Université Mohamed Boudiaf - M'sila

FACULTE DE TECHNOLOGIE

DEPARTEMENT D'ELECTRONIQUE

Numéro de série : ……

Numéro d'inscription : …….

Mémoire

Présentée pour l'obtention du diplôme de

Master

Filière : Electronique
Spécialité : Instrumentation

THEME

Approche Filtre par la sélection de données multi-sensorielles pour l'aide au diagnostic médical

Présenté par

CHARIK Khalissa & CHARIK Loubna

Soutenue le : …./……/……

Devant le jury composé de :

Nom & Prénom	**Grade**	**Qualité**
OUALI Mohammed Assam	MCB	Président
LADJAL Mohamed	MCA	Encadreur
DJERIOUI Mohamed	MCB	Co- Encadreur
KHENNOUF Salah	MCB	Examinateur

Année universitaire : 2019/2020

Résumé:

Les avancées technologiques ont facilité l'acquisition et le recueil de nombreuses données. Ces données peuvent être utilisées comme support de décision, conduisant aux développements d'outils capables de les analyser et de les traiter. Les systèmes d'aide au diagnostic sont considérés comme étant essentiels dans beaucoup de disciplines, ces systèmes reposent sur des techniques issues de l'intelligence artificielle mais les problèmes les plus intéressants sont souvent basés sur des données de haute dimension. Ces problèmes désignent les situations où nous disposons peu d'observations alors que le nombre de variables explicatives est très grand. La sélection de variables est devenue l'objet qui attire l'attention de nombreux chercheurs durant ces dernières années, cette sélection permet d'identifier et d'éliminer les variables qui pénalisent les performances d'un modèle complexe dans la mesure où elles peuvent être bruitées, redondantes ou non pertinentes. De plus, la mise en évidence des variables pertinentes facilite l'interprétation et la compréhension des aspects liés aux applications ; ainsi, elle permet d'améliorer la performance de prédiction des méthodes de classification et de passer outre le fléau de la haute dimensionnalité de ces données. L'approche **filtre** est couramment utilisée à ce jours pour analyser les données biologiques, cette approche consiste à parcourir la sélection des variables avant le processus de l'apprentissage et ne conserve que les caractéristiques informatives.

L'objectif recherché dans le cadre de ce travail est une contribution à l'étude et au développement de systèmes innovants d'aide au diagnostic médical. Consacré à la simulation, ce travail vise l'application des techniques d'apprentissage statistique comme étant une solution dans la conception de ces systèmes par reconnaissance de formes. Dans le cadre de l'apprentissage supervisé, la sélection des caractéristiques permet d'obtenir des classifieurs précis. L'approche filtre est fondée uniquement sur des données, elles permettent à l'utilisateur d'entamer une analyse plus fine de ces données en augmentant la transparence du modèle utilisant la méthode mRMR (Minimum Redondance, Maximum Relevance). Pour la validation de données sélectionnées dans des bases d'apprentissage biomédicales, nous testons leurs capacités et leurs taux de classification avec plusieurs classifieurs. Afin de mener une étude comparative permettant un choix décisif de la méthode la mieux adaptée à l'application proposée, on évaluera pour les méthodes exposées les paramètres liés au taux de reconnaissance, au temps d'apprentissage et à l'erreur d'entraînement.

Mots clés : Diagnostic médical, Sélection des variables, mRMR, Classification, RNA, RBF, k-PPV, Simulation.

Remerciements

Nous remercions tout d'abord ALLAH tout puissant qui nous a donné la santé, le courage durant ces
langues années d'étude, et la patience afin de pouvoir accomplir ce modeste travail.
Nous tenons à remercier grandement notre Encadreur et Co- Encadreur
Dr.LADJAL et Dr.M.DJERIOUI
pour ses grande disponibilité et ses précieux conseils.
Nous remercions également tous les enseignants du
département d'électronique d'université de M'sila
plus spécialement les membres de jury de notre travail.
Merci à tous.

Dédicaces

je dédie ce modeste travail :

- ✓ *Nos chers parents.*
- ✓ *Nos chers frères et sœurs.*
- ✓ *Toute la famille* ***charik****.*
- ✓ *tous les enseignants qui m'ont aidé .*
- ✓ *Tous nos amis.*
- ✓ *Tous les collègues de notre promotion*

(2019/2020).

Enfin à tous ceux et celles qui m'ont encouragé et soutenu.

Abréviations

mRMR : minimum Redondance Maximum Relevance.

RNA : les réseaux de neurones artificiels.

RBF : les réseaux de neurones à base radiale.

k-PPV : k plus proches voisins.

PMC : Perceptron Multi Couches.

RN : Un réseau de neurones.

MLP: multi-layer perceptron.

RBF : les réseaux de neurones à bases radiales.

$\boldsymbol{f}$**:** est la fonction d'activation.

$\boldsymbol{S_M}$**:** La couche de sortie.

$\mathbf{R^n}$ **:** muni de la distance euclidienne.

***K-NN*:** le nom *K-Nearest Neighbor.*

***K* :** Le paramètre de méthode k plus proches voisins.

ACP : L'Analyse en Composantes Principale.

mg/dl: Mili gramme par décilitre.

GHz : Giga Hertz.

Go : Giga Octet.

RAM : Random Acces Memory.

IA : L'Intelligence Artificiel.

T : Taux de classification.

EQM : Erreur quadratique moyenne de généralisation.

NI : le nombre d'itérations.

***T_appr* :** le temps d'apprentissage.

***Er_appr* :** l'erreur d'apprentissage.

***Taux_test* :** le taux de test.

MRE : Minimisation du Risque Empirique.

Liste des Tableaux

Liste des figures

SOMMAIRE

Chapitre II: Les techniques d'apprentissage

Chapitre III : Simulation et Evaluation

INTRODUCTION GÉNÉRALE

Dans le domaine médical, la résolution des problèmes d'aide au diagnostic se base sur le traitement de données extraites à partir des données acquises dans le monde réel, l'application de l'intelligence artificielle à la médecine offre une perspective essentielle à l'essor de ces nouvelles technologies, qu'il s'agisse de renforcer le lien entre patients et médecins, de poser des diagnostics plus rapides et plus précis, ou encore d'optimiser la création de nouveaux traitements [1][2]. Les avancées technologiques ont facilité l'acquisition et le recueil de nombreuses données, notamment dans le domaine médical lors d'examen des patients. Ces données peuvent être utilisées comme support de décision médicale, conduisant aux développements d'outils capables de les analyser et de les traiter; dans la littérature, nous trouvons régulièrement la notion d'aide au diagnostic, ces systèmes sont même considérés comme étant essentiels dans beaucoup de disciplines, ces systèmes reposent sur des techniques issues de l'intelligence artificielle mais les problèmes les plus intéressants sont souvent basés sur des données de haute dimension. Ces problèmes désignent les situations où nous disposons peu d'observations alors que le nombre de variables explicatives est très grand. Cette situation est de plus en plus fréquente dans plusieurs les applications [3].

La sélection de variables consiste à choisir parmi l'ensemble global de variables, un sous-ensemble de variables pertinentes pour le problème étudié. Cette problématique peut concerner différentes tâches de fouille de données, elle regroupe des méthodes permettant de sélectionner un sous-ensemble de variables parmi un ensemble de départ, en utilisant divers critères et différentes techniques [4].

Des systèmes de diagnostic médical permanents doivent alors assurer le contrôle des divers paramètres biomédicaux. Les méthodes traditionnelles sont basées sur la connaissance de différents paramètres effectuées en laboratoire, pour décider après sur son état, et chercher les méthodes adéquates pour le soin. L'intérêt donc est de disposer d'un contrôle efficace des différents patients pour une meilleure décision. Les techniques de l'intelligence artificielle (IA) qui servent comme outil de base pour l'aide à la décision.

Leur réponse est plus élaborée et peut être obtenue soit à partir de données brutes venant directement des variables de contrôle. Il est judicieux de supposer que le problème de diagnostic médical peut être vu comme un problème de reconnaissance de formes, où les formes représentent l'ensemble des mesures biomédicaux, et les sorties correspondent aux différents états des patients. Ces informations biomédicales sont ensuite utilisées comme des données d'entrée dans la procédure de diagnostic de la partie amont d'un système d'aide à la décision. La construction de ce type de ce système peut être vue aussi comme un problème d'apprentissage de relations entre variables à partir de données observées. Ils sont spécialement programmés avec des algorithmes dérivés de l'intelligence artificielle. Le problème devient dans ce cas là, comme un problème de classification. Parmi les techniques d'IA utilisées, on trouve les réseaux de neurones artificiels (RNA), les réseaux de neurones à base radiale (RBF) et les k plus proches voisins (k-PPV). Celles-ci se démarquent des autres outils par leur capacité d'apprentissage et de généralisation, notamment dans les applications de grande dimension. Ces techniques peuvent être utilisées en raison de leurs robustesses et de leurs capacités à tenir en compte de la nature dynamique et complexe du procédé [5]. Ces techniques sont de plus en plus acceptées dans le domaine médical en tant qu'outil de modélisation et de diagnostic.

Le travail présenté dans le cadre de ce mémoire a pour objectif l'étude et la mise en œuvre de trois modèles d'apprentissage statistique (RNA, RBF et k-PPV) appliqués dans le diagnostic médical en classification. Une étude en simulation est effectuée pour valider, évaluer et comparer les performances de ces modèles dans un but de choix décisif adapté au problème posé. D'autre part, nous proposerons une technique de sélection des caractéristiques qui s'appelle minimum Redondance Maximum Relevance (mRMR) pour éliminer les variables qui pénalisent les performances d'un modèle complexe dans la mesure où elles peuvent être bruitées et identifier les variables les plus importantes.

Le travail réalisé est axé autour de trois chapitres qui sont présentés comme suit :

Le premier chapitre, nous présentons en détail le contexte de sélection des variables et la technique de prétraitement des données utilisée en vue d'une réduction de dimension des variables d'entrée appelée : Minimum Redundancy Maximum Relevance (mRMR), est aussi présentée.

- Dans le deuxième chapitre, nous détaillons les fondements théoriques des trois méthodes d'apprentissage statistique déjà évoquées. Après une brève introduction, où nous

allons rappeler la notion de neurone formel, nous décrivons son architecture et rappelons les propriétés générales des réseaux de neurones artificiels statiques (perceptrons multicouches) à apprentissage supervisé, ainsi que les réseaux de neurones à fonction de base radiale et les k plus proches voisins (k-PPV).

- Le troisième et dernier chapitre est consacré à la simulation et vise l'application des techniques étudiées précédemment comme étant une solution dans la conception d'un système d'aide à la décision médicale dans les deux plans : prétraitement de données par mRMR et classification. L'objectif est de valider et d'évaluer les performances de chacune des méthodes présentées. Afin de mener une étude comparative permettant un choix décisif de la méthode la mieux adaptée à l'application proposée, on évaluera pour ces méthodes les paramètres liés au taux de reconnaissance, au temps d'apprentissage et à l'erreur d'entraînement. Une discussion des résultats conclura cette étude de simulation pour choisir la technique la mieux adaptée.

Une conclusion générale en fin de ce travail, elle retrace les différentes étapes réalisées et souligne les perspectives envisagées.

CHAPITRE I

SÉLECTION DES VARIABLES

INTRODUCTION

L'analyse de données en grande dimension est devenue extrêmement fréquente et importante dans divers domaines des sciences, allant de la génomique et de la biologie à l'économie, la finance et l'intelligence artificielle [6]. La réduction de la dimension de grande base de données (Big Data) est un problème complexe qui a été largement étudié dans plusieurs domaines, bien que parfois, les deux grandes approches de réduction de l'espace de descripteurs s'opposent du point de vue de leurs objectifs : l'extraction de caractéristiques et la sélection des variables. Notamment, dans la littérature les travaux sur le domaine biologique ou génomique est encore plus récente en tentent les nouvelles méthodes pour aborder l'échelle de sélection de ce type de données [7]. La sélection des variables et des caractéristiques est devenue l'objet de nombreuses recherches dans des plusieurs domaines d'applications, pour lesquels des ensembles de données contenant des dizaines ou des centaines de milliers de variables sont disponibles. L'objectif de la sélection des variables est triple: améliorer les performances des classifieurs, fournir des classifieurs plus rapides et plus rentables, et fournir une meilleure compréhension du processus sous-jacent qui a gêné les données [8].

Dans ce chapitre, nous détalions le processus de la sélection des variables ainsi que leurs avantages et leurs limitations, nous définissons la sélection des variables, les différentes mesures de pertinence et de redondance rencontrées. Enfin, nous présenterons les différentes approches de sélection des variables entre autres notre méthode utilisée : *Minimum Redundancy Maximum Relevance*, "mRMR".

1.1. Aide au diagnostic

Les avancées technologiques ont facilité l'acquisition et le recueil de nombreuses données, notamment dans le domaine médical lors d'examen des patients. Ces données peuvent être utilisées comme support de décision médicale, conduisant aux développements d'outils capables de les analyser et de les traiter ; dans la littérature, nous trouvons régulièrement la notion d'aide au diagnostic, ces systèmes sont même considérés comme étant essentiels dans beaucoup de disciplines, ces systèmes reposent sur des techniques issues de l'intelligence artificielle mais les problèmes les plus intéressants sont souvent basés sur des données de haute dimension. Ces problèmes désignent les situations où nous disposons peu d'observations alors que le nombre de variables explicatives est très grand. Cette situation est de plus en plus fréquente dans plusieurs applications.

Des bases de données réelles différentes de source biomédicale fournissent plusieurs observations. Ces observations correspondent en générale à une seule classe (Etat du patient normal ou non). Les caractéristiques de chaque maladie jouent le rôle des variables.

Aujourd'hui, la difficulté réside non seulement dans l'obtention des données biomédicale mais également dans leurs analyses, l'objectif consiste à développer des méthodes d'analyse permettant d'extraire un maximum d'informations à partir des données récoltées par les biologistes et généticiens, celles-ci a fait émerger un grand nombre de questions, il est claire qu'une bonne procédure de sélection doit en pratique être complètement explicite, simple à implémenter et rapide à calculer.

La sélection des données biologiques contribuent vers le renforcement de l'aide au diagnostic médical, le niveau et le taux de progression de bio-marqueurs mesurés de façon répétitives sur chaque sujet permettant de quantifier la sévérité de la maladie et la susceptibilité de sa progression ; ceci est usuellement intéressant, sur les plans cliniques et scientifiques, d'aider l'expert à prendre ces décisions dans un temps moins tardif que la survie d'un patient. Le domaine typique de telle situation est le domaine biomédical où nous pouvons maintenant faire énormément de mesures sur un individu donné, mais le nombre d'individus sur lequel nous faisons l'expérience est réduit (dans le cas d'étude d'une maladie, le nombre de porteurs de la maladie qui participent à une étude est souvent limité). Le domaine qui concerne le développement de méthodes qui permettent la sélection de variables pertinentes est très actif, peuvent assurer une meilleure prédiction et de sélectionner correctement ces variables est important pour l'interprétation du modèle

(un clinicien sera évidement intéressé de savoir que tel et tel caractéristique sont impliqués dans le développement d'une maladie bien spécifique par exemple).

Les techniques de sélection (ou réduction) de dimension consistent à rechercher des directives informatives et d'éliminer les directives qui ne contiennent que du bruit. Ces techniques se divisent en deux groupes : les approches multi-variées et les approches purement scalaires.

✓ Les approches multi-variées comme l'optimisation combinatoire et la combinaison linéaire, leurs originalité consiste à chercher non pas un seul facteurs explicatifs mais bien une ou plusieurs combinaisons (sous ensemble) parmi un très grand nombre de facteurs potentiels.

✓ Les approches purement scalaires détiennent principalement les méthodes statistiques et probabilistes qui s'imposent par leurs capacités à exploiter ce nouveau type de données, cette approche se focalise uniquement sur l'extraction de variables réelles et non pas sur la construction de nouvelles variables artificielles pour réduire la dimension.

1.2. Problématique

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

Le problème spécifique de la sélection de variables nécessite une approche particulière puisque le nombre de variables est très largement supérieur vis-à-vis du nombre d'échantillons (expériences ou observations) dans quelques applications, dans la littérature du machine Learning trois approches sont envisagées relèvent des méthodes de type filtre, wrapper ou Embedded ces méthodes sélectionnent de façon implicite les variables où la sélection se fait lors de processus d'apprentissage, ces deux approches sont caractérisées par la pertinences des attributs sélectionnées mais un temps de calcul long à l'opposé de la méthode filtre, approche couramment utilisée à ce jours pour analyser les données

biologiques, cette approche consiste à parcourir la sélection des variables avant le processus de l'apprentissage et ne conserve que les caractéristiques informatives.

Le travail que nous présentons dans ce mémoire s'inscrit dans le contexte général de l'aide au Diagnostic médical, qui a pour but de voir l'intérêt de réduire le nombre de variables parmi lesquelles peu sont informatives et les autres constituent essentiellement du bruit par la sélection de variables pour améliorer les modèles de classification à partir des paramètres descripteurs réduites de différentes maladies étudiées telles que: les maladies cardiaques et le cancer du col de l'utérus qui permet de développer le contexte d'aide au diagnostic pour détecter l'état du patient (malade ou sain) et pourrait apporter plus de connaissance sur les caractéristiques de ces maladies. Aussi, nous mettons en évidence l'utilisation de l'approche filtre qui a été extraite de la littérature scientifique.

