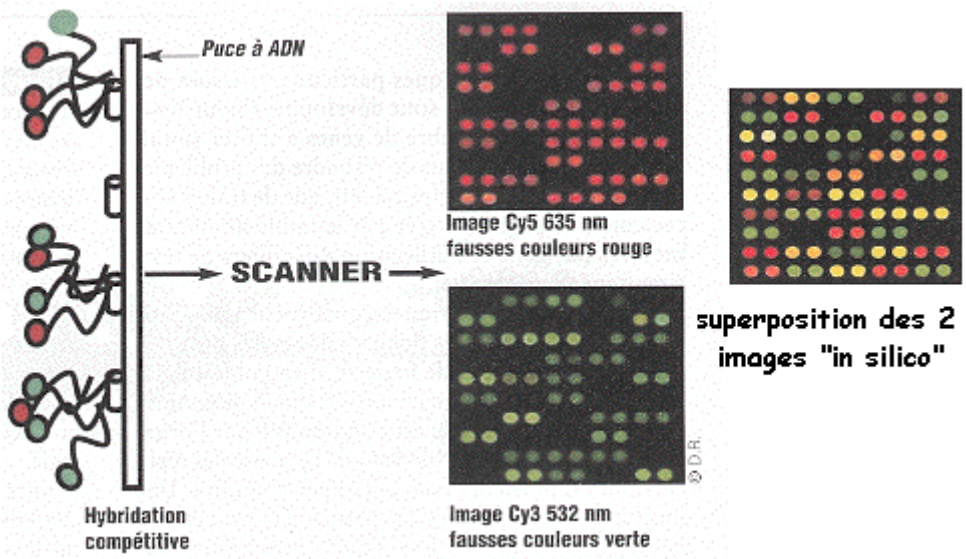


Figure I-5:acquisition du l'image par le scanner

I.4.3.4.Transformation des données

Les rapports des intensités de fluorescences en rouge et vert sont généralement utilisés pour mesurer une variation d'expression d'un gène entre deux conditions (référence et pathologique, par exemple). Les données d'intensité sont rarement manipulées sans transformation et la transformation la plus couramment employée est celle qui utilise le logarithme à base deux. Il existe plusieurs raisons pour justifier cette transformation. D'une part, la variation du logarithme des intensités est moins dépendante de la grandeur des intensités, et d'autre part, cette transformation permet de se rapprocher d'une distribution symétrique et d'obtenir une meilleure dispersion avec moins de valeurs extrêmes.

La normalisation consiste à ajuster l'intensité globale des images acquises sur chacun des deux canaux rouge et vert, de manière à corriger les différences systématiques entre les échantillons sur la même lame, qui ne représentent pas de variations biologiques entre les échantillons et qui tendent à déséquilibrer le signal de l'un des canaux par rapport à l'autre.

Cette procédure de normalisation est définie par les gènes de référence. Les gènes de référence en moyenne ne doivent pas changer d'expression entre deux conditions. La normalisation est effectuée à partir de toutes les sondes présentes sur le support pour éliminer les différences entre les différentes puces liées aux variations de quantité de départ, aux biais de marquage ou d'hybridation et aux variations du bruit de fond [14].

I.4.4.Plateformes

Il existe actuellement deux types de puces à ADN qui dominent le marché :

- Les puces à ADNc qui fonctionnent avec des micros points contenant des fragments d'ADN sur un support de verre. La société Agilent est l'une des plus grandes industries qui les commercialisent.
- Les puces à oligonucléotides qui reposent sur le principe de synthèse in situ de milliers de séquences distinctes d'oligonucléotides. La société Affymetrix est l'unique détenteur de cette technologie [15].

I.4.4.1.Technologie Agilent

Les puces à ADNc de la technologie Agilent ont été les premières puces à être développées. Le pionnier en la matière fut Patrick Brown et ses associés de l'université de Stanford. Elles sont construites grâce à des machines robots qui déposent des points appelés spots contenant des fragments d'ADN (50- 150 m) dans une lamelle de verre.

I.4.4.2.Technologie Affymetrix

Elles dérivent à l'origine d'un projet de séquençage par hybridation. Les sondes sont des oligonucléotides synthétisés par une technique de photolithographie. Cette technique consiste à diriger une lumière sur des sites spécifiques de la puce ce qui active la réaction d'oligosynthèse. On ajoute également des oligonucléotides dont la séquence varie pour une seule base pour confirmer que le signal obtenu pour chacun des gènes est bien spécifique. On hybride une seule expérience par puce et l'intensité de fluorescence mesurée par un scanner permet une mesure de l'abondance relative de chacun des ARNm présent dans l'échantillon biologique étudié [16].

I.4.5.Gestion et partages des données

La mise en œuvre, au sein d'un laboratoire, de la technologie des puces à ADN requiert la mise en place de moyens pour la gestion des données : il faut assurer le stockage, la sécurisation et la manipulation de ces données [17].

En effet, la technologie, qualifiée de « haut débit », engendre très rapidement de larges volumes de données à traiter. De plus, ces données sont de différents types, elles concernent à la fois les résultats : les images acquises par le scanner, les tableaux de données brutes et

transformées mais également les informations sur les différentes étapes des expériences menées : procédés d'obtention des échantillons et des lames, conditions d'hybridation etc...[17].

Le besoin en bases de données s'est très rapidement accompagné d'un besoin en définition de standards afin de normaliser et ainsi de pouvoir échanger et diffuser les données issues de la technologie des puces à ADN entre les différents laboratoires [17].

I.4.5.1.Microarray gène expression data society (La MGED)

C'est la société MGED (Microarray Gene Expression Data), organisation de biologistes et d'informaticiens développée sous l'influence d'Alvis Brazma et d'Alan Robinson de l'EBI, qui a établi ces standards d'annotation d'expériences de puce à ADN [17].

La MGED a initié le développement et la promotion de standard pour le stockage et le partage des données de puces à ADN basées sur l'expression des gènes et du résultat des études effectuées sur ces données. Parmi ces standards l'on peut citer le MIAME (Minimum Information About a Microarray Experiment).

Le MIAME est un standard conceptuel décrivant l'information minimum requise pour une interprétation et une vérification propre des expériences des puces à ADN tandis que MAGEML et MAGE-TAB sont des standards définissant le format MIAME, Le projet MIAME vise à décrire les informations minimales et nécessaires que les chercheurs doivent fournir pour la description d'expériences de type puce à ADN. Dans la pratique, ces informations sont requises pour toute publication [17].

I.4.5.2.Gene Expression Omnibus (GEO)

Gene Expression Omnibus [19] est un entrepôt public à haute capacité de traitement des données génomique et protéomique, essentiellement MIAME. Il a été établi en 2000 au National Center for Biotechnology Information (NCBI). Les données expérimentales peuvent être soumises en remplissant un formulaire sur le web ou comme un paquet de fichiers, feuille de calcul, fichier texte SOFT (Simple Omnibus Format in Text) ou fichier XML MINiML(MIAME Notation in Markup Language).

Les fichiers sont stockés sous la forme de 3 types d'enregistrement basiques :

- Plateform : Description du tableau
- Sample : Description d'un échantillon biologique et les résultats de son hybridation
- Series : Description de l'expérience réalisée sur un groupe d'échantillon Basées sur les

études expérimentales soumises, les données dans GEO sont organisées dans des objets de haut niveau représentés par le type Dataset (Jeu de données), qui est une collection d'échantillons biologiques comparables ayant été traités sur la même plateforme et dont les mesures sont les résultats de ce traitement et de calculs cohérents sur ce jeu de données, et Profils, qui correspondent au niveau d'expression d'un gène dans tous les échantillons d'un jeu de données [18].

I.4.6.Prétraitement des données et Normalisation

La technique utilisée avec les puces à ADN est soumise à de nombreuses variations expérimentales qui rendent impossible l'exploitation directe des résultats. Pour ne garder que Les variations réelles entre les différents échantillons dues aux différences de traitement qu'ils sont subis, le seul moyen est de procéder à un prétraitement des données et une normalisation des données pour éliminer ces différences. Cette étape permet d'adapter les données au type d'analyse souhaité [19].

I.4.6.1.Etapes du prétraitement des données

I.4.6.1.1 Correction du bruit de fond (Background Correction)

Après l'hybridation, une puce à ADN est scannée pour pouvoir générer des fichiers où les résultats de l'hybridation sont traduits numériquement (Fichiers CEL). On obtient dans ces fichiers une quantité énorme d'information. On a pour chaque gène : la moyenne des intensités de tous les pixels sur la zone correspondante au gène, la médiane de ces intensités, l'écart-type de ces intensités et le nombre de pixels dans la zone considérée.

Différentes méthodes ont été proposées pour cette étape (Correction par Robust Multi-Array Analysis (RMA), Correction par Gene Chip RMA (GCRMA)) [19].

I.4.6.1.2 Normalisation

Il est nécessaire d'effectuer une normalisation afin de s'assurer que les différences observées dans les intensités sont dues à des différences réelles d'expression et non à des artefacts expérimentaux. Lors de la fabrication de puces à ADN, les sources de variabilité sont nombreuses. On peut citer, l'amplification des sondes par la technique PCR et leur positionnement sur la puce, l'hybridation sonde/cible, le nettoyage et le séchage de puces ect…

Le but de la normalisation est de corriger les différences systématiques entre les mesures sur la même puce qui ne représentent pas de véritables variations biologiques. Elle permet la comparaison de plusieurs réplicas d'une même expérience et se focalise sur les erreurs systématiques, qui contribuent à sur ou sous évaluer les valeurs mesurées, plutôt que sur les erreurs stochastiques [19].

Dans le cas des puces à oligonucléotides [20] (Normalisation puces Affymetrix), comme les puces Affymetrix, la normalisation est réalisée entre des répétitions de lames ou l'ensemble des lames d'une ou de plusieurs expériences. On parle souvent de normalization between array. La normalisation la plus utilisée est la normalisation des quantiles.

I.4.6.1.2 .1Normalisation des quantiles (quantiles normalization)

Pour cela, il existe une méthode complète dite de centralisation permettant à la fois de normaliser et de calibrer les données de façon à permettre les comparaisons inter-lames.

Cette méthode non paramétrique appelée aussi " normalisation des quantiles " suppose que la distribution de l'abondance des gènes est presque la même dans tous les échantillons [20].

L'algorithme comporte plusieurs étapes :

On trie les gènes par colonnes selon leurs intensités.

1. On calcule la moyenne de chaque ligne.
2. On remplace les valeurs de chaque élément ligne par la moyenne correspondante.
3. On redistribue les valeurs nouvelles selon l'ordre d'origine des intensités.

I.4.6.1.3 Le filtrage

Le filtrage a pour but de supprimer les gènes dont l'expression ne varie pas ou peu, dans une série d'expérience. Ces gènes ne présentent pas d'intérêt pour l'analyse et leur élimination permettra de simplifier les classifications ultérieures, cette procédure permet généralement de soustraire tous les gènes ayant une faible variation, dont l'intensité est trop proche du bruit de fond ou pour lesquels on suspecte une hybridation non spécifique [21].

I.4.6.1.4 Sélection des attributs pour traitement des données

Le principe de la sélection des attributs consiste à évaluer chaque attribut pour lui assigner un score de pertinence qui permet un classement des attributs. Les attributs les mieux classés c'est-à-dire les plus pertinents seront sélectionnés pour la phase du traitement.

L'avantage de la sélection est qu'elle peut être utilisée lorsqu'on travaille avec un très grand nombre d'attributs car elles sont de complexité raisonnable [19].

I.4.6.1.4.1 T-test

Le test t de student est un test paramétrique permettant de comparer deux groupes d'échantillons. Il permet d'identifier des gènes exprimés différentiellement sur les niveaux des intensités d'hybridation.

$$t = \frac{m_1 - m_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$m_1$et m_2 : moyennes des intensités des signaux d'un gène donné dans chaque condition 1 et 2.

S_1^2et S_2^2: variances des intensités des signaux d'un gène donné dans chaque condition 1 et 2.

$n_1$et n_2: nombre d'analyses pour les conditions 1 et 2 correspondant au nombre de réseaux (variabilité technique) ou au nombre d'individus (variabilité biologique), analysés par condition.

I.4.6.1.4.2 SAM

C'est un outil statistique développé par l'université de Stanford [22] dans le but d'analyser des résultats provenant de puce à ADN. Pour identifier les gènes différentiellement exprimés entre 2 échantillons, il implémente une version modifiée du test statistique t pour un gène particulier en attribuant à ce gène une note d (pour "relative différence")

$$d = \frac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} + S_0}$$

Il diffère du test statistique t pour les variances égales, par l'ajout de la constante S_0 pour minimiser la variation de *d*. Il permet de ne pas favoriser les gènes qui auraient une petite variance et du même coup une valeur de *d* grande.

I.4.7.Présentation des données de puces à ADN

Après le prétraitement des données et normalisation [19], les donné recueillies pour l'étude d'un problème donné sont regroupées sous forme de matrice avec une ligne par couple (gène, sonde) et une colonne par échantillon Chaque valeur de m_{ij} est la mesure du niveau d'expression du i−ème gène dans le j−èm eéchantillon, où i = 1, . .. , M et j = 1, . . . , N(Figure I-6).

$Gène_{id}$	$échantillon_1$	$échantillon_2$	……..	$échantillon_j$
Gène 1			……	
Gène 2			……	
Gène 3			……	
:			…….	
Gène N	M_{11}	M_{22}	……	M_{ij}

Figure I-6:Matrice d'expression des gènes

I.4.8.Analyse des données

La dernière étape d'une expérience de puce à ADN est certainement la plus Difficile. Il s'agit d'exploiter les meilleurs de valeurs numériques produites afin d'extraire une information biologique pertinente.

Il existe très grand nombre de méthode d'analyse de données. Permis elles se trouvent les techniques classique de classification supervisée ou non, ainsi que de nombreuses méthodes issues d'autres disciplines, transposées aux données de puces à ADN [23].

I.5.Conclusion

Nous avons présenté dans ce chapitre des notions élémentaires en biologie et la notion du bioinformatique, qui sont les bases de notre sujet de recherche dans ce mémoire. Ensuite Nous avons présenté les différentes étapes d'une analyse par puce à ADN, telles que la préparation des cibles et l'hybridation, acquisition et analyse des images et transformation des données.

