N° d'ordre: 94

Année: 2014

# UNIVERSITE ABDELMALEK ESSAADI
# ECOLE NATIONALE DES SCIENCES APPLIQUEES
# TANGER

Centre d'Etudes Doctorales: «Sciences et Techniques de l'Ingénieur»
Formation Doctorale: «Sciences et Techniques de l'Ingénieur»

## THESE DE DOCTORAT

Présentée
Pour l'obtention du

**DOCTORAT EN SCIENCES ET TECHNIQUES DE L'INGENIEUR**

Par :

**CHAIRI IKRAM**

Discipline : Statistique et Informatique Décisionnelle
Spécialité : Apprentissage automatique

Titre : **Sélection des Echantillons pour le Problème de classification en Distributions déséquilibrées**

**Soutenue le 25 Juillet 2014 devant le Jury :**

| | | |
|---|---|---|
| ***Pr. Fouad LAHJOMRI*** | *Ecole National des Sciences Appliquées - Tanger* | *Président* |
| ***Pr. Lotfi CHRAIBI*** | *Ecole National des Sciences Appliquées - Tanger* | *Rapporteur* |
| ***Pr. Aziz ARBAI*** | *Faculté des Sciences -, Tétouan,* | *Rapporteur* |
| ***Pr. Abdeïhadi AKHARIF*** | *Faculté des Sciences et Techniques - Tanger* | *Rapporteur* |
| ***Pr. Noufal RAISSOUNI*** | *Ecole National des Sciences Appliquées - Tétouan* | *Examinateur* |
| ***Pr. Abdelouahid LYHYAOUI*** | *Ecole National des Sciences Appliquées - Tanger* | *Directeur de thèse* |

Laboratoire des Technologies Innovantes

*A la plus noble et gracieuse des femmes,*

*ma mère*

# *Remerciements*

*Ce mémoire n'aurait pu être réalisé sans la précieuse collaboration et le constant soutien de plusieurs professeurs, collègues, parents, époux et amis. Je remercie donc tous ceux et celles qui, durant les dernières années, ont contribué à la réussite de ce travail.*

*Je tiens plus particulièrement à remercier chaleureusement mon directeur de thèse, Monsieur Abdelouahid Lyhyaoui, pour son encadrement, son aide, ses conseils et encouragements tout au long de ce travail. Il a su me transmettre son insatiable curiosité, sa passion pour le travail intellectuel rigoureux, ainsi que pour la recherche universitaire.*

*Je remercie également le comité de correction pour avoir participé à l'évaluation de mon travail.*

*Un remerciement va au Centre National de la Recherche Scientifique pour m'avoir fait bénéficier de la bourse d'excellence dans le cadre du programme des bourses de recherche initiées par le Ministère de l'éducation national, de l'enseignement supérieur, de formation de cadres et de la recherche scientifique.*

*Un spéciale remerciement à ma sœur, qui n'a pas cessé de me donner son soutien avant et tout au long de cette thèse. Et qui grâce à ces brillants conseil j'ai pu rédiger ce manuscrit.*

*Un grand remerciement à mon époux pour son encouragement, patience et son réconfort durant la période de recherche.*

*Je remercie aussi tous les membre passées et présent du laboratoire LTI, pour leurs discussions porteuses d'idées et tous leurs conseils.*

*Je remercie bien sûre ma famille et amis pour leurs soutien pendant toute la durée du travail.*

# *Résumé*

Le domaine de l'intelligence artificielle a pour objectif le développement de systèmes informatiques capables de simuler des comportements normalement associés à l'intelligence humaine. On aimerait entre autres pouvoir construire une machine qui puisse résoudre des tâches liées à la vision (la reconnaissance d'objet), au traitement de la langue (l'identification du sujet d'un texte) ou au traitement de signaux sonores (la reconnaissance de la parole).Une approche développée afin de résoudre ce genre de tâches est basée sur l'apprentissage automatique de modèles à partir de données étiquetées reflétant le comportement intelligent à simuler. Il s'agit de la classification automatique des données.

Or, les techniques de classification montrent une détérioration de performance face à la croissance exponentielle que le domaine d'information a connue. En effet, les bases de données sont de plus en plus grandes et montrent des anomalies qui nuisent ces classifieurs. L'un des aspects le plus répandu est celui du déséquilibre entre les classes.

Néanmoins, plusieurs études ont montré que le prétraitement des données et l'application des techniques de sélection des échantillons permettent d'accroitre la qualité de la classification.

Ainsi, l'idée principale de cette thèse est d'intégrer les techniques de sélection des échantillons dans le processus de classification, afin de résoudre le problème du déséquilibre des données.

Nous présentons dans cette thèse trois méthodes pour traiter cette problématique: La première approche se base sur le critère de l'erreur d'apprentissage d'un réseau de neurones, pour effectuer la sélection des échantillons. C'est une méthode qui réalise un sous-échantillonnage ciblé au fur et à mesure de l'entraînement du classifieur

La deuxième approche permet de sélectionner les clusters les plus proches à la frontière sans avoir recours au classifieur. Elle se base seulement sur le critère de la distance et des paires opposées les plus proches.

Tandis que la dernière approche, utilise une mesure se basant sur la distance (fonction indicatrice) permettant de sélectionner les centres étendus tout au long de la frontière, nous permettant ainsi, de garder la distribution initiale des données autour de la frontière.

De cette manière nous avons réalisé un équilibrage des données en se basant sur des méthodes de sélection des échantillons.



Mot-clef: Intelligence Artificielle, Apprentissage automatique, Classification, Déséquilibre des données, Sélection des échantillons, MLP.

# *Abstract*

Artificial intelligence is a field of computer science that explores computational models of problem solving, where the problems to be solved are of the complexity of problems solved by human beings. Among the well-developed field of artificial intelligence, we find machine learning that provides computers with the ability to learn without being explicitly programmed. It focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data.

A series of challenges have recently emerged in machine learning area, triggered by the rapid shift in status from academic to applied science and the resulting needs of real-life applications. Although existing knowledge discovery and data engineering techniques have shown great success in many real-world applications, the problem of learning from imbalanced data is a real challenge that has attracted growing attention from both academia and industry. The current thesis is concerned with classification tasks and related issues which may appear in real-world scenarios, such as: incomplete records and irrelevant and/or redundant pieces of information, and especially imbalanced class distribution. Many solutions was proposed to solve this problem. Sampling methods still the most widely used method to deal with the problem of imbalanced class.

Summing up, besides being small, a desirable training set must be constructed in a smart way, it must represent correctly the class boundaries by sampling the most discriminative examples. this technique is called samples selection methods.

In this thesis, we present tree methods to deal with the problem of imbalanced distribution of data: The first one use the criterion of training error to uder-sample examples who generate a low error.

The second one, use the distance between examples and the borderline using technique of clustering. And the final approach, is an improvement of the second one by integrating a new function which associate to each sample a value of proximity and importance.

The current thesis ascertains the problem statement and provides an overview of existing approaches by which all the proposed methods was built and experimented.



Keywords: Machine Learning, Classification, Imbalanced Data, Sample Selection techniques, MLP.

# *Sommaire*

# *Liste des Figures*

# *Liste des Tableaux*

## *Liste des abréviations*

| | |
|---|---|
| SGBD | Système de gestion de Bases de données |
| ERP | Enterprise Resource Planning |
| AAAI | Association for the Advancement of Artificial Intelligence |
| SVM | Support Vector Machine |
| MLP | Multilayer Perceptron |
| kNN | k Nearest neighbor |
| VP | Les vrais positifs |
| FP | Faux positifs |
| VN | Vrais négatifs |
| FN | Vrais négatifs |
| ICMP | Internet Control Message Protocol |
| TCP | Transmission Control Protocol |
| UDP | User Datagram Protocol |
| IDS | Intrusion Detection System |
| DARPA | Defense Advanced Research Projects Agency |
| U2R | User to Root attacks |
| R2L | Remote to Local access. |
| ROC | Receiver Operating Characteristics |
| P2P | Peer-to-Peer |
| RBF | Radial Basis Function |
| FF | Feed Forward. |
| RNN | Recurrent Neural Networks. |
| SS | Sample Selection |
| KDD | Knowledge Discovery and Data Mining |
| ACP | principales components analys. |

# *Chapitre I : Introduction générale*

## *I.1. Contexte*

Nous sommes dans un âge souvent désigné sous le nom de l'âge de l'information où l'information représente le pouvoir et le succès. Grâce aux technologies sophistiquées telles que les ordinateurs, les satellites, etc ..., des quantités énormes d'information a été rassemblées. Par l'arrivée des ordinateurs et des moyens pour la mémoire numérique de masse, ces collectes de données massives enregistrées sur les structures disparates sont très rapidement devenues accablantes, chose qui a mené à la création des bases de données et des systèmes de gestion structurés de ces bases (SGBD).

Aujourd'hui, nous avons bien plus d'information que nous pouvons manipuler : des transactions et des données scientifiques, images satellites, des textes .... Il a ainsi fallu développer de nouveaux outils permettant de manipuler et stocker les données avant d'être utilisées. La motivation des recherches élaborées dans ce sens, était de permettre la mise en œuvre de système manifestant une **intelligence artificielle** permettant aux machines d'apprendre à partir d'un ensemble de données, et d'y extraire des concepts et des modèles caractérisant ces données.

La limitation des techniques usuelles d'analyse de données, développées pour des tableaux de petite taille (inférieure à dix milles exemples), devant les données de masse a mené à l'apparition de nouvelles méthodes connues sous le nom de Fouille de Données (ou Data Mining) [1]. En effet, alors que le principal objectif de la statistique est de prouver une hypothèse avancée par un expert du domaine, et donc de confirmer une connaissance déjà connue ou bien présumée, le but de la Fouille de Données est de découvrir, des nouvelles connaissances. Et ceci sans faire appel à des hypothèses établies. Ce nouveau concept, bien qu'il paraît révolutionnaire pour certains, est en fait une autre vision et une autre utilisation de méthodes existantes .

La Fouille de Données est un domaine multidisciplinaire, combinant la technologie des bases de données, l'intelligence artificielle, l'apprentissage automatique, les réseaux de neurones, la statistique, la reconnaissance de formes et la visualisation des données. Elle permet, à partir de données dont on ne sait rien et sur lesquelles on ne fait aucune hypothèse, d'obtenir des informations pertinentes, et à partir de celle-ci, découvrir la connaissance qui permettra à l'entreprise (ou aux scientifiques) de manipuler, comprendre, analyser, prévoir, économiser, bref, de mieux gérer les données.

Parmi les méthodes les plus appliquées dans le domaine de la fouille de Données, on retrouve l'Apprentissage Automatique (Machine Learning). Il existe plusieurs types différents d'apprentissage automatique, qui se distinguent essentiellement par leur objectif. Dans le cadre de cette thèse on parlera surtout de l'apprentissage supervisé et de l'apprentissage non-supervisé [2].

***Apprentissage supervisé***

L'apprentissage supervisé correspond au cas où l'objectif de l'apprentissage est déterminé explicitement via la définition d'une cible à prédire (étiquette). La tâche d'un algorithme d'apprentissage est alors d'entraîner un modèle qui puisse imiter ce processus d'étiquetage par un humain, c.à.d., qui puisse prédire pour une entrée $\boldsymbol{x}$ quelconque, la valeur de la cible $\boldsymbol{y}$ qui aurait normalement été donnée par un humain. Cependant, les algorithmes d'apprentissage ne se limitent pas à la modélisation du comportement de l'humain, et peuvent être utilisés pour modéliser la relation liant des paires d'entrées et de cibles provenant d'un autre phénomène (par exemple, la relation entre une action et son prix à la bourse telle que générée par les marchés financiers). Deux types de problèmes fréquents dans l'apprentissage supervisé sont les problèmes de classification et de régression [3]

- **Régression:**

  Lorsque la valeur à prédire $\boldsymbol{y}$ (sortie) correspond à un ensemble de valeurs continues, nous parlons d'un problème de régression. Cette technique est très utilisée pour analyser la relation d'une variable par rapport à une ou plusieurs autres. Cela permet d'estimer et de prédire les valeurs de *y*.

- **Classification:**

  Ces méthodes ont plusieurs possibilités de combinaison avec d'autres méthodes, en pre- ou en post-processing. Le but de la classification est de diviser un ensemble des données en plusieurs classes distinctes et homogènes. La classification automatique a

plusieurs applications, et dans des domaines très diverses. En économie, la classification peut aider les analystes à découvrir des groupes distincts dans leur base clientèle, et à caractériser ces groupes de clients, en se basant sur des habitudes de consommations. En biologie, on peut l'utiliser pour dériver des types de plantes et d'animaux, pour catégoriser des gènes avec une ou plusieurs fonctionnalités similaires, pour mieux comprendre les structures propres aux populations... etc.

***Apprentissage non-supervisé***

L'apprentissage non-supervisé correspond au cas où aucune cible n'est prédéterminée. Ainsi, c'est plutôt l'utilisateur qui doit spécifier le problème à résoudre. Généralement ce type d'apprentissage utilise la notion soit de dissimilarité et similarité, soit de distance, afin de constituer l'ensemble de sortie souhaité. Nous distinguons plusieurs types d'application de ce type d'apprentissage:

- **Estimation de densité**

  Pour ce problème, l'algorithme doit fournir une estimation de la fonction de densité ou de probabilité de la distribution ayant généré les éléments d'entraînement.

- **Extraction de caractéristiques**

  L'objectif de cette tâche est d'apprendre une nouvelle représentation de l'entrée qui soit plus utile que la représentation vectorielle originale.

- **Réduction de dimensionnalité**

  Pour cette tâche, la fonction retourné par l'algorithme d'apprentissage $f$ doit associer à un vecteur d'entrée $x$ une représentation $f(x)$ de plus petite dimensionnalité tout en conservant l'essentiel de l'information contenue dans l'entrée

- **Groupage**

  Le problème du groupage *(clustering)* correspond à la recherche d'un partitionnement de l'espace de données $X$ en $g$ sous-ensembles (ou groupes) $G_1, \ldots, G_g$. Chacun de ces sous-ensembles $G_i$ peut facultativement être associé à un prototype (ou centroïde) $X_i$ qui résume bien les données contenues dans $G_i$.

Les méthodes d'apprentissage automatique ont donc progressé dans leur fonctionnement, mais aussi dans la complexité des problèmes auxquels elles se sont attaquées. C'est dans ce contexte qu'ont été conduits les travaux de cette thèse.

## *I.2. Problématique*

Dans le contexte vu précédemment, nous nous focalisons durant les travaux présentés dans cette thèse, sur la classification binaire des données.

La tâche de la classification peut s'avérer complexe et délicate lorsque les données disponibles pour l'apprentissage présentent des anomalies. Parmi les problèmes fréquemment interceptés et dont fait le sujet de ce travail, nous trouvons le cas où les données n'ont pas une distribution équilibrée entre les différentes classes. Dans ce cas les classificateurs auront du mal à bien fonctionner et les classes majoritaires finissent par être trop favorisées par rapport à d'autres, ce qui risquerait de biaiser les résultats.

Cependant, une bonne classification ne peut se faire sans avoir réalisé un prétraitement des données permettant d'éliminer les anomalies (les bruits, aberrances...) et de sélectionner les caractéristiques (features) discriminatoires, permettant ainsi de réaliser une réduction de la dimension des données.

Dans le cas de détection des intrusions, ce problème est très prononcé. C'est la raison par laquelle ce domaine à été choisi comme une application pour notre thèse. Dans la sous-section suivante, nous allons donner un aperçu rapide de notre contribution dans ce domaine.

## *I.3. Contribution*

Tout au long de ces années de thèse nous avons pu réaliser trois approches permettant d'apporter une contribution au domaine de classification des données non-équilibrées.