1.3. La définition de sélection de variables [9]

1.3.1. La sélection

La sélection est un processus (opération volontaire et méthodique, phénomène inconscient ou automatique) par lequel certains éléments (personnes ou choses) sont choisis en fonction de caractéristiques déterminées, éventuellement impliquées par une certaine fin ou objectif**.** La sélection est une action qui permet de choisir des personnes ou des objets qui conviennent le mieux**.**

1.3.2. La variable

Une variable est sujet à des variations, elle peut changer au cours d'une durée, selon les circonstances elle peut être différente selon les cas.

1.3.3. Sélection de variables

La sélection de variables consiste à sélectionner parmi un ensemble de variables de grande taille un sous-ensemble de variables intéressantes et pertinentes pour un problème donné.

Définition : La sélection de variables est généralement définie comme un processus de recherche permettant de trouver un sous-ensemble pertinent de variables parmi celles de l'ensemble de départ, la notion de pertinence d'un sous- ensemble de variables dépend toujours des objectifs et des critères du système. En général, le problème de sélection de variables peut être défini comme suit : Soit, $F = \{f_1, f_2, f_3, ..., f_n\}$, un ensemble de variables de taille ***N*** où ***N*** représente le nombre total de variables étudiées. Soit $\boldsymbol{Ev}$ une

fonction qui permet d'évaluer un sous-ensemble de variables. Nous supposons que la plus grande valeur de Ev soit obtenue pour le meilleur sous-ensemble de variables. L'objectif de la sélection est de trouver un sous-ensemble ***F' (F'⊆ F***) de taille ***N'(N'≤N)*** tel que $E_V(F) = \max E_V(Z); Z \subseteq F$.

Où $|Z| = N'$, N'peut être un nombre prédéfini par l'utilisateur ou contrôlé par une des méthodes de génération du sous-ensemble [10].

1.4. Processus global de la sélection de variables

Les méthode de sélection des variables cherchent dans les sous-ensembles de variable, et essaient de trouver le meilleur parmi les sous-ensembles candidas $2N$ concurrents selon une fonction d'évaluation. Cependant, cette procédure est exhaustive car elle tente de trouver seulement le meilleur. Il peut être trop couteux et pratiquement prohibitif, même pour un ensemble de caractéristiques de taille moyenne (N). D'autre méthodes basées sur des méthodes de recherche heuristique ou aléatoire tente de réduire la complexité informatique en compromettant les performances ces méthodes on besoin d'un critère d'arrêt pour empêcher une recherche exhaustive des sous-ensembles, il ya étapes de bases dans une méthode typique de sélection des variables (figure 1.1) [11] :

- Une procédure de ***génération de sous-ensembles*** candidats qui détermine l'exploration de l'espace de recherche;
- Une fonction ***d'évaluation*** pour vérifier si le sous ensemble à l'étude ;
- Une condition ***d'arrêt*** pour décider quant arrêter ;
- Un processus de ***validation*** pour vérifier si le sous ensemble est valide.

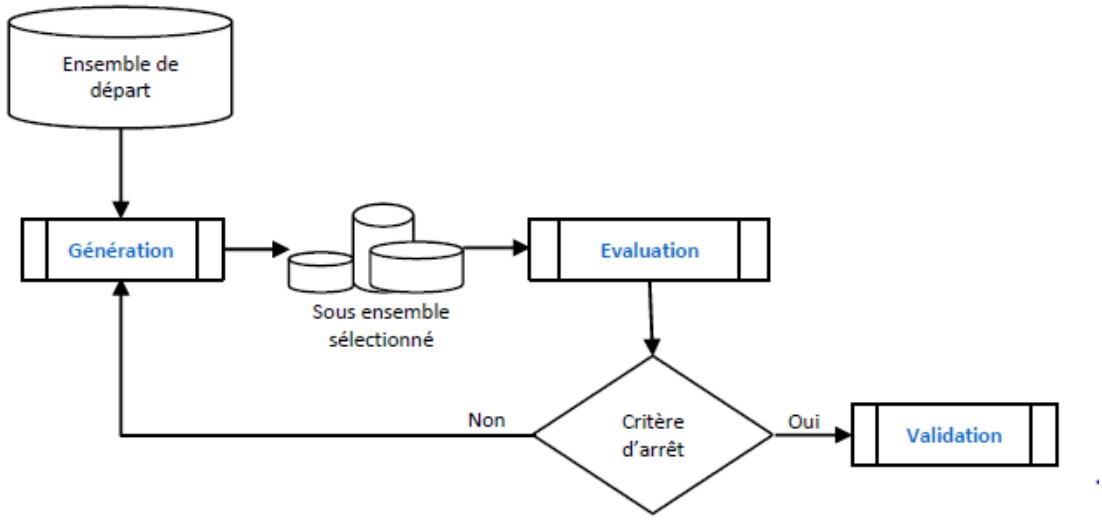


Fig. 1.1. Processus de sélection de variables [10].

Il existe trois types de stratégies de sélection de variables [12] :

Dans la première stratégie : la taille de sous-ensemble à sélectionner est prédéfinie et l'algorithme de sélection cherche à trouver le meilleur sous-ensemble de Cette taille.

La deuxième stratégie : consiste à sélectionner le plus petit sous-ensemble dont la performance est plus grande.

La troisième stratégie : cherche à trouver un compromis entre l'amélioration de la performance et la réduction de la taille du sous ensemble. Le but est de sélectionner le sous-ensemble qui optimise les deux objectifs en même temps.

1.5. Pertinence et redondance de variables

La sélection de variables consiste à choisir parmi un ensemble de variables de grande taille un sous-ensemble de variables intéressantes pour le problème étudié [13].

1.5.1. Pertinence de variables

Kahavi et John définissent les variables pertinentes comme celles dont les valeurs varient systématiquement avec les valeurs de classe. Autrement dit X_i, une variable est pertinente si la connaissance de sa valeur change les probabilités sur les valeurs de la classe Y, selon les défirent définition aussi de variables pertinente, on peut classée les variables comme suivants [13] :

Les variables fortement pertinentes : Ils sont donc indispensables et devraient figurer dans tout sous-ensemble optimal sélectionné, car leurs absences peuvent conduire à un défaut de reconnaissance de la fonction cible (la classe).

La faible pertinence : Elle suggère que la variable n'est pas toujours importante, mais il peut devenir nécessaire pour un sous-ensemble optimal dans certaines conditions.

La non-pertinence : La non-pertinence d'une variable se définit simplement et indique qu'une variable n'est pas du tout nécessaire dans un sous-ensemble optimal de variables.

1.5.2. Redondance de variables

La notion de la redondance de variables se comprend intuitivement et elle est généralement exprimée en termes de corrélation entre variables. On peut dire que deux variables sont redondantes (entre elles) si leurs valeurs sont complètement corrélées, cette définition ne se généralise pas directement pour un sous-ensemble de variables [12].

1.6. Approche de la sélection de variables

La sélection de variables est un dispositif de l'apprentissage.qui permet de évaluer un sous-ensemble et traiter efficacement les valeurs de la variable cible, tout en précisant le type d'approche utilisé. Dans la littérature de la sélection de variables, il existe trois approche principales pour évaluer un sous-ensemble de caractéristiques dans les algorithmes de sélection nous citons:

- Approche filtre (filter),
- Approche wrapper (wrapper approach),
- Approche embedded (embedded approach).

1.6.1. Méthodes de Filtres

Dans les méthodes traditionnelles appelées méthodes filtres ou de filtrage, le principe consiste à évaluer chaque attribut (cas uni-varié) pour lui assigner un score de pertinence. Ce score permet un classement des attributs et in fine la sélection des attributs les mieux classés c'est-à-dire les plus pertinents. Notons que l'on trouve aussi quelques méthodes multi-variées qui attribuent des scores à des groupes d'attributs, l'avantage des méthodes de filtrage est qu'elles peuvent être utilisées lorsqu'on travaille avec un très grand nombre d'attributs car elles sont de complexité raisonnable Elles ne tiennent compte que des informations présentes dans les données et sont indépendantes du processus de classification. Le principal point négatif de ces méthodes est qu'elles évaluent les attributs

individuellement en négligeant les interactions possibles avec les autres attributs. Ces approches ont donc du mal à éliminer les attributs qui sont redondants. Enfin, ces méthodes reposent généralement sur le choix d'un seuil pour le critère de pertinence choisi ou d'un nombre d'attributs à sélectionner qui doit être fixé a priori. Le choix de ces paramètres n'est pas facile à réaliser [14].

Fig. 1.2. L'approche filtre [7] .

1.6.2. Approche enveloppe (wrapper)

Les méthodes wrapper utilisent le prédicateur comme boîte noire et les performances du prédicateur comme fonction objectif pour évaluer le sous-ensemble de variables. Puisque l'évaluation des sous-ensembles $2N$ devient un problème NP-difficile, des sous-ensembles sous-optimaux sont trouvés en utilisant des algorithmes de recherche qui trouvent un sous-ensemble de manière heuristique. Un certain nombre d'algorithmes de recherche peuvent être utilisés pour trouver un sous-ensemble de variables qui maximise la fonction objective qui est la performance de classification. La méthode Branch and Bound a utilisé une structure arborescente pour évaluer différents sous-ensembles pour le numéro de sélection de fonction donnée [15]. De plus, ces méthodes enveloppes sont beaucoup plus coûteuses en temps de calcul puisqu'un classifieur doit être construit chaque fois que l'on doit évaluer un sous-ensemble candidat. Leur complexité de calcul est donc dépendante de la complexité du modèle d'apprentissage utilisé [14].

1.6.3. Approche Embedded (Embedded approch):

Ces méthodes sont proches des méthodes d'enveloppe, car elles combinent le processus d'exploration avec un algorithme d'apprentissage, La différence avec les méthodes enveloppe est que le classifieur sert non seulement à évaluer un sous-ensemble candidats mais aussi à guider le mécanisme de sélection. Selon cette définition, les méthodes de construction d'arbres de décision comme de Breiman relèvent de cette approche puisque la sélection des attributs se fait en même temps que la construction du modèle. Parmi les derniers travaux qui utilisent cette approche, on peut citer Weighted naive Bayes et les travaux dans lesquels le mécanisme d'apprentissage fournit des poids aux attributs pour faire la sélection. On peut citer notamment RFE-SVM, l'avantage de ces

méthodes est que le processus de recherche est guidé par des informations intéressantes fournies par le classifieur, ce qui rend ces méthodes plus efficaces que les méthodes enveloppes [14].

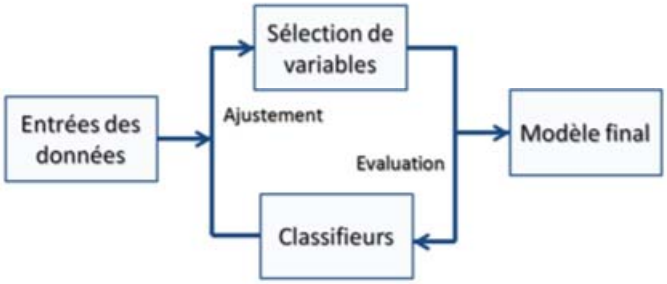


Fig. 1.3. Principe de l'approche wrapper et Embedded [7].

1.7. Notre approche de Sélection de variables (Features Selection)

La sélection de caractéristiques (features selection) est un domaine très actif depuis ces dernières années. Sa particularité s'inscrit dans le cadre de data Manning, en effet, "la fouille de données " dans de très grande base devient un enjeu crucial pour des applications tel que le génie génétique, les processus industriels complexes, il s'agit en fait de résumer et d'extraire intelligemment de la connaissance à partir des données brutes, l'intérêt de la sélection de variables est résumé dans les points suivants [16] :

- Lorsque le nombre de variables est vraiment trop grand, l'algorithme d'apprentissage ne peut pas terminer l'exécution dans un temps convenable, alors la sélection peut réduire l'espace des caractéristiques.
- D'un point de vue intelligence artificielle, créer un classifieur revient à créer un modèle pour les données. Or une attente légitime pour un modèle est d'être le plus simple possible la réduction de la dimension de l'espace de caractéristiques permet alors de réduire le nombre de paramètres nécessaires à la description de ce modèle.
- Elle améliore la performance de la classification : sa vitesse et son pouvoir de généralisation.
- Elle augmente la compréhensibilité des données.

La sélection des caractéristiques est une étape importante lors de la création d'un classificateur sur des données de grande dimension. Comme le nombre d'observations est petit, la sélection des caractéristiques a tendance à être instable. Il est courant que deux sous-ensembles d'entités, obtenus à partir de jeux de données différents mais traitant du

même problème de classification, ne se chevauchent pas de manière significative. Bien qu'il s'agisse d'un problème crucial, peu de travaux ont été réalisés sur la stabilité de la sélection. Le comportement de la sélection de caractéristiques est analysé dans diverses conditions, non exclusivement, mais en mettant l'accent sur les approches de sélection de caractéristiques basées sur le score t et sur de petits échantillons de données. L'analyse se déroule en trois étapes: la première est théorique à l'aide d'un modèle mathématique simple; le second est empirique et basé sur des données artificielles; et le dernier est basé sur des données réelles. Ces trois analyses conduisent aux mêmes résultats et permettent une meilleure compréhension du problème de sélection des caractéristiques dans les données de grande dimension [17].

Au départ, les algorithmes de sélection de fonctionnalités développés sont simples, confortables et informels. Le développement de la technologie a également apporté de nombreux changements dans la sélection des fonctionnalités. De nombreuses approches méthodiques ont été développées. Théoriquement, la sélection des caractéristiques optimales nécessite une recherche exhaustive de tous les sous-ensembles possibles de caractéristiques des critères choisis. Mais une recherche exhaustive de tous les sous-ensembles possibles de fonctionnalités est vraiment une méthode peu pratique si un très grand nombre de fonctionnalités sont disponibles. Ainsi, dans la pratique, il est préférable d'avoir une recherche d'un ensemble satisfaisant de caractéristiques au lieu d'avoir un ensemble optimal de sélection de caractéristiques [18].

Les algorithmes de sélection de fonctionnalités peuvent être classés en deux types: les méthodes de filtrage et les méthodes de wrapper. Les méthodes de filtrage sont beaucoup plus rapides que les méthodes de wrapper et sont donc mieux adaptées aux ensembles de données de grande dimension. La redondance minimale de pertinence maximale (mRMR) est l'un des algorithmes de sélection de fonctionnalités les plus rapides appartenant à la catégorie de méthode de filtrage [19].

1.7.1. Minimum Redundancy Maximum Relevance (mRMR)

Le mRMR est une approche de sélection de variables qui tend pour sélectionner des entités avec une forte corrélation avec la classe (sortie) et une faible corrélation entre eux[20]. L'algorithme de sélection des caractéristiques mRMR est utilisé dans l'étude. Le mRMR est essentiellement un algorithme de filtrage qui tente de sélectionner les fonctionnalités les plus pertinentes pour les étiquettes de classe et pour filtrer le reste. Tout

en identifiant les fonctionnalités les plus pertinentes, l'algorithme tente également de minimiser la redondance entre les fonctionnalités sélectionnées / pertinentes. Plus précisément, l'algorithme mRMR traite chaque caractéristique et le vecteur de classe (variable de réponse ou variable de sortie) en tant que variable aléatoire discrète. Pour mesurer le similitude entre deux entités ou entre un attribut et le vecteur de classe, il utilise mesure d'information mutuelle (I (x, y)). L'information mutuelle est définie comme [21] :

$$I(X,Y)=\sum_{i,j} p(\text{xi},\ \text{yi}) log\frac{p(xi,yj)}{p(xi)p(yj)} \quad (1.1)$$

où p (x) et p (y) désignent respectivement la probabilité de X et Y, p (x, y) est la fonction de densité probabiliste conjointe de X et Y. Supposons que xi représente des caractéristiques individuelles et c représente les classes, max relevance le critère est exprimé comme [22] :

$$max=D(S,c),\ D=\frac{1}{|S|}\sum_{\text{xi}\in S} I(\text{xi}\ ;c) \quad (1.2)$$

où | S | est la dimension de l'espace d'entités S, et I (xi; c) est une information mutuelle entre l'entité individuelle xi et la classe c. Le sous-ensemble de caractéristiques S avec la plus grande corrélation avec les informations de défaut peut être trouvé en utilisant le critère de pertinence maximale, mais les caractéristiques sélectionnées sur la base du critère de pertinence maximale pourraient avoir redondance. Par conséquent, la condition de redondance minimale indiquée dans l'équation est également complétée pour choisir les caractéristiques mutuellement exclusives:

$$min\ R(S)=\frac{1}{|S|^2}\sum_{xi,yj\in S} I(\text{xi}\ ;\text{yj}) \quad (1.3)$$

CONCLUSION

Au cours de ce chapitre, nous avons présenté la problématique de la sélection de caractéristiques pour les problèmes de classification utilisant l'approche filtre. En effet, les outils d'analyse de données moderne doivent travailler dans un espace d'entrée de grande dimension. De plus, ses variables ne sont pas indépendants. Les espaces de grande dimension exhibent des propriétés assez surprenantes et qui vont à l'encontre de l'intuition géométrique et qui ont une grande influence sur les outils dédiés à l'analyse de données. Comme le phénomène de la concentration de norme.