Nous avons aussi présenté les différentes banques de données génomiques publiques et différentes étapes du prétraitement des données y compris : la correction du bruit de fond et

la normalisation, le filtrage et la sélection. Dans le chapitre suivant nous allons parler sur méthodes de la classification qui permet d'analysé les données bio-puces.

Chapitre II : Classification

II.1.Introduction

La classification est une activité mentale qui intervient fréquemment dans la vie courante. En effet, les objets sont souvent répertoriés par rapport à des classes ou des catégories auxquelles ils sont censés appartenir. Cette appartenance est, la plupart du temps, vague et/ou graduelle. Des modificateurs linguistiques tels que très, trop, assez, insuffisamment, traduisent l'incertitude que l'on peut rencontrer concernant les caractéristiques ou les propriétés de certaines entités.

II.2.Classification

La classification repose sur des objets à classer. Les objets sont localisés dans un espace de variables. Il s'agit de les localiser dans un espace de classes [19].

II.3.Les méthodes de classification

Les problèmes de classification s'attachent à déterminer des procédures permettant d'associer une classe à un objet (individu). On peut classer les méthodes de classification en deux grandes classes : la classification supervisée et la classification non supervisée.

- On appelle classification automatique, ou non supervisée, un ensemble de problématique où l'espace des classes n'est pas spécifié à l'avance.il s'agit d'identifier, voire de construire, un système de classes sur la base d'observations dans l'espace des caractéristique [19].
- On appelle classification supervisée un contexte où un ensemble de classes est spécifié à l'avance.

II.4.La classification non supervisée

On distingue classiquement deux grandes familles de méthodes en clustering :

- Les méthodes hiérarchiques
- Les méthodes par partition.

II.4.1.Le clustering hiérarchique

Quant aux méthodes hiérarchiques, elles sont ascendantes ou descendantes. La classification hiérarchique consiste à construire un arbre de classes appelé dendrogramme.

Cette construction se fait par deux manières ascendante et descendante. Les algorithmes ascendants ou encore agglomératifs considèrent chaque objet de l'ensemble de données comme des classes initiales et, à chaque étape, on fusionne deux classes qui optimisent un critère de similarité [24]. A l'inverse, les méthodes descendantes ou divisives partent d'une classe unique formée de tous les objets et le divisent successivement les clusters de manière à ce que les classes résultants soient les plus différentes possibles, et ce jusqu'à obtenir autant de classe que d'objets dans la base.

II.4.2.Le clustering par partitionnement

Les algorithmes de partitionnement construisent directement, en sortie, une partition de l'espace des objets en K classes. Le principe général, par définition d'une partition, cela signifie que chaque classe doit contenir au mois un objet. Et que chaque objet doit appartenir à une classe unique.

Pour ce faire, étant donné le nombre de classes K requises, ces algorithmes génèrent une partition initiale, puis cherchent à l'améliorer en réattribuant les objets d'une classe à une autre.

II.4.2.1.Les méthodes des k-moyennes

L'algorithme k-means, ou l'algorithme k-moyenne est aucun doute la méthode de partitionnement les plus connue et la plus utilisée dans divers domine d'application scientifique ou industrielles. L'algorithme consiste à sélectionner aléatoirement K objets qui représente les controides initiaux. Un objet est assigné au cluster pour lequel la distance entre l'objet et le centroide est minimale [25].

II.4.2.2.L'algorithme des centres mobiles

Le principe de l'algorithme consiste à construire une partition en K classes en sélectionnant K objets au hasard de l'ensemble des objets et les considérer comme centres des classes. Après cette sélection, on affecte chaque objet au centre le plus proche en créant K classes, les centres des classes seront remplacés par des centres de gravité. Ainsi des Nouvelles classes seront crées. Généralement la partition obtenue est localement optimal car elle dépend du choix initial des centres [26].

II.4.2.3.Les méthodes des K-médoides

Les méthodes des k-médoide se différencient de la méthode des k-moyennes par l'utilisation de médoide plutôt que des centroide pour représenter les classes. Dans l'algorithme de k-médoides une classe est représentée par un de ces objets prédominants, Le médoide d'un groupe est l'objet possédant la distance médiane la plus faible avec les autres objets du groupe. C'est un algorithme itératif combinant la réaffectation des objets dans des classes avec une intervention des médoide et des autres objets [25].

II.5.La classification supervisée

L'objectif de la classification supervisée est d'apprendre à l'aide d'un ensemble d'entraînement une procédure de classification qui permet de prédire l'appartenance d'un nouvel exemple à une classe. Les systèmes d'apprentissage permettant d'obtenir une telle procédure peuvent être basés sur des hypothèses probabilistes (classifieur naïf de Bayes),sur des notions de proximité (plus proches voisins) ou sur des recherches dans des espaces d'hypothèses (arbres de décisions, . . .) Nous décrivons dans la suite quelques méthodes de classification supervisée bien connues dans la littérature[27].

II.5.1.Définition de la classification supervisée

Définition 1.1(Exemple)

Un exemple est un couple(x, y), où x∈ Xest la description ou la représentation de l'objet et y ∈Yreprésente la supervision de x [27].

Dans un problème de classification, s'appelle la classe de x. Pour la classification binaire nous utilisons typiquement X pour dénoter l'espace d'entrées tel que X⊆ $R^p et Y$ et l'espace de sortie tel que$Y = \{-1, 1\}$[1].

Définition 1.2 (classification supervisée)

Soit un ensemble d'exemples de n données étiquetées: S = $\{(x_1,y_1), \ldots, (x_n,y_n)\}$. Chaque donnée x_i est caractérisée par p attributs et par sa classe y_i[28]. On cherche une hypothèse h telle que :

(1) .h satisfait les échantillons

$$\forall_i \in \{1, \ldots, n\} h(x_i) = y_i$$

(2) .h possède de bonnes propriétés de généralisation.

Le problème de la classification consiste donc, en s'appuyant sur l'ensemble d'exemple à prédire la classe de toute nouvelle donnée x∈ R^p.

Le problème de la généralisation

L'objectif de la classification est de fournir une procédure ayant un bon pouvoir prédictif c'est-à-dire garantissant des prédictions fiables sur les nouveaux exemples qui seront soumis au système. La qualité prédictive d'un modèle peut être évaluée par le risque réel ou espérance du risque, qui mesure la probabilité de mauvaise classification d'une hypothèse h [27].

Définition 1.3 (Risque réel) Soit h une hypothèse apprise à partir d'un échantillon S d'exemples de X × Y

$$R(h) = \int_{x \in X,Y} l[h(x_i), y_i] dF(x, y)$$

Ou l l est une fonction de perte ou de coût associé aux mauvaises classifications et où l'intégrale prend en compte la distribution F de l'ensemble des exemples sur le produit cartésien de X × Y [27].

La fonction de perte la plus simple utilisée en classification est définie par :

$$l[h(x_i), y_i] = \begin{cases} 0 & si y_i = h(x_i) \\ 1 & si y_i \neq h(x_i), \end{cases}$$

La distribution des exemples est inconnue, ce qui rend impossible le calcul du risque réel.

Le système d'apprentissage n'a en fait accès qu'à l'erreur apparente (ou empirique) qui est mesurée sur l'échantillon d'apprentissage.

Définition 1.3 (Risque empirique)

Soit un ensemble d'apprentissage S = {(x1, y1),…..(x_n, y_n)} de taille n et une hypothèse h. Le risque empirique de h calculé sur S est défini par:

$$\mathrm{R_{emp}} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{l}[\mathrm{h}(\mathrm{x_i}), \mathrm{y_i}]$$

Avec la fonction de perte présentée ci-dessus, le risque empirique ou apparent est simplement le nombre d'exemples de S qui sont mal classés.

On peut montrer que, lorsque la taille de l'échantillon tend vers l'infini, le risque apparent converge, en probabilité si les éléments de S sont tirés aléatoirement- vers le risque réel [27].

Malheureusement on ne dispose que d'un échantillon limité d'exemples; le risque empirique est très optimiste et n'est pas un bon indicateur des performances prédictives de l'hypothèse h.

II.5.2.Les techniques de la classification supervisée

Pour présenter les techniques de la classification supervisée, nous avons repris la répartition formulée par Weiss et Kulikowski (Weiss & Kulikowski, 1991) qui sépare ces techniques en deux catégories :

- Les techniques statistiques.
- Les techniques d'apprentissage automatique.

Les techniques statistiques regroupent une panoplie de méthodes. Nous présentons les techniques basées sur l'apprentissage bayésien, l'analyse discriminante et la méthode du k- plus proches voisins (KNN). Dans la catégorie apprentissage automatique, nous présentons les réseaux de neurones, les arbres de décision, et les Séparateurs à Vaste Marge SVM (Support Vector Machines).

II.5.2.1.K-PPV

Appelée aussi knn (k nearestneighbor) consiste à observer les k plus proches voisins d'une nouvelle observation afin de décider de la classe d'appartenance de cette nouvelle observation. Pour une nouvelle observation à classer, cet algorithme calcule la distance de cette nouvelle observation à chaque observation pressente dans l'ensemble d'apprentissage. Cette distance est souvent la distance euclidienne alors que d'autres distances sont utilisées telle la distance tangente, distance de Manhattan etc. Ce mécanisme peut devenir extrêmement coûteux en calcul et très demandeur en termes de mémoire de stockage [28].

Avantage de K-PPV :

- Elle ne pose aucune hypothèse sur la forme des classes à apprendre.
- La méthode est simple puisqu'il n'y pas besoin d'apprentissage d'un modèle de classification [28].

Inconvénient de K-PPV

- La performance de cette méthode est simple lorsque la dimension augmente, puisque pour chaque nouvelle classification, il est nécessaire de calculer toutes les distances de x à chacun des exemples d'apprentissage.
- La performance dépend fortement de K, le nombre des voisins choisi et il est nécessaire d'avoir un grand nombre d'observation pour obtenir une bonne précision des résultats [28].

II.5.2.2.L'apprentissage bayésien (classifieur bayésien naïf*)*

Ce sont des modèles graphiques dans lequel les connaissances sont représentées sous forme de variables. Chaque variable est un nœud du graphe. Le graphe est toujours dirigé et acyclique. Il est possible de réaliser des classifieurs performants grâce aux réseaux bayésiens. Il existe plusieurs structures permettant d'employer les réseaux bayésiens comme classifieurs : réseau bayésien naïf, réseaux bayésiens naïfs augmentés par un arbre, réseaux bayésiens semi naïfs condensés etc.

L'avantage du raisonnement bayésien est de générer les données manquantes. En outre le résultat de la classification est facilement interprétable. Le choix d'un classifieur est souvent difficile et il se fait en considérant plusieurs critères [29].

II.5.2.3.Les arbres de décision

Les arbres de décision ont pour objectif la classification et la prédiction. Leur fonctionnement est basé sur un enchaînement hiérarchique de règles exprimées en langage courant. Un arbre de décision est composé d'un nœud racine par lequel entrent les données, de nœuds feuilles qui correspondent à un classement de questions et de réponses qui conditionnent la question suivante.

La mise en place d'un arbre de décision consiste à préparer les données par la suite à créer et valider l'arborescence. Il s'agit d'abord de définir la nature, le format des variables et leur méthode de traitement. Ces variables peuvent être non ordonnées ou encore continues. Dans le cas de l'existence d'une base de règles simple et limitée, la construction de l'arbre se fait en interaction avec le décideur, en validant les arborescences une à la fois jusqu'à la détermination de l'affectation. C'est un processus interactif d'induction de règles qui permet d'aboutir à une affectation bien justifiée. Mais, en général la création et la validation de l'arborescence se passe selon l'algorithme de calcul choisi. Il existe différents algorithmes

développés pour appliquer cette technique : CART, C4.5 et CHAID (Quinlan, 1993) ; (Breimann, et al., 1984) ; (Henriet, 2000).

Les avantages des arbres de décision

- leur rapidité et, surtout, leur facilité quant à l'interprétation des règles de décision
- La clarté des règles de décision facilite le dialogue homme-machine.
- Ils peuvent traiter des ensembles d'apprentissage avec des données manquantes.

Les inconvénients des arbres de décision

- les arbres de décision ont une faiblesse au niveau de la performance et le coût d'apprentissage. Ils deviennent peu performants et très complexes lorsque le nombre de variables et de classes augmente.

II.5.2.4.Les réseaux de neurones

Les réseaux de neurones sont inspirés à partir de la physiologie de l'organisation du cerveau. Ils reposent sur une modélisation discriminante. Un neurone permet de définir une fonction discriminante linéaire g dans l'espace de représentation E des entrées. Cette fonction réalise une combinaison linéaire du vecteur de caractéristiques de l'entrée e.

$g(e) = wte + w0$, où w est un vecteur de poids de la combinaison linéaire et w0 est le biais. Ainsi $g(e) = 0$ définit un hyperplan permettant de diviser E en deux régions de décision.

Les réseaux neurones peuvent effectuer des taches de classification et de régression avec ou sans supervision. Ils accomplissent ceci par des méthodes appropriées de réglages de poids(w), en espérant que les sorties du réseau mènent aux valeurs cibles [30] (Figure II-1).

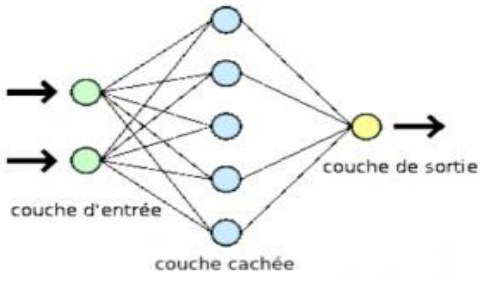


Figure II-1:Architecture de réseau de neurone

Les avantages des réseaux de neurones

- Les réseaux de neurones sont souples, ils sont capables de traiter une gamme très étendue de problèmes.
- Ils donnent de bons résultats, même dans des domaines complexes.
- ils sont beaucoup plus puissants que les techniques statistiques ou les arbres de décision en termes de résistance au bruit et au manque de fiabilité des données

Les inconvénients des réseaux de neurones

- Les réseaux de neurones ont des problèmes au niveau du codage des entrées. Toutes les entrées doivent se trouver dans un intervalle défini, en général, entre 0 et 1. Ce qui entraîne des transformations et risquent de fausser les résultats.
- pour assurer de bons résultats, le nombre d'exemples doit être très grand puisqu'il tient compte du nombre d'entrées, du nombre de couches et du taux de connexion.