Les trois approches présentées ont le même objectif: Former un ensemble d'entraînement équilibré et qui permet de garder la distribution de l'ensemble initial, spécialement autour de la frontière, vu que c'est l'endroit où les exemples les plus discriminants sont localisés.

La première approche se base sur le critère de l'erreur d'apprentissage d'un réseau de neurones, pour effectuer la sélection des échantillons. C'est une méthode qui réalise un sous-échantillonnage ciblé au fur et à mesure de l'entraînement du classifieur [4]

La deuxième approche permet de sélectionner les clusters les plus proches à la frontière sans avoir recours au classifieur. Elle se base seulement sur le critère de la distance et des paires opposées les plus proches [5], [6].

Tandis que la dernière approche, utilise une mesure se basant sur la distance (fonction indicatrice) permettant de sélectionner les centres étendus tout au long de la frontière, nous permettant ainsi, de garder la distribution initiale des données autour de la frontière.

De cette manière nous avons réalisé un équilibrage des données en se basant sur des méthodes de sélection des échantillons.

## *I.4. Organisation du manuscrit*

Cette thèse débute par une exposition du problème du déséquilibre entre les classes, la présentation de ces enjeux ainsi que l'état de l'art relié à ce problème. Par la suite, dans le chapitre 3, nous présentons le domaine d'application choisi, qui est le système de détection des intrusions, ainsi que nous présentons les données et réalisons un prétraitement qui précédera l'application des méthodes proposées. Puis sont présentés dans le chapitre 4, nous exposons les techniques d'édition des données permettant de réduire l'espace des variables ainsi que l'espace des échantillons (d'où le terme de sélection des échantillons). Dans le chapitre 5, nous présentons les contributions apportées dans le cadre du sujet de cette thèse ainsi que les résultats les résultats des trois approches. Nous concluons notre travail dans le chapitre 6, par une synthèse générale et la présentation des perspectives.

# *Chapitre II: Données Déséquilibrées, Problématique et solutions*

## *II.1. Introduction*

Les données sont une image de la réalité, tout phénomène vécu ou observé peut être décrit, expliqué ou même prévu grâce à une bonne exploitation des données. Les données réelles sont connues par la difficulté de leur exploitation et cela, à cause de plusieurs anomalies qui les caractérisent comme la redondance, les aberrances, le manque de données ou le déséquilibre. Dans notre sujet de thèse nous nous focalisons sur l'anomalie du déséquilibre

Le déséquilibre est un problème qui a émergé avec la croissance des données résultat du développement scientifique et technologique. A son apparition, ce problème a eu un grand intérêt de la part de la communauté scientifique pour deux raisons principales: son large spectre d'impact (multiples domaines d'applications) ainsi que, son influence directe sur la performance des différentes techniques d'analyse de données spécialement, la classification.

Ce chapitre, vise à caractériser en profondeur le problème du déséquilibre entre les classes et son influence sur l'apprentissage automatique. Ainsi qu'il présente les différentes solutions suggérées dans la littérature pour faire face à cette anomalie.

## *II.2. Enjeux de l'apprentissage à partir des données non équilibrées*

### *II.2.1. Présentation du déséquilibre entre les classes*

Le déséquilibre entre les classes représente dans un langage technique, tout ensemble de données ayant une inégalité de répartition entre ses classes (variables étudiées) est considéré déséquilibré. Néanmoins dans l'apprentissage automatique, on l'exprime, lorsque l'un ou plusieurs de ses types d'entités constituants de données dépassent (en nombre) les autres. Dans la classification binaire, on désigne une classe minoritaire toute classe ayant le

minimum d'éléments [7]. Ce type de déséquilibre est nommé " Déséquilibre entre les classes". Un autre type de déséquilibre existe, nommé « déséquilibre du coût des erreurs » et qui se produit, lorsque les erreurs de classification de tous les exemples n'ont pas le même impact, par exemple, dans le domaine de la détection de la fraude, les faux négatifs (les fraudes qui ont été considéré normales ) sont nettement plus dangereuses que les faux positifs (les fausses alertes) [8]. Dans ce travail de thèse on s'intéresse au déséquilibre des distributions entre les classes.

Le problème traité, présente dans nos jours un vrai obstacle face au développement des méthodes et techniques de la fouille de données. En effet, la fréquente apparition du déséquilibre dans les bases de données engendrées par des systèmes de collecte ( Intégration de données dans des ERP, SGBD , Datamarts...), a attiré l'attention de la communauté scientifique pour bien étudier ce problème, mesurer son impact sur le système d'apprentissage et proposer des solutions. Depuis l'année 2000, L'intérêt donné à l'apprentissage déséquilibré s'est concrétisé par la mise en œuvre de deux ateliers de travail lors de deux conférences internationales [9], [10]:

- AAAI, American Association for Artificial Intelligence, (connue maintenant sous le nom de Association for the Advancement of Artificial Intelligence) en 2000.
- ICML, International Conference on Machine Learning en 2003.

Ces ateliers de travail se sont focalisés sur l'influence du déséquilibre sur la performance des machines d'apprentissage automatique, ainsi que la discussion des métriques utilisées pour le calcul de cette performance. Depuis ces conférences le taux de publications dans le domaine de données déséquilibrées a connu un accroissement exponentiel vu l'ample spectre concerné par cette anomalie : management des risques, diagnostiques [9], [11] sans oublier les domaines submergeant comme la classification des textes, recherche d'information [12], détection des munitions non explosées [13] ou encore le déminage [14].

### *II.2.2. Le problème de l'apprentissage à partir des données non-équilibrées*

Les jeux de données déséquilibrés constituent un problème important de l'apprentissage supervisé vu que la plupart des modèles sont conçus pour des données équilibrées. Leur utilisation sur des données déséquilibrées, conduit souvent à ignorer la classe minoritaire et à considérer tous les exemples appartenant à la classe majoritaire [10] [15].

Prenons l'exemple du domaine de l'analyse des transactions bancaires. La proportion de transactions frauduleuses est habituellement très faible au regard du volume global des transactions. Les taux de fraude ne dépassent guère le 1% et, fort heureusement, restent souvent en dessous de ce seuil. Dans ce cadre, on parle généralement de classes déséquilibrées.

Il est admis que le principe d'affectation d'un individu à une classe donnée repose sur la règle de la minimisation d'erreurs. Ainsi, on affecte généralement un individu à la classe majoritaire vu que la probabilité de son appartenance à cette classe est beaucoup plus élevée que celle d'appartenance à la classe minoritaire. Du fait du déséquilibre des classes, la stricte application de la règle de minimisation du nombre d'erreurs conduit à considérer que toutes les transactions sont non frauduleuses, ce qui évidement n'est pas très utile dans une optique de détection des fraudes.

Afin de mieux concrétiser l'influence du déséquilibre sur la performance du classificateur, nous reprenons les résultats d'une étude réalisée en [7], et qui montre la variation de la performance de plusieurs classificateurs selon le taux du déséquilibre ( Imbalance Ratio *IR*), la complexité ainsi que la dimension des données.

Le tableau suivant représente les taux des vrais positifs (TP ratio) pour six méthodes de classification :

- Les *k* plus proches voisins (*k* Nearest Neighbor),
- Arbre de décision C4.5
- Support Vector Machines SVM,
- Réseau de neurones artificiel MLP (Multi-layer Perceptron)
- Apprentissage Bayesian NB (Naïve Bayes),
- Adaptive Boosting – AB (AdaBoost.M1).

| Taille des données | IR | Complexité | kNN | C4.5 | SVM | MLP | NB | AB |
|---|---|---|---|---|---|---|---|---|
| *Très petite* | $< 9$ | *Faible* | .53 | .5 | .5 | .61 | **65** | .57 |
| | | *Moyenne* | **.72** | **.71** | .3 | .61 | .65 | .65 |
| | | *Grande* | 73 | .72 | .79 | .76 | **.8** | **.81** |
| | $\geq 9$ | *Moyenne* | .52 | .6 | .15 | .59 | **.83** | .4 |
| *Petite* | $< 9$ | *Moyenne* | .88 | **.89** | **.89** | **.9** | **.89** | .83 |
| | | *Grande* | .81 | .77 | **.85** | .81 | .62 | .67 |
| | $\geq 9$ | *Moyenne* | **.98** | .94 | **.98** | **.99** | **.98** | **.99** |
| | | *Grande* | .24 | .09 | .47 | **.65** | .09 | .0 |
| *Moyenne* | $< 9$ | *Grande* | .74 | **.97** | .92 | **.98** | .69 | .85 |
| | $\geq 9$ | *Moyenne* | .6 | **.91** | .5 | .86 | .78 | .89 |
| | | *Grande* | .57 | **.88** | .04 | .73 | .84 | .82 |
| *Grande* | $< 9$ | *Grande* | **1** | **1** | **1** | **1** | .92 | .98 |
| | $\geq 9$ | *Très grande* | .06 | .0 | .01 | .0 | **.39** | .0 |

*Tableau II-1: Influence du IR, taille et complexité sur la performance de la classification*

En analysant les résultats exposés ci-dessus, nous pouvons affirmer que l'influence du déséquilibre de la distribution des données sur la performance de la classification est plus notable que celles des autres facteurs comme la taille des données ou leur complexité

Ce résultat, déjà confirmé en 2002 par [16], peut être expliqué par le fait que le manque de données adéquates pour soutenir des vecteurs de la classe minoritaire éloigne la limite de séparation ( la frontière) vers la région de la classe majoritaire, de cette manière, les instances de la classe minoritaire sont donc classés comme appartenant à la classe majoritaire [7].

Un autre aspect de l'influence du problème du déséquilibre sur les méthodes d'apprentissage automatique, est l'inappropriation des métriques souvent utilisées pour la détermination de la performance du classifieur.

La plupart des mesures souvent utilisées pour évaluer la performance d'un modèle de classification sont basées sur la matrice de confusion qui croise la classe réelle des individus du jeu d'apprentissage avec la classe prédite par le modèle, et permet de calculer les taux de vrais positifs (VP), faux positifs (FP), vrais négatifs (VN) et faux négatifs (FN).

Certaines mesures évaluent les performances sur une modalité spécifique comme le taux de rappel $R = VP/(VP + FN)$, le taux de précision $P = (VP + VN)/(VP + VN + FP + FN)$ ou la F-mesure qui est la moyenne harmonique du rappel et de la précision [17]. D'autres mesures, ne distinguent pas les classes comme le cas du taux d'erreur global, la sensibilité et la spécificité.

L'usage de ces métriques pour une base de donnée déséquilibrée, mènera à une estimation biaisée de la performance de l'apprentissage. En effet, en estimant le taux de précision de

classification sur une base de données où la classe majoritaire représente 99%, le taux de précision atteindra 99% même si le classifieur considère que tous les exemples sont de la classe majoritaire.

## *II.3. Etat de l'art des méthodes de classification des données non-équilibrées*

Parmi les méthodes proposées pour faire face au déséquilibre des classes, nous pouvons distinguer plusieurs grandes familles suivant le niveau auquel agissent ces méthodes :

- Au niveau des données grâce au rééquilibrage du jeu de données initial avant l'apprentissage par le ré-échantillonnage;
- Au niveau algorithmique, en modifiant les algorithmes pour les rendre moins sensibles au déséquilibre;
- Au niveau de l'évaluation, en recourant à de nouvelles métriques adaptées aux données déséquilibrées.

### *II.3.1. Solutions au niveau des données*

Les techniques de ré-échantillonnage sont considérées parmi les techniques plus simples et les plus intuitives. L'équilibrage de la base d'entrainement a pour effet de ré-calibrer le nombre d'exemples utilisés pour chaque classe, de telle manière que chaque classe possède un nombre d'exemples équivalents. Cet équilibrage est élaboré soit en diminuant le nombre d'exemples de la classe majoritaire (sous-échantillonnage), soit en augmentant le nombre d'exemples de la classe minoritaire (sur-échantillonnage). Plusieurs études ont montré que l'usage de ces techniques d'échantillonnage sur l'ensemble d'entraînement permet d'avoir une meilleure performance de la classification que d'utiliser les données déséquilibrées [18] [19].

***Sous/ Sur- échantillonnage aléatoire:***

Comme son nom l'indique, le sur-échantillonnage aléatoire, consiste à rajouter un ensemble d'exemples sélectionnés aléatoirement à partir de la classe minoritaire. Soit $S$ l'ensemble des données déséquilibrées, et $E_{min}$ l'ensemble des exemples aléatoirement choisit de la classe minoritaire. L'application d'un sur-échantillonnage aléatoire consiste donc à constituer un nouveau ensemble d'entraînement $S_{sur}$ en rajoutant $E_{min}$ à $S$ ( $S_{sur} = S \cup E_{min}$). Par conséquent, l'équilibre de la distribution des classes est ajusté. Cette solution a le mérite

d'être simple et facile à mettre en œuvre. Cependant, elle présente une manière superficielle pour traiter le problème du déséquilibre [9] [20]. En effet, certains individus de la classe minoritaire se retrouvent plusieurs fois dans l'échantillon équilibré, d'autant plus que le déséquilibre est important [21] . Cela risque de forcer le classifieur à apprendre sur des zones très spécifiques de l'espace de représentation, et introduit par conséquent un fort biais d'apprentissage.

Une alternative au sur-échantillonnage est le sous-échantillonnage où la solution de base est la version symétrique du sur-échantillonnage aléatoire. Tandis que le sur-échantillonnage ajoute des exemples à l'ensemble d'entraînement initial, le sous-échantillonnage, consiste à éliminer de la classe majoritaire un nombre d'individus $E_{maj}$ sélectionné aléatoirement de la classe minoritaire pour constituer un nouvel ensemble $S_{sous}$ équilibré ($S_{sous} = S \setminus E_{maj}$). Or, la sélection aléatoire des exemples à éliminer peut être un inconvénient de cette méthode, bien évidement, cela peut mener à la perte des exemples importants à la définition du classifieur ou même à perdre de l'information de la base traité [9], [22].

Les inconvénients de ces deux méthodes de ré-échantillonnage ont mené à la création de plusieurs autres méthodes permettant de sélectionner les exemples de la classe majoritaire ( ou minoritaire) de façon plus intelligente.

### *Le sous-échantillonnage informé*

*Lui et al.* [23] a introduit deux méthodes de sous-échantillonnage informé appelés: the "*EasyEnsemble"* et "*BalanceCascade"*. Ces deux méthodes ont comme objectif de faire face au problème de perte d'information provoquée par le classique sous-échantillonnage aléatoire. La méthode "*EasyEnsemble"* sélectionne plusieurs sous-ensembles à partir de la classe majoritaire et forme des classifieurs en utilisant chacun d'eux avec la classe minoritaire, pour avoir à la fin de l'apprentissage, une sortie qui est tout simplement la combinaison des sorties de ces classifieurs. Cette technique est ainsi considérée comme un algorithme d'apprentissage non supervisé qui explore la classe majoritaire à l'aide de l'échantillonnage aléatoire et indépendant. Le principe de *Easyensemble* est inspiré de la méthode AdaBoost ( Adaptive Boosting), d'ailleurs, plusieurs travaux ont combiné cette technique avec le Boosting ou le Wagging [24] [25]. La figure suivante illustre le comportement du classifieur avec la technique de *Easyensemble.*

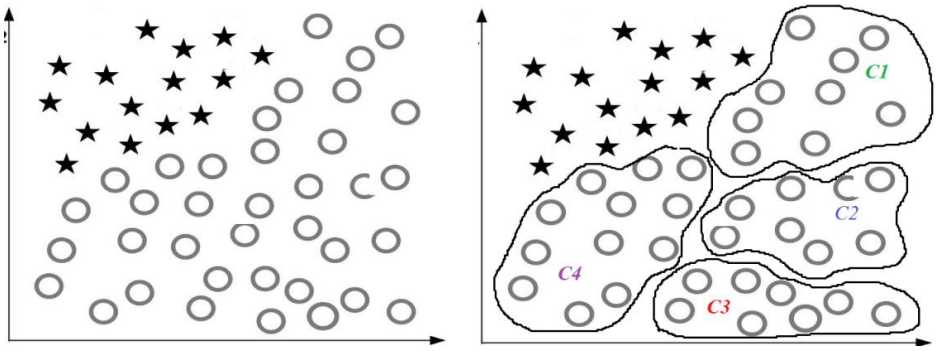



*Figure II.1: Illustration du Easyensemble*

La figure à gauche représente la distribution initiale des données. A droite, la classe majoritaire est partitionnée en quatre sous-ensembles $C1, C2, C3\ et\ C4$. La technique du *Easyensemble* consiste à former un classifieur composé des quatre "sous-classifieur", issues de l'entraînement avec les quatre sous-ensembles de la classe majoritaire.