Nous allons introduire quelques méthodes utilisées dans le cadre de classification de données telles que les réseaux de neurones artificiels (RNA), à base radiale (RBF) et k plus

proches voisins (k-PPV) considérées comme des méthodes complémentaires à la réduction de dimensions, qui sont des fonctions de décision appliquée à un système de diagnostic médical à fusion multi sensorielle. Une étude détaillée des mécanismes de ces méthodes hybrides appliquées à la classification de donnée fera l'objet du chapitre suivant.

CHAPITRE II

LES TECHNIQUES D'APPRENTISSAGE

INTRODUCTION

L'évolution technologique durant les dernières années a permis aux scientifiques d'élaborer et de perfectionner des méthodes pour différents domaines. L'évolution des ordinateurs en particulier et la capacité d'intégration de composants formidable atteintes à nos jours ont permis une grande vitesse de calcul et une grande capacité mémoire. Parmi ces méthodes, il existe des méthodes qui sont utilisées dans plusieurs domaines de recherches et de différentes manières, ainsi elles peut être utilisées d'une manière complètement soft en utilisant uniquement l'ordinateur ou d'une manière hard en utilisant les circuits intégrés. Ces méthodes sont basées sur l'apprentissage automatique (Machine Learning) comme un sous-domaine de l'intelligence artificielle qui se concentre sur l'élaboration de modèles capables de représenter certaines caractéristiques du monde qui nous entoure, d'apprendre certaines propriétés statistiques des distributions des données qu'ils traitent, afin d'accomplir diverses tâches [23]. Parmi ces méthodes en trouves : Les réseaux de neurones artificielles (RNA), les réseaux de neurones à bases radiales (RBF) et les k plus proches voisins (k-PPV). Ces méthodes sont des outils puissant capables d'être utilisés dans presque tous les domaines technologiques, et on peut citer : le traitement du signal, vision, parole, prévision, modélisation, aide à la décision, robotique, évaluation des écosystèmes, identification des bactéries, commande des processus, modélisation des systèmes physiques, reconnaissance des formes, mesure, instrumentation,…[24]. Ces méthodes sont "formés" pour modéliser certains systèmes avec l'utilisation de données existantes contenant des affichés spécifiques d'entrées et des sorties du système à modéliser [25].

Dans ce chapitre, nous allons donc pouvoir passer en revue des méthodes d'apprentissage statistique appliquées à la classification. Après une brève introduction, où nous allons la notion de modèle biologique et neurone formel, nous décrivons définition réseaux de neurones artificiels, modélisation générale, son architecture des réseaux, Les types

d'apprentissage des réseaux de neurones et rappelons les réseaux de neurones les plus utilisés (Perceptron Multi Couches (PMC)), ainsi que algorithme de retro-propagation du gradient et formalisation d'apprentissage. Les réseaux de neurones à base radiale est aussi présenté. Nous décrivons les différentes applications de méthodes neuronales. Enfin la méthode des k plus proches voisins est présenté dans un but de comparaison.

2.1. Réseaux de neurones artificiels

2.1.1. Définition

Le réseau de neurone est un composant essentiel du domaine de l'intelligence artificiel qui permet de faire la modalisation des phénomènes statistiques et dynamiques.

Un réseau de neurones (RN) est un système informatique qui a des caractéristiques semblables aux réseaux de neurones biologiques. Il est constitué de plusieurs unités (**neurones**) organisées sous forme de niveaux différents appelés **couches** du réseau. Les neurones appartenant à la même couche possèdent les mêmes caractéristiques et utilisent le même type de **fonction d'activation**. Entre deux couches voisines les connexions se font par l'intermédiaire **de poids** qui jouent le rôle des synapses. L'information est portée par la valeur de ses poids, tandis que la structure du réseau de neurones ne sert qu'à traiter l'information et l'acheminer vers la sortie. La structure ou la topologie d'un réseau de neurones est la manière dont les neurones sont connectés [26].

Les réseaux de neurones se composent trois couches principales :

- couche d'entrée (*input layer*).
- couche de sortie (*output layer*).
- les couches intermédiaires qui s'appellent les couches cachées (*hidden layer*).

L'apprentissage est la propriété le plus important des réseaux neuronaux (pas tous les modèles mais la plupart).

2.1.2. Historique

Les premiers modèles de réseaux de neurones ont été introduits en 1943 par les neurologues McCulloch et Pitts. Toute fois, la technologie de l'époque ne leur a pas permis d'obtenir beaucoup des progrès. D'autres chercheurs comme Donald Hebb, qui a présenté en 1949 une série d'idées sur la structure et le fonctionnement des systèmes biologiques de neurones et Frank Rosemblatt, qui a développé entre 1957 et 1959 le perceptron, un algorithme neuronal simple, ont contribué au développement de ce type d'algorithmes.

Toutefois on a dû attendre le milieu des années 1980 pour que cette approche acquiert une nouvelle force, grâce à l'algorithme d'apprentissage de rétro-propagation (Back Propagation) introduit par Rumelhart et McClelland en 1986, à partir du quel ils ont montré que les réseaux de neurones de multiples couches ont une capacité exceptionnelle de discrimination en étant capables d'apprendre des patrons complexes [27].

2.1.3. Modèle biologique

Le neurone est une cellule de base du système nerveux central qui composé d'un corps cellulaire et d'un noyau. Il est considéré comme unité de traitement d'information en raison des éléments qu'il contient. Chaque neurone contient les éléments suivants :(le noyau, l'axone les synapses et les dendrites) (figure 2.1). Un neurone biologique est une cellule qui se caractérise par [28]:

- **des synapses**, les points de connexion avec les autres neurones, fibres nerveuses ou musculaires.
- **des dendrites** ou entrées des neurones.
- **les axones**, ou sorties du neurone vers d'autres neurones ou fibres musculaires
- **le noyau** qui active les sorties en fonction des stimulations en entrée.

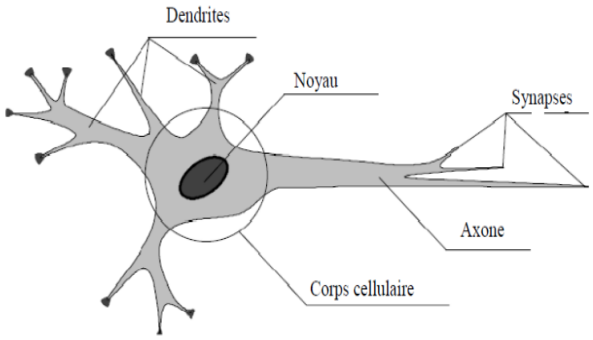


Fig.2.1. Structure d'un neurone biologique [24].

2.1.4. Neurone formel

Un "*neurone formel*" (ou simplement "neurone") est une fonction algébrique non linéaire et bornée, dont la valeur dépend de paramètres appelés coefficients ou poids. Les variables de cette fonction sont habituellement appelées "entrées" du neurone, et la valeur de la fonction est appelée sa "sortie" [29].

Le neurone formel est l'élément essentiel d'un réseau de neurones. C'est un operateur mathématique très simple, dont on peut facilement calculer la valeur numérique. Son

fonctionnement est schématisé sur la figure 2.2. Un neurone formel est une fonction algébrique paramétrée, non linéaire en ses paramètres, et à valeurs bornées. Ses entrées peuvent être les sorties d'autres neurones ou des entrées de signaux extérieurs. Sa sortie est une fonction non linéaire *f* d'une combinaison linéaire Aj des entrées (Pn). Le potentiel Aj le plus fréquemment utilisé est la somme pondérée des entrées Pn pondérées par les coefficients (Wnj) également appelés "**poids de connexions**" [30] :

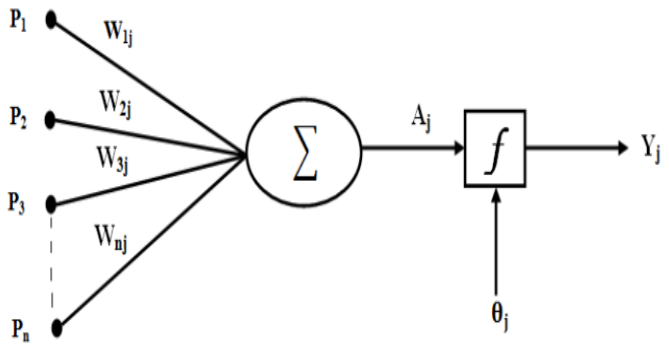


Fig. 2.2. Schématisation d'un neurone formel.

$$A_j = \sum_{i=1}^{n} W_{nj} P_{nj} + \theta_j$$

$$A_J = \sum_{i=1}^{n} W_{nj} P_{nj} + \Theta_j \qquad (2.1)$$

Précisons que le poids Θj est affecté à une entrée constante appelé "biais". Une fonction *f*, appelée fonction d'activation, est appliquée a ce potentiel Aj (Rude, 2008) [30].

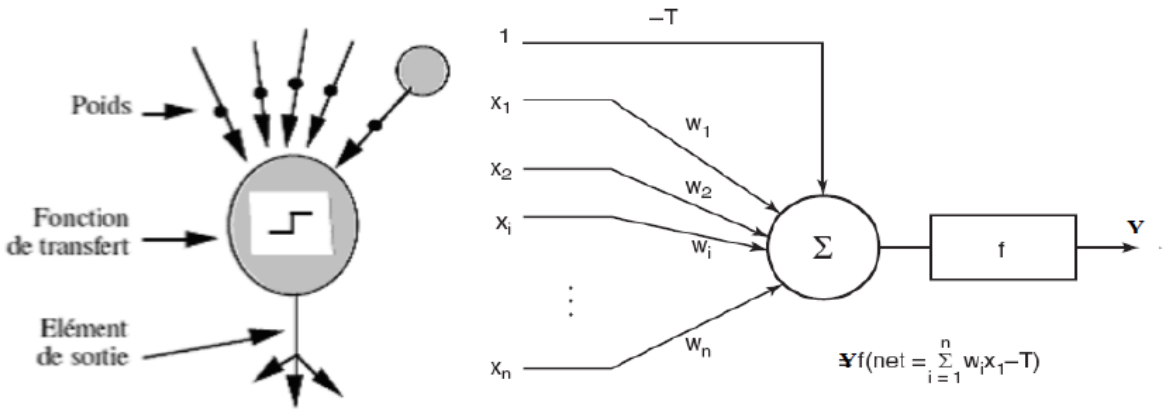
Généralement le modèle de neurone formel constitué de: [31]

- ✓ **Entrées** : sont directement les entrées du système ou peuvent provenir d'autres neurones.
- ✓ **Biais** : sont les entrées qui sont toujours mises à 1 et qui permettent d'ajouter la flexibilité au réseau en variant le seuil de déclenchement du poids de biais lors de l'apprentissage.
- ✓ **Poids** : sont les facteurs multiplicateurs qui affectent l'influence de chaque entrée sur la sortie du neurone.

- ✓ **Noyau** : intègre toutes les entrées et le biais et calcule la sortie du neurone selon une fonction d'activation qui est souvent non linéaire pour donner une plus grande flexibilité d'apprentissage.
- ✓ **Sortie** : la sortie du réseau de neurones peut être distribuée vers d'autres neurones.

On peut comparer la correspondance entre les propriétés respectives de neurones biologiques et neurones artificiels comme le montre le tableau suivant.

Tableau. 2.1: la relation entre le neurone biologique, le neurone formel et le RNA

Neurone biologique	**Neurone formel**
Dendrite	Signal d'entrée
Axone	Signal de sortie
Synapses	Poids de connexion
Corps cellulaire	Fonction d'activation
système nerveux	**réseau neurone Artificiel**
Neurone Le corps cellulaire (soma) Axone Synapse	Traitant élément, noeud, neurone artificiel, neurone abstrait, Niveau de l'activation, fonction de l'activation, fonction du transfert, la fonction de la sortie la communication avec d'autre neurone poids multiplicatifs
	

2.1.5. Fonction d'activation

Cette fonction permet de définir l'état interne du neurone en fonction de ces entrées [32]. Les fonctions les plus utilisées sont la fonction linéaire et la fonction sigmoïde. Leur choix revêt une importance capitale et dépend souvent du type de l'application et du domaine de variation des variables d'entrée/sortie [33].

Il existe plusieurs des formes, nous rappelons quelques formes:

Tableau. 2.2:Différents types de fonctions d'activation pour le neurone formel [34].

Nom de la fonction	Relation d'entrée/sortie	Icône	Nom Matlab
seuil	$a = 0$ si $n < 0$ $a = 1$ si $n \geq 0$		hardlim
seuil symétrique	$a = -1$ si $n < 0$ $a = 1$ si $n \geq 0$		hardlims
linéaire	$a = n$		purelin
linéaire saturée	$a = 0$ si $n < 0$ $a = n$ si $0 \leq n \leq 1$ $a = 1$ si $n > 1$		satlin
linéaire saturée symétrique	$a = -1$ si $n < -1$ $a = n$ si $-1 \leq n \leq 1$ $a = 1$ si $n > 1$		satlins
linéaire positive	$a = 0$ si $n < 0$ $a = n$ si $n \geq 0$		poslin
sigmoïde	$a = \frac{1}{1+\exp^{-n}}$		logsig
tangente hyperbolique	$a = \frac{e^n - e^{-n}}{e^n + e^{-n}}$		tansig
compétitive	$a = 1$ si n maximum $a = 0$ autrement		compet

2.1.6. Le Principe de fonctionnement

Le fonctionnement d'un réseau est un peu plus simple, et dépend uniquement du fonctionnement de ses neurones. Celui-ci doit être considéré comme un circuit électrique contrôlé par une horloge. A chaque cycle, tous les neurones vont, de manière parallèle, calculer une valeur de sortie en fonction de la somme de leurs valeurs d'entrées, sachant que, pour un neurone *A*, les valeurs en entrée au cycle *n* sont les valeurs en sortie au cycle *n*- des neurones connectés à *A* [35].

On va donc assister, à chaque "clockage" du réseau, à une propagation de l'influx nerveux de ses entrées vers ses sorties (figure 2.3)**.**

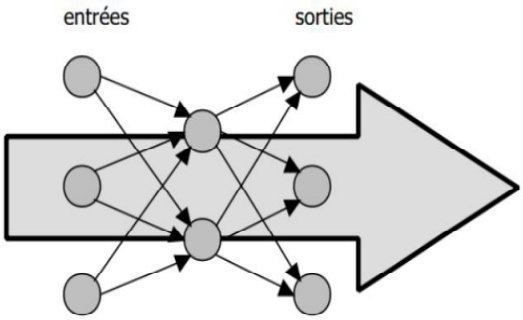


Fig. 2.3: Propagation de l'influe neveu dans un réseau de type perceptron.

2.1.7. Architecture des réseaux de neurones artificiels:

Il existe deux grands types d'architectures de réseaux de neurones : les réseaux de neurones non bouclés et les réseaux de neurones bouclés.

- **Réseaux de neurone bouclé** (*back-forward*)**:**

Un réseau bouclé (récurrent) (figure 2.4), régi par une ou plusieurs équations différentielles, résulte de la composition des fonctions réalisées par chacun des neurones et des retards associés à chacune des connexions. Ces réseaux sont utilisés pour effectuer des taches de modélisation des systèmes dynamiques, de commande de processus ou de filtrage **[36].**

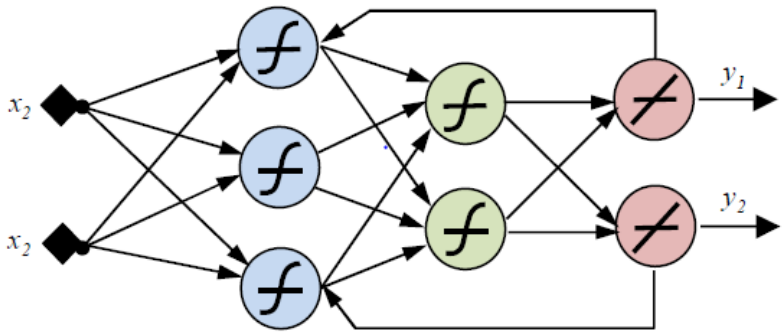


Fig. 2.4. Schéma de Réseau de neurones bouclé [37].

