REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE DE TECHNOLOGIE
DEPARTEMENT D'ELECTRONIQUE
N° : ESEM12/2019

DOMAINE : SCIENCES TECHNOLOGIE
FILIERE : ELECTRONIQUE
OPTION : Système Embarqué

Mémoire présenté pour l'obtention
Du diplôme de Master Académique

Par :

SOUYEB Nadjet et TAHMI Houda

Intitulé

Étude de la stabilité de la sélection de variables pour la classification de données

Soutenu devant le jury composé de :

Mr. ATTALLAH Bilal	Université de M'sila	Président
Mr. Mohamed Ladjal	Université de M'sila	Encadreur
Mr. Mohamed Djerioui	Université de M'sila	Co-encadreur
Mr. BRIK Youcef	Université de M'sila	Examinateur

Année universitaire : 2018 /2019

Remerciements

Avant tout, Nous remercions DIEU miséricorde de nous avoir donné la volonté, le courage et la patience qui nous ont permis de réaliser ce travail, et de poursuivre nos études supérieures et de les réussir et d'avoir la chance d'atteindre le niveau MASTER II.

Nous tenons à remercier en particulier notre encadreurs ***Dr. LADJAL Mohamed*** *et* ***Dr. DJERIOUI Mohamed****de nous avoir aidé par ces conseils, ces remarques pertinentes et par sa collaboration effective pour l'élaboration de ce mémoire.*

Nous remercions également ***Dr. BRIK Youcef*** *pour ses efforts et le dévouement de son temps pour nous, en particulier pour ses précieux conseils.*

Nos remerciements vont également à tous les enseignants du département d'électronique sans oublier toutes les personnes de ce même département.

Nous remercions également, les membres de jury d'avoir accepté d'honorer notre soutenance de leur prudence, qu'il se trouve ici l'expression de notre gratitude.

SOUYEB NADJET & TAHMI HOUDA

Dédicaces

Je voudrais dédier cet humble travail :

A ma très chère mère " FATIHA" pour tous ses sacrifices, son amour, sa tendresse, son soutien et ses prières tout au long de mes études,

Je la remercie pour mes encouragements.

A mon cher frère "ABD ALLAH" et ma chère sœur "KHADIDJA", pour leur appui et leur encouragement.

A Toutes ma famille.

A mon binôme " HOUDA ".

A toutes mes chers amis surtout : RAYHANA, SARA, NAWAL, SIHEM, AMEL,TOURKIYA, RABAB, AMI NA, HASSIBA, SARA, NADJWA, RADHYA.

SOUYEB NADJET

Dédicaces

Je voudrais dédier cet humble travail

à toute ma famille, a ma chère maman et mon cher père

qui nous a quitté à jamais, Qui ont veillé à ce que

je sois ce que je suis devenu maintenant.

A mon binôme NADJET .

A mes belles sœurs .

A mes frères

Atout mes amis

TAHMI HOUDA

Table des matières

CHAPITRE I

Sélection des variables

CHAPITRE II

Techniques D'apprentissage

CHAPITRE III

Simulation et Évaluation

Liste des figures

Liste des tableaux

INTRODUCTION GENERALE

INTRODUCTION GENERALE

Depuis l'aube des temps, l'homme pratique la classification dans sa vie quotidienne, quand il essaie de répondre aux problèmes et questions sur la catégorie des objets , c'est-à-dire d'affectation d'objets a leur classe (en observant leurs formats , couleurs , tailles . . .etc.), La classification est une discipline relié de près ou de loin a plusieurs domaines, elle est connue aussi sous noms variés (classification, clustering, segmentation,. . .) selon les objets qu'elle traite et les objectifs qu'elle vise à atteindre. En mathématique, On appelle classification, la catégorisation algorithmique d'objets. Elle consiste à attribuer une classe ou catégorie à chaque objet à classer, en se basant sur des données statistiques. Elle fait couramment appel aux méthodes d'apprentissage et est largement utilisée en reconnaissance de formes. Dans un classement on affecte les objets à des groupes préétablis, c'est le but de l'analyse discriminante que de fixer des règles pour déterminer la classe des objets. La classification est donc, en quelque sorte, le travail préliminaire au classement, savoir la recherche des classes "naturelles" dans le domaine étudié, en anglais « Cluster Analysis » [1] et un autre idée. La classification est un opération de structuration qui vise à organiser un ensemble d'observation en groupes homogènes et contrastés afin de faciliter l'analyse des informations et d'effectuer des prédictions.

En effet, le monde scientifique d'aujourd'hui fournit des données qui sont chaque jour plus nombreuses et de plus grande dimension. Les données à haute dimension et leur hétérogénéité, ont motivé le développement de méthodes statistiques pour la sélection de variables. La préparation de la base de données dans des applications de développement des outils de supervision/diagnostic à l'aide des modèles basés sur les techniques d'apprentissage statistiques consiste à retenir les variables les plus représentatives des données observées. L'utilisation de ces techniques augmente dans l'industrie de production puisqu'ils permettant le développement de robustes modèles non-linéaires d'unités de procédés industriels complexes. ces données contiennent un très grand nombre de variables. Cette association d'une dimensionnalité élevée à une petite taille d'échantillon fait de la sélection de variables une étape préalable indispensable pour diminuer les temps de calcul, et améliorer l'interopérabilité des modèles. Dans le cadre de ce travail de mémoire, nous avons tout d'abord cherché à déterminer quels sont les facteurs, au niveau des données, qui influencent le plus la stabilité de la sélection, sur tout de type de données. Nous avons ensuite travaillé sur des méthodes de classification et de régression, en nous focalisant tout d'abord sur l'influence de la stabilité de la sélection sur les ces méthodes étudiées. Par conséquent, l'utilisation de

l'Analyse Discriminante Linéaire (LDA) et l'Analyse en Composant Principale (PCA) est très utile pour réduire le nombre de variables et donc le nombre de données, pour que LDA est une technique permet de traiter un nombre très important de données et de dégager les aspects les plus intéressants de la structure et comme une étape de prétraitement pour apprentissage machine et pattern applications de classification et PCA réduisez la dimensionnalité tout en préservant le plus possible la variance dans les grands espaces dimensionnels.

Et pour ce travail on utilises techniques d'apprentissage pour la classification, Nous allons présenter une étude générale sur l'apprentissage automatique et nous nous connaissons sur l'algorithmes utilisés dans l'apprentissage d'entre eux (les réseaux de neurones par exemple), et nous allons présenter le protocole d'apprentissage, les différentes architectures de ces réseaux, et aussi présenter des exemples usuel de Réseau de Neurones Artificiels (RNA) tel que les machine d'apprentissage extrême (ELM), leur principe de fonctionnement, et leur caractéristiques principales. aussi nous allons présenter une généralité sur Radial Basis Functions (RBF) et Réseau de neurones de type MLP (Multi Layer Perceptron - MLP) .

Dans ce travail, un étude d'évaluation de plusieurs techniques de classification issues de l'intelligence artificielle et appliquées au domaine de traitement des eaux propres, telles que : MLP, RBF et ELM, est présentée. L'objectif est de mettre en œuvre une architecture de système de surveillance fondée sur l'emploi de capteurs logiciels à base de ces techniques et pouvant être intégrés au sein de plateformes de capteurs intelligents. D'un usage plus économique, ces capteurs hybrides permettent de prendre une décision adaptée dans le contrôle et le suivi des processus pour une meilleure qualité de l'eau [2].

Le travail réalisé dans ce cadre, est structuré autour de trois chapitres: Le premier chapitre présente une introduction au domaine de sélection de variables pour la classification et la régression de données haute dimension. Des généralités, des définitions de caractéristiques des techniques de prétraitement (LDA et PCA) et employées dans le domaine sont décrites à cet effet. Le deuxième chapitre est particulièrement dédié aux méthodes d'apprentissage statistiques on utilise trois méthodes (MLP, RBF et ELM) à base de la conception de ces capteurs sont donc étudiés et évalués dans ce cadre. Les aspects théoriques et fondements de l'apprentissage statistique de ces techniques sont ainsi décrits, et nous allons présenter le protocole d'apprentissage, les différentes architectures de ces réseaux, les principes de fonctionnements, et les caractéristiques principales de ces techniques. Le troisième et dernier chapitre est consacré à la simulation et vise l'application des techniques étudiées précédemment comme étant une solution dans classification. L'objectif est de valider et d'évaluer les performances de chacune des méthodes présentées. Afin de mener un étude

comparative permettant un choix décisif de la méthode la mieux adaptée à l'application proposé. On évaluera les paramètres liés au taux de reconnaissance. Au temps d'apprentissage. Une discussion des résultats conclura cette étude de simulation pour choisir la technique la mieux placée. Deux méthodes d'application sont prévues à cet effet, à savoir. Les techniques d'apprentissage sans sélection des variables pour ces performances, et avec la sélection des variables en utilisent les technique LDA et PCA de prétraitement.

Une conclusion générale en fin de cette mémoire est prévue, elle retrace les différentes étapes réalisées et souligne les perspectives futures envisagées.

Chapitre I

Sélection desvariables

Introduction

Grâce aux progrès technologiques, l'acquisition de données devient de plus en plus facile techniquement et des bases de données gigantesques sont collectées quasi-quotidiennement. Par conséquent, le nombre de variables présentes dans les problèmes statistiques actuels peut maintenant atteindre des dizaines voire des centaines de milliers. Dans le même temps, pour de nombreuses applications, le nombre d'observations se trouve réduit et peut n'être que de quelques dizaines. Dans cette thèse, nous dirons que les données considérées sont de grande dimension, et nous écrirons p>>n, quand le nombre p de variables est très grand devant le nombre n d'observations [3].

L'objectif principal de cechapitre est de proposer une procédure de sélection des variables où de minimisé le nombre de propriétés observées en sélectionnantla variable appropriée, pour ce faire; il existe plusieurs méthode, notamment: LDA, ACI (L'analyse en composantes indépendantes), PCA, ….

Dans ce chapitre, nous examinerons la méthode de LDA et PCA qui permet le traitement initial de la base de données.

I.1. Sélection des variables

La sélection de variable est un problème difficile qui a été étudié depuis les années 70. Il revient dans l'actualité scientifique avec l'apparition des grandes bases de données et les systèmes de fouille de données «Data Mining » [4-7].

Le domaine de la fouille de données est actuellement caractérisé par la présence de bases de données de taille relativement importante. En effet, la collecte d'information devient de plus en plus facile et donc, rapide. Cependant, la totalité de l'information collectée n'est pas forcément pertinente à la vue du problème considéré. Aussi, appliquer, sur l'ensemble des données récupérées, les techniques de fouille de données est bien trop coûteux. Il paraît, ainsi, nécessaire de distinguer, avant tout apprentissage, l'information pertinente de l'information inutile et/ou redondante. Cette distinction peut s'effectuer à l'aide du processus de sélection de variables lors de la phase de prétraitement des données.

La sélection de variables est un processus très importanten apprentissage supervisé. Nous disposons d'une série de variables candidates, nous cherchonsles variables les plus pertinentes pour expliquer et prédire les valeurs prises par la variable àprédire. Les objectifs sont bien souvent multiples : nous réduisons le nombre de variables àrecueillir pour le déploiement du système ; nous améliorons notre connaissance du phénomène de

Causalité entre les descripteurs et la variable à prédire, ce qui est fondamental si nous voulons interpréter les résultats pour en assurer la reproductibilité; enfin, mais pas toujours, nous améliorons la qualité de la prédiction, le ratio nombre d'observations et dimension de représentation étant plus favorable.

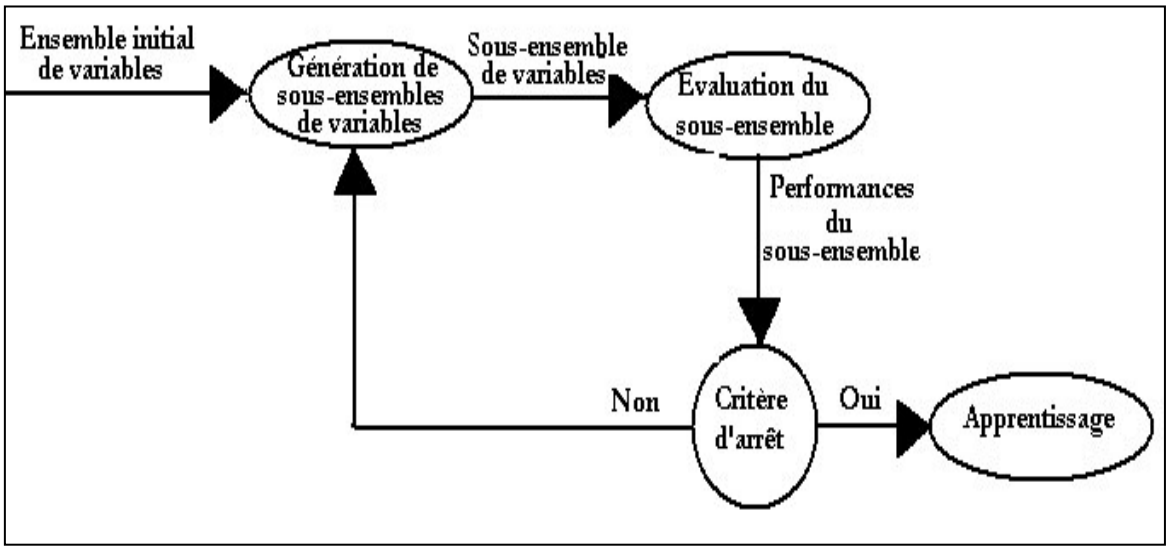

Figure I.1. Processus de sélection de variables.

I.2. Analyse discriminante

Proposée par Ronald A. Fisher en 1936 [8], l'Analyse Factorielle Discriminante - *Fisher Discriminant Analysis (FDA)* - appelée aussi analyse discriminante linéaire de Fisher, s'applique lorsqueles classes des individus sont connues. Elle consiste à chercher un espace vectoriel de faible dimensionqui maximise la variance inter-classe. Une base de cet espace est obtenue en appliquant une Analyseen Composantes Principales sur les centroïdes des différentes classes pondérées par l'effectif de la classe correspondante avec$\sum^{-1}$comme métrique. On conservera, au plus, (C − 1) axes discriminants où C'estle nombre de classes [4].

L'analyse discriminante est le nom donné à la classification dans le cadre supervisé. La classificationsupervisée se distingue de la classification non supervisée par le fait que des observations donton connaît l'appartenance aux classes sont disponibles pour apprendre la règle de décision (on par le raaussi parfois de classifieur). Ces observations, dites d'apprentissage, « supervisent » la construction du classifieur. Après avoir rappelé les objectifs et le problème de la discrimination, nous présenteronsles principales méthodes génératives dont la très connue analyse discriminante linéaire. Enfin, nousdresserons un

panorama des méthodes discriminatives dont certaines présentent des performances deprédiction remarquables [9]

I.2.1. Problème de la discrimination

Le problème de l'Analyse Discriminante, également connue sous le nom de classificationsupervisée, estde prédire l'appartenance d'un individu x à une classe parmi k. On distingue classiquement deux objectifs principaux en Analyse Discriminante :

(i) *descriptif* : l'aspect descriptif vise à trouver une représentation qui permette l'interprétation des groupes grâce aux variables explicatives.

(ii) *décisionnel* : dans ce cas, on cherche à définir la bonne affectation d'un nouvel individu dont on ne connaît que les valeurs des variables explicatives. Cet aspect est particulièrement apprécié dans des domaines où l'aspect diagnostic est essentiel.

On distingue classiquement deux objectifs principaux en analyse discriminante : l'aspect descriptif et l'aspect décisionnel. L'aspect descriptif vise à trouver une représentation qui permette l'interprétation des groupes grâce aux variables explicatives. Cette tâche est rendue difficile quand le nombre de variables explicatives est plus grand que 3. Toutefois, des techniques existent pour visualiser la classification de données ayant un grand nombre de dimensions. On peut citer par exemple la méthode de visualisation hiérarchique de Bishop et Tipping [9,10];Dans le cas de l'aspect décisionnel, on chercheà définir la meilleure affectation d'un nouvel individu dont on ne connaît que les valeurs des variablesexplicatives. Cet aspect est particulièrement apprécié dans des domaines où la notion de diagnostic estessentielle [9].