II.5.2.5.Méthode des SVMs

SVM (Support Vector Machine) est une technique d'apprentissage automatique utilisée pour faire la classification. Elle consiste à minimiser simultanément l'erreur empirique de classification et maximiser la marge géométrique entre les classes, «Séparateur à Vaste Marge». Parmi les avantages de SVM, son habilité d'apprentissage qui peut être indépendante de la dimension du vecteur d'entrées. La force de SVM tient à sa simplicité de mise en œuvre pour résoudre des problèmes difficiles et à ses fondements mathématiques solides. Nous retenons que les modèles SVM ont déjà montré leurs preuves dans plusieurs domaines tels que la classification du texte, d'images et la reconnaissance de locuteurs…etc [31].

II.6.Conclusion

Ce chapitre permet de donner une idée générale sur la classification et les méthodes de classification, dont nous avons défini la classification en générale et les différents types et méthodes de classifications.

Dans le chapitre suivant, nous allons détailler la méthode de classification choisie SVM pour classifiée les données de puces ADN.

Chapitre III : Machine à Vecteur de Support

III.1.Introduction

Depuis quelques années, de nouvelles méthodes d'apprentissage se développent sur la base de la théorie de l'apprentissage statistique de Vladimir Vapnik. L'une de ces méthodes, est la Machine à vecteur de support (ou SVM : l'acronyme de Support Vector Machines en anglais).

Le SVM est une méthode de classification supervisée, qui fut introduite par Vapnik en 1995, elle est basée sur la recherche de l'hyperplan optimale, lorsque c'est possible, pour classer ou séparer correctement les données tout en étant le plus éloigné possible de toutes les observations. Le principe est donc de trouver un classificateur dans un espace approprié. Puisque c'est un problème de classification à deux classes. Cette méthode est basée sur l'utilisation des fonctions dites noyau (Kernel) qui permettent une séparation optimale des données.

Ce chapitre sera consacré au système d'apprentissage SVM : définition des cas linéairement séparable et non linéairement séparable, les différents types d'apprentissage multi classes, les avantages et les inconvénients de ce type d'apprentissage, et enfin les domaines d'application.

III.2.SVM principe de fonctionnement général

III.2.1.Notions de base: Hyperplan, marge et support vecteur

Pour deux classes d'exemples donnés, le but de SVM est de trouver un classificateur qui va séparer les données et maximiser la distance entre ces deux classes. Avec SVM, ce classificateur est un classificateur linéaire appelé hyperplan.

Dans le schéma qui suit, on détermine un hyperplan qui sépare les deux classes.

III.2.1.1.Hyperplan optimal

Un hyperplan va séparer les deux ensembles de donnés en deux classes. Il existe une multitude d'hyperplan valide, mais la propriété de SVM est que cet hyperplan doit être optimal, c-à-d qui permet de classer bien les nouvelles données, donc il faut chercher parmi les hyperplans valide, celui qui passe « au milieu » des points des deux classes des donnés où la distance minimale aux donnés d'apprentissage est maximale. Les points de l'hyperplan vérifier aussi l'équation $w \cdot x + b = 0$ (Figure III-1). [32][33].

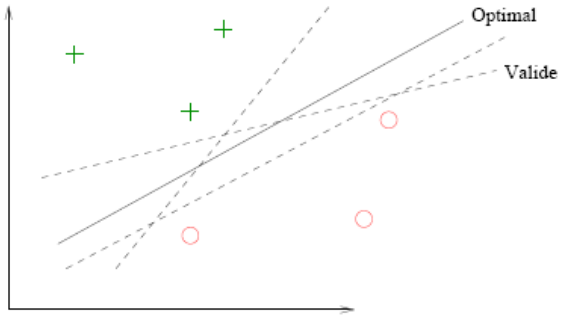


Figure III-1: Choisir hyperplan optimal

III.2.1.2.Les Support de vecteurs

Les Support de vecteurs est un ensemble d'exemples d'apprentissage se trouvant sur la marge. Sont les points les plus proches, qui seuls sont utilisés pour la détermination de l'hyperplan [34] (Figure III-2).

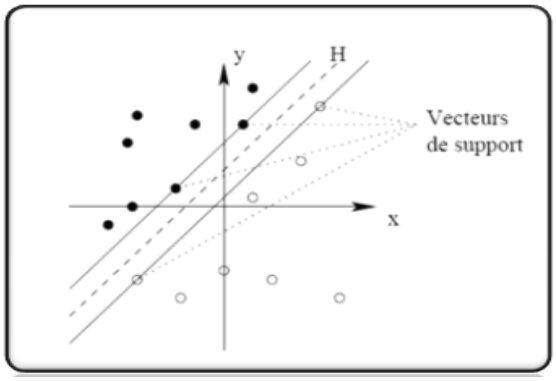


Figure III-2: Les supports de vecteur

III.2.1.3.La marge

La marge est la distance entre la frontière de séparation et les échantillons les plus proches. Ces dernières sont appelées vecteurs supports [35]. La marge est calculée à partir du produit scalaire entre les vecteurs situés à la frontière de chaque classe et le vecteur unitaire normal w de l'hyperplan séparateur (Figure III-2). [36]

Dans un modèle linière, on à $f(x) = w \cdot x + b$. L'hyperplan séparateur (frontière de décision) a donc pour équation $w.x + b = 0$.

La distance d'un point au plan est donnée par $d(x) = |w.x + b|/\|w\|$. Soit $x_1 et\ x_2$ deux points de classes différentes $f(x_1) = +1$ et $f(x_2) = -1$, $w \cdot x_1 + b = +1$ et $w \cdot x_2 + b = -1$ donc $(w.(x_1 - x_2)) = 2$ D'où : $\left(w/\|w\|.(x_1 - x_2)\right) = 2/\|w\|$ [36] [37].

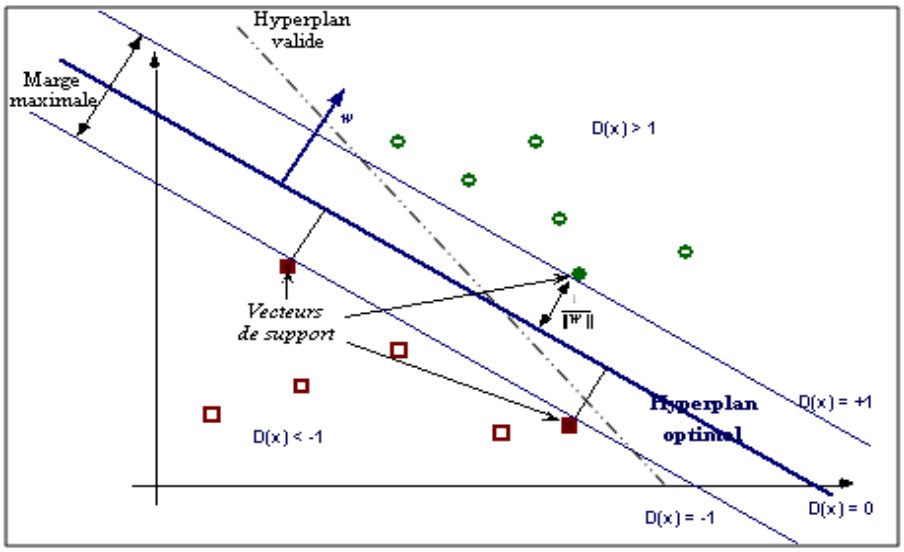


Figure III-3:Hyperplan de séparation des deux classes

III.2.1.4.Pourquoi maximiser la marge ?

Intuitivement, le fait d'avoir une marge plus large procure plus de sécurité lorsque l'on classe un nouvel exemple. De plus, si l'on trouve le classificateur qui se comporte le mieux vis-à-vis des données d'apprentissage, il est clair qu'il sera aussi celui qui permettra au mieux de classer les nouveaux exemples. Dans (Figure III-4), la partie droite nous montre qu'avec un hyperplan optimal, un nouvel exemple reste bien classé alors qu'il tombe dans la marge. On constate sur la partie gauche qu'avec une plus petite marge, l'exemple se voit mal classé.

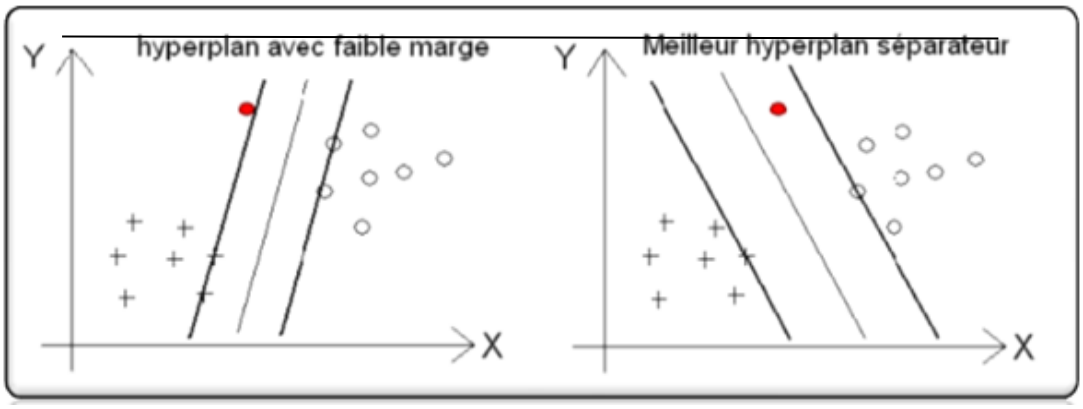


Figure III-4:Hyperplan optimal

En général, la classification d'un nouvel exemple inconnu est donnée par sa position par rapport à l'hyperplan optimal. Dans le schéma suivant, le nouvel élément sera classé dans la catégorie des « + » [38].

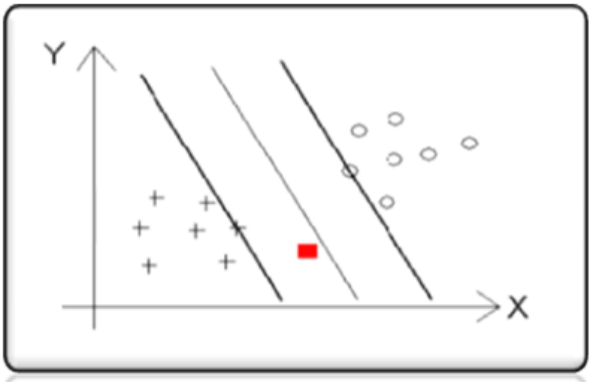


Figure III-5: Classification d'un nouvel exemple

III.2.2.Apprentissage statistique et SVM

L'apprentissage par induction permet d'arriver à des conclusions par l'examen d'exemples particuliers. Il se divise en apprentissage supervisé et non supervisé. Le cas qui concerne les SVM est l'apprentissage supervisé où les exemples particuliers sont représentés par un ensemble de couple d'entrée/sortie. Le but est d'apprendre une fonction qui correspond aux exemples vue, et qui prédit les sorties pour les entées qui n'ont pas encore été vues. Les entrées peuvent être des descriptions d'objets et les sorties les classes des objets donnés en entrée. [38]

III.2.3.Principe du SVM

La méthode des machines à vecteurs de support (SVM) est une alternative récente pour la classification. L'acronyme SVM peut être judicieusement traduit par Séparateurs à Vaste Marge. Le principe théorique des SVM comporte deux points fondamentaux :

- la transformation non linéaire (Φ) des exemples de l'espace d'entrée vers un espace dit de re-description de grande dimension muni d'un produit scalaire (espace de Hilbert). [39]

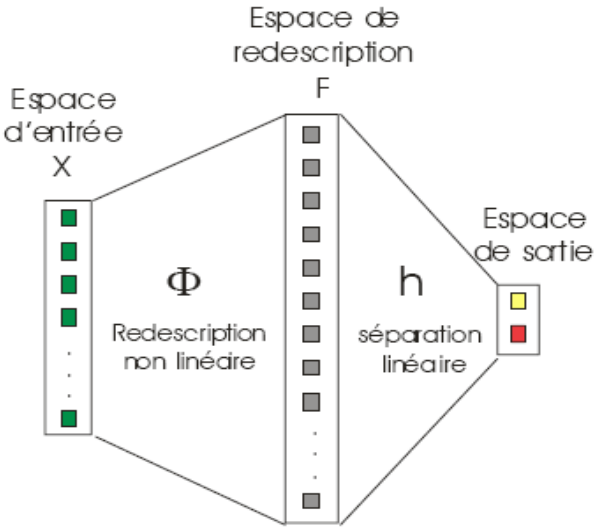


Figure III-6 : La transformation non linéaire des données

- La détermination d'un hyperplan permettant une séparation linéaire optimale dans cet espace de grande dimension. [39]

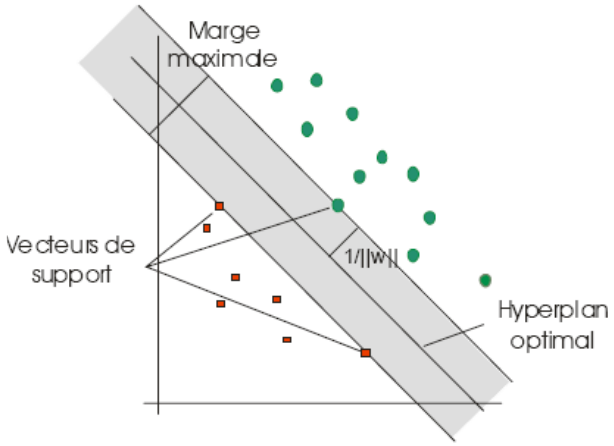


Figure III-7: Hyperplan séparateur optimal

III.3.Linéarité et non-linéarité

Parmi les modèles des SVM, on constate les cas linéairement séparable et les cas non linéairement séparable. Les premiers sont les plus simples puisqu'ils permettent de trouver facilement le classificateur linéaire. En général, la plupart des applications ont besoin de fonctions plus complexes que les fonctions linéaires pour faire de la classification. Une stratégie de prétraitement peut être utilisée pour simplifier la tâche. Il s'agit de changer

l'espace original en un nouvel espace dit de re-description de grande dimension où les données peuvent être linéairement séparables. [36]

III.3.1.Cas linéairement séparable

Le cas linéairement séparable est le modèle le plus simple du SVM, il est également appelé linéaire de marge maximale. Il cherche à séparer des données appartenant à deux classes différentes par un hyperplan optimal, qui est équidistant des frontières de chaque classe. Ce SVM fonctionne seulement si les classes de données d'apprentissage sont linéairement séparables. En suppose que nous avons des donnés empiriques $(x_1, y_1) \dots (x_i, y_i) \in \mathbb{R}\, x\, \{\bar{+}1\}$ Ces donnés conduisent à bien décrire la technique de construction de l'hyperplan optimal.