D'autre part, la deuxième méthode, *BalanceCascade*, se différencie de la première en étant un algorithme d'apprentissage supervisé. Cette méthode consiste à entraîner les classifieurs de façon séquentielle où à chaque itération, l'algorithme élimine de la classe majoritaire les exemples correctement classés. La dépendance entre les classificateurs est principalement utilisée pour la réduction de la redondance de l'information dans la classe majoritaire. Cette stratégie d'échantillonnage permet de réduire l'espace des échantillons en ne gardant que les exemples informatives.

Les deux méthodes citées dessus sont efficace et ont un temps d'entraînement égal à celui du sous-échantillonnage [23].

En 2003, Zhang et Mani [26] ont proposé quatre techniques du sous-échantillonnage informé toutes fondées sur l'algorithme des K plus proches voisins (KNN). Le principe des ces méthodes , nommées NearMiss-1, NearMiss-2, NearMiss-3, et " Most distant", est d'utiliser la technique du KNN pour le sous-échantillonnage de la classe majoritaire. Néanmoins, chaque méthodes a sa propre manière de choisir les éléments à éliminer. Les résultats expérimentaux ont montré que NearMiss-2 permet d'avoir la meilleure performance, cette méthode permet de sélectionner les individus de la classe majoritaire dont la distance moyenne aux K plus proches voisins de la classe minoritaire est la plus faible [9].

### *Génération synthétique d'individus (SMOTE)*

Pour diminuer le biais généré par le sur-échantillonnage aléatoire, [27] propose une technique nommée SMOTE *(Synthetic Minority Oversampling Technique)*, qui permet de

contourner le problème via la génération d'individus synthétiques ayant une représentation différente des individus de la classe minoritaire déjà présents. L'algorithme se déroule comme suit : On cherche les *K* plus proches voisins de chaque individu de la classe minoritaire. Parmi eux, on choisit aléatoirement suffisamment d'individus pour atteindre le taux d'équilibre désiré. Un nouvel individu *N* est alors généré pour chacun de ces voisins $\hat{x}$. Son $i^{ème}$ attribut prend pour valeur :

$$\boldsymbol{N_i = x_i + (x_i - \hat{x}_i).\delta} \qquad \textit{(II.1)}$$

Où:

- $\delta$ est un nombre aléatoire de [0,1];
- $x_i$ l'attribut en question;
- $\hat{x}_i$le voisin de l'attribut $x_i$

Selon la relation *(II.1)*, l'exemple synthétique qui a été générer est un point appartenant au segment reliant l'exemple $x_i$ et son voisin $\hat{x}_i$ sélectionné au hasard. La figure suivante, représente un exemple de l'application du SMOTE.

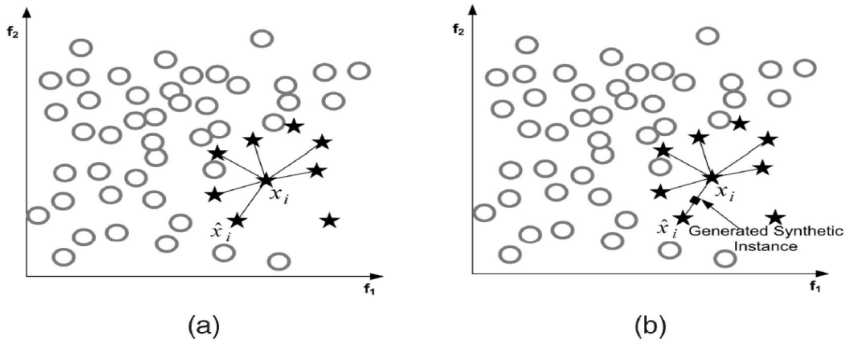



*Figure II.2: Génération d'un attribut par la technique SMOTE dans des donnés déséquilibrés*

La figure *II.2.a.* représente une distribution non équilibrée de données. Les étoiles représentent la classe minoritaire et les cercles, la classe majoritaire. Dans cet exemple, le nombre de voisins *k* est égal à 6 et $\hat{\boldsymbol{x}}_i$ représente un voisin de $x_i$ choisi aléatoirement. La figure II.2.*b* représente la création d'un attribut entre $x_i$ et son voisin sélectionné $\hat{\boldsymbol{x}}_i$.

Bien que cette méthode représente un bon résultat par rapport au sur-échantillonnage aléatoire, elle montre des inconvénients comme le sur-apprentissage ou l'augmentation de la

variance ou encore du chevauchement vu que la génération des données ne tient pas en considération le voisinage à d'autres exemples [28].

### *Echantillonnage synthétique Adaptatif*

Pour faire face aux inconvénients de la méthode classique SMOTE cités dessus, de nouvelles méthodes d'échantillonnage synthétiques, nommées adaptatives, ont été proposées. Deux algorithmes ont été développés dans ce sens: Borderline-SMOTE [29] et Adaptive Synthetic Sampling (ADASYN) [30]

Le principe de Borderline-SMOTE est d'appliquer la génération des données utilisant SMOTE, seulement sur les exemples qui se trouvent au bord de la frontière. L'algorithme de cette technique est comme suit: Pour chaque $x_i$ appartenant à la classe minoritaire, nous cherchons les k voisins les plus proches notés $S_{i-NN}$, ensuite, nous identifions le nombres de ces voisins appartenant à la classe majoritaire $n_i = card(S_{i-NN} \cap S_{maj})$. Finalement, les $x_i$ choisit pour la génération des données sont ceux qui ont plus de voisins de classe à la majoritaire que de classe minoritaire, ca veut dire, ceux qui ont $\frac{k}{2} \leq n_i < k$. Il faut noter que les exemples qui ont $n_i = k$ sont considérés comme étant un bruit. La figure suivante montre les étapes de l'application du Borderline-SMOTE sur des données déséquilibrées.

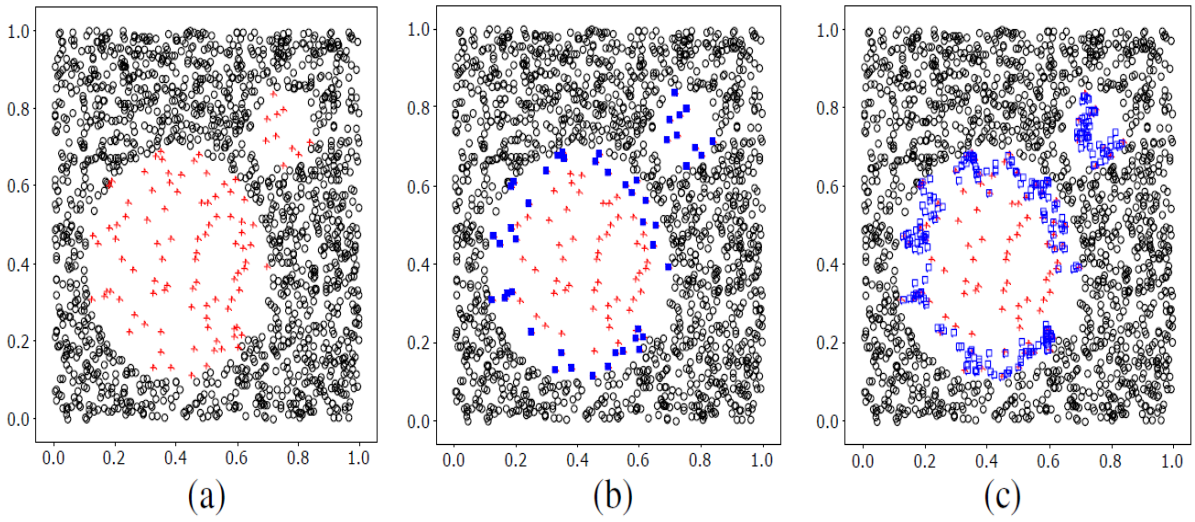



*Figure II.3 Etapes de l'application du Borderline-SMOTE*

La figure *II.3*.a présente une distribution non équilibrée de données. Comme le montre la figure (b), seule les exemples qui se trouvent au bord de la frontière entre les deux classes

sont sélectionnés ( marqués en bleu). Finalement, des données synthétiques sont générés autour des exemples sélectionnés.

D'autre part l'algorithme ADASYN met en œuvre une méthode systématique pour créer différentes quantités de données selon leurs distributions d'une manière adaptative. Le principe de cette méthode est d'affecter à la classe minoritaire une distribution pondérée de tous les exemples selon leurs difficultés à l'apprentissage. Cela est réalisé comme suit: Soit *R* le nombre d'exemple nécessaires à ajouter pour avoir un équilibre entre les classes. Comme le cas du Borderline SMOTE, on détermine $S_{i-NN}$ pour chaque $x_i$ de la classe minoritaire pour qu'ensuite on peut déterminer la distribution des poids de tous les exemples, définie par la relation:

$$\Gamma_i = \frac{S_{i-NN}/k}{Z} \quad (II.2)$$

où

- $Z$ est une constante de normalisation tel que $\sum \Gamma_i = 1$
- $S_{i-NN}$ est l'ensemble des k voisins les plus proches

Enfin, nous déterminons le nombre d'exemples à générer pour chaque $x_i$ de la classe minoritaire selon la relation suivante:

$$r_i = \Gamma_i \times R \quad (II.3)$$

Avec *R* est le nombre d'exemple nécessaires à ajouter pour avoir un équilibre entre les classes.

De cette manière, ADASYN utilise la densité de distribution Γ pour déterminer le nombre d'exemples à générer pour chaque exemple de la classe minoritaire en tenant compte de sa difficulté d'apprentissage.

En 2010, Lebbah et Bennani ont proposé une méthode de sous-échantillonnage adaptatif qui réalise un sous-échantillonnage des données majoritaires guidé par les données minoritaires lors d'un apprentissage semi-supervisé fondé sur les cartes auto-organisatrices [31] . Cette solution guide le choix des données à supprimer dans un voisinage local, en prenant en considération la distribution et la topologie des données. Ceci est réalisé en intégrant les règles de nettoyage de l'algorithme NCR (Neighborhood Cleaning Rule) comme

une troisième étape dans l'algorithme SOM (Self-Organizing Map) qui sera utilisé dans le cas semi-supervisé.

### *Échantillonnage avec des techniques de nettoyage des données*

Tomek links est considérée comme étant la méthode de nettoyage la plus utilisée pour éliminer les exemples créant du chevauchement des données au niveau de la frontière [32]. En 1997, Kubat et Matwin [33] ont eu l'idée d'utiliser Tomek links pour effectuer le sous-échantillonnage. Cette méthode nommée OSS (One Side Selection) consiste à éliminer les exemples de la classe majoritaire de la zone du chevauchement. De cette manière, les classes de l'ensemble d'entraînement sont bien définies, ce qui permet au classifieur d'avoir une meilleur performance.

Cette méthode a déclenché une succession de travaux qui utilisent les méthodes de nettoyage de données pour bien définir la frontière entre les classes. Citons dans ce sens la combinaison entre la méthode "Condensed Nearest Neighbor" (CNN) et Tomek Links [34], " neighborhood cleaning rule" ( NCL) basée sur " edited nearest neighbor (ENN)" [35] ainsi que l'intégration du "Tomek links" dans la méthode SMOTE.

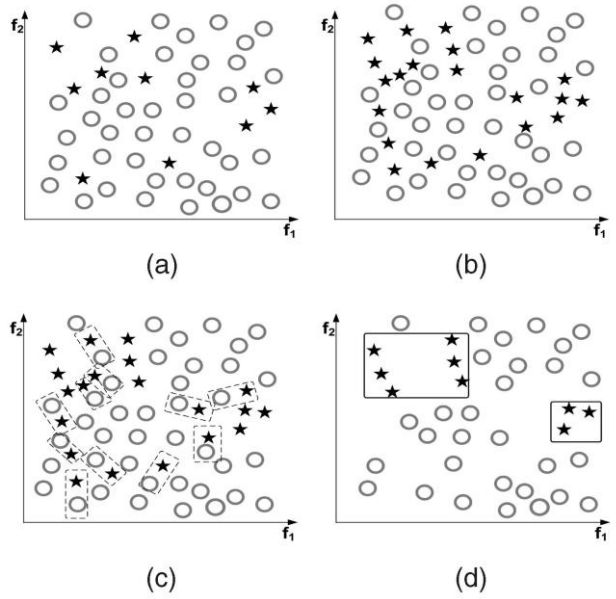



*Figure II.4: Combinaison des techniques Tomek links avec la méthode SMOTE*

La figure II.4.a représente la distribution initiale des données. Apres la réalisation de la méthodes SMOTE on trouve la distribution montrée dans la figure II.4.b. La figure II.4.c, présente l'identification des Tomek links, et enfin la figure II.4.c montre la base finale après

nettoyage. Nous pouvons remarqué que les données ne représentent plus de chevauchement après la génération des données pour la classe minoritaire.

### *Méthodes d'échantillonnage basées sur le clustering*

Les méthodes d'échantillonnage basées sur le clustering sont particulièrement intéressantes car elle permettent de travailler sur des problèmes bien spécifiques en utilisant des éléments de flexibilité supplémentaires que l'on ne retrouve pas dans les méthodes d'échantillonnage simples [9].

En 2004, [36] propose une méthode de sur-échantillonnage basée sur l'algorithme du k-means pour le clustering. La méthode appelée "Clustering-based oversampling" (CBO), suggère de focaliser le problème d'intra-classe plutôt que l'interclasse. Elle permet aussi de régler les problèmes dus aux exemples rares en parallèle de l'apprentissage.

## *II.3.2. Solutions algorithmiques*

Au niveau algorithmique, différents types d'actions sont proposés. Parmi eux, on trouve les méthodes qui prennent en compte les coûts de classification nommées " Cost-sensitive methods". Ces méthodes se basent sur la matrice des coût qui représente le coût d'une mauvaise classification pour chaque classe ( un exemple de cette matrice est présenté dans le tableau suivant). Contrairement au méthodes d'échantillonnage, ces méthodes ciblent à réduire le coût global de classification plutôt que d'essayer de rétablir l'équilibre entre les classes.

| | Vraie classe 0 | Vraie classe 1 |
|---|---|---|
| Prédiction classe 0 | C(0,0) | C(0,1) |
| Prédiction classe 1 | C(1,0) | C(1,1) |

*Tableau II-2: Matrice des coûts de classification*

Comme le montre le tableau *II.3*, la matrice du coût peut être considérée comme étant une représentation numérique de la pénalité du classement d'un exemple appartenant à une classe comme étant de l'autre classe. En général, on n'affecte pas de coût à la prédiction correcte de classe ( C(0,0) = C(1,1) = 0) et le coût d'une mauvaise classification de la classe minoritaire est élevé par rapport à celui de la classe majoritair ( $C(maj, min) >$

$C(min, maj)$). Ces méthodes tâchent à minimiser le risque conditionnel Bayesien défini par [37]:

$$R(i/x) = \sum_j P(j/x).C(i,j) \qquad (II.4)$$

Où

- $P(j/x)$ est la probabilité de la classe *j* sachant *x*
- $C(i,j)$est le coût de classifier un exemple de la classe *j* comme étant de la classe *i*.