I.2.2. Analyse Discriminante de Haute Dimension

Les méthodes classiques d'Analyse Discriminante, présentées au paragraphe précédent, fournissent généralement des résultats satisfaisants pour des données de petite dimension et possèdent l'avantage d'avoir un fondement statistique. Cependant, ces méthodes sont pénalisées en haute dimension car la taille de l'échantillon d'apprentissage devient trop petit devant la dimension de l'espace et les paramètres ne sont plus estimés correctement. En particulier, les matrices de covariance des classes $\sum_i$ ne sont alors pas bien estimées et peuvent devenir singulières[11].

Le phénomène de l'espace vide nous permet de supposer que les données de haute dimension vivent dans des sous-espaces différents et de dimension intrinsèque inférieure à la dimension de l'espace. Afin d'adapter le modèle gaussien de l'Analyse Discriminante

aux données de grande dimension et de limiter le nombre de paramètres à estimer, nous proposons de projeter les données de chaque classe dans leur espace propre que nous décomposerons en deux sous-espaces supplémentaires de dimension inférieure et de faire l'hypothèse que les classes sont sphériques dans ces sous-espaces. Cette hypothèse de sphéricité se traduit par le fait que les nouvelles matrices de covariance des classes n'ont que deux valeurs propres différentes. De manière similaire à l'EDDA, notre méthode comportera plusieurs cas particuliers possédant, pour certains, des interprétations géométriques [11].

I.3. Analyse Linéaire Discriminante (LDA)

L'algorithme LDA est né des travaux de Belhumeur et al. De Yale University (USA), en 1997 [12].Il est aussi connu sous le nom de « Fisherfaces » [13].

- → La LDA a été prouvé qu'elle est équivalente à l'APC plus LDA pour des problèmes des échantillons de petite taille. Dans ces problèmes, le nombre d'échantillons d'apprentissage est inférieur à la dimension du vecteur de caractéristique, de ce fait la matrice de dispersion intra-classe ($\mathbf{S}_w$) est singulière. Comme les problèmes du monde réel sont toujours transformés en problèmes de petite taille d'échantillon par une transformation non linéaire, nous pouvons appliquer le résultat directement aux données de la fonction assignée dans l'espace R^N [14]
- → La LDA analyse les vecteurs propres de la matrice de dispersion des données, avec pour objectif de maximiser les variations entre les images d'individus différents (interclasses) tout en minimisant les variations entre les images d'un même individu (intra-classes) [13]

L'objectif de l'Analyse Linéaire Discriminante (LDA), est de réduire le nombre de dimensions en présentant le maximum des données. Elle cherche les axes tels que la projection des données dans l'espace engendré par ces axes permette une plus grande discrimination entre les classes [15,16].

I.3.1. Extension de LDA

Plusieurs approches permettent de définir l'analyse linéaire discriminante. La règle de Bayes repose sur l'hypothèse de normalité des covariables, l'analyse discriminante de Fisher est fondée sur la maximisation de la variance inter-classes contre la minimisation de

la variance intra-classes. Enfin, on peut définir l'analyse linéaire discriminante comme une régression sur des variables indicatrices dites dummy variables.

Règle de classification de Bayes Pour établir la règle de Bayes, on suppose que les covariables suivent une loi normale :

$$X|Y = y \sim N_m\left(\mu_y, \sum\right) \qquad \textbf{(I.1)}$$

Chaque groupe est caractérisé par une moyenne différente d'un groupe à l'autre et les structures de covariances sont supposées égales entre les groupes. Sous ces conditions et en appliquant le théorème de Bayes, la probabilité d'appartenir au groupe y sachant un profil x s'écrit :

$$P(y|x) = \frac{P_y f_X(x|Y=y)}{f_X(x)} \textbf{(I.2)}$$

où $f_X(.)$(resp. $f_X(.|Y=y)$) est la fonction de densité des covariables X (resp. conditionnelle à Y). En supprimant les termes constants entre les groupes, calculerla log-probabilité $\log\big(P(y|x)\big)$ est en fait équivalent à calculer le score $d(y|x)$:

$$log\big(P(y|x)\big) \infty d(y|x) = log\, p_y - 0.5\mu_y'\sum^{-1}\mu_y + x'\sum^{-1}\mu_y \textbf{(I.3)}$$

Ce score est appelé règle de classification de Bayes ou classifieur de Bayes. On remarque que ce score est linéaire en x. Enfin, la classe attribuée à un individu de profil $X = x$ est calculée en maximisant ce score :

$$\hat{y} = argmax_y d(y|x) \qquad \textbf{(I.4)}$$

En pratique, les scores sont calculés en appliquant la règle de Bayes aux estimations de la matrice de covariance $\widehat{\sum}$, des moyennes $\hat{\mu}_y$ et des probabilités de base $\hat{p}_y$. On peut montrer que cette règle de classification est la meilleure parmi toutes les règles de classification linéaires, c'est-à-dire qu'elle minimise l'erreur théorique de mauvais classement d'un individu. Cette règle de classification, certes simple, a de bonnes propriétés en pratique.

Afin d'étudier l'optimalité de la règle de Bayes, on s'intéresse au cas simple où le nombre de classes K est égal à 2. Sous cette condition, on considère les règles de classement linéaires de la forme :

$$log\frac{P(Y=1|x)}{P(Y=0|x)} = \beta_0 + \beta' x \textbf{(I.5)}$$

où $\beta 0\ \epsilon R$ et β est un vecteur de taille m. Ainsi, la prédiction pour un individu de profil x est:

$$\hat{Y} = 1\ si\ \beta_0 + \beta' x > 0$$

En toute généralité, l'erreur théorique de mauvais classement d'un tel classifieur s'écrit:

$$\pi(\beta_0, \beta) = P\left(\hat{Y} \neq Y\right)$$

$$= p_1 P(\beta_0 + \beta' x < 0 | Y = 1) + p_0 P(\beta_0 + \beta' x > 0 | Y = 0) \textbf{(I.6)}$$

Compte tenu de l'hypothèse de normalité des covariables, cette erreur s'écrit, en fonction de βet β0 :

$$\pi(\beta_0, \beta) = p_1 \left[1 - \Phi\left(\frac{\mu_1'\beta + \beta_0}{(\beta'\sum\beta)^{1/2}}\right)\right] + p_0 \Phi\left(\frac{\mu_0'\beta + \beta_0}{(\beta'\sum\beta)^{1/2}}\right) \textbf{(I.7)}$$

où Φ est la fonction de répartition d'une loi normale centrée réduite. En optimisant cette fonction en βet β_0, on obtient les coefficients β^* et β^*_0 minimisant l'erreur théorique de mauvais classement. Ils s'expriment en fonction de$\mu_0, \mu_1 \; et \sum$:

$$\beta^* = \sum^{-1}(\mu_1 - \mu_0) \textbf{(I.8)}$$

$$\beta_0^* = log\frac{p_1}{p_0} - 0.5(\mu_1'\sum^{-1}\mu_1 - \mu_0'\sum^{-1}\mu_0) \qquad \textbf{(I.9)}$$

Si l'on considère le log ratio des probabilités introduit Expression **(I.3)**, on retrouvebien l'expression des coefficients de la règle de Bayes. Dans ce cas, l'erreur théorique de mauvais classement d'une telle règle de classification est minimale et peut aussi s'écrire :

$$\pi^* = p_1 \left[1 - \Phi\left(log\frac{p_1}{p_0}\frac{1}{\Delta_\sum} + \frac{\Delta_\sum}{2}\right)\right] + p_0 \Phi\left(log\frac{p_1}{p_0}\frac{1}{\Delta_\sum} - \frac{\Delta_\sum}{2}\right) \textbf{(I.10)}$$

où $\Delta_\sum = \sqrt{(\mu_1 - \mu_0)'\sum^{-1}(\mu_1 - \mu_0)}$est la distance de Mahalanobis entre les groupes 0 et 1 pour la métrique $\sum^{-1}$ Cette expression permet d'introduire la fonction d'erreur π suivante :

$$\pi(\mu_0, \mu_1, \sum) = p_1 \left[1 - \Phi\left(log\frac{p_1}{p_0}\frac{1}{\Delta} + \frac{\Delta}{2}\right)\right] + p_0 \Phi\left(log\frac{p_1}{p_0}\frac{1}{\Delta} - \frac{\Delta}{2}\right) \textbf{(I.11)}$$

où$\Delta = \sqrt{(\mu_1 - \mu_0)'\sum^{-1}(\mu_1 - \mu_0)}$C'est l'erreur de classement théorique d'unerègle de classification de Bayes construite à partir des paramètres μ_0, μ_1et $\sum^{-1}$ montre les valeurs que prend cette fonction πpour plusieurs distances de Mahalanobis entre deux groupes et pour plusieurs probabilités de base dans lespopulations. On voit naturellement que plus la distance entre les groupes augmente,plus l'erreur de classement du classifieur de Bayes associée à cette situation estfaible, plus il est facile de classer les individus sans erreur. On peut donc se direqu'on souhaite se ramener à une situation où les paramètres$\pi(\mu_0, \mu_1, \sum)$ sont telsque cette erreur est la plus faible possible. Cette fonction sera utile par la suite pour comparer le classifieur de Bayes et le classifieur de Bayes conditionnel,

L'analyse discriminante quadratique (QDA) étend la LDA au cas où les matrices de covariances différentes d'un groupe à l'autre. La QDA ne sera pas détaillée dansce manuscrit car elle est rarement mise en avant en grande dimension. En effet, le nombre d'individus étant souvent faible, de l'ordre de quelques dizaines, il devient alors difficile d'estimer la matrice de covariance dans chaque groupe.

Analyse discriminante de Fisher Sans hypothèse sur la loi des variables X,l'analyse discriminante de Fisher (Fisher (1936)) peut être vue comme la recherched'une projection des observations Xafin d'atteindre une bonne séparation des groupes. Si µdésigne la moyenne totale sur tous les groupes, la variance intergroupes s'écrit :

$$\sum_b = \frac{1}{K}\sum_{y=1}^{K}(\mu_y - \mu)(\mu_y - \mu)' \textbf{(I.12)}$$

Fisher propose d'étudier la maximisation du ratio de la variance inter-groupes $\sum_b$ sur la variance intra-groupes$\sum$.

On cherchealors des vecteurs discriminants$(\beta_1; \dots; \beta_{K-1})$ orthogonaux tels que :

$$\max_{\beta_y} \frac{\beta_y' \sum_b \beta_y}{\beta_y' \sum \beta_y} \textbf{(I.13)}$$

$\sum_b$ étant de rang K -1, les vecteurs propres de la matrice $\sum^{-1}\sum_b$ sont solution de cette maximisation [17].En pratique, ces vecteurs discriminants sont utiles pour visualiser la séparation des groupes en traçant les nuages de point (Xβ1;Xβ2) etc.

Optimal scoring Une troisième formulation de la LDA peut être faite par analogie avec la régression [18]. La variable catégorielle de groupe Y est transformée en variables quantitatives par des scores. On note $Y^{(d)}$ la matrice $(n \times K)$ de variables design contenant les indicatrices d'appartenance à un groupe : $Y^{(d)}{}_{iy} =$ 1si l'individu appartient à la classe y et 0 sinon. Le problème de classification revient à estimer les paramètres $(\theta_1; \dots; \theta_{K-1}) et (\beta_1; \dots; \beta_{K-1})$ telsque :

$$\min_{\beta_y, \theta_y} \left\| Y^{(d)}\theta_y - X\beta_y \right\| \textbf{(I.14)}$$

$$s.c. \quad \frac{1}{n}\theta_y' Y^{(d)\prime} Y^{(d)} \theta_y = 1, \theta_y' Y^{(d)\prime} Y^{(d)} \theta_{y'} = 0, y' < y \textbf{(I.15)}$$

où$(\theta_1; \dots; \theta_{K-1})$sont des K-vecteurs de scores $et\ (\beta_1; \dots; \beta_{K-1})$les mêmes vecteurs discriminants que dans l'analyse de Fisher [19].

Finalement, la LDA est simple à implémenter et donne en général de bons résultats en termes d'erreur de classement. Cependant, elle fait appel à l'inverse de la matrice de covariance, difficile à estimer en grande dimension, et n'est pas parcimonieuse. De

nombreuses extensions ont donc été proposées pour adapter la LDA au cadre de la grande dimension [20].

I.3.2. Algorithme de LDA

L'algorithme LDA est n´e des travaux de Belhumeur et al. de la Yale University (USA), en 1997.Il est aussi connu sous le nom de Fisherfaces. Contrairement à l'algorithme PCA, l'algorithme LDA effectue une véritable séparation de classes (FigureI.2). Pour pouvoir l'utiliser, il faut donc au préalable organiser la base d'apprentissage d'images en plusieurs classes.

Le LDA analyse les vecteurs propres de la matrice de dispersion des données, avec pour objectif de maximiser les variations inter-classes tout en minimisant les variations intra-classes [21].

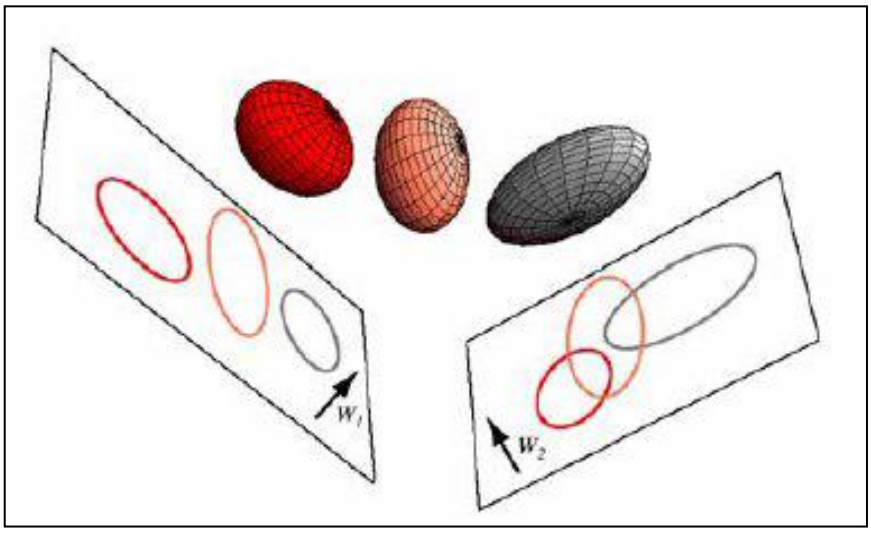

Figure I.2. Illustration du principe de séparation optimale des classes par le LDA.