- **Un réseau de neurones non bouclé** (*feed-forward*) (appelé aussi statique): est représenté comme un graphe dont les nœuds sont les neurones. L'information circule des entrées vers les sorties sans retour en arrière (Figure 2.5). Ce type de réseaux est utilisé pour effectuer des taches d'approximation de fonction non linéaire, de la classification ou de la modélisation de processus statiques non linéaires **[**36**].**

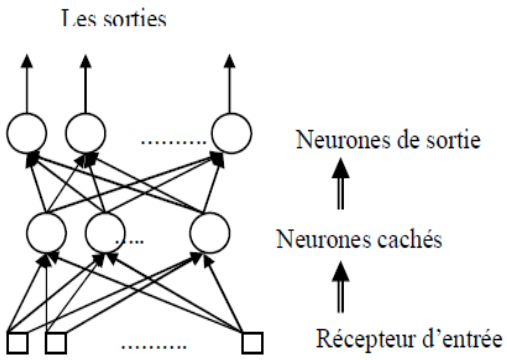


Fig. 2.5: Exemple d'un réseau de neurones non bouclé.

2.1.8. Apprentissage

L'apprentissage des réseaux de neurones artificiels est une phase qui permet de déterminer ou de modifier les paramètres du réseau, afin d'adopter un comportement désiré. Cela est réalisé en adaptant, grâce à certaines règles, les vecteurs de poids. Plusieurs algorithmes d'apprentissage ont été développés depuis la première règle d'apprentissage de Hebb en 1949. L'apprentissage permet aux réseaux de neurones de réaliser des tâches complexes dans différents types d'application (classification, identification, reconnaissance de caractères, de la voix, vision, système de contrôle…) [38,39]. On a déterminé le type d'apprentissage selon la manière dont les paramètres qui sont adaptés.et il existe plusieurs des méthodes et algorithmes pour l'adaptation ces paramètres.

Les méthodes d'apprentissage peuvent être classées en plusieurs catégories : apprentissage supervisé, apprentissage semi-supervisé, apprentissage non-supervisé et apprentissage par renforcement [40].

2.1.9. Types d'apprentissages

Les réseaux à apprentissage supervisé

Dans ce type, une information précise sur la sortie désirée est disponible. Le réseau apprend par présentation de pair d'entrée/sortie. Durant l'apprentissage, les valeurs de sorties désirées sont comparées à celles produites par le réseau. L'erreur résultante est utilisée pour l'ajustement des poids des connexions. La règle du delta en méthode rétropropagtion telle qu'elle est utilisée dans les réseaux multicouches que nous détaillerons dans le chapitre suivant [41].

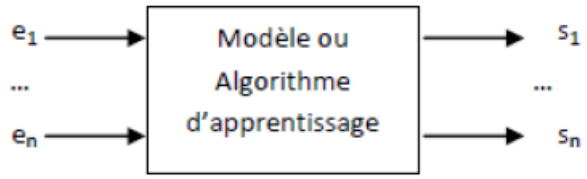


Fig. 2.6. Schéma d'un modèle supervisé. [27]

Les réseaux à apprentissage non supervisé

Dans cet apprentissage, aucune information sur la sortie désirée du réseau n'est disponible.

Ainsi, le réseau manipule des données qui lui sont présentées en entrée et cherche à extraire quelques propriétés qui formeront les sorties du réseau. L'extraction de ces propriétés dépend de la règle d'apprentissage utilisée dans le réseau. [41]

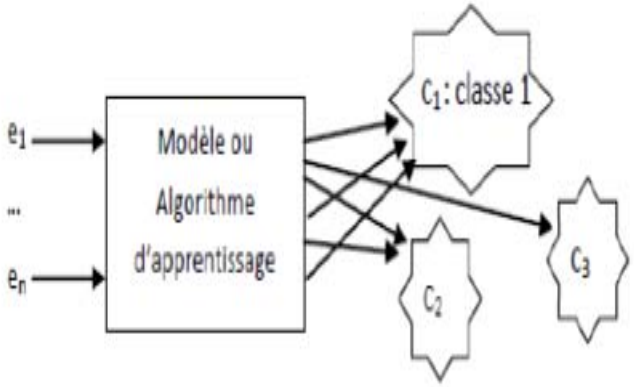


Fig. 2.7. Schéma d'un modèle non-supervisé [27].

2.1.10. Quelques Modèle de réseaux de neurones

2.1.10. 1.Perceptron Multi Couche (PMC)

Le perceptron multicouche ou (PMC, ou *multi-layer perceptron, MLP*).Il est sans doute le plus simple et le plus connu des réseaux de neurones. La structure est relativement simple: une couche d'entrée (Input layer), une couche de sortie (output layer) et une ou plusieurs couches cachées (Hidden layer). Chaque neurone n'est relié qu'aux neurones des couches précédentes, mais à tous les neurones de la couche précédente [39]. Il s'agit d'un réseau de

type (feed-forward) composé de couches successives, ce type de réseau est très performant pour les problèmes de classification [42].

La méthode d'apprentissage utilisée c'est l'apprentissage supervisé.

Les champs d'application des PMC sont très nombreux : discrimination, prévision d'une série temporelle, reconnaissance de forme. . . Ils sont en général bien explicités dans les documentations des logiciels spécialisés [35].

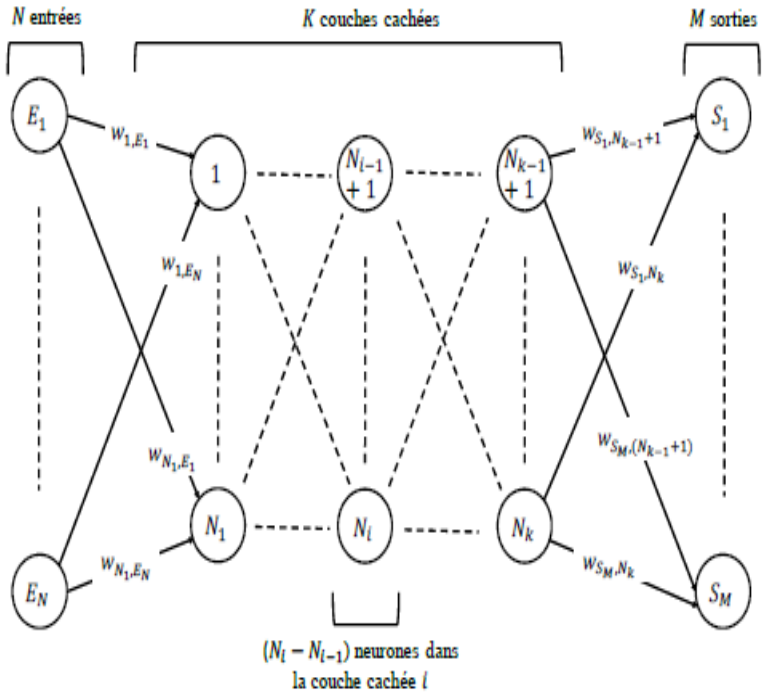


Fig. 2.8. L'architecture d'un réseau de neurones de type MLP [43].

$\{E1 \ldots EN\}$:couche d'entrée.

k: nombre de couche cachée de $(Ni{-}Ni{-}1)$ neurones dans la couche .

$\{S1 \ldots SM\}$:une couche de sortie.

2.1.10. Formalisation d'apprentissage

Pour un réseau multicouche à « m » entrées et « n » sorties, composé de L couches (couches cachées et couche de sortie), la somme de la $j^{éme}$unité cachée est : [44] [45]

$$S_j^L = \sum_{j=1}^{n} w_{ji}^L x_1 + \theta_j^L \qquad (2.2)$$

L : indice de la couche cachée.

W : poids de la connexion$i^{éme}$ entrée.

θ_j^L : bais ou la valeur de seuil interne du neurone

f :est la fonction d'activation de ce neurone j tel que

$$\mathrm{I_j} = \mathrm{f^L}\left(\mathrm{S_j^L}\right) = \sum_{j=1}^{n} \mathrm{w_{ji}^L}\, \mathrm{x_1} + \mathrm{\theta_j^L} \tag{2.3}$$

Les équations de la couche de sortie sont :

$$\mathrm{S_k^R} = \sum_{j=1}^{n} \mathrm{w_{Kji}^R}\, \mathrm{I_j} + \mathrm{\theta_K^R} \tag{2.4}$$

$$R_K = f_K^R\left(\mathrm{S_K^R}\right) \tag{2.5}$$

K : numéro de neurone de la couche de sortie.

L'objectif de rétro-propagation est l'adaptation des paramètres *Wij* de telle façon qu'on minimise la fonction de coût donnée par :

$$E_P = \frac{1}{2}\sum_{K=1}^{T} E = \frac{1}{2}\sum_{K=1}^{T} (\delta)^2 \tag{2.6}$$

Avec :

$$\mathrm{\delta_K} = \mathrm{Y_K} - \mathrm{R_K} \tag{2.7}$$

Où :

$\mathrm{Y_K}$: la sortie désirée

$\mathrm{R_K}$: la sortie actuelle du réseau

T : longueur de l'ensemble d'apprentissage

Adaptation des poids

Après avoir calculé la sortie $\mathrm{R_K}$et l'erreur *E* correspondante a l'ensemble des entrées à partir des équations (2.5) et (2.6), les poids du réseau Sont alors ajuster par la méthode du gradient :

$$\mathrm{w_{ji}^L}(n+1) = \mathrm{w_{ji}^L}(\mathrm{n}) + \mathrm{\Delta w_{jL}^L}(\mathrm{n}) \tag{2.8}$$

$$\mathrm{\Delta w_{jL}^L}(\mathrm{n}) = -\mu \frac{\partial \mathrm{E}}{\partial \mathrm{w_{jL}^L}(\mathrm{n})} \tag{2.9}$$

Avec :

N: numéro de l'itération

M : pas d'apprentissage représentant la vitesse de convergence, sa valeur est généralement choisie expérimentalement (*0 <*μ*<1*)

-Si μest trop petit, la convergence est lente mais la direction de la descente est optimale.

-Si μest trop grand, la convergence est rapide mais la précision est médiocre.

L'application des poids du réseau est faite tout d'abord pour la couche de sortie puis pour les couches cachées on a :

a)pour la couche de sortie

D'après (2-6)

$$E_P = \frac{1}{2}\sum_{K=1}^{T}(Y_K - R_K)^2 \qquad (2.10)$$

La dérive de l'erreur $E_P$par rapport au poids synaptiqueW_k^Rnous donne :

$$\frac{\partial E_P}{\partial w_{ji}^L} = -(Y_K - R_K)\frac{\partial R_k^L}{\partial S_k^L}\frac{\partial S_k^L}{\partial w_{Kj}^L} \qquad (2.11)$$

Le dernier facteur de (II.11) est :

$$\frac{\partial S_k^R}{\partial w_{Kji}^R} = \frac{\partial}{\partial w_{Kj}^L}\left[w_{Kj}^R I_J + \theta_K^L\right] = I_j \qquad (2.12)$$

En combinant (II-10) et (II-11) ; on a pour l'opposé du gradient :

$$\frac{-\partial E_P}{\partial w_{Kj}^R} = (Y_K - R_K) f_k^R\left(S_k^R\right) I_j \qquad (2.13)$$

L'amplitude de la variation du poids étant proportionnelle a l'opposé du gradient, les poids de couche de sortie son renouvelés selon :

$$w_{Kj}^R(n+1) = w_{Kj}^R(n) + \Delta w_{Kj}^R(n) \qquad (2.14)$$

$$\Delta w_{Kj}^R = \mu(Y_K - R_K) f_k^R\left(S_k^R\right) I_j \qquad (2.15)$$

Il y a deux formes de fonctions de sorties qui nous intéressent ici :

- $f_k^R\left(S_k^R\right) = S_k^R$
- $f_k^R\left(S_k^R\right) = [1 + \exp(-S_k^R)]^{-1}$

La première est une fonction linéaire. La deuxième appelée sigmoïde. Pour la fonction linéaire la dérivée est :$f_k^R = 1$

L'équation (2.14) devient :

$$w_{Kj}^R(n+1) = w_{Kj}^R(n) + \mu(Y_K - R_K)f_k^R(S_k^R)I_j \quad (2.16)$$

Pour une fonction d'activation type sigmoïde, la dérivée est :

$$f_k^R = f_k^R(1 - f_k^R) = R_K(1 - R_K) \quad (2.17)$$

L'équation (II.14)devient :

$$w_{Kj}^R(n+1) = w_{Kj}^R(n) + \mu(Y_K - R_K)(1 - R_K)R_K I_j \quad (2.18)$$

En définissant la quantité :

$$S_k^R = (Y_K - R_K)f_k^R(S_k^R) \quad (2.19)$$

Nous pouvons alors écrire l'équation d'adaptation des poids sous une forme indépendante de la fonction de sortie :

$$w_{Kj}^R(n+1) = w_{Kj}^R(n) + \mu\delta_k^R I_j \quad (2.20)$$

b) pour les couches cachées

Nous devons répéter le même calcul que pour la couche de sortie. L'erreur total doit être rapporté d'une façon ou d'une autre aux sorties de la couche cachée :

$$E_P = \frac{1}{2}\sum_{K=1}^{T}(Y_K - R_K)^2 \quad (2.21)$$

$$= \frac{1}{2}\sum_{K=1}^{T}[Y_K - f_k^R(S_k^R)]^2 \quad (2.22)$$

$$= \frac{1}{2}\sum_{K=1}^{T}(Y_K - f_k^R(\sum w_{kj}I_j + \theta_K^R))^2 \quad (2.23)$$

Dans les équations (2-2) et (2-3),on constate que $I_j$dépend des poids de la couche cachée, nous pouvons exploiter ce fait pour calculer le gradient de $E_P$par rapport aux poids de la couche cachée :

$$\frac{\partial E_P}{\partial w_{ji}^L} = \frac{1}{2}\sum_k \frac{\partial}{\partial w_{ji}^k}(Y_K - R_K)^2 \quad (2.24)$$

$$= -\sum_K (Y_K - R_K)\frac{\partial R_K}{\partial S_k^R}\frac{\partial S_k^R}{\partial I_j}\frac{\partial I_j}{\partial S_j^L}\frac{\partial S_j^L}{\partial w_{ji}^L} \quad (2.25)$$

Les facteurs de l'équation (2-20)peuvent être calculés à partir des équations précédentes pour donner :

$$\frac{\partial E_P}{\partial w_{ji}^L} = -\sum_K (Y_K - R_K)\, f_k^R(S_k^R) W_{Kj}^R f_j^L(S_j^L) X_j \quad (2.26)$$

Les poids des couches caches sont adaptés proportionnellement à l'opposé de l'équation (2.26).

$$\Delta W_{Kj}^R = \mu f_j^L(S_j^L) X_i \sum (Y_K - R_K) f_k^R(\delta_k^R) W_{Kj}^R \quad (2.27)$$

Si on utilise la définition de, l'équation (2-27)devient :

$$\Delta W_{Kj}^R = \mu f_j^L(S_j^L) X_i \sum (\delta_k^R) W_{Kj}^R \; (2.28) \; = \mu \delta_j^L X_i$$

Avec :

$$\delta_j^L = f_j^L(S_j^L) X_i \sum \delta_k^R \, W_{Kj}^R \quad (2.29)$$

On consulte que chaque adaptation des poids dans la couche cachée dépend de l'erreur totale de la couche de sortie ce qui conduit a la notion de rétro propagation.
L'équation d'adaptation des poids dans ce cas est :

$$w_{ji}^L(n+1) = w_{ji}^L(n) + \mu \delta_j^L X_i \quad (2.30)$$

2.1.12. Algorithme de la rétro-propagation

Etape 1 : Initialiser les poids W_{ij}et les seuils internes des neurones à des petites valeurs aléatoires.

Etape 2 : Présenter le vecteur d'entrée et de sortie désirée.

Etape 3 : Calculer :

- La somme des entrées des neurones de la couche cachée en utilisant l'expression (2.1) - Les sorties de neurones de la couche cachée en utilisant l'expression (2.2)
- La somme des entrées des neurones de la couche de sortie en utilisant l'expression (2.3)
- Les sorties des réseaux en utilisant l'expression (2.4).

Etape4 : Calculer l'erreur pour les neurones de la couche de sortie en utilisant l'expression (2.19).

Etape5 : Réinjecter l'erreur de sortie en utilisant l'expression (2.29)

Etape6 : Ajuster :

- Les poids de la couche de sortie en utilisant l'expression (2.20)
- Les poids de la couche cachée en utilisant l'expression (2.30)

Etape 7 : Calculer E l'erreur en utilisant l'expression (2.10)

Etape8 : Si la condition sur l'erreur $E - E_p < \varepsilon$ est atteinte, aller a l'étape 9 sinon aller à L'étape 6 et refaire le calcul pour un autre époque.

Etape 9 : FIN.

2.2. Les réseaux de fonction à base radial (RBF)

Les réseaux de neurones RBFs ou (Radial basis Function en anglais). Ils sont principalement utilisés pour résoudre des problèmes d'approximation de fonctions dans des espaces de grandes dimensions. Ils sont lus adaptés, en raison d'apprentissage local. Ce type d'apprentissage peut rendre le processus d'entrainement bien plus rapide que dans le cas d'une MLP, qui apprend de façon globale [35].