L'objectif des SVMs, dans le cas linéaire est, de déterminer un hyperplan qui sépare au mieux les échantillons de deux classes. Dans ce cas, la fonction f est linéaire en x_i et elle prend la forme générale suivante: $f(x_i) = \langle w, x_i \rangle + b$.

Il existe une infinité d'hyperplan capable de séparer parfaitement ces deux classes. Le principe utilisé dans les SVMs est de trouver l'unique hyperplan qui maximise la marge entre les deux classes, c'est-à-dire qui maximise $\frac{2}{\|w\|}$ .Pour ce faire, il faut donc minimiser$\|w\|$. Par la suite, nous chercherons à minimiser $\frac{1}{2}\|\mathrm{w}\|^2$, pour simplifier les calculs. [40][36]

III.3.2.Problème primal

Un point (x_i, y) est bien classé si et seulement si $y.f(x) > 0$. Comme le couple (w, b) est défini à un coefficient multiplicatif près, on s'impose $y.f(x) \geq 1$. On obtient le problème de minimisation sous contraintes suivantes :

$$\begin{cases} min\frac{1}{2}\|w\|^2 \\ \forall i, y_i(w.x_i + b) \geq 1 \end{cases} \qquad \text{(III. 1)}$$

Nous nous retrouvons ainsi face à un problème d'optimisation quadratique convexe sous contraintes linéaires. [33]

III.3.3.Problème dual

Pour résoudre le problème d'optimisation quadratique qui précède, on construit le Lagrangien *L*

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i \left[y_i(\langle w, x_i\rangle + b) - 1\right] \qquad \text{(III. 2)}$$

Où les $\alpha_i, i = 1,\ldots,l$, désignant les multiplicateurs de Lagrange. Pour que w, b et les α_i existent, le problème doit vérifier les conditions de Karush-Kuhn-Tucker (KKT):

$$\frac{\partial L(w,b,\alpha)}{\partial w} = 0 \Leftrightarrow w = \sum_{i=1}^{l} \alpha_i y_i x_i \qquad \text{(III. 3)}$$

$$\frac{\partial L(w,b,\alpha)}{\partial b} = 0 \Leftrightarrow \sum_{i=1}^{l} \alpha_i y_i = 0 \qquad \text{(III. 4)}$$

$$\forall\, i\,, y_i(\langle w, x_i\rangle + b) - 1 \geq 0 \qquad \text{(III.5)}$$

$$\forall\, i, \alpha_i \geq 0$$

$$\forall\, i\,, \alpha_i(y_i(\langle w, x_i\rangle + b) - 1) \geq 0 \qquad \text{(III. 6)}$$

On passe au problème dual en introduisant des multiplicateurs de Lagrange pour chaque contrainte. Dans le problème primal en substituant (III.3) et (III.4) dans (III.2), nous obtenons le problème dual équivalent suivant: [33]

$$\begin{cases} max \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \forall i, \quad \alpha_i \geq 0 \end{cases} \qquad \text{(III. 7)}$$

Ce dernier problème peut être résolu en utilisant des méthodes standards de programmation quadratique et la méthode la plus commune est l'algorithme SMO (Sequential Minimal Optimization). Une fois la solution optimale $\alpha^* = (\alpha_1^*, \ldots, \alpha_l^*)$ du problème (III.7) obtenue, le vecteur de poids de l'hyperplan à marge maximale recherché s'écrit :

$$w^* = \sum_{i=1}^{l} \alpha_i{}^* y_i x_i \qquad \text{(III. 8)}$$

Seuls les αi correspondants aux points les plus proches sont non nuls. On parle de vecteurs de support.

On remarque que pour :

$$\begin{cases} y_i(< w, x_i > + b) = 1 \\ y_i(< w, x_i > + b) > 1 \end{cases} \text{ On a } \begin{cases} \alpha_i \neq 0 \\ \alpha_i = 0 \end{cases} \qquad \text{(III.9)}$$

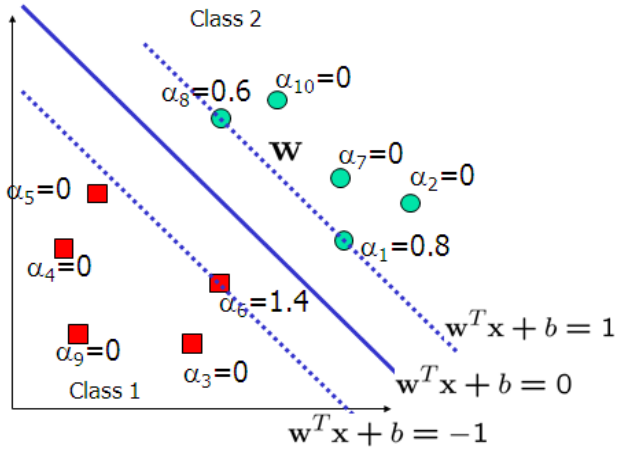


Figure III-8:Une interprétation géométrique

La fonction de décision associée est donc :

$$f(x) =< w, x > +b = f(x) = \sum_{i=1}^{l} \alpha_i{}^* y_i \langle x_i, x \rangle + b \qquad \text{(III. 10)}$$

l : c'est l'ensemble des indices des supports vecteurs. [36] [41]

III.3.4.Cas non séparable :

Comme nous avons déjà mentionné, l'hypothèse dans le cas linéairement séparable conditionne beaucoup la résolution du problème (III.1). En effet, il suffit qu'une observation de deux classes viole la contrainte $y.(w\, x_i + b) \geq 1$ Pour que ce problème n'ait plus de solution. La Figure III-9 montre une telle situation. Pour tenter de résoudre ce problème, l'idée consiste à relâcher les contraintes dans le but d'autoriser quelques erreurs de classification. Cette généralisation de l'hyperplan à marge maximale a été proposée par (Cortes and Vapnik, 1995) en introduisant les variables d'écart à la marge. [40]

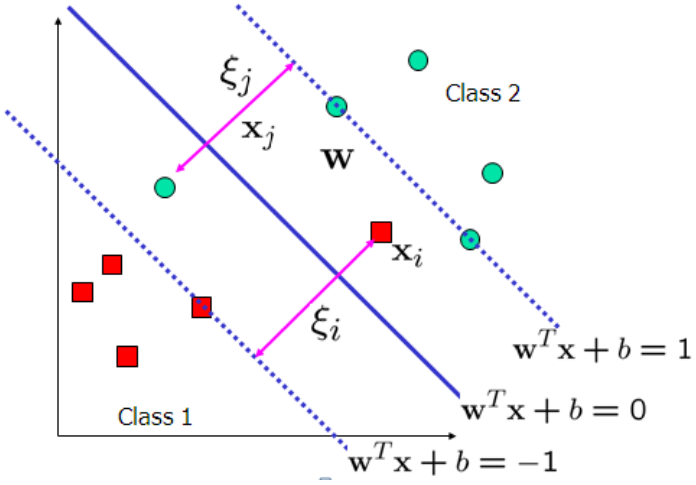


Figure III-9 : Variable relâchement

Si $0 \leq \xi \leq 1$, les exemples se trouvent dans la région de la marge maximale. Si $\xi > 1$, les exemples se trouvent du mauvais coté de l'hyperplan.

On cherche à maximiser la marge en tolérant pour chaque contrainte une erreur positive ξ_i la plus petite possible. [40][41]

On part du problème primal linéaire et on introduit des variables 'ressort' :

$$\begin{cases} \min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \xi_i \\ \forall i, y_i(w.x_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0, \quad i = 1,2 \ldots, L \end{cases} \qquad \text{(III. 11)}$$

Le paramètre supplémentaire C qui apparaît dans le problème d'optimisation qui précède, est une constante positive fixée à l'avance, et qui permet de contrôler l'importance de l'erreur que l'on autorise par rapport à la taille de la marge. Plus *C* est important, moins d'erreurs sont autorisées.

On en déduit le problème dual avec la même démarche du lagrangien que dans le cas séparable:

$$\begin{cases} max \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \forall i, 0 \leq \alpha_i \leq C \end{cases} \qquad \text{(III. 12)}$$

La seule différence est la borne supérieure C sur les α. [33]

III.3.5.Cas non linéairement séparable :

La séparation par un hyperplan est un cadre de classification relativement complexe, car pour beaucoup de cas, les exemples que l'on veut séparer ne sont pas séparables par un hyperplan. Pour surmonter les inconvénients des cas non linéairement séparable, l'idée des SVM est de changer l'espace des données. Où la séparation linéaire des exemples dans un nouvel espace est possible. On va donc avoir un changement de dimension. Cette nouvelle dimension est appelé « espace de re-description ».

La fonction noyau de transfert ϕ (Figure III-10) correspond à un changement de dimension. Une fois ce changement est appliqué, on aboutit à la problématique de la séparation linéaire. Ce qui implique de trouver une fonction $f(x)$ $\phi : \mathcal{R}^n \rightarrow \mathcal{R}^N$[32]

$$x = (x_1, \dots, x_n) \longmapsto \phi(x) = \left(\phi_1(x), \dots \phi_N(x)\right)$$

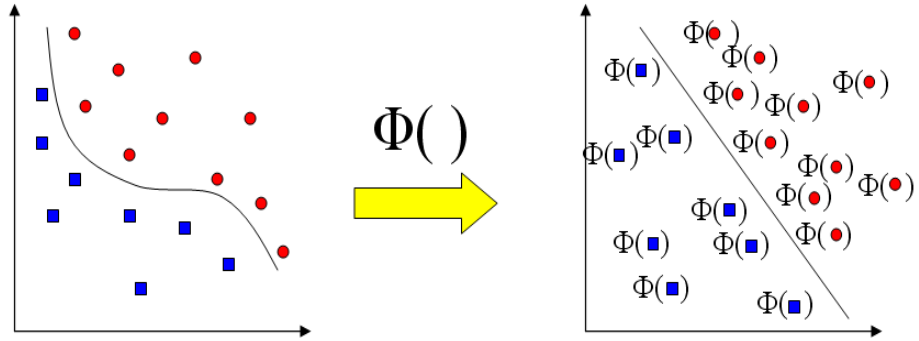


Figure III-10: Un changement de représentation peut simplifier la classification

III.3.6.Fonction de noyau

Dans le cas où les donnés sont non linéairement séparable, on peut transformer les données dans un espace où la classification serait plus aisée par un hyperplan optimal pour séparé les donnés en deux classes. Dans ce cas, l'espace de re-description utilisé le plus souvent est $\mathcal{R}$ (ensemble des nombres réels). Cet espace ne suffit pas pour classer les entrées. On passe donc à un espace de grande dimension [38], espace de Hilbert H, appelé l'espace de re-description d'une cartographie non linéaire . $\phi(x) \in$ H, pour tous les $x \in X$, On transforme les entrées en vecteurs dans un espace $\mathcal{F}$ (feature space) [33].

$\phi : \mathcal{R}^d \rightarrow \mathcal{F}$

$x \rightarrow \phi(x)$

Avec card $(\mathcal{F}) >$ d.

Exemple

f: ϕ $\mathcal{R}^2 \rightarrow \mathcal{R}^3$

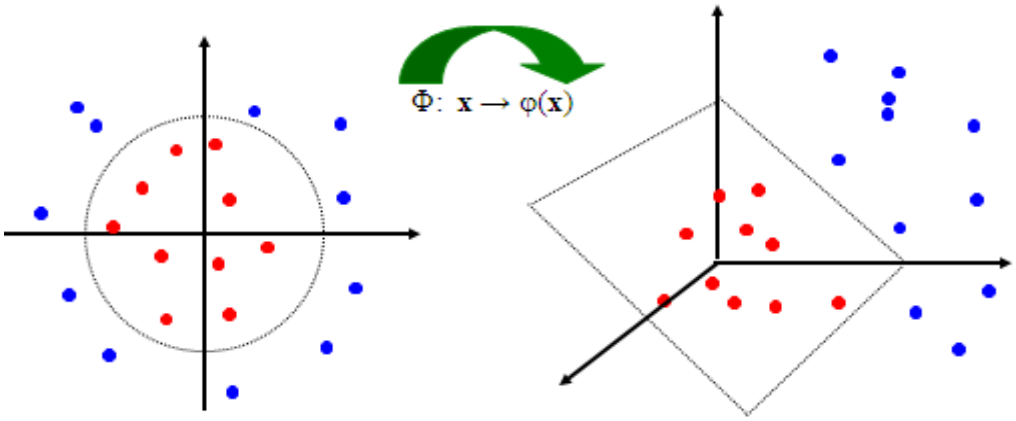


Figure III-11: Illustration de passage à $\mathcal{R}^3$

Le passage dans $\mathcal{F} = \mathcal{R}^3$ rend possible la séparation linière des données. On doit donc résoudre :

$$\begin{cases} max \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle . \\ \sum_{i=1}^{l} \alpha_i y_i = 0 \\ \forall i, 0 \leq \alpha_i \leq C \end{cases} \qquad \text{(III. 13)}$$

Le vecteur de poids de l'hyperplan séparateur et la fonction dans le nouvel espace devient :

$w = \sum_{i=1}^{l} \alpha_i y_i \, \phi(x_i)$

$$f(x) = \sum_{i=1}^{l} \alpha_i{}^* y_i \phi(x_i).\phi(x_j) + b \qquad \text{(III. 14)}$$

Le problème et sa solution ne dépendent que des produits scalaires $\phi(x).\phi(x')$.Plutôt que de choisir la transformation non-linéaire $\phi : x \to \mathcal{F}$ on choisit une fonction $k: X \times X \to R$ (nombres réels) appelée fonction noyau.