I.3.3. Les étapesde développement

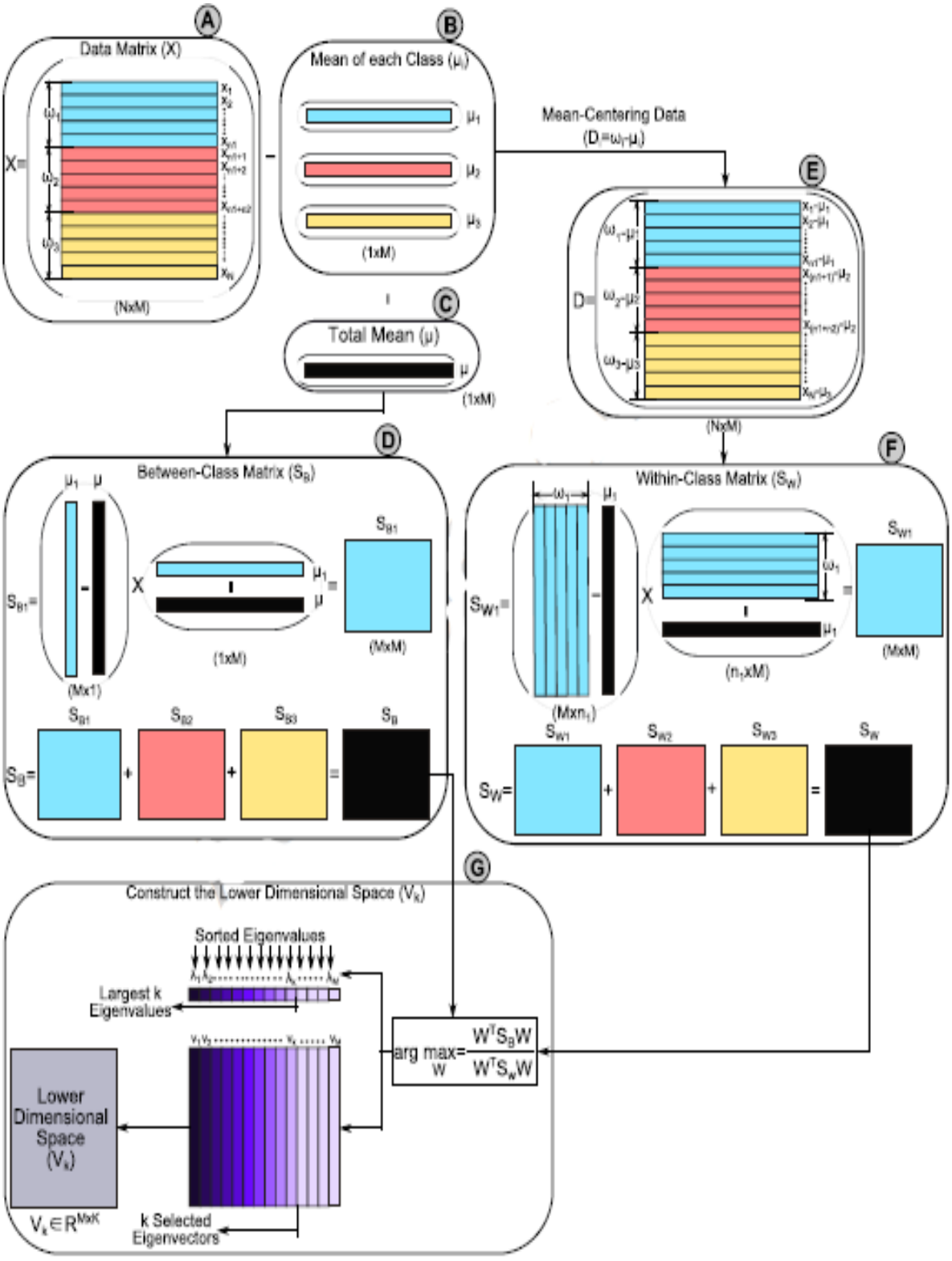

Figure I.3. Etapes visuelles pour calculer un sous-espace avec des dimensions inférieures pour la technique LDA**.**

La figure permette de visualiser les étapes de la technique LDA.

I.3.3.1. Calcul de la variance entre classes (SB)

La variance entre classes de la classes i(SBi) représente la distance entre la moyenne de la classes i (μi) et la moyenne totale (μ). La technique LDA recherche un espace de dimension inférieure, utilisée pour maximiser la variance entre les classes, ou simplement maximiser la distance de séparation entre les classes. Pourexpliquer comment la variance entre classes ou la matrice entre classes (SB) peut être calculée, les hypothèses suivantes sont formulées. Etant donné la matrice de données d'origine $X = \{x1, x2, \ldots, xN\}$, où x_i représente le i^{éme} échantillon, schéma ou observation et N le nombre total d'échantillons. Chaque échantillon est représenté par M caractéristiques ($x_i \in R^M$). En d'autres termes, chaque échantillon est représenté sous la forme d'un point dans l'espace à M dimensions. Supposons que la matrice de données soit partitionnée en c=3 classes comme suit,

X=[w1,w2,w3] comme indiqué dans FigureI.3. (étape(A)). Chaque classe a cinq échantillons (c-à-d:n1=n2=n3=5),où ni représente le nombre de échantillons de la i^{éme} classe. Le nombre total d'échantillons (N) est calculé comme suit,$N = \sum_{i=1}^{3} ni$.

Pour calculer la variance entre classes (SB), la distance de séparation entre différentes classes qui est noté par ($m_i - m$) sera calculé comme suit:

$$(mi - m)^2 = (w^T \mu_i - w^T \mu) x^2$$
$$= w^T (\mu_i - \mu)(\mu_i - \mu)^T W \textbf{(I.16)}$$

où mi représente la projection de la moyenne de la classe i et est calculé comme suit, $mi = w^T \mu_i$,où m est la projection de la moyenne totale de toutes les classes et calculé comme suit , $m = w^T \mu$, w représente la matrice de transformation de LDA ,μ_i(1×M) représente la moyenne de la classes i et est calculée comme dans l'équation (I.17), et μ(1×M) est la moyenne totale de toutes les classes et peut être calculé comme dans équation (I.18).la FigureI.3 montre la moyenne de chaque classe et la moyenne totale aux étapes (B) et (C) , respectivement.

$$\mu_j = \frac{1}{n_j} \sum_{x_i \in w_j} x_i \textbf{(I.17)}$$

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i = \sum_{i=1}^{c} \frac{n_i}{N} \mu_i \textbf{(I.18)}$$

où c représente le nombre total de classes (dans notre exemple, c=3).

Le terme$(\mu_i - \mu)(\mu_i - \mu)^T$ dans l'équation (I.16) représente la distance de séparation entre la moyenne des i^{éme} classe (μ_i)et la moyenne totale (μ), ou tout simplement il représente la variance inter-classe de la i^{éme} classe (SBi).

Remplacez SBi par l'équation (I.16) comme suit:

$$(m_i - m)^2 = w^T S_{Bi} w \qquad \textbf{(I.19)}$$

La variance totale entre les classes est calculée comme suit ,($S_B = \sum_{i=1}^{c} n_i S_{Bi}$).

L'étape(D) montre d'abord comment la matrice inter-classe de la première classe ajoutant toutes les matrices inter-classes de toutes les classes.

I.3.3.2. Calcul de la variance intra-classe (SW)

La variance intra-classe de la classe i (SWi) représente la différence entre la moyenne et les échantillons de cette classe. La technique LDA recherche un espace de dimension inférieure, utilisée pour minimiser la différence entre la moyenne projetée (m_i) et les échantillons projetés de chaque classe $(w^T x_i)$, ou simplement pour minimiser la variance intra-classe .la variance intra-classe de chaque classe (SWj) est calculée comme dans l'équation (I.20).

$$\sum_{x_i \in w_j, j=1,\dots,c} (w^T x_i - m_j)^2$$

$$= \sum_{x_i \in w_j, j=1,\dots,c} (w^T x_{ij} - w^T \mu_j)^2$$

$$= \sum_{x_i \in w_j, j=1,\dots,c} w^T (x_{ij} - \mu_j)^2 \, w$$

$$= \sum_{x_i \in w_j, j=1,\dots,c} w^T (x_{ij} - \mu_j)(x_{ij} - \mu_j)^T \, w$$

$$= \textstyle\sum_{x_i \in w_j, j=1,\dots,c} w^T S_{Wj} \, w \qquad \textbf{(I.20)}$$

A partir de l'équation (I.20), la variance intra-classe pour chaque classe peut être calculée comme suit:

$$S_{Wj} = d_j^T * d_j = \sum_{i=1}^{n_j}(x_{ij} - \mu_j)(x_{ij} - \mu_j)^T \textbf{(I.21)}$$

où x_{ij} représente le i^éme échantillon de la j classe, comme indiqué sur la FigureI.3 (étape (E), (F)),et$d_j$est la donnée de centrage de la j classe, c-à-d: $d_j = w_j - \mu_j = \{x_i\}_{i=1}^{n_j} - \mu_j$. De plus, l'étape (F) de la FigureI.3 illustre le calcul de la variance intra-classe de la première classe (SW1) dans notre exemple. La variance intra-classe totale représente la somme de toutes les matrices intra-classe de toutes les classes (voir la FigureI.3 (étape(F))) et peut être calculée comme dans l'équation (I.22).

$$S_W = \sum_{i=1}^{3} S_{Wi}$$

$$= \sum_{x_i \in w_1} (x_i - \mu_1)(x_i - \mu_1)^T$$

$$+ \sum_{x_i \in w_2} (x_i - \mu_2)(x_i - \mu_2)^T$$

$$+ \sum_{x_i \in w_3} (x_i - \mu_3)(x_i - \mu_3)^T \qquad \textbf{(I.22)}$$

I.3.3.3. Construire l'espace dimensionnel inférieur

Après avoir calculé la variance entre classes (SB) et la variance intra-classe (SW), la matrice de transformation (W) de la technique LDA peut être calculée comme dans l'équation (I.23), appelée critère de Fisher. Cette formule peut être reformulée comme dans l'équation (I.24).

$$\arg\max_W \frac{W^T S_B W}{W^T S_W W} \textbf{(I.23)}$$

$$S_W W = \lambda S_B W \qquad \textbf{(I.24)}$$

oùλreprésente les valeurs propres de la matrice de transformation (W). la solution de ce problème peut être obtenu en calculant les valeurs propres(λ={λ1,λ2,…,λM}) et les vecteurs propres (V={v1,v2,…,vM}) de $W = S_W^{-1} S_B$,Si SW est non singulier.

Les valeurs propres sont des valeurs scalaires, tandis que les vecteurs propres sont des vecteurs non nuls, ce qui satisfait l'équation (I.24) et nous fournit des informations sur l'espace LDA. Les vecteurs propres représentent les directions du nouvel espace et les valeurs propres correspondantes représentent le vecteur de mise à l'échelle, la longueur ou la magnitude des vecteurs propres.

Ainsi, chaque vecteur propre représente un axe de l'espace LDA et la valeur propre associée représente la robustesse de ce vecteur propre. La robustesse du vecteur propre reflète sa capacité à discriminer entre différentes classes, c'est-à-dire à augmenter la variance entre les classes et à réduire la variance intra-classe de chaque classe ; répond donc à l'objectif de la LDA. Ainsi, les vecteurs propres avec les k valeurs propres les plus élevées sont utilisés pour construire un espace dimensionnel inférieur (V_k), tandis que les autres vecteurs propres ($\{v_{k+1},v_{k+2},v_M\}$) sont négligés, comme le montre la FigureI.3 (étape(G)) [22]

I.4. L'analyse en composantes principales (PCA)

Récemment, les auteurs de l'article [23] ont montré que la projection des *i*vecteurs dans le sous-espace de PCA, formé par les axes ayant de fortes variances, compense la variabilité intra-session. De plus, les axes du sous-espace de PCA sont pondérés par la racine carrée des valeurs propres correspondant afin d'accentuer leur importance.

À vrai dire, ce traitement via la PCA est considéré comme local, et ce, du fait que la matrice de covariance de PCA est estimée localement pour chaque conversation en utilisant seulement ses propres segments (i-vecteurs). De par ce caractère local, la normalisation par la projection de PCA a l'avantage de ne solliciter aucun ensemble externe de données de développement.

Les auteurs de[23]recommandent de choisir la dimension de la PCA de manière à conserver une quantité de 50 % de la variance totale des données. Dans nos travaux, nous notons cette quantité par le caractère η. Finalement, la normalisation14 à **1** de la norme euclidienne des i-vecteurs est également souhaitable dans ce contexte tout comme dans le contexte de la vérification du locuteur [24]

L'ACP est une ancienne approche, qui effectue une réduction de dimension par projection des points originaux dans un sous-espace vectoriel de dimension plus réduite. L'ACP détermine des axes de projections orthogonaux, qui maximisent la variance expliquée. Dans la base formée par ces axes, les coordonnées ne sont pas corrélées. L'ACP maximise la variance de la projection dans l'espace de caractéristiques, ce qui est équivalent à minimiser l'erreur quadratique moyenne de reconstruction.

L'ACP se calcule en diagonalisant la matrice de corrélations, le plus souvent en utilisant une décomposition en valeurs singulières (SVD). Elle est très utilisée car elle est simple à mettre en œuvre. Elle est limitée par son caractère linéaire : il est facile d'imaginer des situations dans lesquelles l'ACP n'apporte aucune information utilisable (par exemple, des

données réparties sur un tore en dimension n). A titre illustratif, la figureI.3 présente les Iris de Fisher dans la base obtenue par une ACP sous forme de nuages de points [4].

Le PCA essaye de nous fournir un ensemble d'axes orthogonaux le long desquelles nous pouvons projeter nos données. Elle nous permet si tout va bien d'expliquer la majeure partie des données avec juste les premiers axes dans le nouvel espace. Les tentatives de PCA est de représenter efficacement les données en trouvant les axes orthonormaux qui decorrolent au maximum ces données.

I.4.1. Utilisation de PCA

1. Trouver les vecteurs propres, et arranger les par ordre de valeur propre décroissante.
2. Projeter les points tests sur les vecteurs propres
3. Employer ces coefficients de projection pour faire quelque chose utile (classification, reconstruction d'image, etc.) [14].

PCA essaye de nous fournir un ensemble d'axes orthogonaux le long desquelles nous pouvons projeter nos données. Elle nous permet si tout va bien d'expliquer la majeure partie des données avec juste les premiers axes dans le nouvel espace.

Les tentatives de PCAsont de représenter efficacement les données en trouvant les axes orthonormaux qui décorrèlent au maximum ces données.

I.4.2. L'algorithme standard de PCA

Centrer les Observations : les vecteurs colonnes$X_i \in R^N, i = 1, \dots, m$

Le centrage signifie:

$$\frac{1}{m}\sum_{i=1}^{m} x_i = 0 \textbf{(I.25)}$$

La PCA trouve les principaux axes par diagonalisation de la matrice de covariance :

$$C = \frac{1}{m}\sum_{i=1}^{m} x_i x_i^T \textbf{(I.26)}$$

Noter que C'est défini positif, et peut être diagonalisée ainsi avec des valeurs propres non négatives

$$\lambda v = Cv \textbf{(I.27)}$$

Avec λ: Matrice de vecteurs propres

L'idée est de mettre au point une technique permettant de résumer l'information apportée par « p » variables quantitatives par « n » unités (appelées *individus*) en la détruisant le moins possible. Cette technique utilise des combinaisons linéaires des variables pour la réduction de la dimension d'un ensemble de données en trouvant un nouvel ensemble

constitué d'un petit nombre de facteurs, de dimension plus faible que l'ensemble original des variables. Produire un résumé de cette information, c'est projeter ces points dans un espace de dimension inférieure à « p », le nombre de variables initiales. Toutefois, cette réduction doit garder le maximum d'information. Les axes de ce sous-espace sont dits « *axes factoriels* » ou « *facteurs* ». Chaque variable « p »porte en elle, une part d'information originale ou part d'inertie et une part d'information originale redondante avec les autres, venant des corrélations entre variables.

C'est cette part d'information redondante qui va être regroupée dans le résumé factoriel. Les facteurs sont hiérarchisés de la manière suivante [25] :

- le 1er axe concentre le maximum de l'information : c'est l'axe de la plus grande dimension du nuage de points et il fournit le meilleur résumé dans un espace à une dimension, mais il laisse des résidus d'information ;
- le 2ème axe concentre le maximum de l'information restante, il est orthogonal au premier et c'est le meilleur résumé dans un espace à deux dimensions. Mais, de même il laisse aussi des résidus ;
- le 3ème axe prend encore une part d'information moindre, il est orthogonal aux deux premiers. Et ainsi de suite, pour les axes suivants tant que l'on pense qu'ils apportent encore de l'information.

Le nombre de composantes en théorie est égal au nombre de variables originelles. Mais, en pratique, les premières directions permettent de couvrir un pourcentage élevé (80-90%) de toutes les données originelles et sont donc utilisées pour restreindre l'espace d'observation.

Conclusion

Ce chapitre a fait l'objet de définitions des principaux de quelques méthodes de sélection des variables. Le chapitre suivant fera l'objet d'une mise en œuvre théorique des techniques de l'intelligence artificielle basées sur l'apprentissage statistique et appliquées sur des données multi sensorielles. L'objectif, rappelons-le encore une fois, est l'application de ces techniques comme étant une solution dans la surveillance de la qualité de l'eau par reconnaissance de formes.