Le réseau RBF fait partie des réseaux de neurones supervisés. Il est constitué de trois couches une couche d'entrée qui retransmet les entrées sans distorsion, une seule couche cachée qui contient les neurones RBF qui sont généralement des fonctions gaussiennes et une couche de sortie dont les neurones sont généralement animés par une fonction d'activation linéaire .Chaque couche est complètement connectée à la suivante et il n'y a pas de connexions à l'intérieur d'une même couche. La différence fondamentale par rapport au Perceptron est que le réseau de neurones de type RBF permet d'introduire une contrainte de couverture de la zone d'activation du neurone. Il devient alors possible d'apporter au réseau de neurones, au moment de sa conception, de l'information sur le système considéré. Comme un PMC, un RBF peut être utilisé dans la prédiction, l'identification, la classification …, mais les

réseaux RBF diffèrent des réseaux PMC, du fait que les fonctions d'activation des nœuds de la couche cachée sont des fonctions gaussiennes [46].

2.2.1. Les caractéristiques

Le modèle de réseau RBF est caractérisé par quatre paramètres principaux, qui doivent être réglés, lors de l'étape de construction du réseau. Toute modification d'un de ces paramètres entraîne directement un changement du comportement du réseau [46]. Ces paramètres sont :

- Le nombre de neurones RBF dans l'unique couche cachée ou le nombre des gaussiennes.
- La position des centres des gaussiennes de chacun des neurones.
- La largeur de ces gaussiennes.
- Le poids des connexions entre les neurones RBF et le(s) neurone(s) de sortie**.**

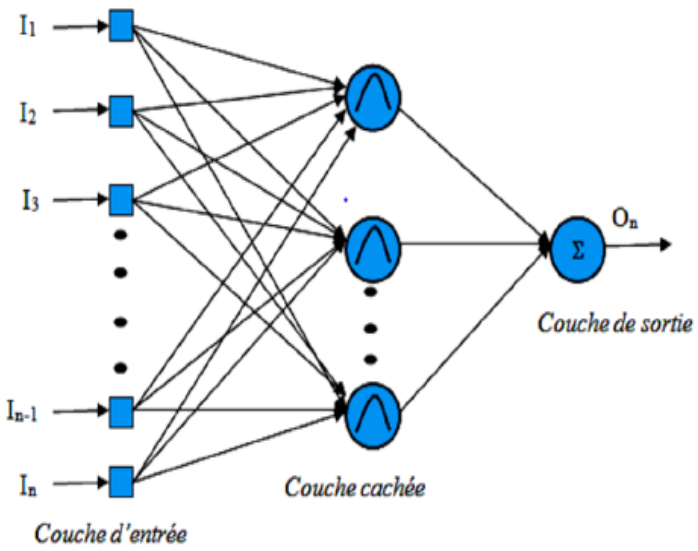


Fig. 2.9. Structure typique d'un réseau de neurones de type RBF.

Ce réseau est constitué de N neurones d'entrée, M neurones cachés et J neurones de sortie. La sortie du m^{ième} neurone de la couche cachée est donnée par [47] :

$$y_m^{(q)} = \exp\left[- \left\| x^{(q)} - v_m \right\|^2 \Big/ (2\sigma_m^2) \right] \qquad (2.31)$$

v_m est le centre du m^{ième} neurone de la couche cachée où du m^{ième} neurone gaussien et $\sigma_m$est la largeur du m^{ième} gaussienne. La sortie du j^{ième} neurone de la couche de sortie est donnée par:

$$Z_j^{(q)} = \left(\frac{1}{M}\right)\left[\sum_{(m=1,M)} w_{mj} y_m^{(q)}\right] \quad (2.32)$$

m =1,.., M et j=1,…, J.

w_{mj} sont les poids reliant la couche cachée à celle de la sortie.

2.2.2. Algorithme d'apprentissage du réseau RBF

L'apprentissage du réseau RBF a été présenté la première fois par Moody et Darken. Il consiste à régler quatre paramètres principaux: le nombre de neurones dans l'unique couche cachée ou le nombre des gaussiennes, la position des centres de ces gaussiennes, la largeur de ces gaussiennes et les poids de connexions entre les neurones cachés et le(s) neurone(s) de sortie. Le réseau RBF consiste à minimiser l'erreur quadratique totale E calculée entre les sorties obtenues du réseau et celles désirées [48] :

$$E = \sum_{q=1}^{Q}\sum_{j=1}^{J} t_j^{(q)} z_j^{(q)^2} \quad (2.33)$$

Pour le réseau RBF, l'ajustement des poids *wmj* reliant la couche cachée à celle de la sortie est réalisé par la règle de Widrow-Hoff. Il se fait comme suit :

$$w_{mj}^{(i+1)} = w_{mj}^{(i)} + \eta(t_j - z_j) y_m \quad (2.34)$$

t_j est la sortie du j^{ième} neurone désirée, z_j est la sortie du j^{ième} neurone calculée, y_m est la sortie du m^{ième} neurone de la couche cachée et η est le pas d'apprentissage dont sa valeur est comprise entre 0 et 1.

En conclusion, mentionnons que la principale difficulté des réseaux RBF concerne la question du nombre de neurones radiaux `a utiliser pour une application donnée. A priori, il n'existe pas de méthode pour fixer leur nombre, et cette architecture souffre de façon particulièrement aigüe de ce qu'on appelle la «malédiction de la dimension», à savoir l'augmentation exponentielle du nombre de neurones cachés requis en fonction de la dimension R de l'espace d'entrée. Lorsque R est grand, une façon d'atténuer ce problème consiste `a remplacer les hyper-sphères qui résultent de l'imposition d'une variance fixe par des hyper-ellipses o`u la matrice de covariance n'est plus contrainte. On peut ainsi réduire le nombre de neurones à positionner au détriment du nombre de paramètres à estimer.

2.3. Les applications des méthodes neuronales

Les grands domaines d'application des réseaux de neurones découlent naturellement des propriétés énoncées précédemment. Nous présentons dans les sections suivantes quelques exemples pour montrer le vaste étendu de leur applicabilité [49, 50, 51]:

- **La régression non linéaire, ou modélisation de données statiques**

Une immense variété de phénomènes statiques peut être caractérisée par une relation déterministe entre des causes et des effets ; les réseaux de neurones sont de bons candidats pour modéliser de telles relations à partir d'observations expérimentales, sous réserve que celles-ci soient suffisamment nombreuses et représentatives.

- **La modélisation de processus dynamiques non linéaires**

Modéliser un processus consiste à déterminer un ensemble d'équations mathématiques qui décrivent le comportement dynamique du processus, c'est-à-dire l'évolution de ses sorties en fonction de l'évolution de ses entrées .Ce problème peut être avantageusement résolu par un réseau de neurones, si le phénomène que l'on désire modéliser est non linéaire.

- **La commande de processus**

La commande d'un processus consiste à concevoir un système comprenant un organe qui calcule la commande à appliquer au processus pour assurer un comportement dynamique spécifié par des cahiers de charges: régulation au voisinage d'un point de fonctionnement, poursuite d'une trajectoire de consigne, commande optimale….

L'ensemble commande / processus peut donc être considéré comme un système qui réalise une fonction (non linéaire) qu'un réseau de neurone peut approcher**.**

- **La classification**

Une autre grande catégorie de problème industriel consiste à attribuer de façon automatique un objet à une classe, parmi d'autres classes possibles. Et en raison de leur propriété d'approximateurs universels, les réseaux de neurones sont capables d'estimer de manière précise la probabilité d'appartenance d'un objet inconnu à une classe parmi plusieurs possibles**.** Aujourd'hui, les réseaux de neurones artificiels donnent de bons résultats dans différant domaines d'applications, mais il existe des problèmes les techniques d'apprentissages classiques ne peuvent pas être résolues.

2.4. La méthode des k plus proches voisins kppv

L'algorithme des *k* plus proches voisins appartient à la famille des algorithmes d'apprentissage automatique (machine learning). L'idée d'apprentissage automatique ne date pas d'hier, puisque le terme de machine learning a été utilisé pour la première fois par l'informaticien américain Arthur Samuel en 1959. Les algorithmes d'apprentissage automatique ont connu un fort regain d'intérêt au début des années 2000 notamment grâce à la quantité de données disponibles sur internet. L'algorithme des k plus proches voisins est un algorithme d'apprentissage supervisé, il est nécessaire d'avoir des données labellisées. À partir d'un ensemble E de données labellisées, il sera possible de classer (déterminer le label) d'une nouvelle donnée (donnée n'appartenant pas à E). L'algorithme des k plus proches voisins est une bonne introduction aux principes des algorithmes d'apprentissage automatique, il est en effet relativement simple à appréhender.

2.4.1. Définition

La méthode des plus proches voisins est une méthode de classification géométrique très utilisée en reconnaissance de formes, en raison de sa simplicité et de sa robustesse. Les caractéristiques sont exploitées dans un espace métrique de représentation, généralement R^n muni de la distance euclidienne [52]. Les k plus proches voisins plus connus en anglais sous le nom *K-Nearest Neighbor* (*K-NN*) est une méthode d'apprentissage non paramétrique qui ne nécessite pas de construction de modèle, C'est l'échantillon d'apprentissage, associé à une fonction de distance et d'une fonction de choix de la classe en fonction des classes des voisins les plus proches, qui constitue le modèle[53].

2.4.2. Principe de l'algorithme

L'algorithme de k plus proches voisins ne nécessite pas de phase d'apprentissage à proprement dite, il faut juste stocker le jeu de données d'apprentissage. Le principe de cet algorithme de classification est associé à une fonction de distance et à une fonction de choix de la classe majoritaire en fonction des classes des voisins les plus proches, qui constitue le modèle. La figure 2.10 montre ce principe qui a pour but de trouver la valeur de la classe où l'inconnu x va être affecté, nous avons donc deux classes, et nous avons pris $k = 3$. Parmi les 3 voisins, nous avons 2 qui appartiennent à ω1 et 1 qui appartient à ω2 donc x sera affecté à ω1, la classe majoritaire [54].

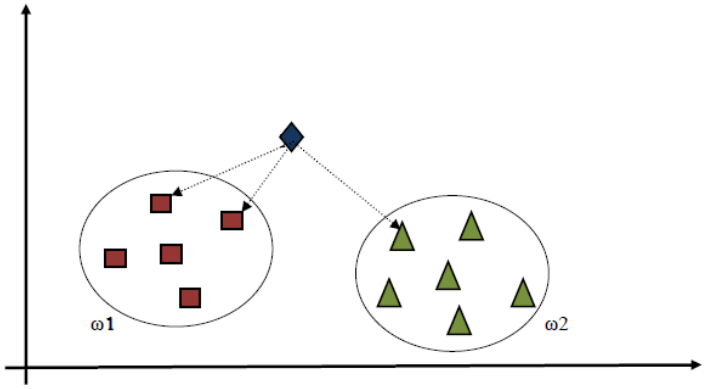


Fig. 2.10. Principe de fonctionnement de l'algorithme K-ppv

2.4.3. *Choix de K*

Le paramètre *k* est l'un des paramètres à déterminer lors de l'utilisation de ce type de méthode, la valeur que l'on choisit pour K va être plus critique, plus déterminante en rapport avec la performance du classificateur (figure 2.11), on peut se permettre de considérer un plus grand nombre de voisins, sachant que plus ils diffèrent du document à classer, moins ils on d'impact sur la prise de décision. Cependant, il demeure nécessaire de limiter le nombre de voisins pour s'en tenir un temps de calcule raisonnable [55].

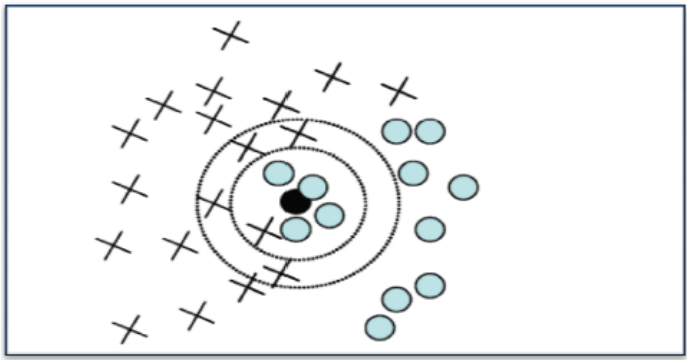


Fig. 2.11. Le choix de «K» influence de décision : pour K=5, la décision est de classer l'objet «noir» dans la classe «rond». Pour k=9, la décision est de classer en tant que «croix».

2.4.4. *Algorithme*

L'algorithme des k-plus proches voisins (k-PPV) est un des plus anciens mais aussi de plus simple algorithme de classification. Il consiste, pour chaque nouvel exemple x à catégoriser, à déterminer les k-plus proches voisins de x d´déjà catégorisées et, à affecter x à la classe la plus représentée dans ces plus proches voisins. Cet algorithme est utilisé dans différentes communautés (base de données, recherche d'information, apprentissage). Plusieurs approches ont été proposées pour améliorer l'algorithme des k-PPV en utilisant la géométrie des données. Ces approches se sont essentiellement concentrées sur des métriques

généralisant la distance euclidienne Or, dans beaucoup de situations, ce ne sont pas des distances qui nous intéressent mais plutôt des similarités [55].

Paramètre : le nombre k de voisins

Donnée : un échantillon de m exemples et leurs classes

-La classe d'un exemple X est c(X)

Entrée : un enregistrement Y

1. Déterminer les k plus proches exemples de Y en calculant les distances

2. Combiner les classes de ces k exemples en une classe c

Sortie : la classe de Y est c(Y)=c

Le choix de la distance est primordial au bon fonctionnement de la méthode.

Les distances les plus simples permettent d'obtenir des résultats satisfaisants .

Fig. 2.12. Forme générale d'Algorithme de k-PPV [56].

CONCLUSION

Ce chapitre a fait l'objet de quelques définitions et généralités sur les techniques d'apprentissage. Nous avons rappelé les concepts les plus importants pour comprendre l'intérêt des réseaux de neurones de type MLP et RBF et K-PPV comme outils de classification dans un but de diagnostic.

L'étude et l'analyse des performances de ces méthodes choisies appliquées au domaine de diagnostic médical, Constituent notre principal objectif. Une étude en simulation ayant pour but d'évaluation des performances de ces techniques associée à une étude statistique à base de réduction de dimension par ACP et mRMR, fera l'objet du chapitre suivant. L'évaluation des résultats, reflétant les performances obtenues, nous conduira au meilleur choix de la méthode hybride la mieux adaptée à l'application.

CHAPITRE III

SIMULATION ET EVALUATION

INTRODUCTION

Ce dernier chapitre est consacré à la simulation et l'application des techniques de d'apprentissage proposées comme étant une solution pour le diagnostic médicale en particulier, à la conception des modèles de classification à partir des paramètres descripteurs de différentes maladies étudiées telles que: les maladies cardiaques et le cancer du col de l'utérus. L'objectif est de valider et d'évaluer les performances des techniques d'apprentissage de chacune des méthodes présentées à savoir MLP, RBF et k-PPV. Cette étude a été précédée d'une analyse statistique « mRMR » pour la sélection des caractéristiques qui nous permettra de déterminer les corrélations existantes entre les variables descripteurs dans un but de réduction de dimension des variables d'entrée. Les exigences principales d'efficacité sont formulées sur deux points essentiels à savoir, les tests de spécification qui vérifient que le programme réalise bien la tâche pour laquelle il a été conçu, et les tests de performances qui vont servir à mesurer l'efficacité avec laquelle cette tâche est remplie.

Une étude en simulation est décrite, permettra de valider et d'évaluer les performances des méthodes présentées. Afin de voir que cette modélisation hybride est adaptable à l'application indiquée, on évaluera pour les méthodes exposées dans une étude comparative, les paramètres liés au taux de reconnaissance, au temps d'apprentissage et à l'erreur d'entraînement. Une discussion des résultats conclura cette étude de simulation pour choisir la technique la mieux adaptée.

3.1. Position de problème

Plusieurs travaux sont effectués afin de développer des outils de classification et d'aide au diagnostic médicale. L'utilisation des méthodes dites intelligentes pour effectuer

cette classification sont de plus en plus fréquentes. Même si la décision du médecin est le facteur le plus déterminant dans le diagnostic, l'aide au diagnostic médical s'est développé et a gagné en popularité. Ces systèmes sont même considérés comme étant essentiels dans beaucoup de disciplines médicales car ils permettent d'assurer un diagnostic préliminaire plus exact et réduisent au maximum les erreurs dues à la fatigue et les doutes du médecin [57].

Nous avons montré dans notre étude les performances des méthodes proposées telles que : les réseaux de neurones artificiels (RNA), les réseaux de fonctions à base radiale (RBF) et la méthode des k plus proches voisins (k-PPV) afin de classifier l'état des patients à travers des paramètres descripteurs des maladies étudiées.