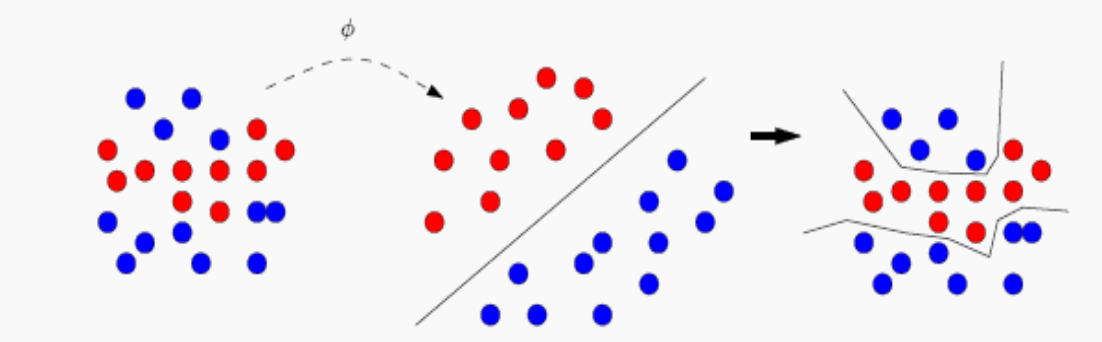
Cette fonction représente un produit scalaire dans l'espace de représentation intermédiaire. Elle traduit donc la répartition des exemples dans cet espace. [33][43]

III.3.6.1.Conditions pour avoir un noyau (théorème de Mercer)

Il existe des conditions mathématiques, appelées théorème de Mercer, qui permettent de dire si une fonction est un noyau ou non, sans construire la projection dans l'espace des caractéristiques. En fait, il faut assurer que pour tout ensemble d'exemples de longueur l, la matrice $\left(k\left(\overrightarrow{x_i}, \overrightarrow{x_j}\right)\right)_{1 \leq i,j \leq l}$ soit définie positive. Dans ce cas, il existe un espace F et une fonction ϕ tels que :

$$K(x, x') = \phi(x).\phi(x')$$

Donc Le changement d'espace dons le cas non linière se fait au moyen d'une fonction répondant au critère de Mercer. Ce critère permet un changement dans les deux sens, ce qui permet à partir de l'expression de l'hyperplan dans l'espace complexe de classer les éléments dans l'espace de description initial.



Espace d'entrée X Espace de caractéristiques *F* Espace d'entrée X

Figure III-12: Fonction de noyau

La fonction $k(x_i, x_j)$ appelée « noyau » (kernel), permet de transformer les vecteurs d'attributs de manière implicite vers l'espace de caractéristiques et d'estimer une fonction de décision linéaire dans cet espace. [44][42]

III.3.6.2. Exemple de noyau

Linéaire $k(x, x') = \langle x, x' \rangle$

Polynomiale $k(x, x') = (x, x')^n ou\ (c + x, x')^n$

Gaussien $k(x, x') = e^{\frac{-\|x-x'\|^2}{2\sigma^2}}$

Laplacien $k(x, x') = e^{\frac{-\|x-x'\|}{\sigma}}$ [33]

III.4. SVM multi-classes

Au premier temps, les SVM sont conçus pour traiter le problème de classification binaire. L'étude théorique des systèmes d'apprentissage s'est concentrée principalement sur des fonctions dont les sorties sont dans $\{-1, +1\}$.

Depuis la première extension proposée par Vapnik, plusieurs chercheurs se sont attachés à utiliser les SVM dans des applications à plusieurs classes. Les approches employées jusqu'à nos jours sont diverses et elles peuvent être réparties en deux méthode. [34]

III.4.1. La méthode Une-contre-reste (one-versus-all)

L'approche " Une-contre-reste " est la plus simple et la plus ancienne des méthodes de décomposition utilisée pour les machines à vecteurs supports multi-classes. Cette extension au cas multi-classe originellement proposée par Vapnik, peut être vue aussi comme une généralisation du cas binaire. À toute classe k est associé un hyperplan $H(w_k, b_k)$ défini par la fonction de décision $f_k(x) = \langle w_k \cdot x \rangle + b_k$ dont le rôle est de discriminer entre les observations de la classe k et de l'ensemble des autres classes.

Elle construit m classifieurs binaires à vecteurs supports où m est le nombre total des classes. L'apprentissage du $k^{\text{ème}}$ classifieur à vecteurs supports s'effectue en considérant tous les exemples de la $k^{\text{ème}}$ classe dans la région positive et tous les autres exemples dans la région négative. [34]

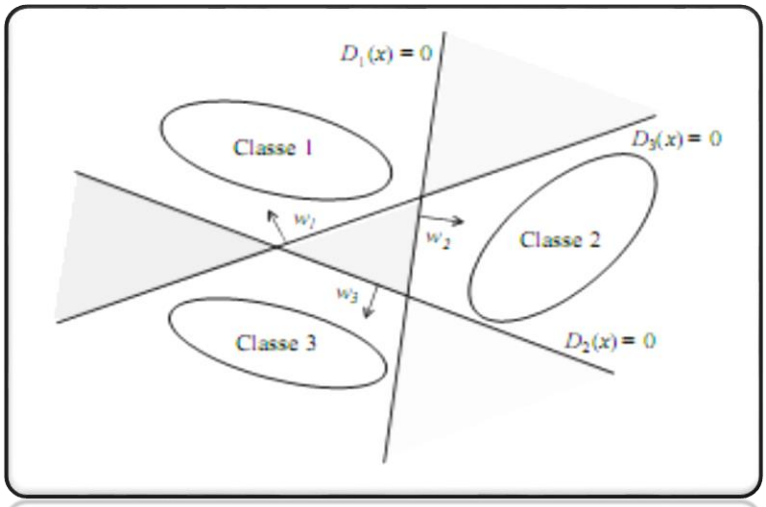


Figure III-13:Approche une-contre-reste

III.4.2.La méthode Une-contre-une (one-versus-one)

Dans la méthode Une-contre-une, pour un problème de classification comportant K classes, on construit pour chaque combinaison possible de deux classes distinctes un classifieur SVM binaire. Donc on aura un total de k (k-1)/2 classifieurs binaires.

Plusieurs alternatives de vote sont envisageables et généralement une classe parmi les classes qui ont un nombre élevé de prédictions est choisie. Dans la stratégie intitulée, le max qui gagne, l'exemple x à prédire est testé par tous les classifieurs binaires possibles entre tous couples de classes c_i et c_j. Pour chaque classifieur, si l'exemple x est assigné à la classe ci donc on incrémente le compteur de la classe ci sinon on incrémente le compteur de la classe cj.

A la fin, la classe gagnante sera la classe ayant la valeur du compteur maximale. En cas de conflit, c'est-à-dire deux classes ou plus ont des valeurs maximales identiques, on choisira la classe du petit indice. [34][45]

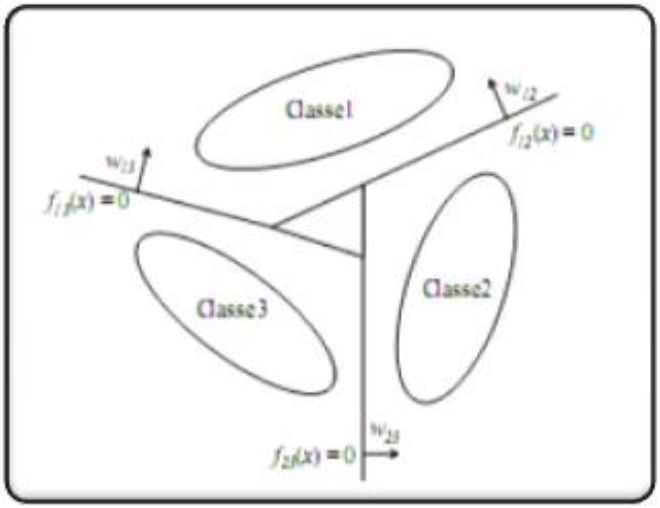


Figure III-14:Approche une-contre-une.

III.5.Les avantages et les inconvénients des SVM

III.5.1.Avantages

- L'hyperplan unique.
- L'algorithme est robuste face aux changements d'échelle.
- contrôler l'importance de l'erreur que l'on autorise par rapport à la taille de la marge.
- Résultats en général équivalents et souvent meilleurs. [41]

III.5.2.Inconvénients

- Il faut trouver la “bonne” fonction noyau
- Problèmes i.i.d. (données indépendantes et identiquement distribuées)
- Deux classes à la fois. [41]

III.6.Les domaines d’applications

SVM est une méthode de classification qui montre de bonnes performances dans la résolution de divers problèmes. Cette méthode a montré son efficacité dans de nombreux domaines d’application tels que le traitement d’images, la catégorisation de textes ou le diagnostique médicale.

III.7.Conclusion

Dans ce chapitre, nous avons présenté le principe de fonctionnent des SVMs. C’est une méthode de classification très performante, qui s’adresse à la fois aux cas linéaire et non-linéaire ainsi qu’aux problèmes de classification binaire, multi-classes et mono-classe.

Le chapitre suivant sera consacré à la conception et l’implémentation d’un système de classification par les SVM pour les données bio-puces.

Chapitre IV :

La Conception du Système

IV.1.Introduction

Les parties mentionnées dans le premier et deuxième et troisième chapitre nous ont permis de comprendre et de situer clairement les notions de base pour la conception et la réalisation de notre projet. Ainsi dans le processus de développement de notre système, nous allons présenter une conception qui va décrire d'une manière non ambiguë notre système. Nous allons donner l'architecture globale de notre système et ceci selon une vue interne, puis nous allons détailler les fonctionnalités de cette architecture avant de présenter sa réalisation.

IV.2.Description du système

L'objectif de notre système est d'analyser des bases de données biopuces par la méthode de classification supervisée SVM. Dans notre cas, on dispose d'une base de données on connaît à priori la classe d'appartenance de chaque exemple de cette base de données.

En premier lieu, la base de données est utilisée pour l'apprentissage du système. C'est-à-dire, l'utiliser pour construire un classifieur. Ce classifieur est exploité par la suite pour classer les exemples de la base de test. Le taux de reconnaissance est calculé à partir du nombre des exemples bien classés.

IV.3.La conception globale

Cette partie définit l'architecture globale de notre système qui se présente à des fichiers GEO comme entrée et avec comme résultat finale des données classée après avoir appliqué la méthode de classification machine à vecteur de support. On peut résumer l'architecture globale de notre système sous le schéma (Figure IV-1).

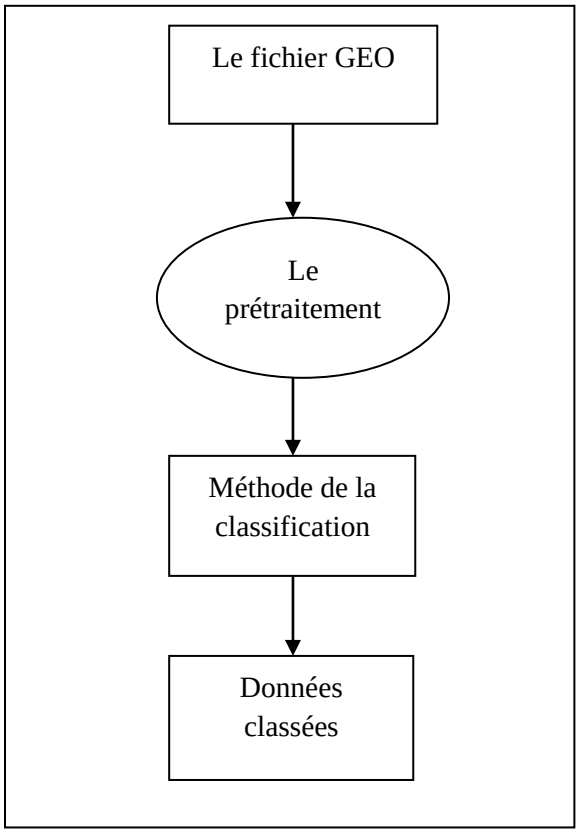


Figure IV-1:Conception Globale du système

IV.3.1.Le fichier GEO d'entrée

L'Entrée de notre système est un fichier GEO choisie par l'utilisateur sous format texte, qui contient une description d'une bio-puces sur une maladie spécifié.

IV.3.2.Prétraitement

Il est nécessaire de procéder à un prétraitement des données et une normalisation des données pour éliminer ces différences artefacts.

Notre prétraitement se compose essentiellement de quatre étapes, la correction du bruit du font, un quantile normalisation, un filtrage du nombre des gènes, à la fin une sélection des gènes.

IV.3.3.Méthode de la classification

La méthode utilisée pour la classification des données bio-puces dans notre travail est une méthode de classification supervisé qui est le SVM, Cette méthode nécessite d'abord la séparation des données en donnés d'apprentissage et donnés du teste. Ensuite la construction d'un model SVM selon les donnés d'apprentissage et le testé sur les donnés du teste.

IV.3.3.1.Séparation des données

Cette étape consiste à séparer les données en base d'apprentissage et base du teste par la méthode de la séparation HOLDOUT ou KFOLD.

IV.3.3.2.Classification SVM

Cette étape consiste sur la construction d'un classifieur selon la base d'apprentissage et l'appliquer sur la base du test, cette classification nous permet du calculé un taux donné pour mesuré la performance du notre model de la classification.

IV.4.Conception détaillé

Grace à la conception détaillée on obtient une description des structures de données utilisées et leur communication entre elles pour parvenir en finale à une donnée classée. Et on peut résumer notre conception détaillée sous le schéma suivant (Figure IV-2).