Chapitre II

Techniques D'apprentissage

Introduction

L'apprentissage automatique est un sous-domaine de l'intelligence artificielle (IA). En général, l'objectif de l'apprentissage automatique est de comprendre la structure desdonnées et de les intégrer dans des modèles qui peuvent être compris et utilisés par les tout le monde. Les RNAs sont composés des ensembles de neurones formels interconnectés permettant la résolution de problèmes complexes tels que la reconnaissance des formes ou le traitement du langage naturel, grâce à l'ajustement des valeurs des connections (ou poids) dans une phase d'apprentissage.

Nous allonsprésenterdanscechapitreuneétudegénéralesurl'apprentissageautomatiqueet nous nousconnaissonssurl'algorithmes utilisés dans l'apprentissage d'entre eux(les réseaux de neurones par exemple),etnous allonsprésenter le protocole d'apprentissage, les différentes architectures de ces réseaux, et aussi présenter des exemples usuel de RNA tel que les ELMs, leur principe de fonctionnement, et leur caractéristiques principales. Et aussi nous allons présenter une généralité sur RBF, on terminer notre chapitre par Réseau de neurones de type MLP.

II.1. Apprentissage automatique

L'apprentissage automatique (en anglais *machine learning*, littéralement « l'**apprentissage machine** ») ou **apprentissage statistique** est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches statistiques pour donner aux ordinateurs la capacité d' « apprendre » à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. Plus largement, cela concerne la conception, l'analyse, le développement et l'implémentation de telles méthodes [26].

L'apprentissage automatique comporte généralement deux phases. La première consiste à estimer un modèle à partir de données, appelées observations, qui sont disponibles et en nombre fini, lors de la phase de conception du système. L'estimation du modèle consiste à résoudre une tâche pratique, telle que traduire un discours, estimer une densité de probabilité, reconnaître la présence d'un chat dans une photographie ou participer à la conduite d'un véhicule autonome. Cette phase dite « d'apprentissage » ou « d'entraînement » est généralement réalisée préalablement à l'utilisation pratique du modèle. La seconde phase correspond à la mise en production : le modèle étant déterminé, de nouvelles données peuvent alors être soumises afin d'obtenir le résultat correspondant à

la tâche souhaitée. En pratique, certains systèmes peuvent poursuivre leur apprentissage une fois en production, pour peu qu'ils aient un moyen d'obtenir un retour sur la qualité des résultats produits [26].

Selon les informations disponibles durant la phase d'apprentissage, l'apprentissage est qualifié de différentes manières. Si les données sont étiquetées (c'est-à-dire que la réponse à la tâche est connue pour ces données), il s'agit d'un apprentissage supervisé. On parle de classification ou de classement si les étiquettes sont discrètes, ou de régression si elles sont continués. Si le modèle est appris de manière incrémentale en fonction d'une récompense reçue par le programme pour chacune des actions entreprises, on parle d'apprentissage par renforcement. Dans le cas le plus général, sans étiquette, on cherche à déterminer la structure sous-jacente des données (qui peuvent être une densité de probabilité) et il s'agit alors d'apprentissage non supervisé. L'apprentissage automatique peut être appliqué à différents types de données, tels des graphes, des arbres, des courbes, ou plus simplement des vecteurs de caractéristiques, qui peuvent être continues ou discrètes[26].

II.1.1.Algorithmes d'apprentissage

Parmi les algorithmes existant dans les littératures on trouve :

- les machines à vecteur de support ;
- le boostant ;
- les réseaux de neurones , dont les méthodes d'apprentissage profond (*deeplearning* en anglais) pour un apprentissage supervisé ou non-supervisé ;
- la méthode des k plus proches voisins pour un apprentissage supervisé ;
- les arbres de décision , méthodes à l'origine des Random Forest, par extension également du boosting (notamment xgboost) ;
- les méthodes statistiques comme le modèle de mixture gaussienne ;
- la régression logistique ;
- l'analyse discriminante linéaire ;
- les algorithmes génétiques et la *programmation génétique*.

Ces méthodes sont souvent combinées pour obtenir diverses variantes d'apprentissage. L'utilisation de tel ou tel algorithme dépend fortement de la tâche à résoudre (classification, estimation de valeurs…)[26].

II.2. Réseau de neurones artificiels

Un réseau de neurones artificiels, ou réseau neuronal artificiel, est un système dont la conception est à l'origine schématiquement inspirée du fonctionnement des neurones biologiques, et qui par la suite s'est rapproché des méthodes statistiques.

Les réseaux de neurones sont généralement optimisés par des méthodes d'apprentissage de type probabiliste, en particulier bayésien. Ils sont placés d'une part dans la famille des applications statistiques, qu'ils enrichissent avec un ensemble de paradigmes permettant de créer des classifications rapides (réseaux de Kohonen en particulier), et d'autre part dans la famille des méthodes de l'intelligence artificielle auxquelles ils fournissent un mécanisme perceptif indépendant des idées propres de l'implémenter, et fournissant des informations d'entrée au raisonnement logique formel (*Deep Learning*)[27]

En modélisation des circuits biologiques, ils permettent de tester quelques hypothèses fonctionnelles issues de la neurophysiologie, ou encore les conséquences de ces hypothèses pour les comparer au réel[27].

II.2.1. Types des réseaux de neurones

On utilise dans ce travail deux types de réseaux de neurones

- Les réseaux de type **MLP** (*multi-layer perceptron*) calculent une combinaison linéaire des entrées, c'est-à-dire que la fonction de combinaison renvoie le produit scalaire entre le vecteur des entrées et le vecteur des poids synaptiques.
- Les réseaux de type **RBF** (*radial basis function*) calculent la distance entre les entrées, c'est-à-dire que la fonction de combinaison renvoie la norme euclidienne du vecteur issu de la différence vectorielle entre les vecteurs d'entrées [27].

II.3.Réseau de neurones de type RBF (Radial Basis Functions)

Les réseaux à fonction radiales de base (RBF)sont des modèles connexionnistes simples à mettre et œuvre et assez intelligible, et sont très utilisés pour la classification. Leurpropriétés théoriques ont été étudiées en détail depuis la fin des années 80 ;il s'agit certainement, avec le perceptron multicouche, du modèle connexionniste le mieux connu[28].

II.3.1.Architecture

Introduit par Powell et Broomhead, le réseau RBF fait partie des réseaux de neurones supervisés. Il est constitué de trois couches (figure II.1): une couche d'entrée qui retransmet les entrées sans distorsion, une seule couche cachée qui contient les neurones RBF qui sont généralement des 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[28].

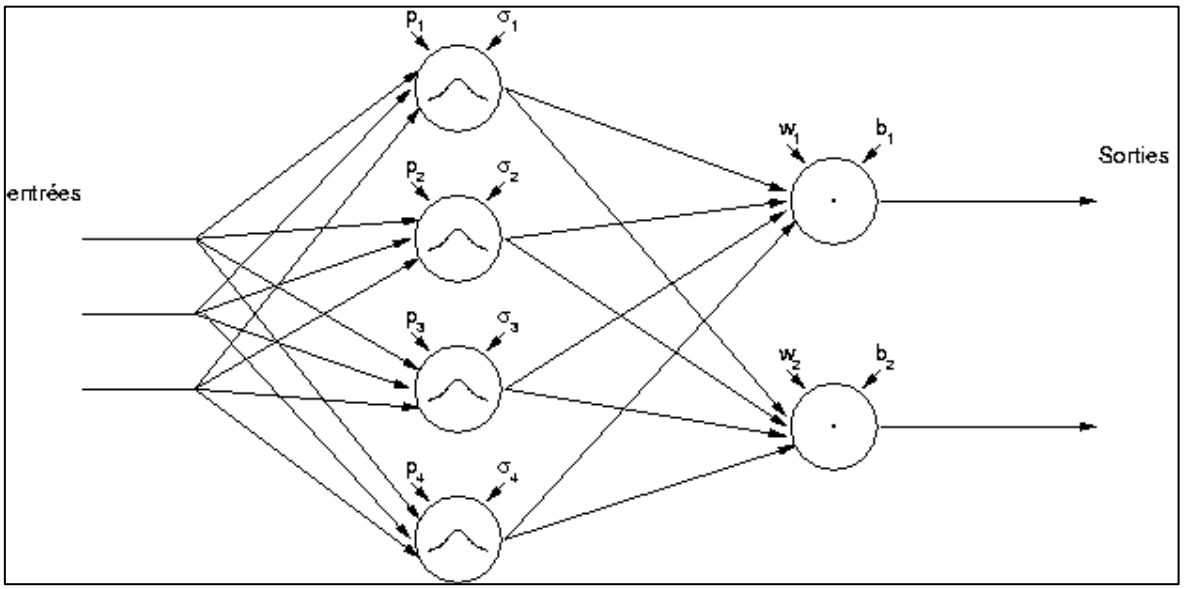

Figure II.1. Présentation schématique d'un réseau RBF.

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[28] :

$$y_m^{(q)} = exp\left[-\left\|x^{(q)} - v_m\right\|^2/(2\sigma_m^2)\right] \qquad \text{(II. 1)}$$

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

$$z_j^{(q)} = \left(\frac{1}{M}\right)\left[\sum_{(m=1,M)} w_{mj} {y_m}^{(q)}\right] \qquad \text{(II. 2)}$$

$m = 1, \dots, M \; et \; j = 1, \dots, J.$

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

II.3.2.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[29]:

$$E = \sum_{q=1}^{Q} \sum_{j=1}^{J} t_j^{(q)} z_j^{(q)} \qquad (\text{II}.3)$$

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

$w_{mj}^{(i+1)} w_{mj}^{i} + \eta(t_j - z_j) y_m t_j$ est la sortie du j$^{\text{ième}}$ neurone désirée, z_j est la sortie du j$^{\text{ième}}$ neurone calculée, $y_m$est la sortie du m$^{\text{ième}}$ neurone de la couche cachée et η est le pas d'apprentissage dont sa valeur est comprise entre 0 et 1[30].

II.3.3.Caractéristiques principales de RBF

- **Le nombre des couches cachées**

Un réseau RBF ne peut contenir qu'une seule couche cachée, son architecture est fixée pour tous les problèmes à étudier.

- **La fonction d'activation**

Le réseau RBF utilise toujours une fonction dite à base radiale centrée d'un point et munie d'un rayon.

- **Les poids synaptiques**

Les poids entre la couche d'entrée et la couche cachée dans les modèles neuronaux de type RBF sont toujours d'une valeur d'unité, c'est-à-dire que l'information inscrite sur la couche d'entrée sera retransmise sans distorsion vers les neurones de la couche cachée.

En ce qui concerne les ressemblances entre un réseau RBF et un PMC, on peut mentionner quelques points

- **La fonction de sortie**

Généralement une simple fonction linéaire qui renvoie une sommation pondérée des valeurs calculées par les neurones de la couche cachée. Bien sûr, ce n'est pas toujours le cas, parfois l'utilisation d'autres fonctions pourrait être plus adéquate dans un problème donné.

- **Le sens des connexions**

Les connexions entre les couches suivent le même sens, on peut dire qu'elles ne sont pas récurrentes, et chaque neurone est entièrement connecté vers les neurones de la couche suivante [31].

II.3.4.Apprentissage des modèles RBF

L'apprentissage d'un modèle RBF consiste à déterminer son architecture (le nombre N de fonctions radiales) et à fixer les valeurs des paramètres. La plupart des utilisateurs déterminent empiriquement la valeur de N en recourant à des techniques de validation croisée[32].

L'apprentissage d'un réseau RBF est de type supervisé : on dispose d'un ensemble d'apprentissage constitué de l couples (vecteur d'entrée, valeur cible) :

$$(x_1, y_1), \dots, (x_m, y_m), x_1 \in R^d, y_1 \in R$$

et du coûtassocié à chaqueexemple :

$$E_i = \frac{1}{2}\left(y_i - f(x_1)\right)^2 \qquad \text{(II. 4)}$$

(auquel on ajouteéventuellement un terme de régularisation).

Une caractéristique intéressante des modèles RBF est que l'on peut diviser lesparamètres en trois groupes : les centres μ, les largeurs σ et les poids w. L'interprétation de chaque groupe permet de proposer un algorithme d'apprentissage séquentiel, simple et performant

II.3.4.1.Approche séquentielle

Cette technique d'apprentissage proposée dès la fin des années 1980 est très couramment utilisée. Elle consiste à optimiser successivement les trois jeux de paramètres(μ_j, σ_j, w_j). Cette technique a l'avantage d'être simple à mettre en œuvre, de demander peu de calculs et de donner des résultats acceptables. La solution obtenue n'est cependant pas optimale[32].

Dans un premier temps, on estime les positions des centres μ_j et des largeurs σ_j à l'aide d'un algorithme non supervisé de type k-moyennes. Une fois ces paramètres fixés, il est possible de calculer les poids w_j optimaux par une méthode de régression linéaire. C'est certainement la simplicité et l'efficacité de cette méthode qui a fait le succès des RBF[32].

a. Calcul des poids

Si l'on suppose les centres et largeurs connus, les poids w optimaux se calculent aisément[32] :

$$Y_{(x)} = \sum_{J=1}^{N} w_j \, \emptyset\left(\|X - \mu_j\|, \sigma_j\right) = \sum_{J=1}^{N} W_J h_J \, (x) \qquad \text{(II. 5)}$$

On cherche la solution w qui minimise la différence e entre la sortie estimée et la sortie désirée. On a donc un système d'équations linéaires qui s'écrit :

$$Y = H\, w + e \qquad \text{(II. 6)}$$

La matrice **H**, de taille **l × N**, donne les réponses des **N** centres **RBF** sur les l exemples, y est un vecteur regroupant les l sorties y_i sur l'ensemble d'apprentissage, et e est le vecteur d'erreur. Le critère à optimiser est :

$$E = e^T e \qquad \text{(II. 7)}$$

Si l'on ajoute un terme de régularisation de type ridgeregression, qui pénalise les solutions avec de grandes valeurs des poids, on écrit :

$$E = e^T e + \lambda w^T w \qquad \text{(II. 8)}$$

La solution s'obtient par un calculclassique de pseudo-inverse, et s'écrit :

$$W = (H^T H + \boldsymbol{\lambda}\mathbf{I})^{-1} H^T \qquad \text{(II. 9)}$$

où **I** est la matrice identité de taille l.

La régression de type ridge est très utilisée en apprentissage statistique. Dans le contexte des réseaux connexionnistes (par exemple les Perceptrons multicouches), on l'appelle *souvent weightdecay***.** Le paramètre **λ** est libre et doit être déterminé par validation croisée ou, de manière plus sophistiquée, en employant des méthodes bayé- siennes de ré estimation[32].

En pratique, il est recommandé de résoudre le système d'équation en utilisant une décomposition en valeurs singulières (SVD), qui résiste bien aux problèmes de mauvais conditionnement numérique[32].

b. Estimation non supervisée des centres et des largeurs

Afin de déterminer les positions et largeurs des centres gaussiens, on les interprète comme représentant la densité de probabilité des données et on cherche une solution locale (chaque fonction va s'activer dans une « petite » région de l'espace d'entrée). On désire qu'au moins un centre soit activé, c'est-à-dire que la valeur de la fonction radiale soit non

négligeable, dans toutes les régions où l'on a des données. La dimension de l'ensemble des points associés à un centre va permettre d'estimer la largeur de ce centre.

II.3.5.Le problème principal que rencontrent les modèles RBF

Le problème principal de RBF est lié à leur comportement lorsque la dimension de l'espace d'entrée augmente (« malédiction de la dimension »). Si l'on veut couvrir l'espace d'entrée avec des sphères placées sur les centres RBF, le nombre de sphères nécessaires augmente exponentiellement avec la dimension d des entrées, affectant non seulement les temps de calcul mais aussi augmentant proportionnellement le nombre d'exemples requis pour l'estimation correcte des paramètres.