Il s'agit dans cette partie de travail d'évaluer les performances des techniques choisies qui est issue, rappelons-le, du domaine de l'intelligence artificielle. Des techniques servant comme outils de base pour l'aide à la décision et présentant une réponse plus élaborée par rapport aux autres techniques se basant sur des données de diagnostic. Le choix effectué sur la base des résultats obtenus, conduira à l'intégration de la technique sélectionnée au niveau d'un système d'aide à la décision médicale. Le processus de diagnostic est vu comme un problème de reconnaissance de formes, où les classes correspondent aux différents états de patient, et les formes représentent l'ensemble de mesures des paramètres liés à ses caractéristiques. La technique d'aide à la décision utilisée effectué la classification et la séparation des données en deux états ou classes bien différentes. Un module d'apprentissage supervisé par un expert, permet de collecter de manière continue les paramètres relatifs aux différents états des patients pour la mise en œuvre d'une base de connaissance complète.

3.2. Approche utilisée dans la surveillance

La solution devant être adoptée par les techniques proposées citées ci-dessus au problème de reconnaissance de formes posé, ne s'applique en fait que si on se trouve dans le cas d'un apprentissage supervisé. Nous procédons donc lors d'une étape préliminaire d'apprentissage, à paramétrer le classificateur pour la reconnaissance. L'étape de test ou de reconnaissance proprement dite, s'effectue une fois le modèle statistique établi. Il y a ici tout l'intérêt pour dire que cette approche se caractérise par sa souplesse et sa généricité. A souligner toutefois que les méthodes de reconnaissance de formes à base d'apprentissage statistique sont les plus utilisées dans les systèmes de classification à fusion multi-

sensorielle. En général l'apprentissage est une étape assez longue, et nécessite plus de temps de calcul. Les techniques partagent ce point commun mais diffèrent sur un certain nombre d'autres points. L'étude effectuée dans les paragraphes suivants en fera la différence. Ce critère (temps d'apprentissage) aussi important dans le choix du modèle de reconnaissance, évoque un traitement hors ligne devant être effectué par le système de surveillance. Le déroulement de cette opération en permanence contribue sans doute à enrichir une base de connaissance qu'on veut qu'elle soit la plus complète possible pour le modèle de surveillance implanté. Le système d'aide à la décision doit donc pouvoir marier à la fois une surveillance permanente des patients et un apprentissage en arrière plan (en différé). Un opérateur (ou système) expert supervisant cet apprentissage permet de collecter de manière continue les paramètres relatifs aux différents états des patients.

3.3. Description des bases de données pour le diagnostic médicale

Nous cherchons à décider sur l'état des patients à travers ses paramètres descripteurs médicales de différentes maladies étudiées telles que: les maladies cardiaques et le cancer du col de l'utérus. Nous n'avons en fait aucune connaissance a priori sur un type de modèle représentant parfaitement ce procédé, par contre nous pouvons porter notre jugement sur l'état des patients à partir de quelques données descriptives. L'objectif qui se trouve derrière la collecte des données relatives à ces paramètres est de trouver un modèle de classification permettant de distinguer deux états bien distincts de patient (malade ou non). Il y a donc intérêt de disposer d'au moins une période assez longue pour archiver des données afin de déterminer une base de connaissance riche en informations capable de fonctionner normalement, et la présence d'un expert.

On utilise dans cette étude, deux bases de données réelles différentes de source biomédicale. Nous présentons par la suite la description de chaque base et leur caractéristique (site http://www.kaggle.com)

3.3.1. Maladies cardiaques (Heart disease dataset)

La base de données des maladies cardiaques est constituée 1026 cas, chaque cas est formée de 13 attributs en plus de sortie:

- 13 représentent les facteurs de risque (paramètres descripteurs d'entrées).
- Le dernier attribut représente la sortie (l'état du patient **-** Target**:** 1 ou 0)

On peut résumée les informations attributs comme suivant [58]:

Age : âge en années.

Sexe: 1 = male; 0 = femelle**.**

Cp : type de douleur thoracique**.**

Trestbps: tension artérielle au repos**.**

Chol : cholestérol sérique en mg/dl**.**

Fbs: glycémie à jeun> 120 mg / dl) (1 = vrai; 0 = faux

Thalach: fréquence cardiaque maximale atteinte**.**

Exang: angine de poitrine induite par l'exercice (1 = oui; 0 = non).

Oldpeak: Dépression ST induite par l'exercice relatif au repos**.**

Slope: la pente du segment ST d'exercice de pointe**.**

Ca: nombre de vaisseaux majeurs (0-3).

Thal: 3 = normal; 6 = défaut fixe; 7 = défaut réversible**.**

3.3.2. Cervical cancer (cancer du col de l'utérus):

Cette base de données a été construire la liste des facteurs de risque de cancer du col de l'utérus pour la biopsie. 33 attributs décrivent les informations démographiques, l'activité sexuelle et la fréquence, les grossesses, l'âge, la dépendance à la fumée, l'utilisation de contraceptifs et les maladies sexuellement transmissibles diagnostiquées de 968 patientes, les valeurs manquantes sont dues à la décision des patients de ne pas divulguer ces détails en raison de problèmes de confidentialité. Nous présenterons un contexte et des clarifications sur la signification des attributs pour décrire la relation avec la maladie [59]:

- **Hormonal Contraceptives**: Les contraceptifs hormonaux indiquent si le patient utilise des contraceptifs hormonaux. DIU indique l'utilisation du contraceptif intra-utérin Dispositif.
- **STDs**: MST indique si le patient a eu au moins un Maladie sexuellement transmissible. C'est positif lorsqu'au moins une des MST: condylomatose, MST: condylomatose cervicale, MST: vaginale condylomatose, MST: vulvo-périnéale condylomatose, MST: syphilis, MST: pelvienne maladie inflammatoire, MST: herpès

génital, MST: molluscum contagiosum, MST: SIDA, MST: VIH, MST: hépatite, MST: HPV les attributs sont positifs.

- **STDs**: le VPH indique si le patient est infecté par le virus du papillome humain.
- **Dx**:dix le VPH révèle si le patient était déjà diagnostiqué avec le VPH dans le passé.
- **Dx**: dix le cancer est positif si le patient Avait un cancer. Il n'est pas clair si cela se réfère à Cancer du sein, tel que diagnostiqua blé par l'on cotype Test DX ou tout type de cancer. L'original l'article présentant cet ensemble de données n'a pas informations à ce sujet .
- **Dx: CIN** indique si le patient a été affecté par Néoplasie intra épithéliale cervicale (CIN).
- **Dx**: est positif si au moins une des variables Dx est positif.
- **Hinselmann and Citology**: sont deux (complémentaires) tests pour détecter le cancer du col de l'utérus.
- **Schiller**: est un autre test capable d'identifier les anomalies zones à biopsies et à examiner.
- **Biopsy:** La biopsie est une procédure médicale fournissant un confirmation du diagnostic par colposcopie, une inspection visuelle agrandie du col de l'utérus.
- Cet attribut a été choisi comme notre cible.

Le matériel utilisé pour réaliser nos expériences de simulation est le suivant: nous avons utilisé un processeur Intel Core TM i7-6820HQ et 2,71 GHz avec une mémoire RAM de 8 Go. Toutes les méthodes utilisées ont été implémentées dans l'environnement Toolbox Matlab.

3.4. Simulation et Evaluation

Le développement des systèmes d'aide à la décision pour le diagnostic médicale joue un rôle important dans la surveillance des maladies des patients. Toutefois, les méthodes utilisées traditionnellement souffrent d'un handicap dû à la non-linéarité et la complexité des relations entre les variables descripteurs de l'état des patients. On peut dire qu'il n'y a pas de méthode générale acceptée à ce jour. Toutefois, les méthodes d'intelligence artificielles souvent considérés comme une solution incontournable aux problèmes de modélisation de processus non linéaires, sont plus sollicités. D'ailleurs, plusieurs travaux de diagnostic médicale se basant sur ce type de méthodes, ont été élaborés.

Les méthodes d'IA ont étés aussi largement utilisée pour résoudre des problèmes de classification dans plusieurs domaines d'application. Dans notre application, la classification est l'approche de diagnostic utilisée pour la surveillance médicale.

3.4.1. Apprentissage et Test

Les méthodes de reconnaissance de formes telle que le MLP (PMC ou MLP en anglais), RBF et K-PPV appliquées à la classification des données, présente l'avantage de couvrir un grand nombre d'applications. Elles sont utilisées pour les systèmes de décision de hauts niveaux, et fondées sur l'analyse de données expérimentales. Il est alors primordial de procéder au choix de la technique la mieux adaptée, afin de pouvoir l'intégrer éventuellement dans un système de diagnostic. Pour une évaluation des performances des deux modèles précédemment testés, définissons quelques mesures statistiques. Soient :

- **Training**: la phase d'apprentissage (75%) 2/3.
- **Testing** :la phase de test (25%) 1/3.
- **Taux de classification**: $T = \frac{NB}{N} * 100\ \%$

N_B: nombre des échantillons les biens classées (cas de 0).

N: nombre totale des échantillons.

- **Erreur quadratique moyenne de généralisation**:

$$EQM = \frac{1}{n}\sum_{i=1}^{n} e^2(i) \tag{3.1}$$

Où : $e(i) = y_r(i) - y_{e/c}(i)$, tel que : y_r: sortie réelle, $y_{e/c}$: sortie calculée (apprentissage) ou estimée (test).

Mise en œuvre des modèles Neuronaux

L'apprentissage supervisé du modèles neuronaux MLP (figure 3.1) et RBF (figure 3.2) consiste à déterminer les poids de celui-ci qui minimisent sur l'ensemble des données de la base d'apprentissage, les écarts entre les valeurs de sortie désirée y_d et les valeurs de sortie calculée y_c. Ceci consiste à trouver le minimum du critère quadratique suivant :

$$C_w = \frac{1}{N}\sum_{i=1}^{N} (y_{c_i} - y_{d_i})^2 \tag{3.2}$$

Où N est le nombre d'exemples de la base d'apprentissage.

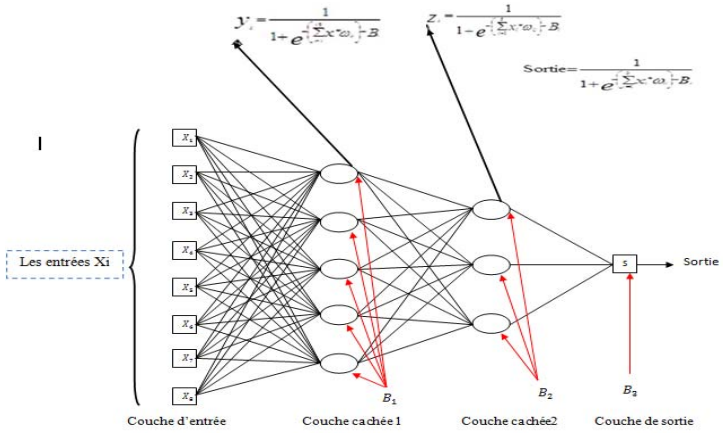


Fig. 3.1. Architecture du MLP.

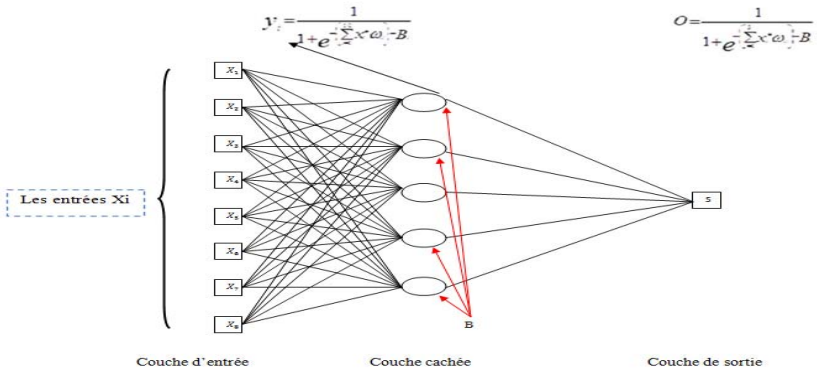


Fig.3.2 .Architecture du réseau RBF.

C'est un problème d'optimisation non-linéaire classique. La fonction d'activation utilisée est la fonction sigmoïde et RBF pour les deux réseaux MLP et RBF respectivement. Ayant la propriété d'être différentiable, condition nécessaire pour certains algorithmes d'apprentissage, elle permet aux fonctions dérivées qui en découlent d'être exprimées facilement à l'aide des mêmes fonctions, ce qui assure un gain en temps de calcul non négligeable.

Différentes architectures sont testées et évaluées pour déterminer le meilleur modèle pour les méthodes utilisées. On présente dans le tableau 3.1, les résultats correspondant aux bases de données choisies : les maladies cardiaques et le cancer du col de l'utérus. Les paramètres tels que le nombre d'itérations (*NI*), le temps d'apprentissage (*T_appr*) et l'erreur d'apprentissage (*Er_appr*) et le taux de test (*Taux_test*) sont indiqués pour les réseaux testés. En matière de reconnaissance, on peut d'ores et déjà supposer qu'un taux plus de 75% est jugé acceptable.

Tableau. 3.1. Résultats d'apprentissage et de test des deux modèles MLP et RBF.

Base de données / Paramètres	Models	Architecture	NI	T_appr (s)	Er_appr	Taux_test (%)
Heart Disease datasets	**MLP**	24-8	1000	183.4	0.494	89,18
	RBF	500	500	117.7	$2.17\ 10^{-29}$	95.61
Cervical Cancer datasets	**MLP**	100-8	100	184.5	0.007	93.12
	RBF	1000	1000	72	$2.07\ 10^{-31}$	95.19

D'après les résultats obtenus, et pour l'ensemble des tests, un taux de reconnaissance supérieur à 89 % est obtenu avec les réseaux testés pour les deux modèles. Les taux de reconnaissances fournis par le réseau RBF sont souvent les meilleurs. Cependant, il faut noter que le temps d'apprentissage et le nombre de neurones cachés pour le réseau RBF sont beaucoup plus importants que ceux des réseaux neuronaux classiques, c'est la complexité du modèle. Une amélioration positive de l'erreur d'apprentissage est constatée lorsqu'il y a association de neurones de plus en plus croissante dans la couche cache. La phase d'apprentissage du réseau RBF est plus rapide que celle du modèle MLP, et il converge pour un nombre réduit d'itérations par rapport aux réseaux MLP. Dans ce cas, le réseau RBF est le mieux placé et semble être adapté pour résoudre des problèmes non-linéaires complexes tels que la classification.

Mise en œuvre du modèle k-PPV

Dans notre cas, nous entraînerons k-PPV pour différentes valeurs de k et dix fonctions de distance (*Euclidean, mahalanobis, cosine, correlation, spearman, hamming, jaccard, minkowski, cityblock, chebychev*) pour déterminer le meilleur modèle avec les deux bases de données médicales utilisées auparavant. On présente dans le tableau 3.2 les résultats d'apprentissage et de test correspondant aux paramètres d'évaluation tels que le temps d'apprentissage (*T_appr*) et le taux d'apprentissage (Taux_*appr*) et de test (*Taux_test*).

Tableau. 3.2. Résultats d'apprentissage et de test du modèle k-PPV.

Base de données / Paramètres	Models	Distance	*K*	T_appr (s)	Taux_appr (%)	Taux_test (%)
Heart Disease datasets	**K-PPV**	Hamming	1	0.0158	100	94.34
Cervical Cancer datasets	**K-PPV**	Euclidean	3	0.0142	96.08	94.59

D'après les résultats obtenus, et pour l'ensemble des tests, un taux de reconnaissance supérieur à 94 % est obtenu. La phase d'apprentissage est plus rapide que celle de deux modèles précédents MLP et RBF.

Les résultats obtenus soulignés jusqu'ici, insistent sur l'intérêt que présentent les modèles de classification utilisés dans les deux bases de données étudiées. Nous proposons dans ce qui suit une validation de ces modèles sur des données réduites utilisant la sélection des caractéristiques d'entrée par mRMR.

3.4.2. Sélection des caractéristiques

Pour des raisons de performances, la préparation et l'analyse de données ont besoin d'un traitement spécifique pour garantir une bonne décision en sortie. A cet effet, l'extraction de caractéristiques est une étape fondamentale du processus de reconnaissance, préalable à l'étape de classification, qui permet d'extraire les caractéristiques pertinentes dans une base de données. Le processus d'analyse de données est utilisé dès lors que les variables d'entrée se présentent en trop grand nombre pour être appréhendées par l'esprit humain. La représentation des données multidimensionnelles (cas de notre application) dans un espace à dimension réduite, fait partie des analyses factorielles où une méthode telle que celle de *Minimum Redundancy Maximum Relevance* (mRMR), est connue et appliquée pour ses performances dans le domaine.

Dans ce qui suit, nous présentons la combinaison de la méthode mRMR avec la deux techniques RBF et k-PPV, et ce toujours dans le but du diagnostic médicale. mRMR est utilisé pour la sélection des caractéristiques comme variables d'entrée pour en extraire celles qui sont pertinentes et non redondantes. L'intérêt d'une telle application, est de montrer la validité de la conception de l'approche proposée.

Dans un premier temps, une analyse par mRMR est effectuée afin de déterminer les paramètres descripteurs ou variables d'entrée les plus représentatives de deux bases de données étudiées. Les deux approches indiquées plus haut (RBF et k-PPV) vont être appliquées dans ce qui suit avec réduction de dimension pour ces meilleures résultats dans la première phase de simulation. Le tableau 3.3 montre les résultats de ce test avec le taux de réduction des paramètres d'entrée après sélection des caractéristiques.