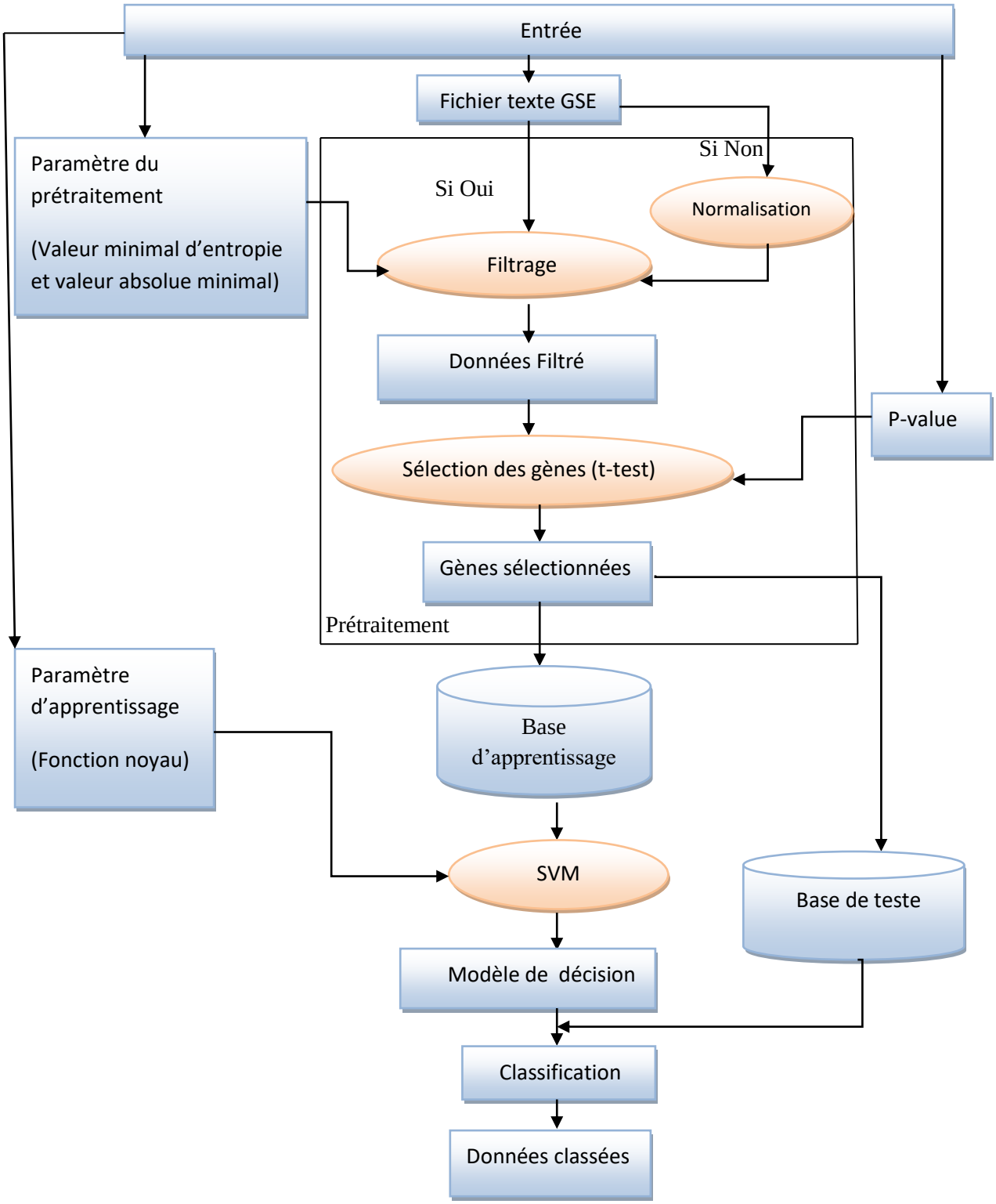


Figure IV-2:Schéma représentant la conception détaillée du système.

IV.4.1.Le fichier d'entrée

Notre application accepte comme une entrée un fichier texte GEO qui contient des informations sur une bio-puces à propos une maladie spécifiée. Ce fichier contient une description de l'expérience réalisée sur un groupe d'échantillon et une description sur les noms des gènes et leur niveau d'expression dans les échantillons (Figure IV-3). Ce fichier peut être passé par une Correction du bruit de fond et peut être normalisé ou non selon notre base du donnée.

On peut télécharger ces fichiers GEO du site :https://www.ncbi.nlm.nih.gov/geo/

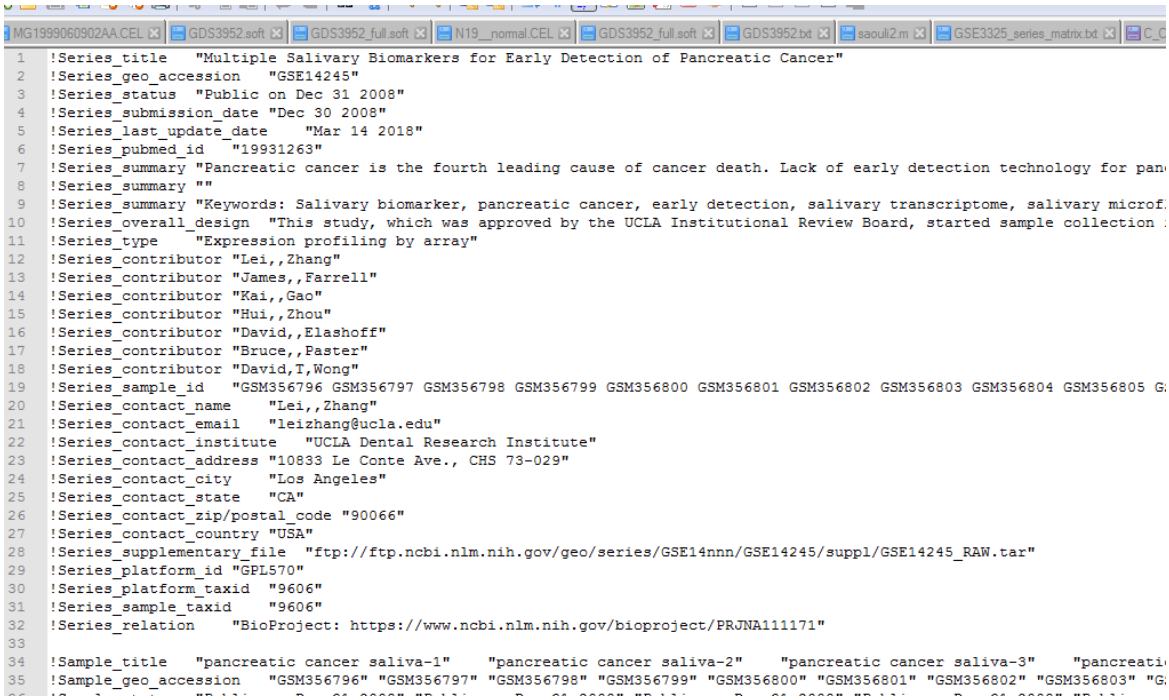


Figure IV-3:morceau d'un fichier GSE d'une séquence génomique

Dans notre application nous avons implémenté trois jeux du données pour appliquer notre méthode de la classification, ses descriptions est la suivantes :

- Le premier jeu du donnée est sur cancer du pancréas qui est réalisé sur 24 échantillons (12 premier échantillons appartient à la 1ére classe et le reste échantillons appartient à la 2éme classe), sa description est dans le lien suivant :
  https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4100#details

- Le deuxième jeu du donnée est sur cancer du poitrine qui est réalisé sur 95 échantillons (47 échantillons appartient à la 1ére classe et 52 échantillons appartient à la 2éme classe), sa description est dans le lien suivant : https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5847

- Le troisième jeu du donnée est sur le cancer du poumon qui est réalisé sur 192 échantillons (90 échantillons appartient à la 1ére classe et 97 échantillons appartient à la 2éme classe et 5 échantillons appartient la 3éme classe) sa description est dans le lien suivant : https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2771#details

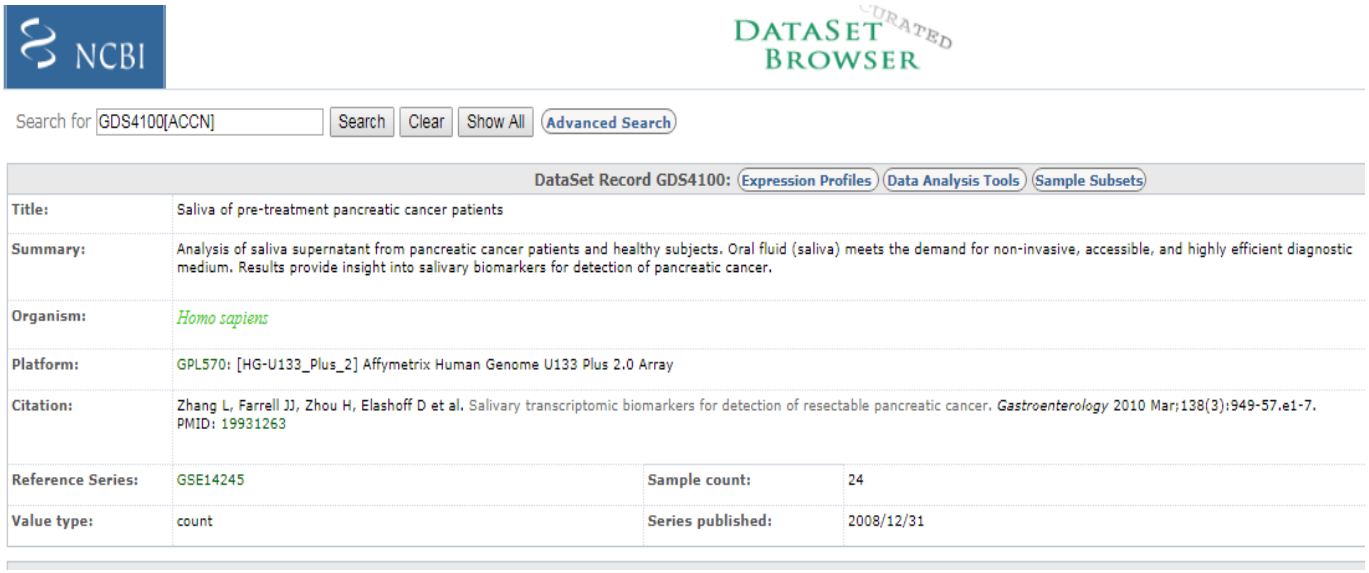


Figure IV-4:description du jeu du donné cancer du pancréas

IV.4.2.Le prétraitement des données

IV.4.2.1. La correction du bruit du fond

La première étape dans notre prétraitement est la correction du bruit du fond soit par la méthode RMA (Correction par Robust Multi-Array Analysis) ou par GCRMA(Correction par Gene Chip RMA).Nos bases des données GEO téléchargés sont déjà passées par une correction du bruit du fond.

IV.4.2.2.Normalisation

Est l'opération qui vise à rendre les puces d'une mémé expérience comparables, Dans notre cas des puces affymetrix, la normalisation que nous avons utilisé est la normalisation des quantiles.

Nos bases des données GEO téléchargés peut être déjà passé par un quantile normalisation, si non, nous devons les normalisé.

IV.4.2.3.Filtrage

Cette étape illustre comment filtrer les données en supprimant les gènes qui ne sont pas exprimés ou qui ne changent pas. Nous avons utilisé un certain nombre de techniques pour réduire le nombre de profils d'expression à un sous-ensemble contenant les gènes les plus significatifs.

- La fonction genelowvalfilter supprime les gènes qui ont des valeurs d'expression absolue très faibles, l'utilisateur doit spécifier la valeur absolue minimale
- la fonction geneentropyfilter élimine les gènes dont les profils ont une faible entropie, L'utilisateur doit entrer la valeur d'entropie minimal.
- la fonction genevarfilter pour filtrer les gènes avec une petite variance dans le temps (la variance minimale utilisée est 10%).

IV.4.2.4. Sélection

Le principe de la sélection des attributs consiste à évaluer chaque attribut pour lui assigner un score de pertinence qui permet un classement des attributs. Les attributs les mieux classés c'est-à-dire les plus pertinents seront sélectionnés pour la phase du traitement.

La méthode utilisée pour la sélection des gènes dans notre travail est une méthode de sélection nommée t-test basée sur p-value, cette méthode est un test paramétrique permettant de comparer deux groupes d'échantillons, Il concerne des données quantitatives, mesurées sur une échelle d'intervalle ou de rapport.

L'utilisateur doit spécifier la valeur maximal du p-value, les gènes qui ont une valeur p-value inférieur à la valeur p-value d'entrés seront sélectionnés pour la phase de la classification.

IV.4.3.La méthode de classification

La construction du classificateur SVM nécessite la spécification de quelques paramètres d'apprentissage tels que :

IV.4.3.1.Les méthodes de la séparation des données

L'utilisateur doit choisir une méthode parmi deux méthodes qui sont :

1. HOLDOUT : L'utilisateur doit spécifier une valeur qui est entre 0 et 1 et qui représente le ratio du données qui doit spécifier pour le test.

2. KFOLD : L'utilisateur doit spécifier une valeur (K) qui représente le nombre du partition, on devise le données original en K partition, puis en sélectionnes un des K partition comme ensemble du test et les K-1 autres partitions constitueront l'ensemble d'entrainement. On calcule le score de performance, puis on répète l'opération en sélectionnant une autre partition de test parmi les K-1 échantillons qui n'ont pas encore été utilisés pour le test du modèle. L'opération se répète ainsi K fois. La moyenne des K erreurs quadratiques moyennes est enfin calculée pour estimer l'erreur de prédiction.

IV.4.3.2. La fonction noyau (kernel)

L'utilisateur doit spécifier la fonction noyau pour la classification SVM parmi les trois fonctions suivantes :

1. Linéar : classification SVM linéairement séparable
2. Fonction quadratique : une fonction noyau pour le cas non linéairement séparable
3. Fonction base radiale gaussienne : une fonction noyau pour le cas non linéairement séparable.

IV.4.3.3. Classification SVM

Après la sélection de la méthode de la séparation et la fonction noyau, Un modèle classificateur SVM est construit selon la base d'entraiment, ensuit ce model est appliqué sur la base du teste afin du classé nos données, la classification du base du teste nous permis de donné des prédictions sur la base du teste, ainsi Le taux de reconnaissance qui est calculé à partir du nombre des exemples bien classés.