Pour ces raisons, on observe facilement en pratique que les performances des modèles RBF se dégradent rapidement lorsque la dimension des entrées augmente. Il est alors nécessaire de faire précéder le système RBF par une phase de réduction de dimension (sélection de variables supervisée ou non)[32].

II.4. Réseau de neurones de type MLP

Les réseaux de neurones de type Perceptrons Multicouches (Multi Layer Perceptron -MLP) sont des réseaux à propagation avant, composes d'une ou plusieurs couches cachées et d'une couche de sortie. Chaque couche du réseau est composée de neurones artificiels. La première couche cachée reçoit l'information provenant des entrées. L'information est traitée et transmise vers les couches suivantes jusqu''a la dernière. Les MLP sont connus comme étant des approximateurs universels et sont très utilises dans des problèmes de régression non linéaire[33]**.** Les signaux se propagent de l'entrée vers la sortie (direction de propagation avant). La fonction de transfert est de type sigmoïdale (log-sigmoïde). Pour effectuer l'apprentissage nous avons utilisé une modification de l'algorithme standard de rétro propagation appelé la "rétro-propagation avec terme de moment". Le moment permet au réseau de répondre non seulement au gradient local mais aussi aux tendances récentes de la surface d'erreur (hyper-surface de la fonction de coût vis-à vis des paramètres libres du réseau, c'est-à-dire les poids synaptiques)[34]**.**

II.4.1.Processus d'apprentissage dans les réseaux MLP

Les réseaux MLP utilisent un mode d'apprentissage supervise. Dans ce mode d'apprentissage, un ensemble de données constitue des entrées du système a modéliser et des sorties correspondantes est présente au réseau qui doit adapter ses paramètres suivant

un algorithme d'apprentissage de façon a ce que la divergence entre la sortie du système et celle du modèle soit suffisamment faible[35].

II.4.2. L'algorithme de retro propagation

Les techniques d'apprentissage les plus utilisées dans les réseaux MLP sont l'algorithmede retro propagation du gradient. L'algorithme de rétro propagation du gradient (RP) est certainement à la base des premiers succès des réseaux de neurones. Sa mise en application a permis au domaine du connexionnisme de sortir de la période de silence qui a régné après la sortie du livre «Perceptrons » de Minsky et Papert.

On considère un réseau à trois couches illustrées par la figure (II.2). Les conventions de notation sont les suivantes :

O_k activation de la unité de sortie, k = 1, …n, M;

T_k activation désirée de la k^e unité de sortie;

C_j activation de la je unité cachée, j = 0, 1, …,n_h ; c_0 =1 : c'est l'entrée du biais pour la couche de sortie;

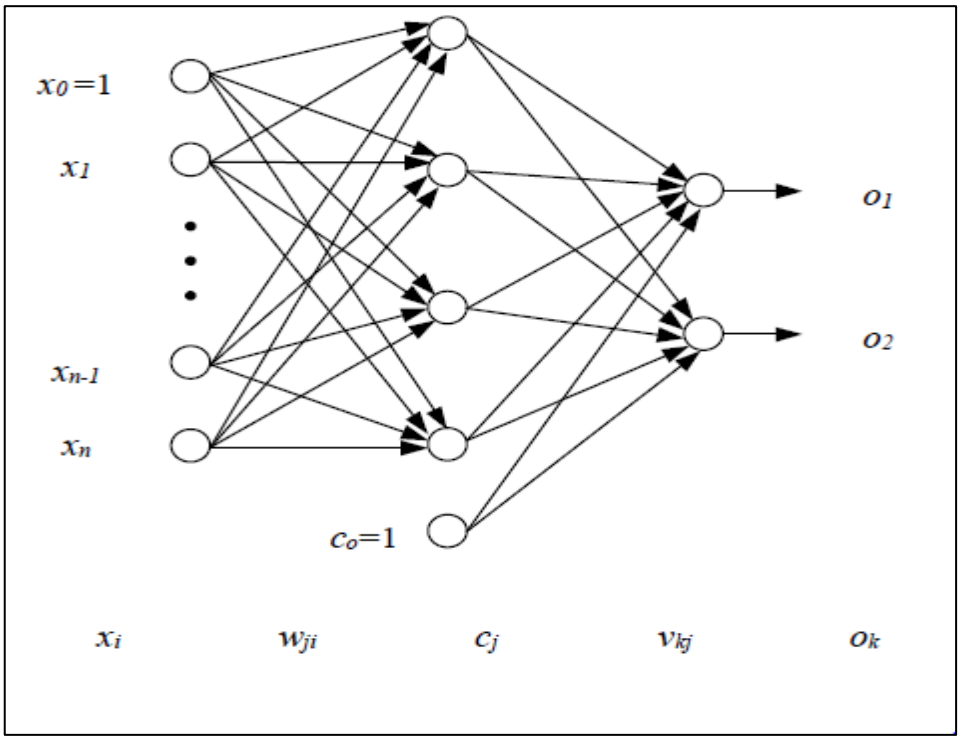

Figure II.2. Exemple d'algorithme de rétro propagation.

X_i ,i^e entrée externe du réseau; , n; = 1 : entrée du biais pour la couche cachée ,w_{ji} poids d'une connexion entre la i^e entrée et la j^e unité cachée,(x_i , $i^e v_{kj}$)poids d'une connexion entre la j^e unité cachée et la k^e unité de sortie.

Les indices i, j et k font référence aux unités d'entrée, aux unités cachées et aux unités de sortie, respectivement. L'exposant p correspond au numéro de l'exemple présenté à l'entrée du réseau : p = 1, . . .,n_A, où n_A est le nombre d'exemples d'apprentissage. Le p^e exemple est noté$x^p = [x_0^p, \dots, x_i^p, \dots, x_n^p]$et la i^e composante $x^p{}_i$ désigne la i^eentrée lorsque le p^e exemple est présenté au réseau. Les valeurs $x^p{}_i$ peuvent être binaires ou continues. Pour un exemple p, la j^e unité cachée a l'entrée résultante $I^p{}_j$:

$$I_j^p = \sum_{i=0}^{n} w_{ji}\, x_i^p \qquad (\text{II.}10)$$

et une activation:

$$c_j^p = h\left(I_j^p\right) = h\left(\sum_{i=0}^{n} w_{ji}\, x_i^p\right) \qquad (\text{II.}11)$$

où h est la fonction d'activation. La k^e unité de sortie reçoit une entrée résultante $I^p{}_k$ définie par :

$$I_k^p = \sum_{i=0}^{n} v_{kj}\, c_j^p \qquad (\text{II.}12)$$

et génère en sortie l'activation $o^p{}_k$

$$O_k^p = h(I_k^p) \qquad (\text{II.}13)$$

Pour l'application du MLP en régression, la fonction d'activation des neurones de sorties est linéaire. L'équation (II.14) devient :

$$O_k^p = I_k^p \qquad (\text{II.}14)$$

On prend une fonction d'activation non-linéaire h uniquement dans le cas de l'application en discrimination. Dans le cadre de cette mémoire, nous nous intéresserons uniquement à l'application de MLP en régression.

La fonction de coût usuelle est l'erreur quadratique moyenne définie comme :

$$E(w) = \frac{1}{2}\sum_{k,p}\left(t_k^p - o_k^p\right)^2 \qquad (II.15)$$

où **w** est le vecteur contenant tous les poids du réseau. La fonction E(**w**) est continue et différentiable par rapport à chaque poids. Pour déterminer les poids qui la minimisent, on

peut donc utiliser l'algorithme de descente du gradient. Pour faciliter la notation, E(**w**) sera notée E dans ce qui suit.

Pour les poids des connexions des unités cachées vers les unités de sortie, le terme d'adaptation des poids au cours de l'apprentissage est défini par :

$$\Delta V_{kj} = -\eta\left(\frac{\partial \mathrm{E}}{\partial \mathrm{V_{kj}}}\right)$$

$$= \eta \sum_{\mathrm{p}} \sigma_{\mathrm{k}}^{\mathrm{p}}\, c_{\mathrm{j}}^{\mathrm{p}} \quad \text{(II. 16)}$$

Avec

$$\sigma_k^p = \left(t_k^p - o_k^p\right)$$

Pour les poids des connexions entre la couche d'entrée et la couche cachée, le terme d'adaptation des poids est :

$$\Delta w_{ji} = -\eta\left(\frac{\boldsymbol{\partial E}}{\boldsymbol{\partial w_{ji}}}\right)$$

$$= -\eta\left(\frac{\boldsymbol{\partial E}}{\boldsymbol{\partial C_j^P}} \times \frac{\boldsymbol{\partial C_j^p}}{\boldsymbol{\partial w}}\right)$$

$$= \eta \sum_{\boldsymbol{k,p}} \left(\boldsymbol{t_k^p - o_k^p}\right) \boldsymbol{V_{kj}} \boldsymbol{h'}\left(\boldsymbol{I_j^p}\right) \boldsymbol{x_i^p}$$

$$= \eta \sum_{\boldsymbol{p}} \boldsymbol{\sigma_j^p}\, \boldsymbol{x_j^p} \quad \text{(II. 17)}$$

Avec

$$\sigma_j^p = h'\left(I_j^p\right) \Sigma_k V_{kj} \sigma_k^p \quad \text{(II. 18)}$$

On peut constater que les équations **(II.18)** et **(II.17)** ont la même forme et ne diffèrentque par la définition de la quantité **б** . Ces formules se généralisent facilement aux cas des réseaux possédant un nombre quelconque de couches cachées. D'après l'équation **(II.18)** le calcul de $\boldsymbol{б_j}$ pour une unité cachée j nécessite les $\boldsymbol{б_k}$ des unités de sortie, qui sont fonctions des erreurs$(t_k - o_k)$ en sortie du réseau Ainsi, pour corriger les poids des connexions entre la couche d'entrée et la couche cachée, on a besoin de rétro-propager l'erreur depuis les sorties vers les entrées, d'où le nom de l'algorithme d'apprentissage : rétro-propagation de 1'erreur[36].

II.5.Machine d'apprentissage extrême (ELM)

L'apprentissage mécanique et intelligence artificielle n'ont apparemment jamais été aussi critiques et importantes pour les applications réelles telles qu'elles sont dans l'ère autonome et de grande taille des données d'aujourd'hui. Le succès de l'apprentissage machine et de l'intelligence artificielle repose sur la coexistence de trois conditions nécessaires : environnements informatiques puissants, riches et / ou de grandes données et des techniques efficaces d'apprentissage (algorithmes). ELM comme une technique d'apprentissage émergente fournit une efficacité solutions unifiées aux réseaux avancés y compris mais sans s'y limiter (à la fois Multi-caché) réseaux de neurones, RBF[37].

II.5.1. Fonctionnement

Le principe du réseau neuronal n'est pas modifié, mais le rôle de l'adaptation est reconsidéré, ainsi que le nombre de couches cachées qui se restreint à une seule couche. Ainsi, plutôt que d'ajuster tous les poids d'un réseau pour émuler une fonction, le réseau est constitué d'un grand nombre de neurones dans la couche interne. Les poids d'entrée sont initialisés aléatoirement une seule fois et restent avec cette valeur. L'adaptation, qui se fait en une seule fois aussi, porte donc uniquement sur les poids de la couche de sortie [37].

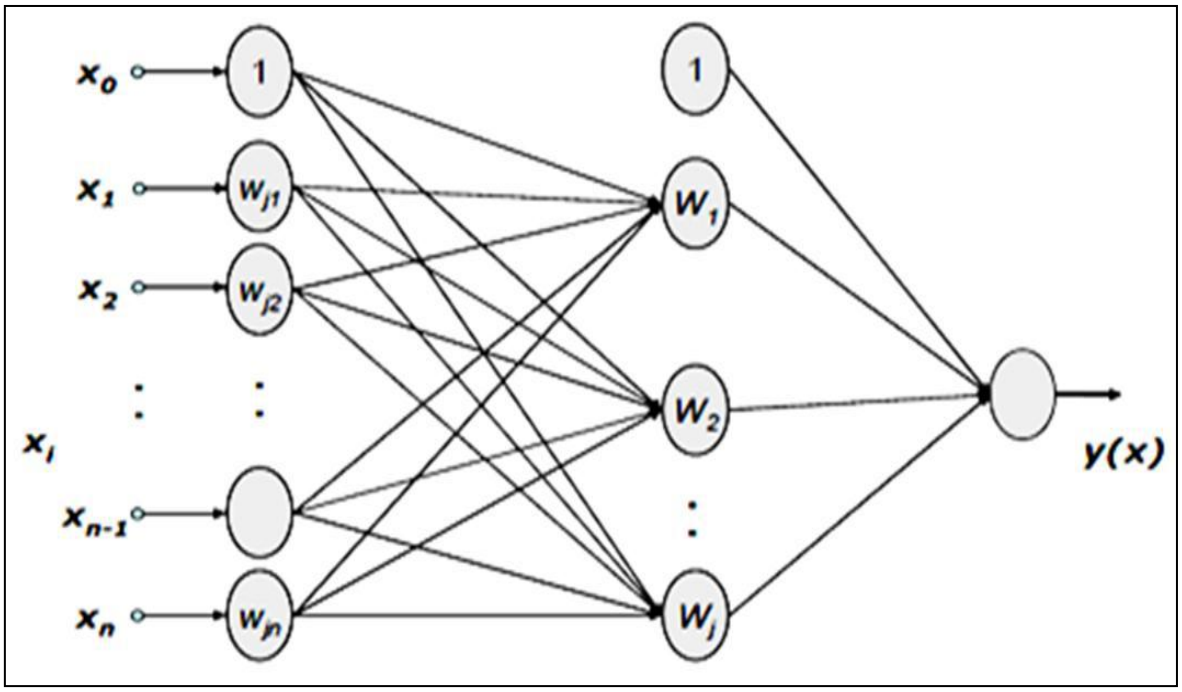

Figure II.3. Architecteur d'ELM.

II.5.2. Formulation mathématique de l'ELM

La fonction de sortie d'ELM est représentée dans l'équation suivant[38] :

$$y = \sum_{i=1}^{m} B_i f\left(w_i x_j + b_j\right), \quad j \in [1, n]. \qquad \text{(II. 19)}$$

où m : est le nombre de nœuds cachées et B est définie de la manière suivante:

$$B = [B_1 \dots \dots \dots \ B_m]$$

B est le vecteur des pondérations de sortie entre la couche cachée de n nœuds et le nœud de sortie, et x est le vecteur de sortie de la couche cachée **:**$\mathrm{X} = [\mathrm{X}_1 \dots \dots \dots \ \mathrm{X}_\mathrm{n}]$

F:Fonction d'activation

$$H\beta = Y, \beta = BT. \qquad \text{(II. 20)}$$

II.5.3. Caractéristiques principales

- La vitesse d'apprentissage de l'ELM est extrêmement rapide.
- Contrairement aux théories de l'existence conventionnelle, les paramètres du nœud caché ne sont pas seulement indépendants des données de formation mais aussi de l'autre. Bien que les nœuds cachés soient importants et critiques, ils n'ont pas besoin être réglé.
- Contrairement aux méthodes d'apprentissage conventionnelles qui doit voir les données de formation avant de générer les paramètres du nœud caché, ELM Pourrait générer les paramètres du nœud caché avant de voir la formation les données.
- Architectures homogènes pour la compression, l'apprentissage des fonctionnalités, regroupement, régression et classification [31].