Plusieurs méthodes ont été suggérées pour l'implémentation de l'algorithme du "cost-sensitive". Par exemple, [38] propose METACOST, une méthode générale qui permet d'introduire les coûts de mauvaise classification dans un algorithme d'apprentissage supervisé en ré-étiquetant chaque individu par la classe qui permet de minimiser le coût final grâce à une approche *bootstrap* pour ensuite construire le modèle final sur l'échantillon ainsi ré-étiqueté. Certains auteurs suggèrent de prendre des coûts proportionnels aux effectifs de la matrice de confusion, par exemple [39] qui proposent une méthode de construction d'arbre fondée sur un critère de coût minimal. On peut se reporter à [40] pour une étude bien-fondé des méthodes opérant sur les coûts et à [41] pour une comparaison des méthodes de coûts et de rééquilibrage.

Une autre action possible pour répondre à la problématique du déséquilibre consiste à modifier les métriques utilisées pour guider l'algorithme dès l'induction d'un apprenant. Dans cette optique, [42] ont proposé une mesure d'entropie asymétrique tenant compte du déséquilibre des classes pour induire un arbre de classification adapté. Parallèlement, [43] ont élaboré une méthode permettant de décentrer n'importe quelle entropie, notamment l'entropie de Shannon et l'entropie quadratique. L'idée directrice de ces différents auteurs est de modifier une des propriétés classiques des entropies qui est la symétrie. Cette dernière assure que le meilleur éclatement d'une population correspond à une répartition des classes la plus éloignée possible d'une répartition équilibrée, c'est à dire où les classes sont équi-réparties. Cette propriété de symétrie doit être abandonnée dans le cas de données fortement déséquilibrées, puisqu'une répartition équilibrée peut être intéressante dans le cas où la répartition initiale des classes est a priori très déséquilibrée. Un autre type de solution a été proposé par [44] qui remplacent le critère entropique usuel par un critère prenant en compte la notion de coût.

Enfin, en apprentissage par arbre, [45] étudie des stratégies de pré-élagage efficaces pour éviter le sur-ajustement lorsque l'on utilise les méthodes fondées sur les coûts en induction par arbre. Dans le cas précis des arbres de décision avec C4.5, Chawla a étudié la qualité des estimations probabilistes, pré-élagage et le prétraitement des données, trois problèmes habituellement considérés de façon séparée [15].

### *II.3.3. Métriques utilisées pour le cas du déséquilibre*

Suite au développement que les méthodes et algorithmes ont connus pour faire face au problème du déséquilibre, il est devenu nécessaire d'adapter les métriques d'évaluation de la performance en dépendance de ces méthodes car, le choix d'une mesure d'évaluation inappropriée peut conduire à des prédictions inattendues qui ne sont pas en accord avec les objectifs du problème. En effet, Les métriques "standard" qui ne prennent pas en compte le déséquilibre des classes, attribuent plus de poids aux classes ordinaires qu'aux classes rares, ce qui rend difficile pour un classifieur d'être performant sur ces dernières.

Le taux de bonne prédiction (ou *Accuracy),* est considérée comme étant la métrique la plus utilisée pour mesurer la performance d'un algorithme. Or, plusieurs travaux ont montré son inaptitude dans un problème de déséquilibre de classes [46], [47]. Au lieu de l'*Accuracy,* plusieurs auteurs ont proposé des métriques qui dépendent de la fréquence de la classe et donc du déséquilibre des données. Ces métriques peuvent améliorer la fouille des données en orientant mieux le processus de recherche et permettent de mieux évaluer le résultat final de la classification. Parmi ces métriques, on trouve principalement la précision et le rappel, la moyenne géométrique ou la F-mesure et finalement la courbe ROC, la courbe de précision-rappel et la courbe des coûts.

La précision, le rappel, F-mesure et la moyenne géométriques font partis de ce que l'on nomme " Les mesures d'évaluation singulières" . Elle sont calculées à partir de la matrice de confusion représentée dans la figure suivante

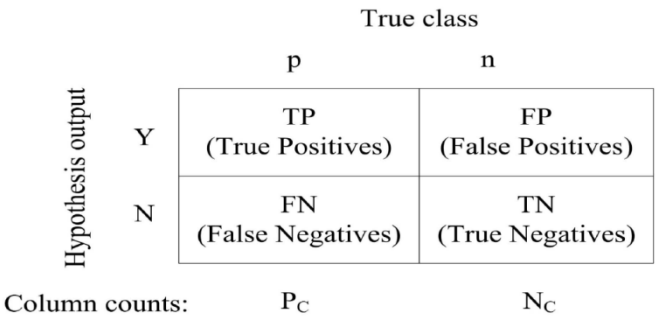



*Figure II.5: Matrice de confusion pour évaluation de la performance*

La matrice de confusion est composée de deux colonnes: A droite, on retrouve les instances négatives et à droite les positives

La précision, définie par

$$P = \frac{TP}{TP+FP} \quad (II.5)$$

représente l'exactitude de l'apprentissage. Selon *(II.5)*, elle permet de compter les exemples identifiés comme étant positive.

Le rappel définie par

$$R = \frac{TP}{TP+FN} \quad (II.6)$$

mesure l'exhaustivité en calculant le nombre d'exemples positifs qui ont été correctement classés.

La F-mesure définie par

$$F_{mesure} = \frac{(1+\beta)^2 \times R \times P}{\beta^2.R+P} \quad (II.7)$$

est une combinaison entre la précision et le rappel. Elle permet de déterminer l'efficacité de la classification en pondérant l'importance donnée au rappel ou à la précision à l'aide d'un paramètre $\beta$ fixé par l'utilisateur.

Finalement, la G-mean ou la moyenne géométrique est trouvée par

$$G_{mean} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \quad (II.8).$$

Cette mesure permet d'avoir une idée à propos du degré du biais inductive de la classification.

En 1978, Metz a développé une méthode d'analyse de performance nommée ‘Receiver Operator Characteristic’ ROC [48]. Cette méthode permet à un classifieur d’être évalué vis-à-vis un ensemble de conditions possibles et la valeur scalaire communément extraite est l’aire sous la courbe ROC, AUC ("Area Under the ROC Curve"). La figure suivante montre un exemple de la courbe ROC.

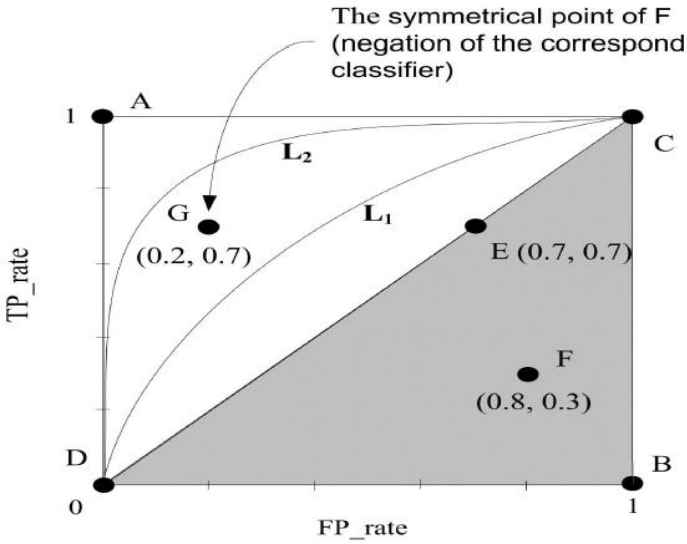



*Figure II.6: Exemple d'application de la courbe ROC*

La figure ci-dessus correspond à un graphique dont l'abscisse indique le taux de faux positifs ( $FP_{\_rate} = \frac{FP}{N_c}$) et l'ordonnée le taux de vrais positifs ($TP_{\_rate} = \frac{TP}{P_c}$). Chaque classifieur fournit un couple ( *TP_rate*, *FP_rate)* est représenté dans l'espace ROC. Dans la figure II.6, A,B,C,D,E,F et G correspondent aux points ROC de sept classifieurs tandis que $L_1$ et $L_2$ représentent deux courbe ROC.

Selon la structure du graphe de ROC, le point A est considéré comme étant la classification parfaite. Du fait, plus un classifieur est proche de A plus il a est performant. Si un classifieur permet d'avoir un point ROC sur la diagonale ( comme le point E), alors il a une classification aussi performante qu'une classification aléatoire. Tandis que si un point se trouve au triangle inférieur de l'espace ( comme le point F) alors il performe moins qu'un classifieur aléatoire.

Dans le cas où le classifieur a une sortie continue, nous représentons les courbes ROC au lieu des points. La comparaison entre les classifieur se fait en utilisant le critère AUC [49], [50]. Dans la figure II.6, la courbe $L_2$ fournit une surface AUC plus grande que celle de la courbe $L_1$ d'où, le classifieur lié à la courbe $L_2$ est plus performant que celui de $L_1$.

Bien que la courbe ROC est une méthode puissante pour visualiser la performance des classifieurs elle connaît quelques limitations quand il s'agit d'une base largement déséquilibrée ou asymétrique. Dans ce cas, la courbe de la précision-rappel (PR) est plus informative que la courbe ROC [51]. La courbe RP est une représentation du taux de précision en fonction du taux du rappel. Un classifieur est dit performant par la courbe ROC s'il l'est par la courbe RP, tandis que le contraire n'est pas vrai [51].

Une dernière métrique qui permet aussi de mesurer la performance des classifieurs dans le cas des données déséquilibrées est la courbe du coût. Cette technique permet de visualiser la performance du classifieur selon la variation du coût de mauvaise classification et la distribution des données [52]. Elle permet ainsi une meilleur visualisation de la performance permettant une comparaison plus nette des classifieurs [53].

## *II.4. Synthèse*

Ce chapitre a été consacré à l'un des problèmes majeurs des bases de données réelles. Nous avons décrit un portrait global de la classification des donnés déséquilibrées.

Cette anomalie a un grand impact négatif sur la performance de classification, notamment en favorisant les classes majoritaires et ignorant les classes minoritaires, la performance des méthodes classiques de l'apprentissage automatique est gravement dégradée.

Une série de méthodes qui traitent le déséquilibre de classe a été proposée dans la littérature au cours des dernières années. Les techniques d'échantillonnage sont importantes car elles peuvent être utilisées comme des stratégies de pré-traitement. Cependant, certaines approches sont plus difficiles à utiliser ou très coûteuses comme le cas des solutions portées au niveau algorithmiques. Le tableau suivant synthétise l'ensemble des solutions proposées

| Type de solution | Nom de la méthodes | Principe de la méthode |
|---|---|---|
| ***Solutions au niveau des données*** | Sous/sur- échantillonnage aléatoire | Choisir d'une manière aléatoire les exemples à éliminer de la classe majoritaire, ou ajouter à la classe minoritaire |
| | Le sous-échantillonnage informé | Eliminer les exemples bien classifiés ou combiner un ensemble de classifieurs |
| | Génération synthétique d'individus (SMOTE) | Générer un ensemble de données selon le critère du voisinage |
| | Echantillonnage synthétique Adaptif | Génération des données synthétiques au bord de la frontière |
| | Echantillonnage avec technique de nettoyage | Eliminer les exemples de la classe majoritaire de la zone de chevauchement |
| | Echantillonnage basé sur le Clustering | Basé sur la technique du k_means pour réaliser un sur-échantillonnage |
| ***Solutions au niveau des algorithmique*** | Cost-sensitive methods | Utiliser la matrice coût afin de réduire le coût globale |
| | Arbre de décision C4.5 | |
| ***Solution pour les mesures utilisées*** | Matrice de confusion | Adapter les métriques du calcule de la précision sur le problème du déséquilibre.<br>Tenir en considération la classification de la classe minoritaire |
| | Précision | |
| | Rappel | |
| | F-mesure et G-mesure | |
| | La courbe et le graphe ROC | |

*Tableau II-3: Synthèse des solutions proposées pour régler le problème du déséquilibre*

D'autre part, le choix des métriques utilisées pour évaluer la performance d'un classifieur joue un rôle très important dans la procédure de la fouille des données, vu que ces outils doivent aussi prendre en considération la distribution des données ainsi que le déséquilibre

entre les classes. Dans le chapitre suivant nous verrons un aperçu sur les systèmes de détection d'intrusion, vu que c'est un domaine caractérisé par le déséquilibre de ses données.

# *Chapitre III: Application : Système de détection d'intrusion*

## *III.1. Introduction*

Les innovations techniques récentes alliées à une demande croissante des utilisateurs ont favorisé un développement fulgurant des technologies et des services mobiles auquel est associée une intégration massive de la technologie communicante. La généralisation des liaisons à haut débit et la multiplication des accès distants (extranet, intranet, télétravail, cybercafé) laissent l'information de l'entreprise accessible à chaque instant, à partir de n'importe quel endroit, grâce aux réseaux virtuels privés *(VNP).* Chaque connexion augmente la vulnérabilité du réseau par rapport aux agressions. Parallèlement, les problèmes de sécurité, en particulier les intrusions par Internet, sont en accroissement transcendant. D'où la nécessité primordiale de protection. C'est dans ce cadre bien réel, celui de la détection des intrusions dans les systèmes informatiques, que nous avons choisi d'appliquer les méthodes de sélection d'échantillon pour l'équilibrage des données que nous avons proposées précédemment. Ces approches sont appliquées à la base de données KDD-Cup 1999, dont la classe à prédire est le statut de la connexion, connexion normale ou attaque. Cette base de données est caractérisée par la présence de plusieurs anomalies, comme les bruits, la redondance et le déséquilibre.

Ce chapitre est organisé comme suit: La section 2 présente le contexte, à savoir les systèmes de détection d'intrusions. Dans la section 3, nous présentons les données KDD-Cup que nous allons utiliser. La section 4 rapporte l'ensemble des techniques de prétraitement appliquées sur les données. Enfin, on conclura par une synthèse du chapitre.

## *III.2. Aperçu sur les Système de Détection d'Intrusion*

La détection d'intrusions sur un système informatique a pour objectif de déceler toute violation de la politique de sécurité en vigueur [54]. Selon Mé *et al.* [55], il faut agir préventivement par l'élaboration d'une politique de sécurité, en termes de confidentialité, d'intégrité et de disponibilité des données du système à protéger. Cependant, l'action préventive ne suffit pas, il faut lui associer une politique de détection d'intrusions.