Conclusion

Dans ce chapitre, nous avons présenté les différentes étapes du fonctionnement de notre système de classification des données biopuces en passons de la conception général à détaillé en mentionnant à chaque étape sa description. Le chapitre suivant est consacrée à l'implémentation de notre système.

Chapitre V : Implémentation

Chapitre V : Implémentation

V.1.Introduction

Dans ce dernier chapitre et après la conception du notre système des chapitres précédents, nous présenterons le côté pratique de l'application .Notre but est la réalisation d'un système robuste qui fais une classification SVM des données bio-puces. Nous commençons par le choix de l'environnement de travail en passons par le langage et logiciel de codage la bibliothèque intégrée ainsi que les étapes fondamentaux de la conception de notre application.

V.2.L'environnement de travail

Pour que notre travail atteint l'objectif qu'on visait, on a pris l'initiative d'exploiter et d'implémenter notre algorithme sur : Windows 7. Ce choix se traduit par l'efficacité de cet environnement en ce qui concerne la structure d'interaction événementielle qu'elle dispose pour communiquer avec des applications actives, ainsi que les ressources de la machines qu'il offre aux différentes applications , enfin , son système d'allocation de mémoire qui est un des meilleurs présents dans ce domaine.

V.3.Le langage de codage

Notre application a été codée en sa globalité par le langage MATLAB ce choix repose sur le fait qu'il est :

- Populaire : en particulier en domaine biologie, le langage MATLAB est un investissement pour celui qui l'apprend,
- Riche : il existe de nombreuses bibliothèques dans tous les domaines ; celles-ci ont l'avantage considérable d'être standardisées,
- Rapide : Programmation infiniment plus rapide pour le calcul et pour l'affichage
- Lisible : Code facile à comprendre et très lisible

V.4.Caractéristique de la Machine :

L'implémentation du projet est réalisé par sur un ordinateur portable qui se caractérisé techniquement par :

- Un processeur Intel® Core(TM) i5-480M.

- Une Mémoire RAM de 4 GO

-Une carte graphique RADEON HD 6370M de 512 MB

V.5.Présentation de l'interface de l'application

Dans notre application nous offrons à l'utilisateur une interface graphique qui facilite la manipulation des paramètres pour la rendre plus compréhensible.

La (Figure V-1) présente La fenêtre d'accueil du notre système.

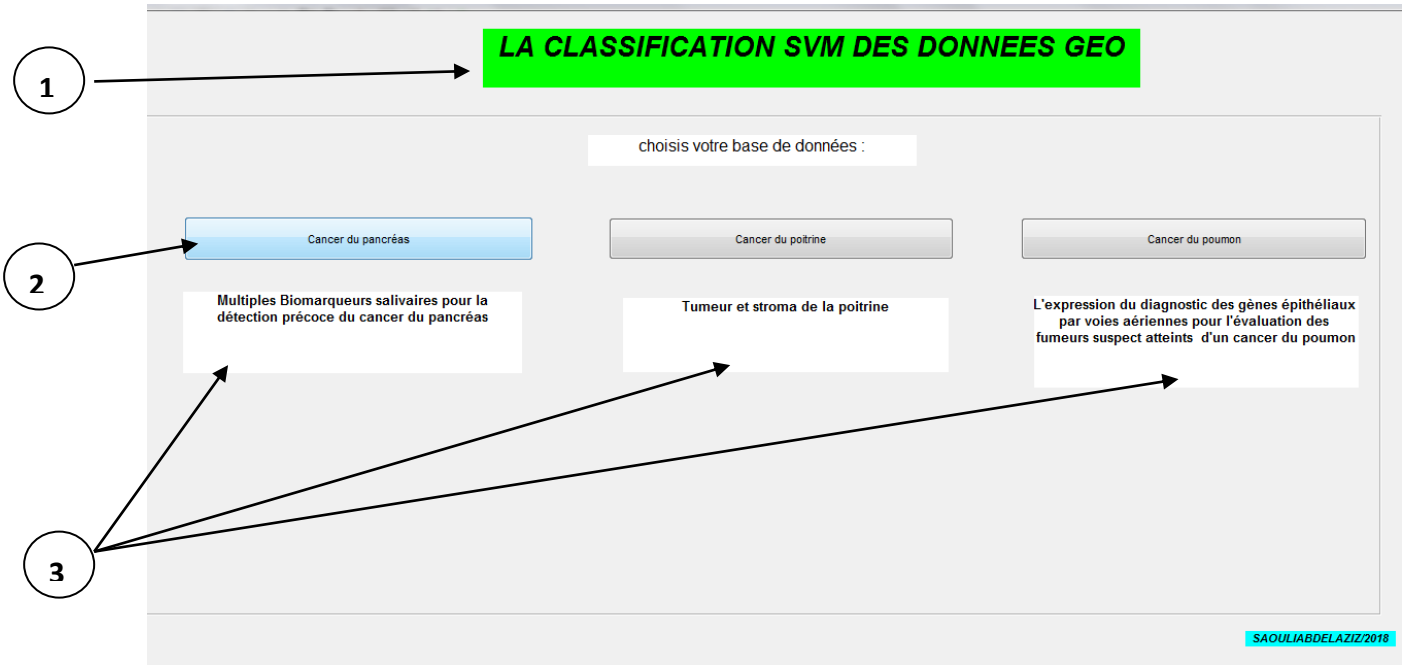


Figure V-1: La fenêtre d'accueil

1. le titre du l'application
2. les jeux des données disponibles
3. la description des jeux de données

En choisissons un jeu du données permis les trois disponible, une fenêtre pour la classification de cette jeu du données apparait (Figure V-2).

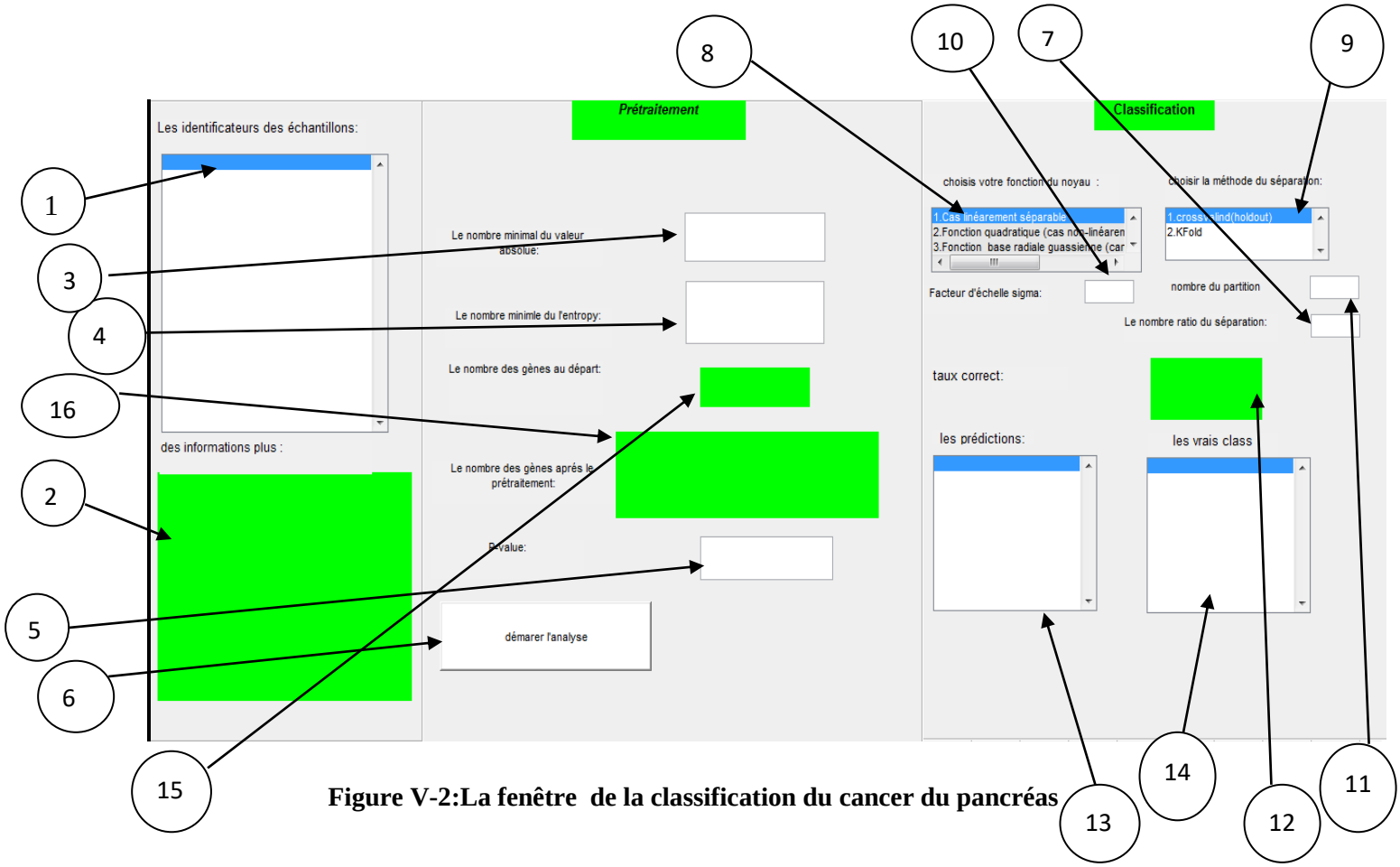


Figure V-2:La fenêtre de la classification du cancer du pancréas

1. les identificateurs des échantillons.
2. des informations plus sur cette jeu du données concernant le prétraitement (normalisation ou la correction du bruit du font).
3. Le nombre minimal de la valeur absolue pour le filtrage qui est spécifié par l'utilisateur.
4. Le nombre minimal de l'entropie pour le filtrage qui est spécifié par l'utilisateur.
5. La valeur du P-value choisis pour la sélection.
6. bouton du démarrage du l'exécution.
7. Le nombre du ratio pour la séparation des données HOLDOUT.
8. Une liste de la fonction noyau disponible pour classification SVM.
9. Les méthodes de la séparation des données disponibles.
10. Le nombre du facteur sigma pour la fonction noyau Gaussien.
11. Le nombre des partitions pour la séparation des données KFOLD.
12. Le taux correct de la classification.
13. Les prédictions du model SVM.

14. Les vrais classes du base de teste.
15. Le nombre du gène au départ avant le filtrage.
16. Le nombre du gène après le prétraitement.

La Figure V-3 représente également quelques résultats de notre algorithme de classification qui est appliqué sur le jeu du donné cancer du pancréas.

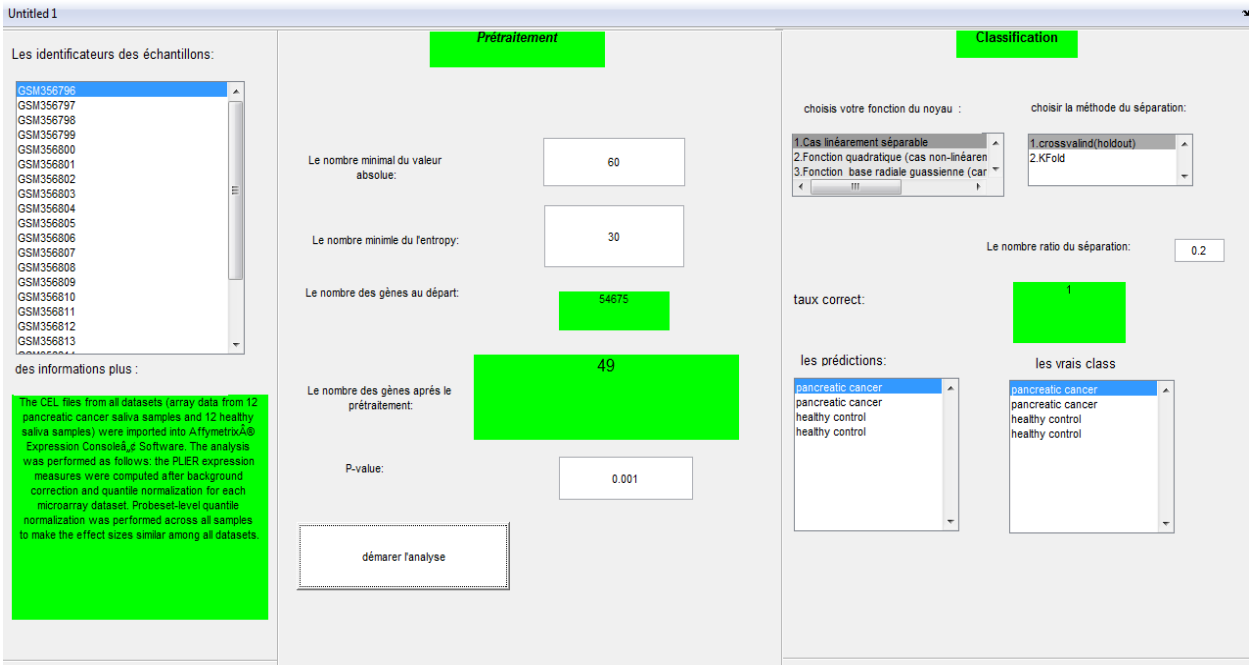


Figure V-3 : Résultat final après la classification

Les résultats obtenue dans la Figure V-3 sont obtenus en choisissons la méthode de la séparation HOLDOUT (avec un ratio de 20%) et une fonction linéar sur notre jeu du donnée cancer du pancréas. Cette classification nous à permet d'obtenir des prédiction avec un taux correcte de 100%.

La Figure V-4 représente le résultat de notre de classification SVM, En choisissons la méthode de la séparation KFOLD (avec 10 partition) et la fonction quadratique sur notre jeu du donnée cancer de la poitrine. Cette classification nous à donné un taux correct de 0.97%.