Tableau. 3.3. Résultats d'apprentissage et de test des modèles RBF et k-PPV.

Base de données / Paramètres	Models	Taux de réduction	T_appr (s)	Taux_appr (%)	Taux_test (%)
Heart datasets	**mRMR_k-PPV** *cityblock, k*=1	61%	0.014	98.84	98.64
	mRMR_RBF		147.5	100	98,25

3.5. Discussion des résultats

3.5.1. Analyse et évaluation

Pour les méthodes neuronales, l'algorithme d'apprentissage influe sur la généralisation qui représente la tâche accomplie par un réseau une fois que son apprentissage est achevé. Celle-ci est aussi influencée essentiellement par quatre facteurs : la complexité du problème, l'algorithme d'apprentissage, la complexité de l'échantillon (le nombre d'exemples) et enfin la complexité du réseau (nombre de poids). La complexité du problème est déterminée en partie par sa nature même : on peut parler de « complexité intrinsèque ». Par ailleurs, l'algorithme d'apprentissage influe sur la généralisation par son aptitude à trouver un minimum local assez profond, sinon le minimum global. Un facteur influent sur la généralisation est la complexité du réseau. On peut constater que le modèle ayant très peu de paramètres n'a pas assez de flexibilité pour réaliser un apprentissage correct des exemples d'apprentissage. Les erreurs d'apprentissage et de test sont toutes deux importantes : c'est la situation de *sous-apprentissage*. En revanche, le modèle constitué de nombreux paramètres, lisse parfaitement les exemples d'apprentissage. Il commet donc une erreur faible sur ses données, mais probablement une erreur plus importante sur les données de test. C'est la situation de *sur-apprentissage.* Finalement, le modèle possédant un nombre de paramètres modérés réalise un bon compromis entre précision d'apprentissage et bonne généralisation.

Le problème de la généralisation est souvent vu sous trois perspectives différentes. Dans la première, la taille du réseau est fixée (en accord avec la complexité du problème). Dans notre cas, plusieurs neurones d'entrée, où chaque neurone représente un paramètre descripteur et un neurone réservé en sortie délivrant la décision. La question qui se pose est de : combien d'exemples d'apprentissage sont nécessaires pour atteindre une bonne généralisation ? Cette perspective est intéressante dans les applications où l'on a la possibilité d'acquérir autant d'exemples que l'on veut. Dans le cas d'un système de

diagnostic multi-sensorielle (notre cas précis), on peut acquérir autant d'exemples que l'on veut, cependant une base de 1000 vecteurs par exemple est-elle suffisante ? La deuxième perspective c'est quand nous supposons que le nombre d'exemples d'apprentissage est fixé ; la question qui se pose dans ce cas est : quelle est la taille du réseau qui donne la meilleure généralisation de ces données ? Pour cette application, quel est le réseau pris parmi les différents réseaux testés qui donne la meilleure généralisation ? Est ce un réseau à une seule couche cachée ? À deux couches cachées ? Ou bien à trois couches cachées ? On est conduit à adopter finalement ce point de vue puisqu'on est devant l'impossibilité d'avoir une base de connaissance aussi complète qu'elle soit. Un enrichissement continuel de cette base avec le temps est pratiquement indispensable. Il importe alors dans cette situation, de déterminer la taille du réseau qu'il faut pour décrire au mieux les données en notre possession. Cependant, tous les différents réseaux validés peuvent acquérir des données d'apprentissage, et l'erreur d'entraînement est acceptable presque dans tous ces réseaux. La variance de l'estimation due à la taille finie de l'échantillon induit un écart entre la capacité réelle de généralisation et la capacité estimée (risque empirique). Dans la troisième perspective, on se donne des complexités d'échantillon et de modèle et on cherche pour une probabilité fixée, l'écart maximum entre la vraie capacité de généralisation et la capacité de généralisation estimée à partir de l'échantillon. La théorie de l'apprentissage statistique, permet de répondre à la première et à la troisième question permettent d'établir un lien entre la complexité de l'échantillon et la complexité du réseau [60]. Si on revient aux principes théoriques de base, les réseaux de neurones sont basés sur le principe de la minimisation du risque empirique (MRE). Ce principe se traduit par les méthodes connues ; pour la classification par exemple, on minimise le nombre d'erreurs en apprentissage. On peut minimiser le risque empirique (par une règle d'apprentissage) après le choix d'une architecture d'un réseau, soit fixer la valeur du risque empirique (idéalement, à la valeur 0). Dans notre cas et avec la diversité des architectures de réseaux de neurones, surtout qu'il n'existe pas de règles bien précises pour fixer le nombre de neurones et de couches cachées dans un réseau ; ce problème de choix est posé et demeure le principal inconvénient. On remarque par ailleurs qu'il n'y a pas eu une amélioration nette du taux de reconnaissance (même avec l'obtention d'une erreur d'entraînement presque nulle) quand on a fait augmenter la base d'apprentissage [61]. Dans ce cadre, l'utilisation de l'algorithme d'apprentissage classique « *descente du gradient* » nécessite plus de nombres d'itérations et l'erreur d'entraînement est moins plus faible que celle obtenue avec l'utilisation de l'algorithme de « *Levenberg-Marquardt* » [61]. Ce

compromis de choix entre le nombre d'itérations (temps d'apprentissage) et l'erreur d'entraînement (pour la phase de généralisation), avec la considération du taux de reconnaissance, rentre dans le choix de l'algorithme d'apprentissage et l'architecture du réseau le plus préférable pour notre application. D'après les résultats obtenus (taux de reconnaissance plus de 89% pour MLP et plus de 95% pour RBF), le nombre d'erreurs en apprentissage qui représente le risque empirique est minimisé, avec l'utilisation de l'algorithme de « Levenberg-Marquardt » de $2^{ème}$ ordre, devenu aujourd'hui l'algorithme le plus performant, mais reste l'inconvénient majeur est l'architecture qui n'est pas toujours standard dans les différents cas.

Quant à l'emploi de la technique k-PPV; Nous avons trouvé que cette méthode fournit de très bons résultats pour le cas d'une classification binaire. Les solutions trouvées dépendent exclusivement des exemples (base de données) présentés en entrée. On peut dire d'après les résultats obtenus, que l'erreur d'entraînement et très faible, ainsi qu'un taux de reconnaissance plus de 98 % après sélection des caractéristiques confirment clairement l'adéquation de la technique avec ce type d'application.

On résume dans le tableau 3.4 un état comparatif des caractéristiques liées aux solutions envisageables dans l'utilisation de l'un des modèles (MLP, RBF ou k-PPV) dans un système de classification.

Tableau. 3.4. Tableau comparatif des modèles (MLP, RBF et k-PPV).

Propriétés	MLP	RBF	k-PPV
Algorithme	Apprentissage et généralisation	Apprentissage et généralisation	Sans apprentissage
Optimisation	Quadratique (1ère et 2ème ordre) non linéaire	Quadratique (2ème ordre) non linéaire	Mesure des distances
Apprentissage	Nombre de poids	Nombre de poids	Par examples
Ajustement des poids	N'est pas stable	N'est pas stable	Stable
Paramètres d'apprentissage	Trop de paramètres	Trop de paramètres	Moins de parameters
Temps d'apprentissage	Long	Assez long	Rapide
Taux de reconnaissance	Plus de 89 %	Plus 93 %	Plus de 94 %
Inconvénients	- Classe les éléments qui n'appartiennent à aucune classe à la classe la plus proche - Le nombre de neurones et de couches caches est indéfinité	- Apprentissage complexe - Nécessite une grande capacité de calcul	- L'algorithme devient beaucoup plus lent à mesure que le nombre d'exemples d'apprentissage augmente. - Le choix de la méthode de calcul de la distance ainsi que le nombre de voisins k peut ne pas être évident
Avantages	- Accepte les données bruitées et la classification non-linéaire - Représentation globale de l'espace - Architecture simple	- Accepte les données bruitées et la classification non-linéaire - Capable de dire « je ne sais pas » - Représentation locale de l'espace - Une grande précision. - Apprentissage rapide	- L'algorithme est simple et facile à mettre en œuvre - Aucune hypothèse sur les données - Apprentissage rapide

3.5.2. Principales caractéristiques

Les résultats caractéristiques correspondant aux trois modèles étudiés sans et avec sélection des caractéristiques sur la base de donnée sélectionnée pour la comparaison (**Heart Disease datasets**), sont résumés dans le tableau 3.5:

Tableau. 3.5. Caractéristiques principales des modèles (MLP, RBF et k-PPV).

Caractéristiques / Modèles	MLP	RBF	mRMR_RBF	k-PPV *Hamming*, k=1	mRMR_k-PPV *cityblock*, k=1
Temps d'apprentissage (sec)	183.4	117.7	147.5	0.0158	0.014
Erreur/Taux d'apprentissage	0.494	$2.17\ 10^{-29}$	100	100	98.84
Taux de reconnaissance (%)	89,18	95.61	98,25	94.34	98.64

Il apparait clairement que les trois modèles étudiés présentent en général de bons résultants spécialement RBF et k-PPV, avec des taux de reconnaissance acceptables sur le plan décisionnel. Le modèle k-PPV est plutôt mieux placé du point de vue temps d'apprentissage, ce qui lui confère l'avantage d'une intégration dans un système de diagnostic dynamique. Cependant, le modèle MLP et RBF souffre particulièrement d'un handicap majeur lié à sa durée d'apprentissage. Cet inconvénient est toutefois contourné par le modèle k-PPV en raison de sa rapidité remarquable. L'amélioration sensible pour ce modèle en matière de taux de reconnaissance après sélection des caractéristiques, laisse envisager son intégration dans un système de surveillance en continu permettant la collecte en permanence des données supervisées par un expert. Un diagnostic médical étendu à une grande échelle, peut être pris en charge de façon dynamique par ce modèle, puisque le temps de calcul réalisé est très faible. Les caractéristiques affichées dans les tableaux 3.4 et 3.5 soulignent l'intérêt pratique du modèle k-PPV pour ce type d'application.

On peut dire enfin que la technique de classification k-PPV même avec réduction de dimension des variables d'entrée, donne de meilleures performances en matière de classification. L'impact de ce résultat est important sur les plans, aussi bien technique (temps d'apprentissage plus faible), qu'économique (nombre de test médical réduit).

CONCLUSION

Ce dernier chapitre a fait l'objet d'une étude en simulation concernant la mise en œuvre de trois techniques d'apprentissage appliquées dans le domaine médical. Cette étude a permis la validation et l'évaluation des performances de chacune de ces méthodes présentées. Une étude comparative dans le but d'un choix décisif de la méthode la mieux adaptée à l'application a été effectuée. Les paramètres liés au taux de reconnaissance, au temps d'apprentissage et à l'erreur d'entraînement, ont été les facteurs pertinents qui ont permis d'évaluer les méthodes étudiées. L'étape de sélection des caractéristiques basées sur deux technique RBF et k-PPV a été alors élaborée. La discussion des résultats obtenus, a permis d'opter pour la technique k-PPV retenue pour ses qualités et avantages adaptés au problème posé. Cette technique a fourni de très bons résultats de simulation et elle a montré l'efficacité de cette approche.

CONCLUSION GÉNÉRALE

Le travail présenté dans ce mémoire a été consacré à la mise en œuvre de trois techniques d'apprentissage appliquées à la reconnaissance de formes dans le domaine médical. Cette étude découle des progrès technologiques importants qui ont été enregistrés ces dernières années, dans le but et l'intérêt d'une surveillance moderne et plus efficace de patients. A cet effet, notre modeste travail peut être considéré comme une contribution aux solutions proposées, pour résoudre des problèmes d'intérêt stratégique à préoccupation nationale, utilisant des outils modernes à base de techniques avancées dans le domaine de la santé.

L'objectif est de valider et d'évaluer les performances des techniques d'apprentissage de chacune des méthodes présentées à savoir MLP, RBF et k-PPV comme étant une solution pour le diagnostic médical en particulier, à la conception des modèles de classification à partir des paramètres descripteurs de différentes maladies étudiées telles que: les maladies cardiaques et le cancer du col de l'utérus. Cette étude a été précédée d'une analyse statistique « mRMR » pour la sélection des caractéristiques qui nous permettra de déterminer les corrélations existantes entre les variables descripteurs dans un but de reduction de dimension des variables d'entrée.

Cette étude a été structurée autour de trois chapitres essentiels. Le premier consacré à la sélection de données, a permis de présenter des généralités ainsi que la méthode de réduction de dimension employées. Le deuxième chapitre a été particulièrement dédié aux mécanismes théoriques des méthodes de classification de données à apprentissage statistique supervisé. Dans ce chapitre, trois modèles (RNA, RBF et k-PPV) fondés sur ce type d'apprentissage ont été exposés. Enfin le troisième et dernier chapitre, a fait l'objet d'une étude en simulation concernant la mise en œuvre de ces trois modèles d'apprentissage statistique appliqués dans le domaine du diagnostic médical. Cette étude a permis la validation et l'évaluation des performances de chacune des méthodes présentées. Une étude comparative dans le but d'une sélection de la méthode la mieux adaptée à l'application a été effectuée. Les paramètres liés au taux de reconnaissance, au temps

d'apprentissage et à l'erreur d'entraînement, ont été les facteurs pertinents qui ont permis d'évaluer les méthodes étudiées. La discussion des résultats obtenus, a permis d'opter pour la technique k-PPV retenue pour ses qualités et avantages adaptés au problème posé.

D'après les résultats obtenus, il apparaît que sur le plan décisionnel, les trois modèles ont présenté des résultats acceptables pour une simulation sur des données biomédicales spécialement RBF et k-PPV. Le modèle k-PPV est plutôt mieux placé du point de vue temps d'apprentissage, ce qui lui confère l'avantage d'une intégration dans un système de diagnostic dynamique. Cependant, le modèle MLP et RBF souffre particulièrement d'un handicap majeur lié à sa durée d'apprentissage. Cet inconvénient est toutefois contourné par le modèle k-PPV en raison de sa rapidité remarquable. L'amélioration sensible pour ce modèle en matière de taux de reconnaissance après selection des caractéristiques, laisse envisager son intégration dans un système de surveillance en continu permettant la collecte en permanence des données supervisées par un expert. Un diagostic médical étendu à une grande échelle, peut être pris en charge de façon dynamique par ce modèle, puisque le temps de calcul réalisé est très faible. L'usage d'un prétraitement des données par mRMR dans un but de réduction de dimensionnalité, a confirmé davantage cet intérêt.

Le principal souci pour l'application de tel modèle est l'obtention d'une base de données « optimale ». Ceci met évidemment en jeu le nombre et le type d'exemples à utiliser dans la base d'apprentissage. Comme souligné auparavant, la présence d'un expert (ou système expert) serait indispensable dans ce cas là. Le temps correspondant à la phase d'entraînement reste relativement important, ce qui laisse envisager d'autres outils de calcul plus puissants afin d'améliorer les capacités et obtenir plus de performances. La décision clinique peut par contre être prise en charge de façon dynamique par le système de surveillance multi-sensorielle, puisque le temps d'exécution est faible.

Les perspectives de l'application de cette approche restent prometteuses. La décision du système peut être améliorée par l'exploitation de nouveaux paramètres d'entrée. Il reste à noter que la sensibilité du domaine à des menaces imprévues, exigent de plus grands efforts pour maximiser l'immunité du système et apporter d'autres améliorations afin de minimiser les risques encourus pour la santé humaine. Enfin, cette application montre une alternative prometteuse pour notre pays dans l'avenir, pour une surveillance intelligente, automatique et efficace de l'état des patients.

Références

[1] M. Souhila et Z. Tabti "sélection de variable de reconnaissance de la maladie de parkinson", Mémoire de Master, université de Belhadj Bouchaib Ain-Temouchent, 2016.

[2] D. MEHIDI et S. MEDJOUDJ, " Application des Méthodes d'Apprentissage dans la Prédiction du Diabète de Type 2", Mémoire de Master, Université A.MIRA de Bejaia, 2018.

[3] N. Settouti & A. Hafa, "Approche Filtre pour la sélection des gènes pertinents des données biopuces du Cancer du Côlon" , Université Abou Bekr Belkaid – Tlemcen, 2013.

[4] N. BOUDIA, " Le recuit simulé pour la sélection de variables des puces à ADN", Mémoire De Master, UNIVERSITE ABDELHAMID IBN BADIS – MOSTAGANEM, 2016.

[5] R. Reyna Rojas, "Conception et intégration VLSI d'un système de vision générique, Application à la détection et la localisation d'objets à l'aide de "support vector machines"", Thèse de Doctorat. Institut National des Sciences Appliquées, Laboratoire LAAS/CNRS, N°02226, Toulouse, France, 2002.

[6] L. Comminges, "Quelques contributions à la sélection de variables et aux tests non-paramétriques", thèse doctorat, Université Paris-Est, 2013.

[7] N. Settouti & A. Hafa, "Approche Filtre pour la sélection des gènes pertinents des données biopuces du Cancer du Côlon", Université Abou Bekr Belkaid – Tlemcen, 2013.