II.5.4.Algorithme ELM

Dans l'algorithme ELM, contrairement à l'approximation de fonction traditionnelle théories qui doivent ajuster les poids d'entrée et cache les biais de couche pendant l'entraînement, les paramètres (les poids d'entrée $\mathbf{a_j}$ et les biais de couche cachés $\mathbf{b_j}$) des PNLs peuvent être aléatoire sas signé puis corrigé sans réglage itératif. Les poids de sortie βi (reliant la couchecachée et lesortie) des SLFN peuvent être déterminée s'analytiquement par la simple opération inverse généralisée du caché matrices de sortie de couche G après les valeurs des poids d'entrée, et les biais de couche cachés sont choisi s'arbitrairement. Le seul paramètre à prendre en compteest le nombre de masques nœuds L. Par conséquent, pour un ensemble de données R de N distinctif arbitraire échantillons d'apprentissage $R = \{(x_i, t_i) \mid \mathrm{i} = 1,2, \dots, \mathrm{N}\}. T \in R^n$

où $x_i = [x_{i1}, x_{i2}, \dots x_{in}]T \in R^n ett_i = [t_{i1}, t_{i2}, \dots, t_{in}]T \in R_m$. les sorties de une SLFN standard avec L nœuds cachés et fonction d'activation f(.) peut être décrit mathématiquement comme

$$O_i = \sum_{j=1}^{L} \beta_j f_j(x_i) = \sum_{j=1}^{L} \beta_j . f(a_j, x_i + b_j) \dots i = 1,2, \dots, N. \quad (II.21)$$

où o_i est le vecteur de sortie des SLFN par rapport à l'entrée $\mathbf{x}_i$, $\beta_j = [\beta_{i1}, \beta_{i2}, \dots, \beta_{im}]^T \in R^m$est le vecteur de poids reliant le j^{ième} nœud caché aux nœuds de sortie, $f(x_i; a_j, b_j)$est la fonction d'activation de la couche cache $a_j = [a_{j1}, a_{j2}, \dots, a_{jn}]^T \in R^n$est le vecteur de poids reliant le jth cache noeud et les nœuds d'entrée, bj est le seuil du j_{th} nœud caché, et $\mathbf{a_j} \cdot \mathbf{x_i}$ désigne le produit intérieur de $\mathbf{a_j}$ et $\mathbf{x_i}$.

Les N équations ci-dessus peuvent être écrites simplement sous la forme de matrice comme suit:

$$\boldsymbol{G\beta = O}$$

où

$$G(a_1, \dots, a_L, b_1, \dots, b_L, x_1, \dots, x_{N)}$$

$$= \begin{pmatrix} f(a_1.x_1 + b_1) & \cdots & f(a_l.x_1 + b_L) \\ \vdots & \ddots & \vdots \\ f(a_1.x_N + b_1) & \cdots & f(a_L.x_N + b_L) \end{pmatrix}$$

$$\beta = [\beta_1^T, \dots, \beta_L^T]_{L\times m}^T \; et \; O = [O_1^T, \dots, O_N^T]_{N\times m}^T$$

où G est la matrice de sortie de la couche cachée et la sixième colonne G est la sortie du nœud caché en ce qui concerne les entrées $\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_N}$.

Les SLFN standard avec L nœuds cachés et activation la fonction $f(x_i; a_j, b_j)$ a la capacité d'approximation universelle (les PNLF peuvent approcher ces N échantillons avec zéro erreur), ce qui signifie que la fonction de coût $E = \sum_{i=1}^{N} |O_i - T_j| = 0$, c'est-à-dire qu'il existe des $\boldsymbol{\beta_j}$, $\mathbf{a_j}$ et $\mathbf{b_j}$ spécifiques telsque

$$\sum_{j=1}^{L} \beta_j . f\left(a_j . x_i + b_j\right) = t_i, \dots i = 1,2, \dots, N. \qquad (II.22)$$

Cette équation peutêtresimplifiéecomme

$$G\beta = T$$

où T est la matrice cible des échantillons d'apprentissage. Différent des algorithmes d'apprentissage traditionnels basés sur des gradients avec des poids d'entrée fixes $\mathbf{a_j}$ et les biais de couche caches $\mathbf{b_j}$, les théories ELM affirment que les paramètres $\mathbf{a_j}$ et $\mathbf{b_j}$ peuvent être attribué au hasard. Ensuite, la question de la formation des PNLs se transforme en trouver une solution du moindre carré du linéaire système $\boldsymbol{G\beta = T}$:

$$\left\|\boldsymbol{G}(\boldsymbol{a_1}, \dots, \boldsymbol{a_L}, \boldsymbol{b_1}, \dots, \boldsymbol{b_L})\widehat{\boldsymbol{\beta}} - \boldsymbol{T}\right\|$$

$$= \boldsymbol{min_\beta} = \|\boldsymbol{G}(\boldsymbol{a_1}, \dots, \boldsymbol{a_L}, \boldsymbol{b_1}, \dots, \boldsymbol{b_L})\boldsymbol{\beta} - \boldsymbol{T}\|. \qquad (\mathbf{II.23})$$

La norme minimale € solution carrée **β** de la précédente système linéaire peut être calculé comme suit $\widehat{\beta} = H^{+}T$[39].

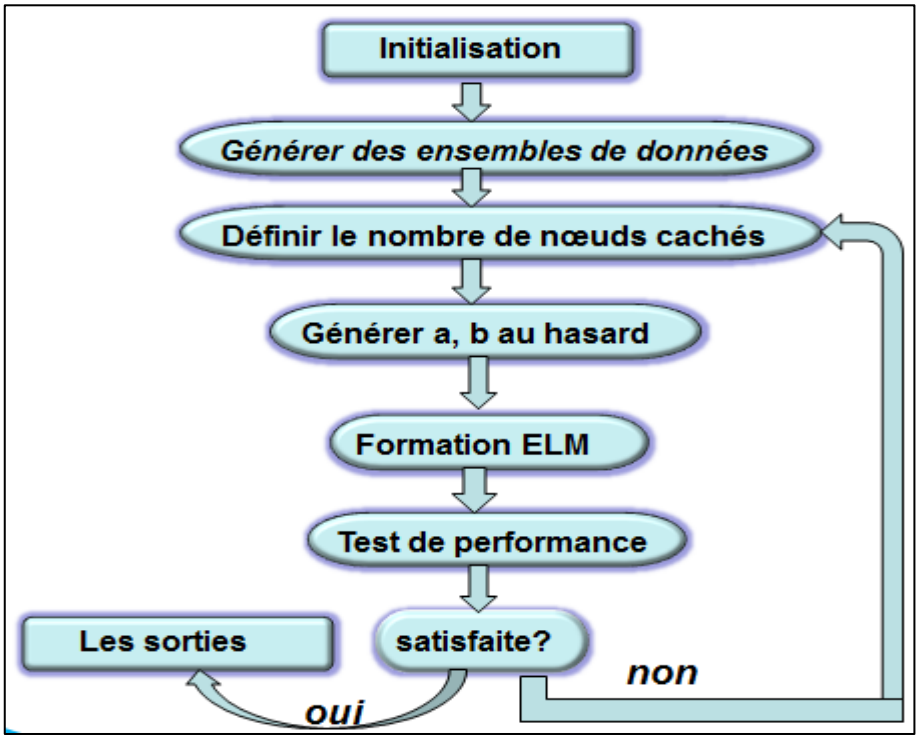

Figure II.4. Organigramme du modèle de machine d'apprentissage extrême (ELM)

Conclusion

Ce chapitre a fait l'objet de rappels des fondements des méthodes et techniques d'apprentissages statistiques appliqués à la classification et à la régression. En premier, nous avons présenté une généralité sur l'apprentissage automatique les RNAs avec des modèles des réseaux de neurones. Par la suite nous avons représenté une description sur les techniques d'apprentissage ELM, RBF et MLP.

Chapitre III

Simulation et Évaluation

Introduction

Dans un domaine d'application donné, la résolution d'un problème de régression et/ou de classification s'effectue en comparant des modèles afin de choisir le plus apte à résoudre le problème posé. L'évaluation des modèles est donc un préalable inévitable à la sélection. L'état de l'art proposé dans le chapitre précédent montre le nombre important d'approches, tant pour la classification que pour la régression.

Ce chapitre est consacré à la simulation et vise l'application des techniques étudiées comme étant une solution dans le développement des capteurs logiciels et de surveillance de la qualité de l'eau par reconnaissance des formes. L'objectif est de valider et d'évaluer les performances des techniques d'apprentissage et de sélection des caractéristiques dechacune des méthodes présentées à savoir LDA, PCA, MLP,RBF et ELM. 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. Afin de mener une étude comparative permettant un choix décisif de la méthode la mieux adaptée à l'application indiquée, on évaluera pour les méthodes exposées les paramètres liés au taux de reconnaissance et temps d'apprentissage. Une discussion des résultats conclura cette étude de simulation pour choisir la technique la mieux adaptée.

III.1. Système proposée

Il s'agit dans cette partie de travail d'évaluer les performances des cinq techniquesétudiées précédemment qui sont issues, rappelons-le, du domaine de l'intelligence artificielle à savoir, L'Analyse Discriminante Linéaire (LDA), L'analyse en composantes principales (PCA) pour la sélection des caractéristiques et les réseaux neuronal-MLP, les réseaux neuronal-RBF et les réseaux neuronal-ELM pour la classification. 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 brutes, venant directement des variables de surveillance, ou à partir de données traitées venant des sorties de traitements de bas niveau. Le choix effectué sur la base des résultats obtenus, conduira à l'intégration de la technique sélectionnée au niveau d'un système de surveillance assurant uncontrôle permanent de la qualité de l'eau.

Le schéma présenté dans la figure 3.1,présente l'architecture du système proposé de surveillance de la qualité de l'eau. Il se compose d'une entrée des données obtenues à partir de capteurs physico-chimiques, d'un système d'acquisition de données et de logiciels permettant le traitement des données et la prise de décision quant à la qualité de l'eau surveillée.

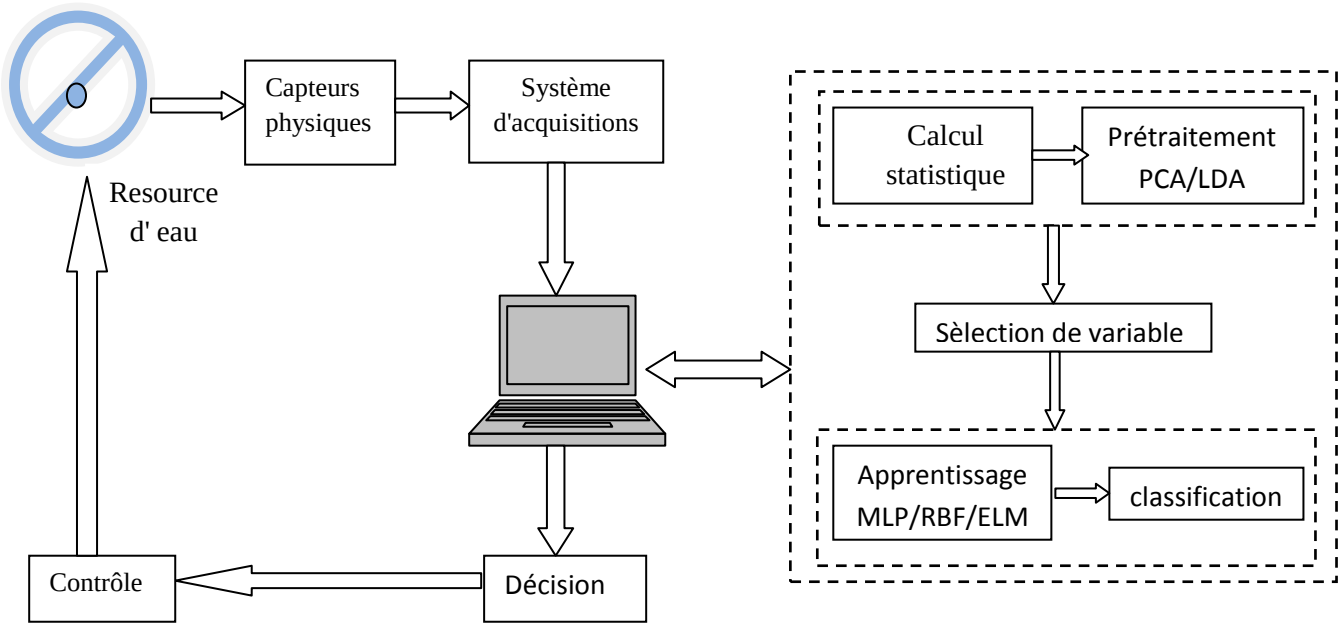

Figure III.1. L'architecture du système proposé de surveillance de la qualité de l'eau.

Le système à base de surveillance par reconnaissance de formes est fondé sur une approche multi-sensorielle (réseau de capteurs) capable de qualifier l'état de l'eau à contrôler. Les caractéristiques remarquables des données acquises dans les installations de traitement des eaux sont redondantes et possible mentin signifiantes, cela après la sélection des données. Les différents paramètres de scripteurs de l'eau devraient être transformés en signaux électriques à l'aide de capteurs qui permettent de capter et collecter les informations nécessaires et de transmettre les données ainsi recueillies. Ces informations récoltées sont acheminées grâce à des communications sans filversune station PC de contrôle et de traitement. Les étapes de traitement (apprentissage, classification, décision,…etc.) assurées par le système dans le but du contrôle et de surveillance du processus évoqué.

III.2. Description des données d'entrées

Nous cherchons à décider sur la qualité de l'eau à travers ses paramètres descripteurs.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 la qualité de cette eau à partir de quelques données descriptives. La base de données d'entrée constituée est composée d'un ensemble de 10 paramètres descripteurs de la qualité de l'eau brute qui sont,*température,conductivité,potentiel Hydrogène, Oxygène dissous, Matières en suspension, demande bio-Oxygène, Calcium, chlorures, magnésium, Bicarbonates.*

L'objectif qui se trouve derrière la collecte des données relatives à ces paramètres est de trouver un modèle de classification. La qualité de cette eau reflétée par sa potabilité repose en fait sur une corrélation qui ne peut être identifiée que statistiquement. Des données descriptives expérimentales recueillies sur une longue période (plusieurs années) pourraient atteindre cet objectif. Il y a donc intérêt de disposer d'au moins une année pour archiver des données afin de déterminer une base de connaissance assez complète capable de fonctionner normalement. D'où la nécessité d'une base de connaissance riche en informations exigeant d'abord une collecte des données sur une longue période, et la présence d'un expert. Dans ce travail, un total de 774échantillons est obtenu à partir de 10 variables de données sur la qualité de l'eau. Des statistiques descriptives simples de ces données d'entrée sont illustrées dans le tableau III.1. La figure III.2.présente l'évolution temporelle de ces paramètres de scripteurs.

Tableau III.1. Les paramètres statistiques d'eau brute.