Un système de détection d'intrusion (IDS) rassemble et analyse les informations de différentes régions, au sein d'un ordinateur ou d'un réseau, afin d'identifier les violations de sécurité possible, qui comprennent à la fois les intrusions (attaques de l'extérieur de l'organisation) et l'utilisation abusive (attaques de l'intérieur de l'organisation). Il n'élimine pas l'utilisation d'un mécanisme de prévention mais il fonctionne comme un second mécanisme de défense derrière un pare-feu qui peut surveiller le réseau, tout en laissant inchangeable sa performance [56]

### *III.2.1. Architecture d'un système de détection d'intrusions IDS*

Selon Bidan *et al.* [57], un système de détection d'intrusions est constitué classiquement de trois composants. La figure suivante illustre les interactions entre ces trois composants :

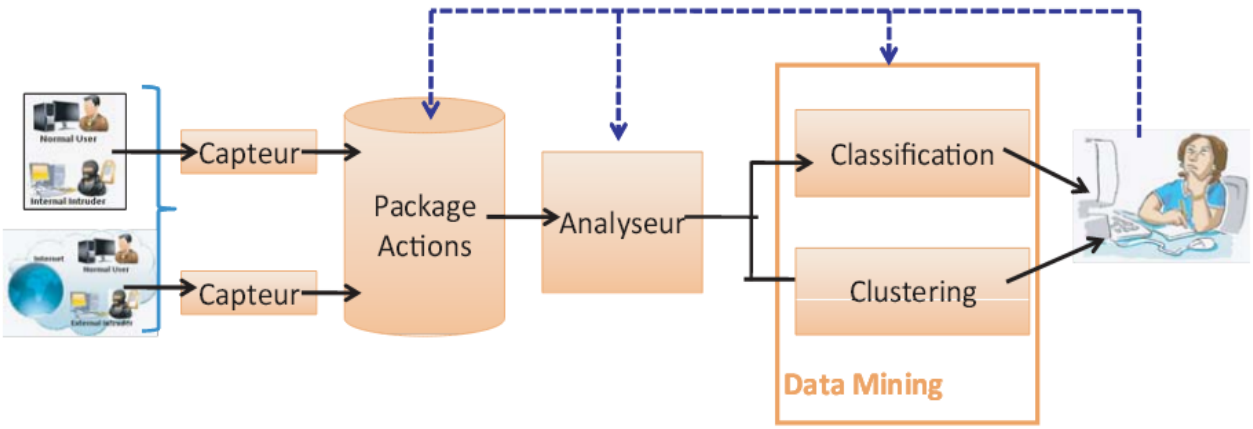



*Figure III.1: Architecture d'un système de détection d'Intrusion*

Comme le montre la figure, un IDS est constitué des trois composantes principales :

- **Le capteur:** Rassemble les informations sur l'évolution de l'état du système et fournit une séquence d'événements qui rendent compte de cette évolution. Plusieurs capteurs peuvent être installés dans divers point stratégiques de la zone sous surveillance.
- **L'analyseur:** Détermine si un sous-ensemble des événements fournis par le capteur est caractéristique d'une activité malveillante.
- **Le manager:** Réunit les alertes en provenance du capteur, il les met en forme et les présente à l'opérateur. Il peut aussi avoir la responsabilité de réagir.

L'analyseur est la composante qui nous intéresse dans cette thèse vu qu'il doit détecter de manière automatique les intrusions.

### *III.2.2.L'apprentissage automatique et les IDS*

En 1986, *Dr Dorothy Denning* a mentionné plusieurs modèles de développement commercial des IDS basés sur des statistiques, chaînes de Markov, séries chronologiques, etc [58]. Dans le modèle de *Denning*, le comportement de l'utilisateur qui s'écarte suffisamment du comportement normal est considéré comme anormal.

En 1988, les IDS basés sur des anomalies statistiques ont été proposés [59], ils ont utilisé à la fois l'utilisateur et le groupe des stratégies fondées sur la détection d'anomalies. Dans ce système, une gamme de valeurs a été considérée comme normale pour chaque attribut. Lors d'une session, si un attribut a une valeur en dehors de la fourchette normale alors une attaque est soupçonnée et une alerte est déclenchée.

En 1996, Forrest et *al.* ont proposé une analogie entre le système immunitaire humain et la détection d'intrusion qui a permis de créer un programme d'analyse de séquences pour construire un profil normal [60], exécuté sous UNIX, en utilisant des programmes comme Sendmail, DPI, etc. Si les séquences sont déviées de la séquence d'un profil normal, alors elles sont considérées comme une attaque. Le système qu'ils ont mis au point, est basé sur l'utilisation des données recueillies précédemment et utilise pour apprendre les profils normaux àun algorithme de recherche qui se base sur un tableau.

En 2000, Valdes and Skinner ont développé dans [61] une approche comportementale basée sur la détection d'intrusions qui utilise le réseau bayésien naïf à partir du flux de circulation. En 2003, Kruegel *et al.* ont proposé une approche de fusion multi-capteurs utilisant le classifieur bayésien pour la prédiction et la suppression des fausses alertes [62].

## *III.3. Présentation de la base de donnée traitée*

### *III.3.1.Présentation des données KDD-Cup 1999*

Les données utilisées pour nos expérimentations sont des données réelles issues de la base KDD-Cup 1999. Elles ont été préparées et contrôlées par les laboratoires MIT Lincoln pour le programme d'évaluation de détection d'intrusion DARPA 1998. Ces données ont aussi été utilisées pour le concours de détection d'intrusions de KDD 1999 [63]. Chaque connexion est étiquetée en tant que connexion normale ou attaque, avec le type spécifique d'attaque. Les attaques trouvées sont classées selon quatre catégories principales comme le montre le tableau suivant:

| Types d'attaques | Catégories |
|---|---|
| neptune, back, land, pod, smurf, teardrop | DOS |
| buffer overflow, loadmodule, perl, rootkit | U2R |
| ftp write, guess passwd, imap, multihop, phf, spy, | R2L |
| ipsweep, nmap, portsweep, satan | Probe |

*Tableau III-1: Types et catégories d'attaques existantes dans KDD-Cup 1999*

- La catégorie DOS provoque un déni de service via des requêtes d'écho ICMP, manipulées à une adresse de diffusion d'un réseau.
- U2R (User to Root attacks*):* l'attaquant essaie d'avoir les droits d'accès au système par le biais d'un poste
- R2L (Remote to Local access): Ce type d'attaque essaie d'exploiter la vulnérabilité du système afin de contrôler la machine distante
- *Probe* (sondage et surveillance*):* Ces actions ne sont pas vraiment des attaques puisqu'elles ne sont pas " destructrices "elles n'empêchent pas une entité de fonctionner correctement, mais permettent d'acquérir des informations parfois cruciales pour mener une attaque de plus grande envergure plus tard.

Ces données sont décrites au moyen de différentes variables explicatives. Pour une meilleure compréhension, celles-ci ont été classifiées en cinq types. Les variables d'une même machine décrivent seulement les connexions faites durant les deux dernières secondes et ayant le même destinataire que la connexion courante. Les variables de même service décrivent seulement les connexions faites durant les deux dernières secondes et ayant le même service que la connexion courante. On trouve aussi des variables de connections TCP individuelles. Enfin, il existe des variables qui indiquent un comportement anormal dans les données, ainsi le nombre de tentatives d'ouverture échouées. Il s'agit des variables de contenu. Les variables explicatives sont décrites dans le tableau en annexe (*Les variables de la base KDD-Cup 1999)*

Les données KDD-Cup 1999 sont construites à partir des données collectées par le programme d'évaluation de détection d'intrusions de DARPA en 1998. Ces données qui correspondent à environ quatre giga-octets de données binaires *TCPdump* compressées, contiennent sept semaines du trafic de réseau. Ceci a été transformé en environ cinq millions de connexions. Les données d'apprentissage KDD-Cup 99 possèdent 4,900,000 connexions étiquetées normale ou attaque. Chaque connexion contient 41 variables descriptives (*Tableau*

*V-2*). Le tableau suivant, donne la répartition exacte d'un échantillon de 10% des données utilisé lors de la compétition, nommé LS-10%, selon les différentes étiquettes de connexions.

| | Données d'entrainement | | Données test | |
|---|---|---|---|---|
| Catégories | Nombre de connexions | Pourcentage | Nombre de connexions | Pourcentage |
| Normal | 79 270 | 19.69 % | 60 593 | 19.48 % |
| DOS | 391 458 | 79.24 % | 229 851 | 74.90 % |
| U2R | 59 | 0.01 % | 228 | 0.07 % |
| R2L | 1 126 | 0.23 % | 16 189 | 5.21 % |
| Prob | 4 107 | 0.83 % | 4 168 | 1.34 % |
| Total | 494 021 | 100% | 311 029 | 100% |

*Tableau III-2– La distribution des données pour la base d'entrainement et de test*

A partir du tableau ci-dessus, nous remarquons que la classe "Normal" ne représente qu'environ 20% de l'ensemble de la base.

### *III.3.2.Anomalies de la base KDD-Cup 1999*

Comme il a été mentionné dans la section précédente, KDD'99 est construite à partir des données de DARPA'98. De ce fait, La base de données KDD' 99 souffre de quelques défauts majeurs dans la distribution de données qui peut fausser les expériences [64]. Citons les anomalies les plus fréquemment trouvées:

- L'ensemble de données KDD'99 présente un déséquilibre entre les classes. Environ 80% de tous les cas, correspondent à des attaques.
- Les types des attaques ne sont pas répartis d'une façon équivalente. Un type d'attaque est dominant.
- Le passage de DARPA'98 à KDD'99 a créé un problème de perte de données [64]
- La redondance de données est importante.
- Le niveau de difficulté d'apprentissage des classifieurs à partir de cette base est élevé

Toutes ces anomalies contribuent à la détérioration de la performance de la classification. Pour cette raison nous étions amenés à faire un prétraitement des données avant d'appliquer les méthodes proposées dans cette thèse.

## *III.4. Prétraitement des données*

Le prétraitement des données décrit tout type de traitement effectué sur les données brutes pour les préparer à une meilleure exploitation. Communément utilisé comme une tâche préalable au data Mining, il transforme les données en un format qui sera plus facilement et plus efficacement traitées par l'utilisateur. Le nettoyage des données, la transformation et l'extraction de l'information comprend la majorité des travaux de construction d'un entrepôt de données, nous pouvons résumer les différentes tâches du prétraitement de données comme suit:

- Traiter le problème des valeurs manquantes par les différentes méthodes statistiques
- Supprimer le bruit dans les données, causé par des erreurs aléatoires ou par la variance d'une variable mesurée,
- Identifier et supprimer les valeurs aberrantes.
- Eliminer la redondance causée par l'intégration de données.
- Appliquer une normalisation des valeurs des variables présentant un grand écart-type
- Effectuer un codage des données afin d'uniformiser les notations et éviter la perte de l'information
- Réduire la dimension des données, quand il s'agit d'une grande base de données

En ce qui suit, nous allons exposer les résultats de la réalisation du prétraitement de la base KDD-Cup'99 avec laquelle nous travaillerons dans cette thèse.

### *III.4.1.Analyse descriptive des données traitées*

Avant de commencer le traitement de la base étudiée, il est toujours recommandé de réaliser une études descriptive des données afin d'avoir une idée globale sur les variables explicatives et expliquées, leurs variations, les aberrances et incohérences ... ect.

Dans cette partie on n'exposera que l'analyse uni-variée, nous garderons l'analyse multi-variée pour la partie de réduction de dimension des données.

***Variables quantitatives:***

La base de données traitée contient 38 variables quantitatives. Les différentes mesures statistiques qui permettent de décrire ces variables sont synthétisées dans le tableau Annexe 1 (voir annexe).

Comme le montre le tableau, nous avons calculé la plage, la moyenne et l'écart-type pour chacune des variables considérées quantitatives.

On remarque qu'il y a une grande variation de la dispersion des données entre les variables. En effet, l'écart-type change de 0,11 pour la variable V18 jusqu'à atteindre 988 219,101 pour la variable V5. Cela montre une hétérogénéité au niveau des données.

### *Variables qualitatives*

La base des données contient un ensemble formé de neuf variables qualitatives. Nous réaliserons pour chacune de ces variables une analyse de sa répartition sur l'ensemble des connexions.

- Variables protocol_type : Elle représente les différents types de Protocol de connexion. Il existe trois modalités pour cette variable: ICMP, TCP et UDP

  La répartition des attributs de cette variable est représentée dans la figure suivante

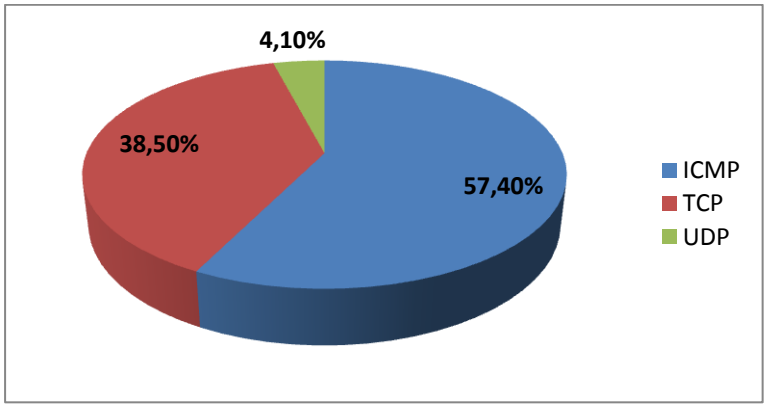



*Figure III.2: Distribution des modalités de la variable V2*

  Nous remarquons que la modalité UDP ne représente que 4% du total des types de protocoles. Cela paraît logique vu qu'il s'agit d'un Protocol simple peut fiable par rapport aux autres.

- Variable 3: Service du réseau. Cette variable a 66 modalités. Nous représentons dans Annexe 3 le tableau descriptif de cette variable contenant les proportions ainsi que la proportion cumulée de chaque modalité.

  Le diagramme suivant résume la répartition des modalités de cette variable

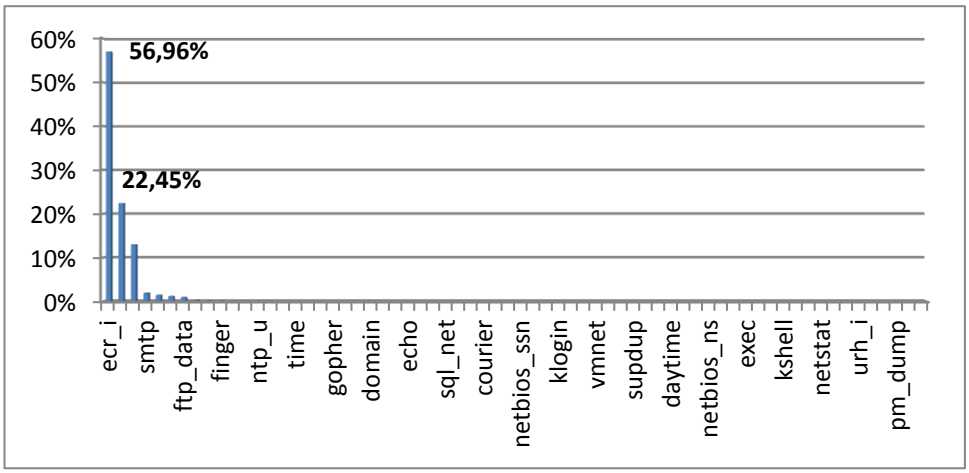



*Figure III.3: Répartition des modalités de la variable 3*

On remarque que les dix premières modalités constituent 98,6% de l'intégrité de la base de données.

- Variable 6: le statut normal ou erreur de la connexion. Cette variable a 11 modalités.

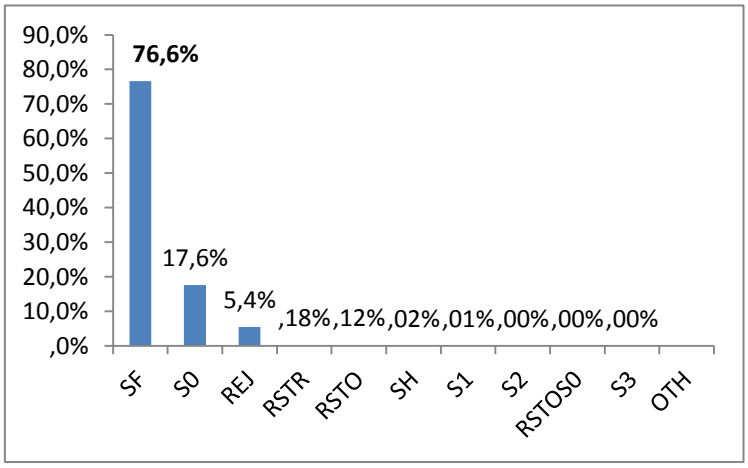



*Figure III.4: Répartition des modalités de la variable 6*

On constate que les quatre premiers types de la variable V6 constituent 99,8% de la totalité de la base de données, alors que les 7 modalités restantes ne constituent que 0,2%.

Le reste des analyses descriptives (contenant la représentation de la distribution des données qualitatives) est énoncé dans l'Annexe.