[8] I. Guyon & A. Elisseeff, "An Introduction to Variable and Feature Selection",Journal of the Max Planck Institute for Biological Cybernetics, vol.3, mars2003, p.1157-1182.

[9] A. ABDELMALEK & W. HEBBAR, "Evaluation des méthodes de Sélection de Variables en Apprentissage supervisé", mémoire de master, université Abdelhamid ibn badis Mostaganem, 2014.

[10] H. CHOUAIB, " Sélection de caractéristiques: méthodes et applications", thèse de doctorat, Université Paris Descartes, Juillet 2011.

[11] M. Dash et H. Liu, "selection for classification ", Intelligent Data Analysis, vol.1, n°3, p.131-156, 1997.

[12] S. Mansour & Z. Tabti, "Sélection de variables pour la reconnaissance de la maladie de parkinson", Mémoire de master, université de Belhadj Bouchaib d'Ain-Temouchent, 2016.

[13]A. El Akadi, " contribution à la sélection de variable pertinente en classification supervisée: Application à la sélection des gènes pour les puces à ADN et des caractéristiques faciales" , thèse de doctorat, 2012.

[14] J C. HERNÁNDEZ "Algorithmes méta-heuristiques hybrides Pour la sélection de gènes et la Classification de données de biopuces", thèse de doctorat, 2008.

[15] G. Chandrasekhar & F. Sahin, "A survey on feature selection methods" , Computers and Electrical Engineering, vol. 40 n°1, p. 16-28, 2014.

[16] A. HAFA, "Sélection de Variables Biologiques par l'approche FILTER" , Mémoire de Master, Université Abou Bekr Belkaid, 2012.

[17] D. Dernoncourt, "Stabilité de la sélection de variables sur des données haute dimension : une application à l'expression génique", Thèse de doctorat, universite pierre et marie curie, 2014.

[18] chapitre 3, "feature selection"

[19] S. Gunasekaran & K. Revathy. Content-based Classification and Retrieval of Wild Animal Sounds using Feature Selection Algorithm. Second International Conference on Machine Learning and Computing . IEEE , 2010, pp. 272-275.

[20] M. Radovic, M. Ghalwash, N. Filipovic Z. Obradovic, "Minimum redundancy maximum relevance feature selection approach for temporal gene expression data", BMC Bioinformatics, Vol. 18, n°1, 2017, p. 1-18.

[21] M. Peker, B. Şen and D. Delen, " Computer-Aided Diagnosis of Parkinson's Disease Using Complex-Valued Neural Networks and mRMR Feature Selection Algorithm" , Journal of Healthcare Engineering · Vol. 6, n°3, 2015, p. 281–302.

[22] Intelligent fault diagnosis of rotating machinery using improved multiscale dispersion entropy and mRMR feature selection , Knowledge-Based Systems, Vol.163, 2019, p.450-471.

[23] B. Valentin, "Algorithmes d'apprentissage pour la recommandation", Université de Montréal, Faculté des études supérieures et postdoctorales, Septembre, 2012.

[24] A. DJAIDJA, " Etude de la classification supervisée des données environnementales à l'aide de réseaux de neurones de fonctions à base radiales", Mémoire de master, Univ de M'sila, juin 2016.

[25] Ferentinos, K. P, (2018) .Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture, 145, 311–318.

[26] O. Guenounou, " Méthodologie de conception de contrôleurs intelligents par l'approche génétique- application à un bioprocédé", Thèse de doctorat**,** l'Université Toulouse III, 2009.

[27] B. Abassi, "Approche de sélection des données biomédicales pour l'identification de pathologies en utilisant les supports vecteur machine", Mémoire master, Univ Muhamed Boudiaf, M'sila. 2019.

[28] wikistat [online] : http://wikistat.fr/pdf/st-m-app-into.pdf consulté 20.

[29] G. DREYFUS, "Les réseaux de neurones". Mécanique Industrielle et Matériaux 51 (1998).

[30] A. EL Hmaidi, H. EL Badaoui, A. Abdallaoui, & EL Moumni, "Application des réseaux de neurones artificiels de type PMC pour la prédiction des teneurs en carbone organique dans les dépôts du quaternaire terminal de la mer d'alboran". European Journal of Scientific Research, vol.107 n° 4 ,2013, p.400-413.

[31] L. Amimer, "Modélisation et Commande des Systèmes Non Linéaires Fractionnaires par des Réseaux de Neurones Fractionnaires" .Mémoire de Magister, Université Mouloud Mammeri, Tizi-Ouzou, 2015.

[32] H. Kaddour, A. Djedid Amar, " Évaluation des performances des techniques RNA et ELM utilisées dans le développement des capteurs logiciels pour la surveillance de la qualité de l'eau", Mémoire de Master, Univ de M'sila, 2017.

[33] Parizeau, M. (2004). " Réseaux de neurones". GIF-21140 et GIF-64326, 124.Université LAVAL, 2004

[34] Y. Daoud & K. Souyeh, "système de prediction de la vitesse du vent base sur la selection de caracteristique et les reseaux de neurones artificiels ", mémoire de master universite mohamed boudiaf - m'sila, 2018.

[35] A. Belhou, "LDA et RNA pour la prédiction de la vitesse du vent", Mémoire de Master, Univ de M'sila, 2016 /2017.

[36] M. Nouressadat, " Etude des performances des réseaux de neurones dynamiques à représenter des systèmes réels : une approche dans l'espace d'état", Mémoire de Magister, Univ de Sétif 1Algerie, **2009.**

[37] Djeriri ,Y., **"**Les Réseaux de Neurones Artificiels**" .** Univ de Sidi-Bel-Abbès**,** Septembre 2017**.**

[38] N. Ifrek et L. Boussaid, "Etude et application du réseau ELM (Extreme Learning Machine) pour la classification de donnes", Mémoire de Master, Univ de Mouloud Mammeri, Tizi-Ouzou, 2017.

[39] A. Belkadi, "Approche neuronale pour la classification des images hyper spectrale" N, Mémoire de Magister, Université des sciences et de la technologie d'Oran Mohamed Boudiaf.

[40] F. Gueniaa et I. Tourqui, "Comparaison de méthodes de classification appliquées à la détection d'objets", Mémoire de Master, UNIVERSITÉ Echahid Hamma Lakhdar - EL-OUED, 2015/2016.

[41] M. Kalakh, "Modélisation avec les réseaux de neurones d'un canal UWB dans un environnement minier souterrain", université du QUEBEC EN ABITIBI TEMISCAMINGUE, Mars, 2013.

[42] M. Zribi et Y. Boujelbene, " Les réseaux de neurones un outil de sélection de variables : Le cas des facteurs de risque de la maladie du cancer du sein", Faculté des Sciences Economiques et de Gestion, Université de Sfax, Tunisie, 2012.

[43] A. Kawthar, "Contribution au diagnostic et a l'analyse de défauts d'une machine synchrone à aimants permanents", Thèse de doctorat, l'Université de Rouen Normandie, 2018.

[44] A. ASSOUM, « étude de la tolérance aux aléas logiques des réseaux de neurones Artificiels », thèse de doctorat, institut national polytechnique de Grenoble, France.1997.

[45] Guillaume B.« Contrôle sensori-moteur par réseaux neuromimetiques modulaires - Approche pour le pilotage réactif en atelier flexible-» Thèse de Doctorat, Institut national des sciences appliquées de Lyon ,1995.

[46] El Badaoui, H., Abdallaoui, A., & Chabaa, S. Perceptron Multicouches et réseau à Fonction de Base Radiale pour la prédiction du taux d'humidité. International Journal of Innovation and Scientific Research(ISSN), Vol. 5, n° 1, 2014, p. 55-67.

[47] M. LADJAL « Contribution au développement de systèmes de surveillance innovants dédiés au contrôle de la qualité des eaux potables » thèse de doctorat, université de m'sila 2013.

[48] Sylvain Tertois, «Réduction des effets des non linéarités dans une modulation à l'aide de réseaux de neurones», Thèse de Doctorat, Université de Rennes 1, France, N° d'ordre : 2924, 2003.

[49] L.BELOUAFI & K. SAIDI, " Estimation de l'énergie d'un capteur solaire en utilisant les réseaux de neurones artificiels" .Mémoire de Master, Université Ahmed Draia - Adrar , 2018.

[50] S. KHELLOUT& H. LAKHDARI, " Etude et application du réseau ELM-LRF en classification des images", Mémoire de Master, Université Mouloud Mammeri, Tizi-Ouzou, 2017.

[51] Luo, J., Vong, C., & Wong, P.K. " Sparse Bayesian Extreme Learning Machine for Multi-classification" , IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 25, n° 4, APRIL 2014, p. 836-843.

[52] B.Taconet, A. Zahour, S. Ramdane & W. Boussellaa, "Classification des k-ppv par sous-voisinages emboîtés" , Novembre 2006.

[53] H. SENOUSSI, " Sélection de Données pour l'Apprentissage des Réseaux de Neurones, Arbres de Décision et les k-Plus Proches Voisins : Application en Diagnostic de Pannes", Thèse de doctorat, Université d'Oran Mohamed BOUDIAF, 2015.

[54] S. BENADLA & H. OUAHRANI, " Affectation de nœuds dans un réseau de radio cognitive: Approche K-plus proches voisins parallèle" , Mémoire de Master, Université Abou Bakr Belkaid– Tlemcen 2018.

[55] A M. Qamar& E. Gaussier, " Apprentissage de diff´erentes classes de similarit´e dans les k-PPVs " . XVlième Rencontres de la Société Français de Classification, Grenoble, France, 2009.

[56] A. BOUALI, " Reconnaissance automatique des crises d'épilepsie par techniques de classification supervisée (SVM-KNN)" , Université Abou Bakr Belkaïd de Tlemcen, 2016.

[57] M. Bekhti, "La Sélection de Variables Neuronale pour la Reconnaissance du Diabète" , Université Abou Bekr Belkaid – Tlemcen, 2012.

[58] http://www.kaggle.com/zeeshanmulla/heart-disease-dataset .

[59] cervical cancer risk exploratory study, universita degli studi di milano-bicocca, june 2020.

[60] N. Valentin, "Construction d'un capteur logiciel pour le contrôle automatique du procédé de coagulation en traitement d'eau potable", Thèse de doctorat, UTC, Centre International de Recherche sur l'Eau et l'Environnement, CNRS, France, 2000.

[61] M. Ladjal, "Traitement et fusion multi-sensorielle appliques a la surveillance des eaux potables", Mémoire de Magister, Université de M'sila, Algérie, Juin 2006.

Résumé

Les avancées technologiques ont facilité l'acquisition et le recueil de nombreuses données. Ces données peuvent être utilisées comme support de décision, conduisant aux développements d'outils capables de les analyser et de les traiter. Les systèmes d'aide au diagnostic sont considérés comme étant essentiels dans beaucoup de disciplines, ces systèmes reposent sur des techniques issues de l'intelligence artificielle mais les problèmes les plus intéressants sont souvent basés sur des données de haute dimension. Ces problèmes désignent les situations où nous disposons peu d'observations alors que le nombre de variables explicatives est très grand. La sélection de variables est devenue l'objet qui attire l'attention de nombreux chercheurs durant ces dernières années, cette sélection permet d'identifier et d'éliminer les variables qui pénalisent les performances d'un modèle complexe dans la mesure où elles peuvent être bruitées, redondantes ou non pertinentes. De plus, la mise en évidence des variables pertinentes facilite l'interprétation et la compréhension des aspects liés aux applications ; ainsi, elle permet d'améliorer la performance de prédiction des méthodes de classification et de passer outre le fléau de la haute dimensionnalité de ces données. L'approche **filtre** est couramment utilisée à ce jours pour analyser les données biologiques, cette approche consiste à parcourir la sélection des variables avant le processus de l'apprentissage et ne conserve que les caractéristiques informatives.

L'objectif recherché dans le cadre de ce travail est une contribution à l'étude et au développement de systèmes innovants d'aide au diagnostic médical. Consacré à la simulation, ce travail vise l'application des techniques d'apprentissage statistique comme étant une solution dans la conception de ces systèmes par reconnaissance de formes. Dans le cadre de l'apprentissage supervisé, la sélection des caractéristiques permet d'obtenir des classifieurs précis. L'approche filtre est fondée uniquement sur des données, elles permettent à l'utilisateur d'entamer une analyse plus fine de ces données en augmentant la transparence du modèle utilisant la méthode mRMR (Minimum Redondance, Maximum Relevance). Pour la validation de données sélectionnées dans des bases d'apprentissage biomédicales, nous testons leurs capacités et leurs taux de classification avec plusieurs classifieurs. Afin de mener une étude comparative permettant un choix décisif de la méthode la mieux adaptée à l'application proposée, on évaluera pour les méthodes exposées les paramètres liés au taux de reconnaissance, au temps d'apprentissage et à l'erreur d'entraînement.

Mots clés : Diagnostic médical, Sélection des variables, mRMR, Classification, RNA, RBF, k-PPV, Simulation.

Abstract

Advances in technology have facilitated the acquisition and collection of extensive data. These data can be used as decision support, leading to the development of tools able to analyze and process them. Diagnostic support systems are considered essential in many disciplines, based on artificial intelligence techniques, but the most interesting problems are often based on high-dimensional data. These problems refer to situations where we have few observations while the number of explanatory variables is very large. The selection of variables has become the object that attracts the attention of many researchers in recent years, this selection makes it possible to identify In this way, it improves the predictive performance of classification methods and overcomes the scourge of high dimensional data. The filter approach is commonly used today to analyze biological data, this approach consists of browsing the selection of variables before the learning process and retains only informative characteristics.

The aim of this work is to contribute to the study and development of innovative systems to aid medical diagnosis. Dedicated to simulation, this work aims to apply statistical learning techniques as a solution in the design of these systems by pattern recognition. Within the framework of supervised learning, the selection of characteristics makes it possible to obtain precise classifiers. The filter approach is based solely on data, allowing the user to begin a more detailed analysis of this data by increasing the transparency of the model using the mRMR method (Minimum Redundancy, Maximum Relevance). For validation of selected data in biomedical learning databases, we test their capabilities and classification rates with several classifiers. In order to carry out a comparative study allowing a decisive choice of the method best suited to the proposed application, the parameters related to recognition rate, learning time and training error will be evaluated for the methods described.

Keywords: Medical diagnosis, Variable selection, mRMR, Classification, RNA, RBF, k-PPV, Simulation.

الملخص

وقد يسرت التطورات في التكنولوجيا الحصول على البيانات المستفيضة وجمعها .ويمكن استخدام هذه البيانات كدعم لاتخاذ القرارات، مما يؤدي إلى تطوير أدوات قادرة على تحليلها ومعالجتها .تعتبر أنظمة الدعم التشخيصي أساسية في العديد من التخصصات، استنادًا إلى تقنيات الذكاء الاصطناعي، ولكن غالبًا ما تستند أكثر المشاكل إثارة إلى بيانات عالية الأبعاد .وتشير هذه المشاكل إلى حالات لا توجد فيها ملاحظات قليلة بينما عدد المتغيرات التوضيحية كبير جدا. قد أصبح اختيار المتغيرات هو الهدف الذي يجذب انتباه العديد من الباحثين في السنوات الأخيرة، وهذا الاختيار يجعل من الممكن تحديد هذه المتغيرات وبهذه الطريقة، فإنه يحسن الأداء التنبؤي لأساليب التصنيف ويتغلب على آفة البيانات عالية الأبعاد .يستخدم نهج التصفية اليوم لتحليل البيانات البيولوجية، ويتألف هذا النهج من استعراض اختيار المتغيرات قبل عملية التعلم والاحتفاظ فقط بخصائص مفيدة.
والهدف من هذا العمل هو المساهمة في دراسة وتطوير نظم مبتكرة للمساعدة في التشخيص الطبي .ويهدف هذا العمل المخصص للمحاكاة إلى تطبيق تقنيات التعلم الإحصائي كحل في تصميم هذه الأنظمة من خلال التعرف على الأنماط .وفي إطار التعليم الخاضع للإشراف، يتيح اختيار الخصائص الحصول على المصنفات الدقيقة .يستند أسلوب التصفية فقط إلى البيانات، مما يسمح للمستخدم ببدء تحليل أكثر تفصيلا لهذه البيانات من خلال زيادة شفافية النموذج باستخدام أسلوب mRMR الحد الأدنى للتكرار، الحد الأقصى للصلة .للتحقق من صحة بيانات محددة في قواعد بيانات تعلم الطب الحيوي، نقوم باختبار قدراتهم ومعدلات تصنيفهم مع العديد من المصنفين .ومن أجل إجراء دراسة مقارنة تتيح الاختيار الحاسم للطريقة الأنسب للتطبيق المقترح، سيجري تقييم العوامل المتصلة بمعدل التقدير ووقت التعلم والخطأ في التدريب على الطرائق الموصوفة.

الكلمات المفتاحية : التشخيص الطبي، الاختيار المتغير، mRMR ، التصنيف، RNA ، RBF، k-PPV ، المحاكاة