Propriétés / Paramètres	Maximum	Minimum	Moyenne
T° (°C)	27.2	0.1	11.89755
Cà25°C(ms/cm)	176	140	147.616
pH	8.6	7.8	7.653093
OD	13	13.8	10.44333
OBD	19	6	22.88907
SM	4.9	4.5	3.280133
$Ca2^{+}$(mg/l)	25	17.9	17.41253
Cl^{-}(mg/l)	13.4	21.5	11.15723
$Mg2^{+}$(mg/l)	6.3	4.7	5.1284
$HCO3^{-}$(mg/l)	102	78	69.30667

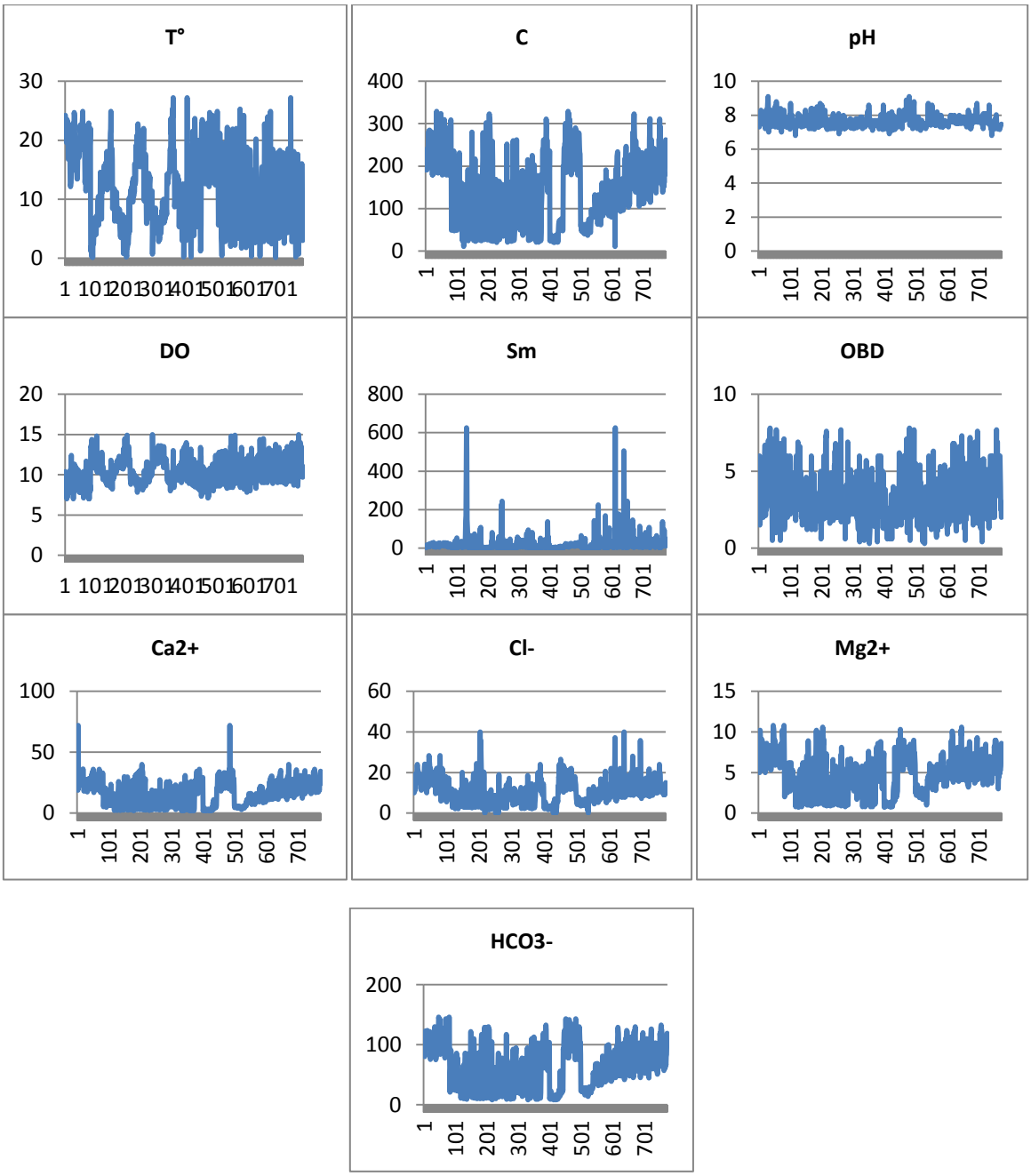

Figure III.2. Évolution temporelle de ces paramètres descripteurs.

III.3. Choix du modèle

Les méthodes de reconnaissance de formes telle que le MLP(PMC ou MLP en anglais),neuronal-RBF et ELM appliquées à la classification des données, présente l'avantage de couvrir un grand nombred'applications. Elles sont utilisées pour les systèmes de décision de hauts niveaux, etfondées 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 contrôle et de surveillance.

III.3.1.Résultat sans sélection des variables d'entrée

On présente d'abord un ensemble de données extrait d'une base réelle et constitué de 774 échantillons ou vecteurs, correspondant aux dix paramètres physico-chimiques (pH, T°, C,...) L'ensemble des échantillons est séparé en deux, à savoir, 500 échantillons utilisés pour l'apprentissage, le reste (274 échantillons) est utilisé pour test.

a. Perceptron Multicouches

Différentes architectures sont testées pour déterminer le nombre adéquat de neurones dans la couche cachée du modèle Neuronal. Le tableau 3.2 montre pour plusieurs architectures de réseaux testés, les résultats correspondants aux différents paramètres d'apprentissage, tels que : le nombre de neurones dans la couche caché (NNCC), temps / Taux d'apprentissage et de test.

Tableau III.2.Résultats d'apprentissage et test (Modèle neuronal-MLP).

Paramètres		**Apprentissage**		**Test**	
MLP	**NNCC**	**t_app**	**T_app**	**t_test**	**T_test**
	8_4_1	65.6719	99	0	99
	12_8_4_1	139.7813	87,6	0.0313	87
	16_12_8_4_1	468.6250	**100**	0.0313	**99**

On remarque que quand il y a associations le nombre de neurones dans les couches cachés, le taux de classification augmenté et le temps d'apprentissage augmenté.

b. Radial Basis Function (RBF)

Différentes architectures sont testées pour déterminer le nombre adéquat de neurones dans la couche cachée du modèle Neuronal-RBF. Les résultats d'entraînement utilisant la base de données estprésentée dans le tableau III.3.

Tableau III.3. Résultats d'apprentissage et test (Modèle neuronal-RBF)

Paramètres		**Apprentissage**		**Test**	
	NNCC	**t_app**	**T_app**	**t_test**	**T_test**
RBF	**100**	8.1406	74	0.0313	100
	200	20.2188	74	0.0156	100
	500	105.2500	**100**	0.0313	**99**
	1000	105.3906	100	0.0469	99
	2000	107.9531	100	0.0469	99

Quand il y a augmenté le nombre de neurones dans les couches cachés le taux de classification est augmenté aussi que le temps d'apprentissage.

c. Machine d'apprentissage extrême (ELM)

Les machines à apprentissage extrême sont des réseaux neuronaux feed-forward pour la classification ou la régression avec une seule couche de nœuds cachés, où les poids reliant les entrées aux nœuds cachés sont répartis au hasard et ne sont jamais mis à jour. Les poids entre les nœuds cachés et les sorties sont appris en une seule étape, ce qui revient essentiellement à l'apprentissage d'un modèle linéaire.

On a déterminé les paramètres tels que NNCC, temps /Taux d'apprentissage et de test correspondant aux ELM; comme ce qui est montré dans le tableau.III.4.

Tableau.III.4. Résultats d'apprentissage et test (Modèle neuronal-ELM)

Paramètres		**Apprentissage**		**Test**	
	NNCC	**t_app**	**T_app**	**t_test**	**T_test**
ELM	**200**	1.4375	98,2	0.0313	99
	300	3.6094	99,2	0	99
	400	4.1406	**100**	0.0625	**99**
	500	4.1406	100	0.0313	99

Quand il y a augmenté le nombre de neurones dans les couches cachés on remarque des meilleurs résultats en termes de taux de reconnaissance et le temps est augmenté dans l'apprentissage, dans le test on remarque que le taux est similaire.

d. Évaluation des performances

Pour une évaluation des performances de trois modèles précédemment testés, définissons quelques mesures statistiques telles que le temps et le taux d'apprentissage, et de test, qui sont présentés dans le tableau III.5 dans le but de fournir une comparaison entre les résultats de trois techniques d'apprentissage.

Tableau III.5.Évaluation des performances

Paramètres		**Apprentissage**		**Test**	
	NNCC	**t_app**	**T_app**	**t_test**	**T_test**
MLP	**16_12_8_4_1**	468.6250	**100**	0.0313	**98,9051095**
RBF	**500**	105.2500	**100**	0.0313	**98,9051095**
ELM	**400**	4.1406	**100**	0.0625	**98,540146**

On remarque que les taux de classification presque similaires mais l'apprentissage avec la technique ELM est très rapide qu'avec les autres modèles (ELM plus rapide 25 fois que RBF et 113 fois que MLP).Donc, nous concluons que les réseaux des neurones de type ELM est meilleur et plus efficace dans cette application.

III.3.2.Résultats avec sélection des variables

On présente d'abord le même ensemble de données extrait d'une base réelle et constitué de 774 échantillons ou vecteurs, mais correspondant aux quelque paramètres physico-chimiques (après la sélection des variables – moins de paramètres d'entrée), L'ensemble des échantillons est séparé en deux, à savoir, 500 échantillons utilisés pour l'apprentissage, le reste (274 échantillons) est utilisé pour test.

III.3.2.1. L'Analyse Discriminante Linéaire (LDA)

Le prétraitement des données en utilisant l'Analyse Discriminante Linéaire consiste à recueillir les différentes informations suivantes : la matrice de corrélations,les valeurs propres et le cercle de corrélation.

→ La matrice de corrélations (voir tableau III.6).

Dans le tableau III.6, ont été reportées les corrélations entre variables pour les 3 premières facteurs. On peut observer qu'il existe une forte corrélation entre des variables de deux groupes différents. Un groupe de 5 variables (C,Ca,Cl, Mg et HCO) et l'autre

variable Sm.la température, Le *pH*, le DO,DBO ne sont pas corrélés avec d'autres variables. La dose de coagulant est une variable passive.

Tableau III.6. Corrélations entre variables pour les 3 premières facteurs.

Variable	Fact1	Fact2	Fact3
T	0,1981838	0,18763587	0,16028538
C	**0,98445721**	**0,99714819**	0,56906985
PH	0,41935477	0,41788216	0,25532195
DO	0,04196348	0,020514	0,10180889
Sm	0,19092616	0,01588787	**0,83742666**
DBO	0,45568871	0,40982339	0,44433272
Ca	**0,91844841**	**0,91396313**	0,57918354
Cl	**0,83703633**	**0,86462044**	0,41859825
Mg	**0,88240262**	**0,8937057**	0,4993232
HCO	**0,95337788**	**0,9571706**	0,57102577

→ **Les valeurs propres**

L'Analyse Discriminante Linéaire appliquée sur l'ensemble de cesdonnées a fourni le tableau. L'ensemble des 10 variables estsusceptible d'être simplifié et remplacé par les 3 nouvelles variablesreprésentées par les 3 premiers axes principaux.

Tableau III.7. Valeurs propres

	valeur_propre	valeur_percent	valeur_Cum
1	**951,12064**	**99,87904**	**99,87904**
2	**1,040171**	**0,10923**	**99,98827**
3	**0,0899583**	**0,009447**	**99,99772**
4	0,0217449	0,002283	100
5	2,28E-13	2,39E-14	100
6	1,23E-13	1,29E-14	100
7	2,59E-14	2,72E-15	100
8	3,50E-15	3,67E-16	100
9	8,28E-15	8,69E-16	100
10	8,28E-15	8,69E-16	100

a.Modèle Perceptron multi couches (MLP)

On applique la technique d'Analyse Discriminante Linéaire sur le modèle neurone MLP, déterminer quelques mesures statistiques et on observe les résultats dans le tableau III.8.

Tableau. III.8.Résultats d'apprentissage et test (Modèle neuronal-MLP) sélection par LDA

Paramètres		**Apprentissage**		**Test**	
	NNCC	**t_app**	**T_app**	**t_test**	**T_test**
MLP	**8_4_1**	27,24003	96,8	0,012853	93
	12_8_4_1	49,19444	**98,2**	0,014456	**96**
	16_12_8_4_1	151,9606	96,2	0,016125	94

On observe que dans le nombre des neurones des couches cachées(12_8_4_1)donne un meilleur résultat en termes de taux de reconnaissance dans l'apprentissage et le test.

b. Radial Basis Function (RBF)

On applique la technique d'Analyse Discriminante Linéaire sur le modèle neurone RBF, déterminer le temps et le taux d'apprentissage et de test et on observe les résultats dans le tableau III.9.

Tableau.III.9. Résultats d'apprentissage et test (Modèle neuronal-RBF) sélection par LDA

Paramètres		**Apprentissage**		**Test**	
	NNCC	**t_app**	**T_app**	**t_test**	**T_test**
RBF	**100**	8,296875	74	0,046875	100
	200	15,79688	76,8	0,046875	59
	500	98,0625	**100**	0,078125	**99**
	1000	98,8125	100	0,09375	99
	2000	99,25	100	0,078125	99

Quand il y a augmenté le nombre de neurones dans les couches cachés le taux de classification est augmenté, à partir de 500 neurones dans les couches cachés et plus, le taux reste similaire.

c. Machine d'apprentissage extrême (ELM)

On applique la technique d'Analyse Discriminante Linéaire sur le modèle ELM, déterminer quelques mesures statistiques et on observe les résultats dans le tableau 3.10.

Tableau. III.10.Résultats d'apprentissage et test (Modèle neuronal-ELM) sélection par LDA

Paramètres		**Apprentissage**		**Test**	
	NNCC	**t_app**	**T_app**	**t_test**	**T_test**
ELM	**200**	1.8125	96,4	0.0312	95
	300	2.5315	97,2	0.0312	95
	400	3.6093	**98**	0.0312	**96**
	500	4.6406	97,2	0.0313	95

On observe que dans le nombre des neurones des couches cachées(400) donne un meilleur résultat en termes de taux de reconnaissance dans l'apprentissage et le test.

d. Évaluation des performances

Pour une évaluation des performances des trois modèles précédemment testés en utilisant la technique de sélection LDA, définissons quelques mesures statistiques telles que le temps et le taux d'apprentissage, et de test, qui sont présentés dans le tableau III.11 dans le but de fournir une comparaison entre les résultats de trois techniques d'apprentissage.

Tableau III.11.Évaluation des performances

Paramètres		**Apprentissage**		**Test**	
	NNCC	**t_app**	**T_app**	**t_test**	**T_test**
MLP	**12_8_4_1**	49,19444	**98,2**	0,014456	**96**
RBF	**500**	98,0625	**100**	0,078125	**99**
ELM	**400**	3.6093	**98**	0.0312	**96**

On remarque une amélioration positive de taux de reconnaissance quand y a uneassociation des neurones ou des couches cachées donne un meilleur résultat en termes detaux de reconnaissance. Le réseau de RBF présenté des meilleurs résultats en termes detaux de reconnaissance mais le temps d'apprentissage de technique ELM est très rapide (ELM plus rapide 27 fois que RBF et 13 fois que MLP).

III.3.2.2. L'analyse en composantes principales(PCA)

La technique ACP est utilisé pour sélectionner les variables qui ont le plus d'influence sur le capteur logiciel en utilisant la matrice de corrélation, ses valeurs propres et leur histogramme. En outre, nous pouvons comprendre qu'il y a un changement de caractéristiques de données à des composants qui ne sont pas corrélés. A souligner toutefois, que l'ensemble des 10 variables d'entrée de cette base de données est retenu en raison de l'importance de ces paramètres pour la qualité de l'eau et la continuité de leurs mesures dans le temps. L'Analyse par ACP appliquée à l'ensemble des données de la base.Après la sélection des variables en utilisant la technique PCA, les paramètres C, pH, OD et MES sont considérés comme des variables d'entrée pour la classification. Afin de comparer l'efficacité des modèles : MLP, RBF et ELM en termes de temps d'entraînement et d'autres paramètres, l'ensemble de données a été subdivisé en quatre bases.

a. Modèle Perceptron multi couches (MLP)

On applique la technique d'Analyse en Composant Principale sur le modèle neurone MLP, déterminer quelques mesures statistiques et on observe les résultats dans le tableau III.12.

Tableau.III.12. Résultats d'apprentissage et test (Modèle neuronal-MLP) sélection par PCA

Paramètres		**Apprentissage**		**Test**	
	NNCC	**t_app**	**T_app**	**t_test**	**T_test**
MLP	**8_4_1**	50.2188	**97,8**	0.0625	**96**
	12_8_4_1	100.0156	87,8	0.0313	90
	16_12_8_4_1	314.3594	82,2	0.0313	88

On remarque que dans le nombre des neurones des couches cachées(8_4_1) donnent un meilleur résultat en termes de taux de reconnaissance dans l'apprentissage et le test.

b. Radial Basis Function (RBF)

On applique la technique d'Analyse en Composant Principale sur le modèle neurone RBF, déterminer le temps et le taux d'apprentissage et de test et on observe les résultats dans le tableau III.13.