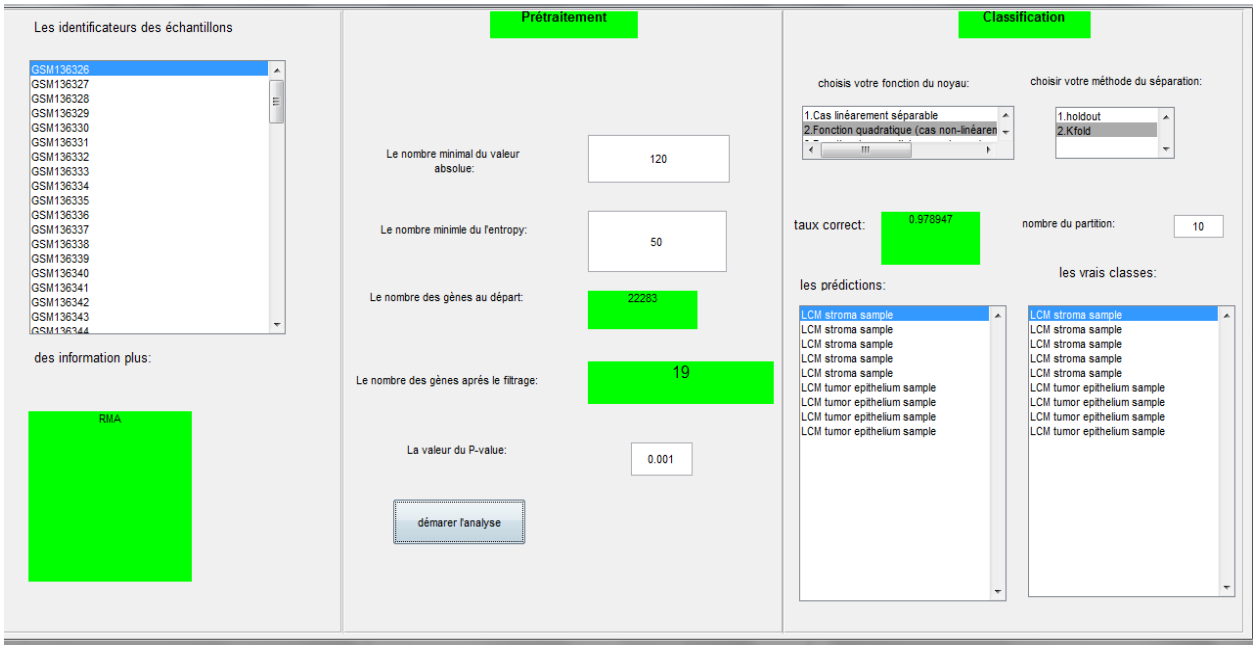


Figure V-4:Résultat final après la classification

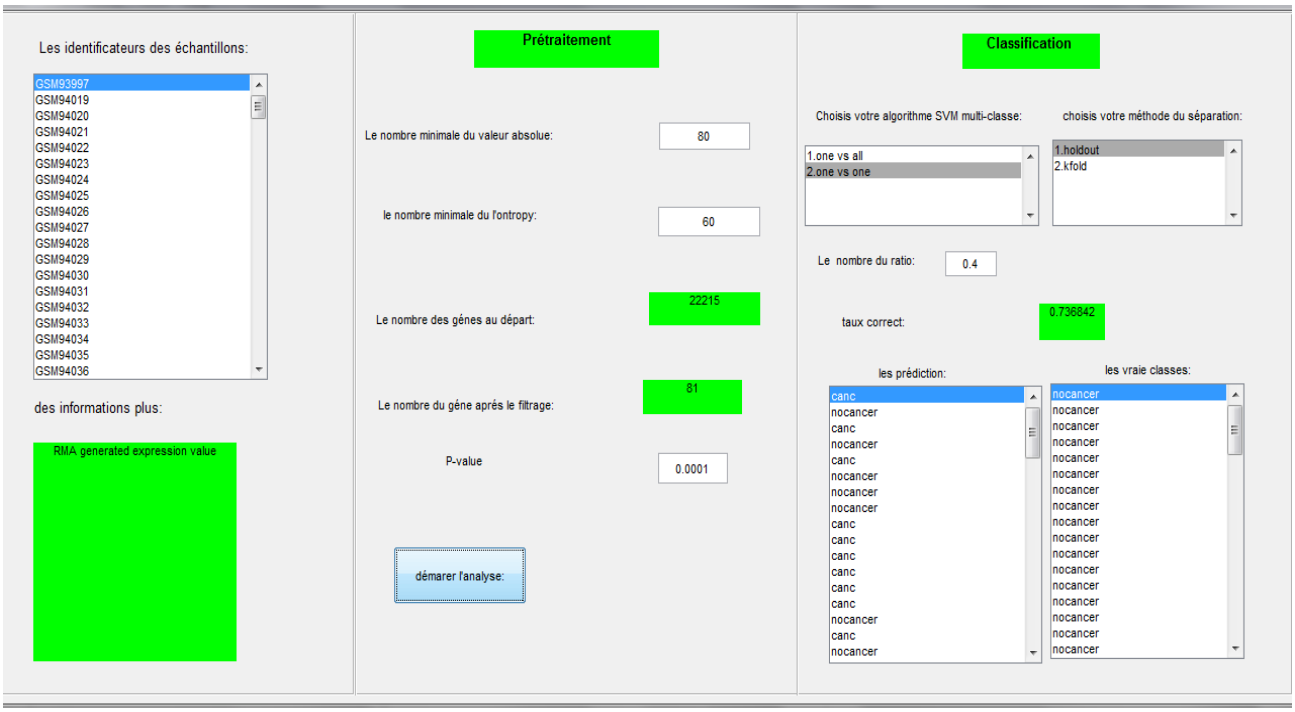


Figure V-5:Résultat final après la classification multi-classe (cancer du poumon).

La Figure V-5 représente le résultat de notre de classification SVM sur une base de données à 3 classe, En choisissons la méthode de la séparation HOLDOUT (avec un ratio de 0.4) et la fonction multi-classe un contre un sur notre jeu du donnée cancer du poumon.

Cette classification nous à donné un taux correct du 0.73%.

V.6.Conclusion

Dans ce chapitre, nous avons défini l'étape de fonctionnement de notre application et tous les composants qui produisent cette application et nous avons présenté également quelques résultats de notre méthode de classification.

Conclusion Générale

Conclusion générale

Dans ce mémoire, nous avons abordé la problématique de la classification des données bio-puces avec une méthode de classification supervisée, parmi la multitude des méthodes supervisée pouvant être utilisées, nous avons opté pour la méthode de classification machine à vecteur de support qui une méthode très populaire et très performante par rapport aux autres

Notre méthode de classification optimisé a été testé sur des jeux du données diverses, Nous avons pu constater à travers nos différents tests que les résultats obtenus montrent que la méthode de classification SVM apporte des résultats satisfaisantes.

Nous avons implémenté une application qui permet la classification des données de bio-puces. Et offert une interface interactive qui facilite l'utilisation de ces différents composants. Avec les quel nous aient permis d'obtenir des résultats compétitifs, notre méthode pourrait être encore améliorées.

Nous pouvons encore envisager d'autres idées, puisque le sujet à traiter rentre dans le cadre de la bioinformatique, nous pensons donc à la validation biologique des résultats, c'est à- dire étudier les gènes choisis par la méthode de sélection au niveau du laboratoire.

Les résultats du laboratoire va nous guider à construire notre système de diagnostic par des méthodes simples et rapides.

La bibliographie

La bibliographie

[1] http://pbil.univ-lyon1.fr/databases/oldacnuc/acnuc.html (consulté le 25/08/2017).

[2] Jacques van Helden. Introduction to Bioinformatics. Support de cours. France : Université Aix-Marseille, Technological Advances for Genomics and Clinics, 2012, 35 p.

[3] http://www-helix.stanford.edu/people/altman/bioinformatics.html#one (consulté le 30/08/2017).

[4] http://campus.cerimes.fr/genetiquemedicale/enseignement/genetique28/site/html/1.html (consulté le 30/08/2017).

[5] khatir, Nadjia, Bahlou, Safia nait. Clustering dans les bases de données. Universitéd'oran 2012.

[6] Pichot, André. Histoire de la notion de gène.Flammarion paris 1999.

[7]Brown ,Terence A, Irène Mowszowir, Alain Raisonnier, Françoice wright. Génomes.Flammarion médecine-science paris 2004

[8] A, Poitras E. Houde. Principe de l'amplification par PCR.2002

[9] http://www.sfbi.fr/sites/default/files/jobim/jobim2002/papiers/P-p081_048.pdf (consulté le 22/01/2018)

[10] Brown P. O. Exploring the new world of the genome with DNA microarrays(1999). Nat Genet , 21,33-7.

[11] Lokhart,w, a . Genomics, gene expression and DNA arrays, Nature ,405,825-36. 2000

[12]lander E. Array of hope, Nature Genetics, 21 :3-4, 1999.

[13] E. M. Southern. DNA Arrays methods and protocols, chapter DNA Microarrays, pages 1–15. Humana Press, 2001.

[14] Genome Resource Facility GRF, Microarray section, London School Of Hygiene and Tropical, Article technique, Medecine. 2006

[15] Genome Resource Facility GRF, Microarray section, London School of Hygiene and Tropical, Article technique, Medecine. 2006.

[16] Moussa A. Vannier B. , Workflow d'analyse de données des puces à ADN, Spectra Analyse n291 p 48, revue scientifique, Mai 2013

[17] Emilie Guérin. Intégration de données pour l'analyse de transcriptome : mise en œuvre par l'entrepôt GEDAW. Thèse de doctorat. Informatique .Rennes. Université de Rennes 1

[18] Hardin,J,et al robust measureof Correlation between two gens on a microarray .BMC

Bioinformatics 2007.

[19] Mr. MOUSSATI Omar .Classification des données de biopuces. Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf

[20] Barrett T., T. O. Suzek, D. B. Troup, S. E. Wilhite, W.-C. Ngau, P.Ledoux, NCBI GEO: mining millions of expression profiles—databaseand tools 2005.

[21] Lockhart,w,a,. Genomics,gene expression and DNA array,Nature ,405,827-36.2000

[22] Tusher, V., R. Tibshirani and G. Chu, Significance analysis of microarraysapplied to transcriptional responses to ionizing radiation.Proceedings of the National Academy of Science USA, 2001.

[23] Sébastien Rimour. Méthode et outils logiciels pour la conception de sondes oligonucléotidiques pour puces à ADN. Application aux biopuces transcriptomiques et aux biopuces de type phylogénétique. Universite Blaise pascal-clermont 2

[24] Guillaume Bouchard. Les modèles génératifs en classification supervisée etapplications à la catégorisation d'images et à la fiabilité industrielle. Thèse de doctorat. University Joseph Fourier â Grenoble 1.2005.

[25] Nadjia Khatir. Clustering dans les bases de données. Mémoire de magistère.
Informatique. Université d'oran Es Senia

[26]Forgy, E .Cluster analysis of multivariante data : efficiency vs interpretability of classification. Biométrics, 21, 768-769.

[27] https://tel.archives-ouvertes.fr/tel-00447684/document (consulté le 22/01/2018)

[28].José Crispín HERNÁNDEZ HERNÁNDEZ. ALGORITHMES MÉTAHEURISTIQUES HYBRIDES POUR LA SÉLECTION DE GÈNES ET LA CLASSIFICATION DE DONNÉES DE BIOPUCES. ÉCOLE DOCTORALE STIM

[29] ZAABOT Zohra.Les Réseaux Bayésiens. Application en Reconnaissance de Formes à partir d'Informations Complètes ou Incomplètes.Thèse de doctorat Université Mouloud Mammeri, Tizi-Ouzou.2012.

[30] F. Slimane, R. Ingold, M. Alimi & J. hannebert , « Duration Models for arabic Text recognition using Hidden Markov Models »University of Fribourg ,Suisse ,2008.

[31] International Conference on Web and Information Technologies Sidi Bel Abbes, Algeria 29-30 June, 2008. http://www-inf.univ-sba.dz/icwit.

[32] Hala NAJMEDDINE : « *Méthode d'identification et de classification de la consommation d'énergie par usages en vue de l'intégration dans un compteur d'énergie électrique* », Université blaise pascal clermont II. Thèse de doctorat 2009.

[33] Olivier Bousquet : « *Introduction aux Support Vector Machines (SVM)* » Centre de Mathématiques Appliquées Ecole Polytechnique, Palaiseau. Orsay 2001.

[34] Regueb Saleh, Rais Houssem Eddine « *Sélection des dattes par la méthode SVM* »Mémoire d'ingénieure d'état en informatique, Faculté des Sciences exactes et des Sciences de la nature et de la vie. Université de Biskra Juin 2009.

[35] Bounneche Meriem Dorsaf : « *réduction de données pour le traitement d'image* » Mémoire de Magistère, Option Contrôle. Université mentouri constantine faculté de la science de l'ingénieur département d'électronique 2009.

[36] Anibal Arias Aguilar : « *Méthodes à vecteurs de support et Indexation sonore* », IRIT 2003/2004.

[37] [6] Antoine Cornuéjols *« Méthodes à noyaux et SVMs (Séparateurs à Vastes Marges)*» Équipe TAO (INRIA/CNRS) - Université de Paris-Sud, Orsay & ENSIIE

[38] M. Zaize Faouzi « *les supports vecteurs Machine (SVM) pour la reconnaissance des caractères manuscrits arabes* » Mémoire de magistère en informatique, Faculté des Sciences exacte et des sciences de la nature et de la vie, Université de Biskra 2010

[39] Alctra R&D : « introduction aux Machine a Vecteurs de support *(SVM)* »

[40]M. Tarhouni, K. Laabidi, S. Zidi, M. Ksouri «Surveillance des systèmes complexes par séparateurs a vaste marge (SVM) » Université des Sciences et Technologies de Lille 2010

[41] Martin Law et Antoine Cornuéjols : *« Une introduction aux machines à vecteurs supports (ou séparateurs à vastes marges - SVM*) »

[42] Reda Jourani « *Reconnaissance de visages* » Université Mohammed V-Agdal, Faculté des Sciences Rabat. DESA Informatique et Télécommunications UFR SYSCOM2. 11 Novembre 2006

[43] Padhraic Smyth *« CS 277: Data Mining Notes on Classification* » Department of Computer Science University of California, Irvine

[44] Hervé Frezza-Buet, Supélec : « *Machines à Vecteurs Supports Didacticiel* » école supérieure d'électricité. Support de cours 29 avril 2010

[45] A. Statnikov, F. Aliferis, I. Tsamardinos, D. Hardin, S. Levy, «A *Comprehensive Evaluation of Multicategory Classification Methods for Microarray Gene Expressio* », 2004.