***La variable expliquée:***

La variable à expliquer ou à prédire est le type de connexion, elle prend deux types de modalités: Attaque ou Normale. On présente dans la figure suivante, la répartition des modalités de cette variable

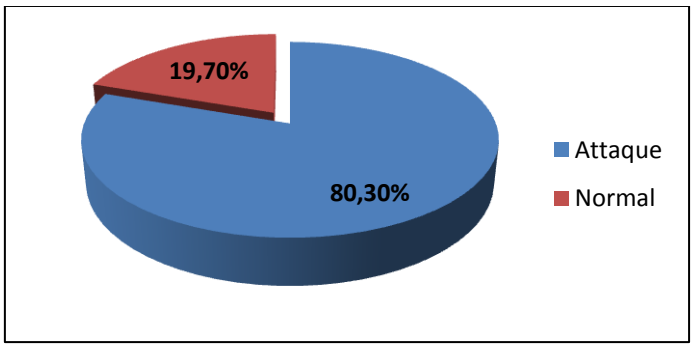



*Figure III.5: Répartition des deux classes de connexion*

Le diagramme de répartition des classes montre clairement le problème de déséquilibre entre les classes que nous sommes en train de traiter dans cette thèse, puisque les transactions frauduleuses (Attaques) représentent 80% de la totalité des connexions.

### *III.4.2.Nettoyage et normalisation des données*

Apres l'analyse descriptive exposée ci-dessus, nous avons réalisé un ensemble d'opérations permettant de nettoyer la base des redondances et d'aberrances et de normaliser les valeurs des variables.

Nous résumons les opérations réalisées sur les données dans les points suivants:

- Les attributs de type symboliques comme : Type du Protocol (Tcp, Udp, tcmp), Service et Drapeau ont été cartographiés à des valeurs entières allant de 0 à N-1, où N est le nombre des modalités de ces variables. Ces attributs ont été réduit de façon linéaire dans l'intervalle [0, 1].
- Les variables ayant la plage de variation très petite sont traités de manière à avoir une variation normalisée par rapport aux autres variables. Pour cela, ils ont été mis à l'échelle linéaire pour l'intervalle [0.0, 1.0].
- Les variables 4 et 5 ont une plage de variation très grande. Nous avons appliqué une échelle logarithmique (base 10) à ces attributs pour réduire l'intervalle de variation de ces données.

- Les exemples redondants ont été tous supprimé vu qu'ils ont une influence négative sur la classification. le tableau suivant montre le nombre des connexions des données d'entraînement avant et après suppression des doublons

| Les données d'entrainement | | |
|---|---|---|
| **Les classes** | **Avant** | **Après** |
| Normal | 97 270 | 87 731 |
| Dos | 391 458 | 54 572 |
| Prob | 4 107 | 2 129 |
| R2L | 1 126 | 999 |
| U2R | 59 | 52 |
| Total | 494 020 | 145 483 |

*Tableau III-3: Le nombre des connexions avant et après suppression des doublons*

- Enfin, nous ramenons la base en un problème de classification binaire en groupant les différents types d'attaques existantes, en une seule classe représentée par l'étiquette -1 et celle normal constitue une autre classe représentée par l'étiquette 1 (connexion normal)

### *III.4.3.Réduction de la dimension*

Pour simplifier la présentation, nous considérons que l'objectif est d'étudier la relation entre la variable dépendante Y (dans ce cas la variable discriminante Attaque/Normal) et les variables explicatives ainsi que leurs inter-corrélation. Nous allons maintenant présenter brièvement les stratégies de réduction réalisées ainsi que les résultats trouvés.

#### *Usage de l'analyse descriptive multi-variée*

Afin de déterminer le degré de relation entre les variables quantitatives qui décrivent notre réponse, on a choisi de calculer la matrice de corrélation entre ces variables. La matrice est représentée dans l'annexe (Annexe 5), vu sa grande taille.

La matrice de corrélation est un outil statistique très utilisé dans l'analyse multi-variée des données. Elle permet de déterminer le type de liaison qui existe entre toutes les variables quantitatives ainsi que l'intensité de cette relation selon le degré d'augmentation du coefficient de corrélation.

Depuis la matrice nous pouvons dans un premier temps, déterminer les variables indépendantes. On considère une variable comme étant indépendante, si la variation de ses valeurs n'influence pas d'autres variables, et donc les coefficients de corrélation entre cette variable et les autres variables sont très petits.

On déduit alors à partir de la matrice de corrélations toutes les variables qui ont des coefficients de corrélation inférieure à un seuil fixé dans ± 0,40 (marqués en bleu). Les variables indépendants trouvés sont :

V1 - V4 - V5 - V8 - V9 - V10 - V11 - V17 - V18 - V29 - V30 - V31 - V34 - V37.

Apres extraction des variables indépendantes, la matrice de corrélation formée par les variables restantes est la suivante :

| | V13 | V16 | V19 | V23 | V24 | V25 | V26 | V27 | V28 | V32 | V33 | V35 | V36 | V38 | V39 | V40 | V41 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| V13 | 1 | ,994 | ,412 | -,009 | -,007 | -,003 | -,003 | -,001 | -,001 | -,008 | -,005 | ,001 | -,005 | -,002 | -,003 | -,001 | -,001 |
| V16 | ,994 | 1 | ,414 | -,009 | -,007 | -,003 | -,003 | -,001 | -,001 | -,012 | -,008 | ,002 | -,005 | -,003 | -,003 | -,001 | -,001 |
| V19 | ,412 | ,414 | 1 | -,043 | -,032 | -,012 | -,013 | -,007 | -,007 | -,021 | -,001 | -,006 | -,026 | -,012 | -,013 | -,005 | -,006 |
| V23 | -,009 | -,009 | -,043 | 1 | ,944 | -,297 | -,306 | -,215 | -,200 | ,533 | ,515 | -,221 | ,684 | -,298 | -,308 | -,196 | -,172 |
| V24 | -,007 | -,007 | -,032 | ,944 | 1 | -,512 | -,521 | -,276 | -,278 | ,402 | ,718 | -,370 | ,753 | -,512 | -,527 | -,256 | -,255 |
| V25 | -,003 | -,003 | -,012 | -,297 | -,512 | 1 | ,950 | -,107 | -,097 | ,149 | -,745 | ,456 | -,448 | ,923 | ,950 | -,099 | -,088 |
| V26 | -,003 | -,003 | -,013 | -,306 | -,521 | ,950 | 1 | -,109 | -,107 | ,152 | -,756 | ,465 | -,457 | ,937 | ,972 | -,101 | -,100 |
| V27 | -,001 | -,001 | -,007 | -,215 | -,276 | -,107 | -,109 | 1 | ,945 | -,092 | -,313 | ,260 | -,201 | -,105 | -,109 | ,886 | ,840 |
| V28 | -,001 | -,001 | -,007 | -,200 | -,278 | -,097 | -,107 | ,945 | 1 | -,085 | -,318 | ,251 | -,204 | -,096 | -,110 | ,849 | ,845 |
| V32 | -,008 | -,012 | -,021 | ,533 | ,402 | ,149 | ,152 | -,092 | -,085 | 1 | -,027 | ,096 | ,181 | ,148 | ,157 | -,081 | -,027 |
| V33 | -,005 | -,008 | -,001 | ,515 | ,718 | -,745 | -,756 | -,313 | -,318 | -,027 | 1 | -,538 | ,568 | -,747 | -,771 | -,294 | -,322 |
| V35 | ,001 | ,002 | -,006 | -,221 | -,370 | ,456 | ,465 | ,260 | ,251 | ,096 | -,538 | 1 | -,323 | ,457 | ,474 | ,246 | ,254 |
| V36 | -,005 | -,005 | -,026 | ,684 | ,753 | -,448 | -,457 | -,201 | -,204 | ,181 | ,568 | -,323 | 1 | -,437 | -,463 | -,178 | -,194 |
| V38 | -,002 | -,003 | -,012 | -,298 | -,512 | ,923 | ,937 | -,105 | -,096 | ,148 | -,747 | ,457 | -,437 | 1 | ,956 | -,098 | -,088 |
| V39 | -,003 | -,003 | -,013 | -,308 | -,527 | ,950 | ,972 | -,109 | -,110 | ,157 | -,771 | ,474 | -,463 | ,956 | 1 | -,102 | -,101 |
| V40 | -,001 | -,001 | -,005 | -,196 | -,256 | -,099 | -,101 | ,886 | ,849 | -,081 | -,294 | ,246 | -,178 | -,098 | -,102 | 1 | ,813 |
| V41 | -,001 | -,001 | -,006 | -,172 | -,255 | -,088 | -,100 | ,840 | ,845 | -,027 | -,322 | ,254 | -,194 | -,088 | -,101 | ,813 | 1 |

*Tableau III-4: Matrice de corrélations des variables quantitatives non-indépendantes*

Les coefficients marqués en rouge représentent les corrélations qui sont considérés fortes. A partir de cette matrice on peut former trois groupes de variables (orange, vert et violet) d'après lesquels on extraira les variables qui peuvent représenter leur groupe selon le coefficient de corrélation. On retient alors les variables suivantes : V13 – V19 – V33 – V24 – V28 – V40.

Nous résumons dans le logigramme suivant les étapes de sélections des variables.

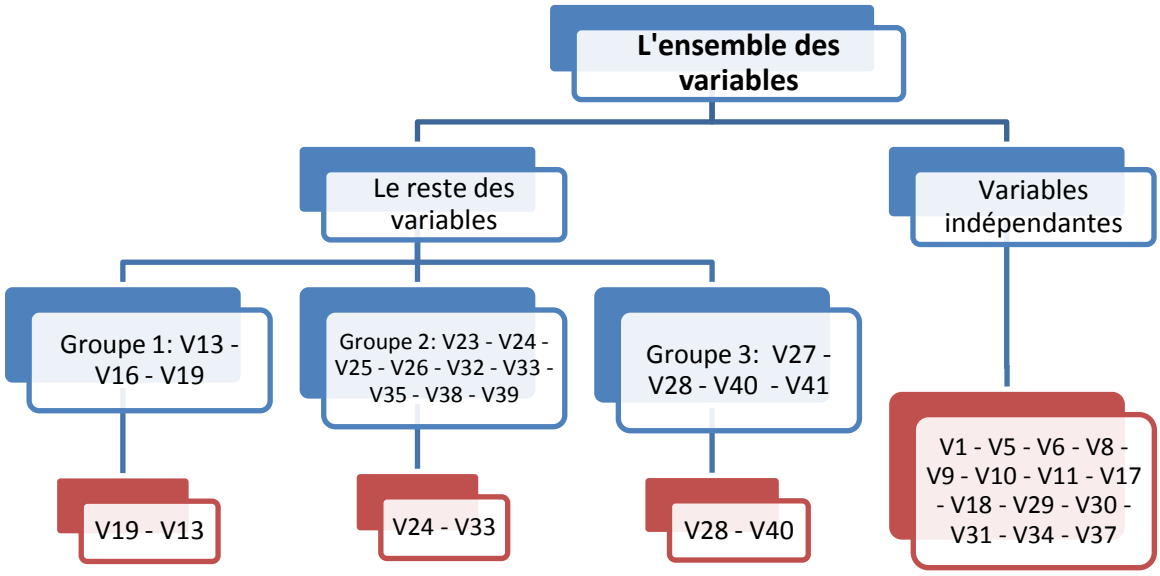



*Figure III.6: Logigramme de sélection des variables*

***Usage de la statistique inférentielle***

Pour mesurer l'influence des variables de type qualitative sur la réponse, on applique le test de Chi-deux à chacune des variables. Le test de chi-deux, comme nous l'avons déjà présenté dans la section III.3.1, permet de tester l'existence d'une influence d'une variable qualitative X sur une autre variable Y.

- Variable V2:

  Le test de chi-deux appliqué sur la variable V2, nous montre qu'il existe une dépendance entre cette variable et la variable " classe " étudiée. Le tableau suivant montre la statistique de chi-deux calculé, ainsi que le degré de liberté ( ddl) et le degré de significativité. Le risque permis est de 5%.

| | |
|---|---|
| Khi-Chi-deux | 189468,753 |
| ddl | 2 |
| Sig. | ,000* |

*Tableau III-5: Test de chi-deux réalisé entre V2 et la variable à expliquer " Classe"*

Le degré de significativité (Sig.) est inférieur à 5%, d'où la variable "classe" est considérée dépendante de V2.

Nous présentons dans l'annexe (Annexe 6), les différents tests de chi-deux réalisés sur l'ensemble des variables qualitatives trouvées dans la base.

Enfin, l'ensemble des variables qui influence la variable " classe " sont: V2 - V3 - V6 - V12 - V14 - V15 - V22

***Usage de l'analyse des composantes principales (ACP)***

Une autre méthode pour réaliser la réduction de la dimension de la base est de d'appliquer l'analyse des composantes principales sur la base KDD-Cup'99. L'ACP permet de créer de nouvelles variables à partir des variables principales ( voir chapitre IV).

En appliquant cette méthode sur l'ensemble des variables quantitatives, nous avons pu réduire l'espace de représentation de 41 à seulement 9 variables tout en gardant quasiment la totalité de l'information apporté par la base originale, ce qui constitue une amélioration remarquable de la qualité des données et donc de l'apprentissage en général.

Les détails de l'application de cette méthode sont fournis dans l'Annexe 7

## *III.5. Synthèse*

Dans ce chapitre, nous avons donné un aperçu général sur les systèmes de détection d'intrusion (IDS), qui représente le domaine choisi pour l'application des méthodes proposées dans cette thèse. Le choix de ce type de données n'était pas au hasard, mais plutôt bien ciblé, vu les anomalies intéressantes qui se trouvent dans les bases de données des IDS dont la plus connue est le déséquilibre.

Nous avons aussi exposé les données avec lesquels cette étude sera réalisée, accomplie un diagnostic et une analyse descriptive de données, pour enfin réaliser un prétraitement de la base.

Le prétraitement réalisé, comprend le nettoyage des données des doublons et aberrances, ainsi que la réduction de la dimension de la base à l'aide des techniques de sélection des variables précédemment exposées.

Le chapitre suivant introduira un axe de recherche particulier dans lequel s'inscrivent nos travaux, à savoir l'application des techniques de sélection des échantillon pour améliorer la performance de la classification.

# *Chapitre IV: Les méthodes d'édition de données*

## *IV.1. Introduction*

Dans de nombreux domaines, l'acquisition massive de données est désormais possible, mais le traitement de ces données est trop souvent confronté au bruit et à la redondance de l'information. Plusieurs travaux ont démontré l'existence d'une relation étroite entre la qualité des données et la performance achevée par un système d'apprentissage automatique. Parmi les aspects qui jouent un rôle primordial dans la détermination du niveau de la qualité d'un ensemble de données, nous retrouvons la taille de l'ensemble étudié ainsi que la pertinence de ces composantes. En effet, plus la taille des données augmente plus la redondance des exemples est élevée, ce qui mène souvent à des mauvaises prédictions du classifieur entraîné. Plusieurs solutions ont été présentées pour faire face à cette anomalie: Certaines, appelé "Méthodes de Sélection des Echantillons", proposent de réduire la taille des exemples en gardant ceux qui sont considérés comme étant informatives ou critiques. D'autres, interviennent au niveau des variables et tâchent à réduire la dimension de la base de données en utilisant les relations qui existent entre ses variables. Ces dernières sont connues sous le nom de " Méthodes de Sélection des Variables".

Dans ce chapitre, nous détaillerons chacune de ces méthodes en proposant les différents algorithmes trouvés tout au long des travaux réalisés dans ce domaine.