Tableau.III.13.Résultats d'apprentissage et test (Modèle neuronal-RBF) sélection par PCA

Paramètres		**Apprentissage**		**Test**	
	NNCC	**t_app**	**T_app**	**t_test**	**T_test**
RBF	**100**	8.0313	84,2	0.0313	98
	200	18.9844	92,2	0.0469	93
	500	101.7344	**100**	0.0469	**99**
	1000	102.0625	100	0.0469	99
	2000	102.6094	100	0.0469	99

Quand il y a augmenté le nombre de neurones dans les couches cachés le taux de classification est augmenté, à partir de 500 neurones dans les couches cachés et plus, le taux reste similaire.

c. Machine d'apprentissage extrême (ELM)

On applique la technique d'Analyse en Composant Principalesur le modèle ELM, déterminer quelques mesures statistiques et on observe les résultats dans le tableau III.14.

Tableau.III.14. Résultats d'apprentissage et test (Modèle neuronal-ELM) sélection par PCA

Paramètres		**Apprentissage**		**Test**	
	NNCC	**t_app**	**T_app**	**t_test**	**T_test**
ELM	**200**	1.8593	95,6	0.0312	98
	300	2.7500	99	0.0313	98
	400	3.5781	99,8	0.0313	98
	500	4.250	**100**	0.0313	**99**

On remarque que quand il y a augmenté le nombre de neurones dans les couches cachés le taux de classification est augmenté ainsi que le temps d'apprentissage.

d. Évaluation des performances

Pour une évaluation des performances des trois modèles précédemment testés en utilisant la technique de sélection PCA, définissons quelques mesures statistiques telles que le temps et le taux d'apprentissage, et de test, qui sont présentés dans le tableau III.15 dans le but de fournir une comparaison entre les résultats de trois techniques d'apprentissage.

Tableau III.15. Évaluation des performances

Paramètres		**Apprentissage**		**Test**	
	NNCC	**t_app**	**T_app**	**t_test**	**T_test**
MLP	**8_4_1**	50.2188	**97,8**	0.0625	**99**
RBF	**500**	101.7344	**100**	0.0469	**99**
ELM	**500**	4.250	**100**	0.0313	**99**

Le réseau de ELM présenté des meilleurs résultats en termes de taux de reconnaissance et le temps d'apprentissage qu'avec les autres modèles (ELM plus rapide 24 fois que RBF et 11 fois que MLP).et le taux de test est similaire dans tous les modèles.

III.3.3.Discutions des résultats

Après la simulation de différentes techniques d'apprentissage étudié sans sélection de variable et avec la sélection on peut comparer entre les résultats de cette étude.

Tableau.III.16. Résultats d'apprentissage et test (tous les modèles)

Modèles		**Paramètres**		**Apprentissage**		**Test**	
	Number des variables	**Technique**	**NNCC**	**t_app**	**T_app**	**t_test**	**T_test**
Sans sélection de variables	**10 variables**	**MLP**	16_12_8_4_1	468.6250	**100**	0.0313	**99**
		RBF	500	105.2500	**100**	0.0313	**99**
		ELM	400	4.1406	**100**	0.0625	**99**
Avec sélection de variables (LDA)	**03 variables**	**MLP**	12_8_4_1	49,19444	**98,2**	0,014456	**99**
		RBF	500	98,0625	**100**	0,078125	**99**
		ELM	400	3.6093	**98**	0.0312	**96**
avec sélection de variables (PCA)	**04 variables**	**MLP**	8_4_1	50.2188	**97,8**	0.0625	**96**
		RBF	500	101.7344	**100**	0.0469	**99**
		ELM	500	4.250	**100**	0.0313	**99**

- Pour la comparaison des résultats avec sélection de variable on remarque que le réseau ELM présente de meilleurs résultats en termes de taux de reconnaissance et de temps d'apprentissage en utilisant la technique de sélection LDA.

- La technique ELM donne demeilleurs résultatsavec ou sans sélection de données en termes detaux de reconnaissance et le temps d'apprentissage ;mai**s** on remarque que le meilleur résultat après la sélection en termes de temps d'exécution.
- Le prétraitement de données au sens de sélection des caractéristiques est très important pour la classification de données dehaute dimension.

Conclusion

Ce dernier chapitre a fait l'objet d'une étude en simulation concernant la mise en œuvre de trois techniques d'apprentissage statistique appliquées dans le domaine du contrôle et de surveillance des eaux potables. 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 et au temps d'apprentissage, 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 ELM retenue avec sélection de données pour ses qualités et avantages adaptés au problème posé.

CONCLUSION GENERALE

CONCLUSION GENERALE

Le travail présenté dans ce mémoire a été consacré à la mise en œuvre de trois techniques d'apprentissage statistique appliquées à la reconnaissance de formes dans le domaine du contrôle et de surveillance des eaux potables. 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 la qualité des eaux propres. 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.

Cette étude a été structurée autour de trois chapitres essentiels. Le premier consacré à une introduction au domaine de sélection des caractéristiques a permis de présenter des généralités ainsi que les différentes méthodes de sélection de données. Dans le second a été particulièrement dédié aux mécanismes théoriques des méthodes de classification de données à apprentissage statistique supervisé. Dans ce chapitre, trois modèles (MLP, Neuronal-RBF et ELM) fondés sur ce type d'apprentissage ont été exposés. Enfin le trosième 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 contrôle et de surveillance des eaux potables ainsi que les méthodes de sélection des caractéristiques. 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 et au temps d'apprentissage, 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 ELM retenue pour ses qualités et avantages adaptés au problème posé. Plusieurs scénarios ont été alors effectués.

D'après les résultats obtenus, il apparaît que sur le plan décisionnel, les trois modèles ont présenté de bons résultats pour une simulation sur des données réelles Les résultats obtenus ont montré que ces trois modèles sont rapides, ce qui leur confère l'avantage de s'intégrer dans un système de surveillance dynamique. Cependant, les modèles MLP et RBF ont montré particulièrement une invalidité majeure liée au temps d'entraînement. Un désavantage qui a été heureusement levé par le modèle ELM, choisi à cet effet pour sa rapidité en cette phase d'apprentissage et bien adapté à l'apprentissage de grandes bases de données, ce modèle a

présenté en fait de nombreux avantages, que ce soit en termes de classification. L'usage et l'application de celui-ci pourrait avoir un impact direct touchant aussi bien les aspects de conception d'environnement et d'économie. L'usage d'un prétraitement des données dans un but de réduction de dimensionnalité, a confirmé davantage cet intérêt.

Les performances ainsi obtenues peuvent être alors améliorées. En effet, une base de données réelle plus importante et plus significative, contribue sans doute à augmenter la précision de reconnaissance. Toutefois, 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. Le contrôle de potabilité peut par contre être pris en charge de façon dynamique par le système de surveillance multicapteur, puisque le temps d'exécution est faible.

Les horizons de cette application restent prometteurs. La décision du système peut être améliorée par l'exploitation de nouveaux paramètres d'entrée. Les capteurs logiciels peuvent dans ce cas jouer un rôle primordial en se substituant davantage à des paramètres descripteurs chimiques ne pouvant être mesurés en continu. 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 publique. Enfin, cette application montre une alternative prometteuse pour notre pays dans l'avenir, pour une surveillance intelligente, automatique et efficace de la qualité des eaux potables.

REFERENCES

[1] [http://dspace.univ-tlemcen.dz/bitstream/112/1045/4/Memoire.pdf].

[2] **DJERIOUI Mohamed** "Contribution au Développement de Systèmes Multi capteurs Intelligents Dédiés à la Surveillance et au Contrôle de la Qualité des Eaux Propres." Thèse de Doctorat Science, Université - M'SILA 2018/2019.

[3] Sélection de variables pour la classification non supervisée en grande dimension Caroline Meynet , Synthèse.

[4] **Sébastien GUÉRIF,** "Réduction de dimension en Apprentissage Numérique Non Supervisé", Docteur de l'Université Paris 13, *11 décembre 2006*.

[5] **H.Liu** and **H.Motoda**. Feature Selection for Knowledge Discovery and Data Mining.Kluwer Academic Publishers, 1998.

[6] **Cakmakov** and **Younes Bennani**. Feature Selection for Pattern Recognition. Informa Press, Ed., 2002.

[7] **I. Guyon, S. Gunn, M. Nikravesh,** and **L. Zadeh**. Feature Extraction, Foundations and Applications, Editors. Series Studies in Fuzziness and Soft Computing, Physica-Verlag. Springer, 2006, to appear.

[8] **Ronald A. Fisher**. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7 :179–188, 1936.

[9**] Charles Bouveyron.** "Modélisation et classification des données de grande dimension : application à l'analyse d'images.. Mathématiques [math]." Université Joseph-Fourier - Grenoble I, 2006. Français.

[10] **C. Bishop** and **M. Tipping**. A hierarchical latent variable model for data visualization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3) :281–293, 1998.]IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3) :281–293, 1998.

[11] **Charles Bouveyron, Stéphane Girard, Cordelia Schmid**. "Analyse Discriminante de Haute Dimension".[Rapport de recherche] RR-5470, INRIA. 2005, pp.46.

[12] **K. Etemad, R. Chellappa**, "Discriminant Analysis for Recognition of Human Face images", Journal of the Optical Society of America A, Vol. 14, No. 8, August 1997, pp. 1724-1733.

[13] **MERAMRIA Nabila**, "Reconnaissance de visages par Analyse Discriminante Linéaire(LDA)" Université de ANNABA, M aster 2016.

[14] **NASRI Mustapha**, "Transformation non linière KPCA et KLDA pour l'authentification de visages" Présenté pour l'obtention du diplôme de Magister Université M'SILA 17 /12/ 2013.

[15] [Sabrina Tollari. Indexation et recherche d'images par fusion d'informations textuelles et visuelles. Thèse préparée dans le laboratoire LSIS - UMR CNRS 6168, octobre 2006.].

[16] Guermoudi Mohammed el Amine.Fekih Mohammed el Amine, devant la commission composé de MM Fusion des classifieurs supervisés: Application sur la classification pixellaire des images microscopiques.master Présenté le 01 Juillet 2013].

[17] (Hardle and Simar (2007)).

[18] (Hastie et al. (1994)).

[19] (Clemmensen et al. (2011)).

[20] **Emeline Perthame**. "Stabilité de la sélection de variables pour la régression et la classification de données corrélées en grande dimension. Statistiques [math.ST]." Université Rennes 1, 2015. Français.

[21] **Nicolas Morizet, Thomas EA**, **Florence Rossant**, **Frédéric Amiel**, **Amara Amara** , "Revue des algorithmes PCA, LDA et EBGM utilisés en reconnaissance 2D du visage pour la biométrie" , Institut Supérieur d'Electronique de Paris (ISEP), Département d' Electronique 21, rue d'Assas 75270 Paris Cedex 06.

[22] **Alaa Tharwat** , **Tarek Gaber** , **Abdelhameed Ibrahim** , and **Aboul Ella Hassanien** , "Linear discriminant analysis: A detailed tutorial", AI Communications 30 (2017) 169–190 169 DOI 10.3233/AIC-170729 IOS Press.

[23] (Shum, *et al.*, 2011)

[24] **Mohammed Senoussaoui,** "Amélioration de la robustesse des systèmes de reconnaissance automatique du locuteur dans l'espace des I-vecteur"**,** École de technologie supérieure université du Québec **,** le 10 Juin 2014.

[25] **S. BAZI**, "Contribution à la Détection et au Diagnostic des Défauts dans un Système Machine à Induction-Convertisseur", Thèse Doctorat en Sciences, Univ de Batna 2, 2016.

[26] https://fr.wikipedia.org/wiki/Apprentissage_automatique.

[27] https://fr.wikipedia.org/wiki/R%C3%A9seau_de_neurones_artificiels.

[28] **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'sila2013.

[29] **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.

[30] **Mr.Ali DJAIDJA** ''Etude de la classification supervisée des données environnementales à l'aide de réseaux de neurones de fonctions à base radiales.'' Mémoire de master, Université - M'SILA, Juin 2016.

[31] **Hadj Kaddour Aissa ,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, Université - M'SILA, Mai 2017.

[32] https://hal.archives-ouvertes.fr/hal-00085092 Submitted on 11 Jul 2006.

[33] **THIAW .L «**Identification de systèmes dynamiques non-linéaires par réseaux de neurones et multimodales », thèse doctorat, Université Paris, 2008.

[34] **LAROUCHE .E** «exploration de différentes architectures de réseau de neurone pour la prédiction de la glace atmosphérique sur les conducteurs des réseaux de neurone » mémoire master, Université de Québec, 2002.

[35] **DJERARDA Souheyla,HEDLI Khaoula**''Traitement automatique des eaux basées sur les techniques d'apprentissage statistique.'' diplôme de Master, Université - M'SILA, Année universitaire :2017 /201om8.

[36] **N. VALENTIN** «Construction d'un capteur logiciel pour le contrôle automatique du procédé de coagulation en traitement d'eau potable» thèse de doctorat 2000.

[37] **S. BAZI**, "Contribution à la Détection et au Diagnostic des Défauts dans un Système Machine à Induction-Convertisseur", Thèse Doctorat en Sciences, Univ de Batna 2, 2016.

[38] **S. Sojasi**, "Caractérisation de minéraux indicateurs par imagerie hyperspectrale et traitement de l'image dans l'infrarouge proche et l'infrarouge lointain", Thèse de doctorat, Univ de Laval, Québe , Canada, 2016.

[39] Concrète dam déformation prédiction model for Heath monitoring based on extreme learning machine DOI: 10.1002/stc.1997.

Résumé : La classification des données, ont motivé le développement de méthodes statistiques pour la sélection de variables. La préparation de la base de données dans des applications de développement des outils de supervision/diagnostic à l'aide des modèles basés sur les techniques d'apprentissage statistiques consiste à retenir les variables les plus représentatives des données observées. L'utilisation de ces techniques augmente dans l'industrie de production puisqu'ils permettant le développement de robustes modèles non-linéaires d'unités de procédés industriels complexes, ces données contiennent un très grand nombre de variables. Cette association d'une dimensionnalité élevée à une petite taille d'échantillon fait de la sélection de variables une étape préalable indispensable pour diminuer les temps de calcul, et améliorer l'interopérabilité des modèles. Dans le cadre de ce travail de mémoire, nous avons tout d'abord cherché à déterminer quels sont les facteurs, au niveau des données, qui influencent le plus la stabilité de la sélection, sur tout de type de données. Nous avons ensuite travaillé sur des méthodes de classification, en nous focalisant tout d'abord sur l'influence de la stabilité de la sélection sur les ces méthodes étudiées.

Abstract: The classification of the data motivated the development of statistical methods for the selection of variables. The preparation of the database in development applications of supervisory / diagnostic tools using models based on statistical learning techniques consists in retaining the most representative variables of the observed data. The use of these techniques increases in the production industry since they allow the development of robust nonlinear models of complex industrial process units, these data contain a very large number of variables. This combination of high dimensionality and small sample size makes variable selection an essential precondition for reducing computation time and improving interoperability of models. As part of this memory work, we first sought to determine which factors at the data level most influence the stability of the selection on any type of data. We then worked on classification methods, focusing first of all on the influence of the stability of the selection on these studied methods.

ملخص :حفز تصنيف البيانات على تطوير أساليب إحصائية لاختيار المتغيرات. إن إعداد قاعدة البيانات في تطبيقات تطوير الأدوات الإشرافية / التشخيصية باستخدام النماذج المعتمدة على تقنيات التعلم الإحصائي ، يتمثل في الاحتفاظ بالمتغيرات الأكثر تمثيلا للبيانات المرصودة. يزداد استخدام هذه التقنيات في الإنتاج لأنها تسمح بتطوير نماذج غير خطية قوية لوحدات العمليات الصناعية المعقدة ، وتحتوي هذه البيانات على عدد كبير جدًا من المتغيرات. هذا المزيج من الأبعاد العالية وصغر حجم العينة يجعل اختيار المتغير شرطا مسبقا ضروريا لتقليل وقت الحساب وتحسين قابلية التشغيل البيني للنماذج. كجزء من عمل الذاكرة هذا ، سعينا أولاً لتحديد العوامل التي تؤثر بدرجة أكبر على استقرار التحديد على أي نوع من البيانات على مستوى البيانات. ثم عملنا على أساليب التصنيف ، مع التركيز أولاً على تأثير استقرار الاختيار على هذه الأساليب المدروسة.