## *IV.2. Méthodes de sélection des échantillons*

En 1968, le travail réalisé par Hart [65] a été un déclenchement de plusieurs travaux dans les méthodes de sélection des échantillons (SE). Dans son travail, Hart, a considéré que la qualité de l'apprentissage s'améliore s'il est réalisé sur un ensemble d'exemples bien sélectionnés. Ainsi, le principe des méthodes de sélection des échantillons est de former le plus petit ensemble d'apprentissage constitué par les éléments les plus informatives. D'une manière explicite, vu un ensemble d'entraînement *E*, l'objectif est de trouver un sous-ensemble $E_{critique}$ ( $tq.\ E_{critique} \subset E$ ) formé des échantillons les plus significatifs, c'est à dire, les exemples

qui contiennent l'information la plus pertinente aidant à la résolution du problème de classification. Dans ce sens, les méthodes de sélection des échantillons divisent implicitement l'ensemble de données en deux parties: La première, formée des exemples nommés " Critiques" et qui sont les données qui participent à la formation de la frontière de décision, tandis que la deuxième partie est celle formée par le reste des exemples redondant, qui ne fournissent aucune information durant l'entraînement du classifieur.

A l'apparition de cette idée, le domaine de l'apprentissage automatique a connu le commencement d'une nouvelle ligne de recherche fructueuse dans des divers champs d'application [66]. Plusieurs propositions ont été faites dans le cadre des méthodes de sélection des échantillons dont le point commun entre eux est de définir le critère sous lequel un exemple est considéré "critique".

Généralement, deux grands aspects sont tenus en compte dans le choix des exemples critiques: L'erreur de l'apprentissage [67], [68], [69] et la proximité à la frontière de décision [66], [70], [71], [72]. D'autres algorithmes basés sur les méthodes Boosting proposent de considérer les deux critères dans le choix des exemples critiques [73], [74], [75].

En ce qui suit, nous proposons faire une révision de certains travaux réalisés durant cette dernière décennie, en se limitant sur ceux qui représentent un point de départ pour le développement de nouvelles méthodes, ou ceux qui sont en relation avec cette thèse.

### *IV.2.1. Application de sélection des échantillons sur les réseaux de neurones*

Vers la fin des années 80, de nouveaux travaux sur la SE sont apparus dans la littérature des réseaux de neurones (RN). De tels travaux prenaient comme point de départ la capacité de l'apprentissage des réseaux de neurones à partir des exemples. En effet, les RN ont tendance à utiliser tout l'ensemble d'entraînement afin d'adapter les poids du réseau, sans prendre en considération la contribution de chacun de ses éléments dans le processus de l'apprentissage. Pourtant, dans les cas pratiques, dans lesquels, parfois, le volume des données est grand, l'utilisation de tout l'ensemble d'entraînement s'avère coûteux [76], [77].

Le processus de sélection d'échantillons peut avoir divers objectifs selon le type du réseau de neurones utilisé (MLP, RBF) : Dans le cas d'un MLP, la sélection d'échantillons s'emploie pour entraîner. Elle consiste à trouver le sous ensemble optimal de l'ensemble d'entraînement qui garantit une bonne généralisation. Cependant, dans le cas du réseau RBF,

en plus de ce qui a été antérieurement mentionné, la sélection peut être appliquée afin de construire les centres des RBF.

***SE appliqué au perceptron multicouche:***

Le perceptron multicouche (Multi-Layer Perceptron MLP) est parmi les structures de réseau les plus utilisées au domaine de l'apprentissage automatique. Sa structure est constituée de trois couches essentielles: Entrée, Sortie et couches intermédiaires (dites cachées) [78]. Durant son entraînement, le MLP utilise l'algorithme d'apprentissage nommé rétro-propagation du gradient. Il s'agit toujours de minimiser l'erreur quadratique, en propageant la modification des poids de la couche de sortie jusqu'à la couche d'entrée, donc cet algorithme passe par deux phases [79]:

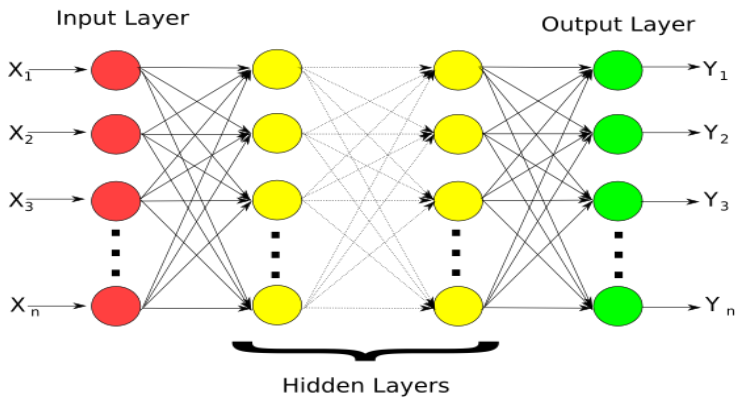



*Figure IV.1: Illustration de la structure d'un MLP*

- Phase 1: Les entrées sont propagées de couche en couche jusqu'à la couche de sortie.
- Phase 2: Si la sortie du MLP est différente de la sortie désirée alors l'erreur est propagée de la couche de sortie vers la couche d'entrée en modifiant les poids durant cette propagation

Les différents travaux concernant l'application des SE au MLP se fait au niveau de la constitution de l'ensemble d'apprentissage. En 1990, Wann fut le premier à proposer l'utilisation du critère du "voisin le plus proche" afin de distinguer les échantillons qui induisent des confusions [80]. La proximité à la frontière est mesurée à partir du nombre des plus proches voisins de la classe opposée. Wann a montré que la taille de l'ensemble d'entraînement, ne garantit pas une bonne généralisation, et qu'il existe un sous ensemble, formé par des échantillons proches aux frontières de classification, qui assure une bonne généralisation. D'autre part, les échantillons proches des frontières ne sont pas tous

nécessaires pour atteindre les meilleures prestations. En suivant la même idée, Ohnishi [81] a montré que la généralisation peut se dégrader à l'usage des exemples de confusion uniquement. Ces résultats montrent bien que la stratégie est très sensible à la complexité du problème ce qui, raisonnablement, ne devrait pas se produire.

En 1992, d'autres travaux ont proposé l'entraînement du MLP en utilisant les exemples les plus difficiles à apprendre. Parmi ces méthodes nous citons Cheung [82] qui a identifié ces exemples comme étant ceux qui ont une erreur quadratique moyenne qui décroit plus rapidement que les autres, il a donc utilisé les exemples les plus difficiles à apprendre sous deux formes: l'ensemble dynamique d'entraînement, et le facteur de pondération. Or, il s'est avéré que ces propositions ont un coût de calcul très élevé. Une autre façon d'entraîner le MLP en utilisant les exemples difficiles à apprendre fut proposé dans [83]. Cette méthode permet de constituer l'ensemble d'apprentissage en éliminant les exemples déjà appris par le réseau. Munro [67] propose aussi dans la même année une stratégie semblable aux précédentes qui consiste à répéter le processus d'apprentissage jusqu'à ce que l'erreur atteigne un certain seuil. Cette méthode reste aussi sensible à la complexité du problème et à la valeur du seuil d'erreur toléré.

Un autre type d'application du SE est l'usage de la méthode de sélection croissante proposée par Zhang [84], [85], [86], [87]. Elle consiste à rajouter au cours de l'entrainement les exemples qui ont une erreur quadratique élevée à un ensemble choisi aléatoirement.

Finalement, en 1994 Cachin a proposé un ensemble de stratégies nommées " Les stratégies Pédagogiques de Selection des Echantillons" [68] qui ont comme principe d'éviter l'entraînement seulement avec les échantillons qui génèrent une grande erreur et de rafraîchir l'apprentissage avec les exemples qui produisent de faibles erreurs vu l'importance de leurs présence dans l'ensemble d'entraînement.

### *Construction du réseau RBF à l'aide de SE*

Les Réseaux de fonctions de base radiaires (RBF) est un réseau de neurones supervisé. Il s'agit d'un cas spécial du MLP. Un RBF est constitué uniquement de 3 couches

- La couche d'entrée : elle retransmet les entrées sans distorsion.
- La couche RBF : couche cachée qui contient les neurones RBF.
- La couche de sortie : simple couche qui contient une fonction linéaire.

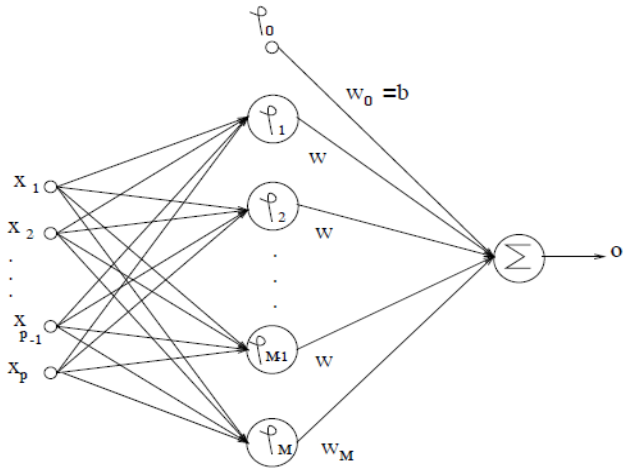



*Figure IV.2: schéma de la structure du réseau RBF*

Chaque neurone RBF contient une gaussienne qui est centrée sur un point de l'espace d'entrée. Pour une entrée donnée, la sortie du neurone RBF est la hauteur de la gaussienne en ce point. La fonction gaussienne permet aux neurones de ne répondre qu'à une petite région de l'espace d'entrée, région sur laquelle la gaussienne est centrée. Donc il y a quatre paramètres principaux à régler dans un réseau RBF.

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

Toute modification d'un de ces paramètres entraîne directement un changement du comportement du réseau, ainsi que sa capacité à la généralisation. Les méthodes de SE interviennent au niveau de la détermination du deuxième paramètre du RBF. En effet, les échantillons sélectionnés s'utilisent comme centres du réseau. Dans [88], Chang, propose une méthode de SE afin de construire progressivement le réseau RBF qui utilise les exemples situés au bord de la frontière de décision comme centres.
Lyhyaoui et *al.* [71] offrent dans leurs travail une alternative pour construire un classifieur RBF en développant un nouveau concept de SE et cela en se basant sur le principe d'agroupement afin de déterminer les centres du RBF.

### *IV.2.2. Sélection des échantillons appliquée au SVM*

Les machines à vecteurs de support (Support Vector Machine, SVM) appelées aussi séparateurs à vaste marge sont des techniques d'apprentissage supervisées destinées à

résoudre des problèmes de classification. Ces techniques est une méthode de classification à deux classes qui tente de séparer les exemples positifs des exemples négatifs. La méthode cherche alors l'hyperplan qui sépare les deux classes, en garantissant que la marge entre le plus proche des positifs et des négatifs soit maximale ( voir figure suivante). L'intérêt de cette méthode est la sélection des vecteurs supports qui représentent les vecteurs discriminants grâce auxquels l'hyperplan est déterminé. Les exemples utilisés lors de la recherche de l'hyperplan ne sont alors plus utiles et seuls ces vecteurs supports sont utilisés pour classer un nouveau cas, ce qui peut être considéré comme un avantage pour cette méthode [89].

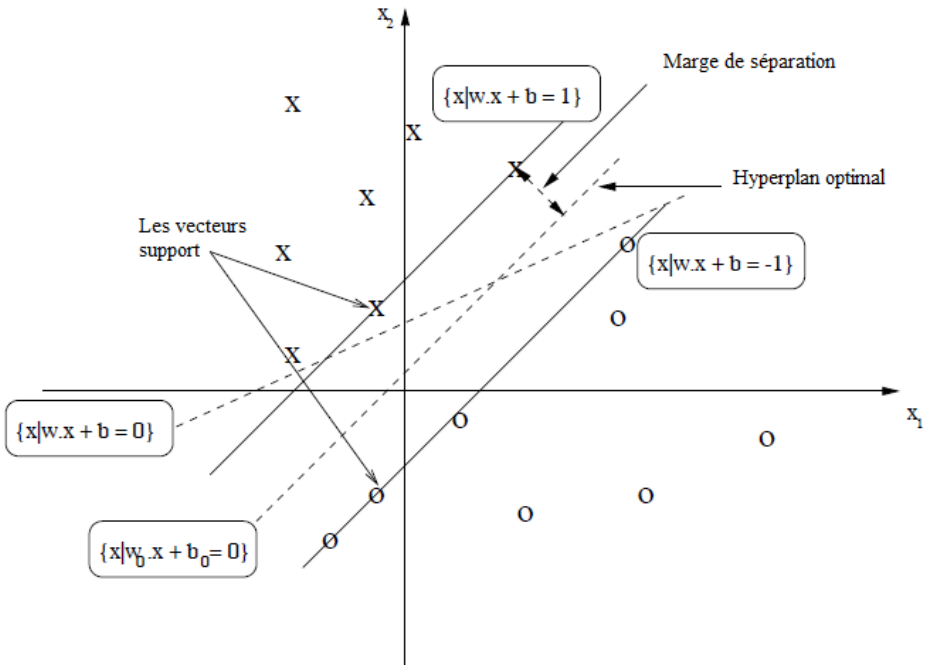



*Figure IV.3: Recherche d'hyperplan optimal*

Plusieurs travaux ont proposé l'application de SE pour réduire la complexité de calcul du classifieur SVM. Almeida [90] a développé une procédure de sélection des échantillons basée sur la technique d'agroupement k-means. Dans son travail, il propose d'utiliser les centroids des clusters qui ne contiennent qu'une seule classe au lieu de travailler avec tous les exemples. De façon, le coût de calcul du SVM est nettement réduit.

Une autre procédure de SE, proposée par Shin en 2002, se base sur la technique des k plus proches voisins (KNN) [91] où la sélection des exemples se fait selon deux critères: La proximité à la frontière et la probabilité de la bonne classification. Les exemples sélectionnés sont par la suite utilisés comme des vecteurs supports du classifieur. En 2003, le même auteur

a développé son propre travail dans [92] où il a proposé de se limiter aux exemples les plus proches à la frontière pour construire l'ensemble d'entraînement du SVM.

### *IV.2.3. Application de SE sur le Boosting*

L'idée principale du *Boosting*, en apprentissage automatique, était d'améliorer les compétences d'un classifieur, supposé a priori instable, appelé *weak learner*. La méthode originale de Shapire [93], améliorée par Freund et lui même [94], décrit AdaBoost *(Adaptative boosting)*, l'algorithme de base du *Boosting*, pour la prédiction d'une variable binaire. Le principe du *Boosting* est de construire un ensemble de classifieurs qui sont ensuite agrégés par une moyenne pondérée des résultats ou un vote. Les classifieurs sont construits d'une façon récurrente et itérative de manière à ce que chaque classifieur soit une version adaptative du précédent en donnant plus de poids aux exemples mal prédits. Ce procédé permet à l'algorithme de se concentrer sur les exemples les plus difficiles à classifier. L'agrégation de classifieurs permet au *Boosting* d'échapper au sur-apprentissage. Cette méthode réduit à la fois la variance et le biais. L'algorithme du AdaBoost peut être résumé comme suit:

**Entrés** $S_1 = \{(x_1, y_1), ...., (x_N, y_N)\}$.

**Initialisation:** $D_1(i) = \frac{1}{N}$ Tous les exemples ont le même poids

**Pour** $t$ =1,...,$T$ **faire :**

- Entraîner le classifieur faible $h$ selon la distribution des poids $D_t$
- Calculer l'erreur de classification du classifieur $h_t$ hypothesis

  $e_t = \sum_{i:h_t(x_i) \neq y_i} D_t(i)$
- Calculer $\alpha_t = \frac{1}{2} \ln (\frac{1-e_t}{e_t})$
- Mettre à jour la distribution des poids $D_{t+1}(i) = \frac{D_t(i)}{Z_t} \exp(-\alpha_t y_i h_t(x_i))$ . Où $Z_t$ est le facteur de normalisation.

**Fin Pour**

**Sortie:** $H(x) = sin(\sum_{t=1}^{T} \alpha_t \, h_t(x))$

Initialement, chaque exemple possède le même poids, égale à *1/N*. Après la première classification, le poids des exemples évoluent à chaque itération et pour chaque nouvelle

classification. Le poids d'un exemple $D_t$ est inchangé s'il est bien classé, au cas contraire, son poids croît d'une façon exponentielle.

Dans [74] et [75] les auteurs ont développé une nouvelle fonction de mise à jour de la distribution des poids dans la procédure du Real AdaBoost, cette fonction est définie par

$$D_{t+1,\lambda}(i) = \frac{1}{Z_{t,\lambda}} \exp\left[\lambda(f_t(x_i) - y_i)^2 + (1-\lambda) f_t(x_i)^2\right] \quad \textit{(IV.1)}$$

Où

- $f_t(x_i)$ est la sortie du classifieur partiel
- $Z_{t,\lambda}$ la normalisation de la distribution.

A partir de la formule *(IV.1)* précédemment décrite, nous pouvons remarquer que cette méthode permet de définir le niveau d'attention prêtée à chacun des deux critères : La proximité à la frontière exprimée par le terme $f_t(x_i)^2$ et l'erreur de classification (difficulté d'apprentissage) représenté par $(f_t(x_i) - y_i)^2$. Et cela, en utilisant le paramètre de pondération $\lambda$ qui varie entre 0 et 1.

Une autre approche de l'application du SE avec l'algorithme du Boosting, est d'utiliser la combinaison de plusieurs classifieurs classiques comme SVM, KNN et C4.5 dans la technique de AdaBoost [95]. La méthode proposée consiste à commencer l'apprentissage avec tout l'ensemble d'entraînement et d'éliminer au fur et à mesure les exemples les moins difficiles à apprendre en se basant sur la nouvelle distribution des poids. De cette manière, l'entraînement est focalisé sur les exemples qui permettent de mieux minimiser l'erreur pondérée.

## *IV.3. Méthodes de sélection des variables*

Après avoir vu, la possibilité de manipuler l'espace des individus en modifiant l'échantillonnage pour améliorer la performance des classifieurs, nous nous intéressons dans cette partie à l'espace de représentation (ou dimension). La qualité d'un apprentissage est entre autres choses liée à la présence de variables discriminantes. Or, dans le cas d'une qualité d'apprentissage insuffisante, il est nécessaire de trouver un moyen qui, à partir de l'information disponible, permet de ré-décrire les données d'entrée du problème en obtenant

un ensemble de variables discriminantes. Les méthodes de sélection des variables résolvent ce problème. Elles permettent d'arriver, à partir d'un ensemble de variables explicatives, à un modèle final qui retiendrait le plus grand nombre de variables explicatives qui sont significatives dans l'explication de la variable dépendante. Cela permettra une meilleure prise en compte des facteurs de confusion potentiels.

Les méthodes de sélection des variables (ou sélection des caractéristiques) peuvent être divisées en deux catégories: La première, rassemble les méthodes **d'élimination de variables** non pertinentes de l'ensemble de données. Elle rassemble des techniques permettant de supprimer les variables qui sont considérées comme non-discriminantes tous en évitant la perte de l'information globale. La deuxième catégorie est celle des méthodes de la **construction de variables**. Elle permet la création de nouvelles variables synthétiques. Ces variables synthétiques sont issues de la découverte des relations entre les variables initiales. Cependant, aucune information extérieure aux données initiales n'est ajoutée lors du processus de construction.

Nous présentons par la suite une partie des différentes méthodes développées dans ce sens, ayant une relation avec notre sujet traité.

### *IV.3.1. Réduction de dimension par élimination des variables*

Cet ensemble de méthodes de sélection des variables se base sur l'étude uni-variée des caractéristiques de manière à pouvoir distinguer ceux qui sont considérées les plus pertinentes de ceux qui n'apportent aucune information utile pour la construction du classifieur. Cette distinction est réalisé en utilisant différentes techniques statistiques citant, le calcul du coefficient de corrélation ou de l'information mutuelle, distance ou indépendance [96], [97].

Avant d'évoquer les différentes méthodes de sélection de variables, commençons par définir le concept de la pertinence. Selon [98] une variable est considérée pertinente si ses valeurs varient systématiquement avec les attributs de la variable à expliquer (classe), dans [99], l'auteur considère une variable pertinente si la probabilité conditionnelle d'une classe dépend de cette variable. C'est à dire que X est dite pertinente si:

$$\exists\, \boldsymbol{x, y\ où\ P(X = x) > 0, t.q. P(Y = y/X = x) \neq P(Y = y)} \quad \textit{(IV.2)}$$

Où, Y est la variable classe, et x et y sont respectivement des valeurs de X et Y.

### *Technique de mesure de distance inter-classe*

Cette technique considère la mesure de la pertinence d'une variable X en utilisant la distance calculée entre les attributs provenant de différentes classes [100], les meilleurs variables sont ceux qui maximisent la distance inter-classe J

$$J = \sum_k P(y_k) \sum_l P(y_l) D(y_k, y_l) \qquad (IV.3)$$

Où $y_k, y_l$ représente respectivement la k$^{ème}$ et la $l^{ème}$ étiquette de la variable classe Y, et

$$D(y_k, y_l) = \frac{1}{N_K N_l} \sum_i^{N_k} \sum_j^{N_l} d(x_{i,k}, x_{j,l}) \qquad (IV.4)$$

Avec

- $N_K \; et \; N_l$ sont le nombre des attributs de la variable X appartenant respectivement aux classes *k* et *l.*
- $x_{i,k}$ est le i$^{ème}$ attribut appartenant à la classe *k*
- $x_{j,l}$ est le j$^{ème}$ attribut appartenant à la classe *l*
- $d(x_{i,k}, x_{j,l})$ est la distance euclidienne entre les deux attributs

### *Calcul de la corrélation*

Cette mesure statistique permet de chiffrer l'intensité de la liaison qui peut exister entre chaque variable X et la classe Y. Ce paramètre n'est utilisé que dans le cas des variables quantitatives, il s'agit d'une mesure symétrique de manière que plus elle est proche de 1 (en valeur absolue), plus la relation entre les variables en question est forte. La mesure du coefficient de corrélation la plus utilisée est celle définie par:

$$Cor(X, Y) = \frac{\mathbb{E}\left[(X - \mu_x)(Y - \mu_y)\right]}{\sigma_x \sigma_y} \qquad (IV.5)$$

Avec,

- $\mu_x$ et $\mu_y$ les moyennes des distributions des variables X et Y respectivement,
- $\sigma_x$ et $\sigma_y$ sont les écart-types de X et Y respectivement.

La corrélation peut être estimée depuis l'ensemble d'entraînement (supposons de taille *n*) par la relation suivante:

$$r_{XY} = \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{(n-1)\varsigma_X \varsigma_Y} \qquad \text{(IV. 6)}$$

Où

- $x_i$ et $y_i$ sont les attributs des variables X et Y respectivement
- $\overline{X}$ et $\overline{Y}$ sont les moyennes de X et Y
- $\varsigma_X$ et $\varsigma_Y$ représentent les écart-types des deux variables étudiées.

La mesure de la corrélation peut être utilisée de plusieurs façons, en effet, à part son usage traditionnel pour classer les variables, une deuxième application est étudiée dans [101], où une mesure heuristique pour un sous-ensemble de variables est proposée, selon laquelle l'ensemble qui contient les variables montrant une grande corrélation avec la classe et une faible inter-corrélation est considéré comme pertinent. La mesure de la pertinence d'un sous-ensemble de variables F' est calculée comme suit:

$$M_{F'} = \frac{k\,\overline{r_{F'Y}}}{\sqrt{k + (k-1)\overline{r_{F'F'}}}} \qquad \text{(IV. 7)}$$

Où

- $\overline{r_{F'Y}}$ est la moyenne de tous les coefficients de corrélation entre les variables de F' et la classe Y
- $\overline{r_{F'F'}}$ est la moyenne des coefficients de corrélation des variables de F' entre eux (inter-corrélation)
- k est le nombre des variables sélectionnées ($k = \mathrm{card}(F')$).

***Méthode de Chi2 ($\chi^2$)***

Il s'agit d'une mesure statistique bien connue, qui s'adapte bien à la sélection des variables. Elle évalue le manque d'indépendance entre une variable X et la variable classe Y.

Le calcul de Chi2 nécessite de construire le tableau de contingence pour chaque variable X.

| | Attribut $x_1$ | Attribut $x_2$ | Total |
|---|---|---|---|
| Classe $y_1$ | A | c | a+c |
| Classe $y_2$ | B | d | b+d |
| Total | a+b | c+d | N=a+b+c+d |

*Tableau IV-1: Tableau de contingence d'une variable X et la variable classe Y*

La statistique du $\chi^2$ peut être calculée sous la forme :

$$\chi^2 = \frac{N(ad - cb)^2}{(a + c)(b + d)(a + b)(c + d)} \quad \text{(IV.8)}$$

Le test de $\boldsymbol{\chi^2}$ se base sur la comparaison entre les valeurs théoriques déduites du modèle et les valeurs observées dans l'échantillon. Il permet ainsi de tester l'hypothèse $H_0$ de l'indépendance entre les deux variables X et Y.

### *IV.3.2. Réduction de la dimension par la construction de variables*

L'inconvénient des méthodes de réduction de dimension par élimination de variables, est qu'elles ne parviennent pas à détecter les caractéristiques redondantes qui peuvent détériorer la qualité de l'apprentissage d'un classifieur [102]. Par conséquent, la plupart des techniques de sélection de variables se concentrent sur la recherche du meilleur sous-ensemble de caractéristiques pertinentes ou de former de nouvel variables en combinant plusieurs autres. En effet, il faut noter que des caractéristiques individuellement peu pertinentes peuvent être très informatives si on les utilise conjointement.

Par la suite, nous citons quelques méthodes qui permettent de réduire la dimension des données en créant de nouvelles variables à partir de celles existantes.

***Analyse en Composantes Principales (ACP)***

L'analyse en composantes principales (ou ACP), est une technique d'analyse de données utilisée pour réduire la dimension de l'espace de représentation des données. Le principe de cette méthode est de construire de nouvelles variables appelées composantes à partir des caractéristiques initiales permettant de maximiser la variance [103]

Soient p variables quantitatives $X_1, X_{2,\dots,} X_p$ mesurées sur un ensemble de n individus. Les données obtenues sont présentées sous la forme d'un tableau ou matrice M à n lignes et p colonnes :

$$M = \begin{bmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{bmatrix}$$

Le problème posé consiste à réduire les *p* variables initiales en un nombre *q* plus petit de variables « composées », ou facteurs appelées encore composantes principales. Il s'agit donc de passer de la matrice des données initiales M (*n* individus et *p* variables) à une matrice réduite *M'* (*n* individus et *q* variables). Ces facteurs doivent répondre aux deux conditions: La linéarité et l'indépendance. Ils doivent aussi restituer le maximum de l'information contenue dans le nuage de points, c'est à dire l'inertie du nuage projeté, qui est la moyenne pondérée des carrées des distances des points. La maximisation de l'inertie permet de préserver au mieux la répartition des points. Il s'agit donc d'un problème classique d'optimisation sous contrainte résolue par la méthode de Lagrange.

Les composantes des vecteurs propres (issues de la matrice de variance-covariance empirique des *p* variables) sont alors les coefficients générateurs des *q* nouvelles variables composées, appelées aussi les *q* premières composantes principales des variables *X,* qui peuvent être utilisées en tant que nouvelles variables du modèle.

Depuis les années 2000, les méthodes d'analyse de composantes principales a connu beaucoup de développement spécialement dans le domaine de reconnaissance facial [104] où les données sont caractérisées par un grand nombre de caractéristiques.

### *Analyse Factorielle des Correspondances (AFC)*

L'analyse des correspondances, présentée sous ce nom et développée par Benzécri en 1969, a un certain nombre de précurseurs, parmi lesquels il faut citer Guttman (1941) et Hayashi (1956).

L'AFC peut être présentée selon divers points de vue. Il est d'ailleurs difficile de faire l'historique précis de cette méthode. Les principes théoriques remontent probablement aux travaux de Fisher sur les tables de contingences, dans un cadre de statistique inférentielle classique. Depuis les travaux de Benzécri [105] et de Escoér-Cordier (1965), on utilise surtout les propriétés algébriques et géométriques de l'outil descriptif que constitue l'analyse. On peut présenter l'AFC comme un cas particulier de l'analyse canonique lorsque les données ont un codage "disjonctif" et également comme un cas particulier de l'analyse discriminante.

On peut enfin montrer qu'il s'agit de la recherche de la meilleure représentation simultanée de deux ensembles constituant les lignes et les colonnes d'un tableau de données positives. Dans le cas d'une AFC les données sont transformées afin de mettre en évidence la répartition relative de l'individu par rapport aux variables et d'établir les corrélations entre les profils obtenus. En fait, l'AFC correspond à une ACP réalisée sur un tableau qui a subi un

traitement supplémentaire, en prenant en compte la marginale des variables et des individus. Ce traitement permet de représenter les individus et les variables dans le même espace (les vecteurs propres de la matrice $n \times p$ et sa transposée sont identiques). Elle permet donc d'observer la typologie des individus par rapport aux variables ou la typologie des variables par rapport aux individus [105].

## *IV.4. Synthèse*

Dans ce chapitre nous nous sommes focalisés sur le problème de traitement des données à grandes masses qui présentent certaines anomalies comme la redondance, le manque d'information, le bruit ou la non-pertinence. De nombreux travaux ont montré la relation étroite qui existe entre la qualité des données et la performance de la classification. Dans ce sens, plusieurs méthodes de traitement de données ont été développées afin de faire face à ces anomalies retrouvées dans les données réelles.

Ces techniques peuvent être classées en deux types: les méthodes de sélection des échantillons (ou attributs), et les méthodes de sélection des variables (ou caractéristique). Le premier ensemble traite les exemples, et réduit la taille des attributs en sélectionnant ceux qui sont considérés critiques, ou de caractère discriminant. tandis que le deuxième type de méthode, se concentre sur la réduction de dimension au niveau des variables en gardant les plus pertinentes, ou en créant de nouvelles variables informatives à partir d'un ensemble moins pertinent.

Dans les chapitres suivants, nous allons présenter les méthodes proposées dans cette thèse ainsi que leurs application dans le domaine de détection des intrusions.

# *Chapitre IV: Sélection des échantillons pour l'équilibrage des données*

## *V.1. Introduction*

Comme nous avons précédemment évoqué, plusieurs techniques ont été développées pour traiter l'anomalie de la dégradation de performance de la classification des données non-équilibrées. D'une part, il y a les méthodes qui s'intéressent uniquement au problème du déséquilibre, et d'autre part, ceux qui, en utilisant des méthodes d'édition, permettent d'accroitre la qualité de la classification.

Au cours de cette thèse, nous avons réalisé plusieurs approches qui seront exposées dans ce chapitre. Généralement, ces méthodes ont un point commun qui est l'utilisation de la sélection des échantillons comme moyen d'équilibrage des données d'entraînement, pour un but précis: accroitre la performance du classifieur utilisé. Dans un premier temps, nous avons utilisé l'erreur d'entraînement comme critère qui permettra de choisir les exemples considérés critiques. Ensuite, nous avons développé des techniques géométriques en utilisant le Clustering afin de déterminer les échantillons les plus proches à la frontière de décision.

## *V.2. Méthode 1: Le sous-échantillonnage à l'aide de l'erreur d'apprentissage*

### *V.2.1. Principe du sous-échantillonnage ciblé sous critère d'erreur*

Les méthodes du sous-échantillonnage (ou sur-échantillonnage) sont considérées parmi les solutions les plus utilisées pour faire face au problème du déséquilibre entre les classes. En général, le sous-échantillonnage consiste à éliminer un certain nombre d'exemples appartenant à la classe majoritaire afin de rétablir l'équilibre entre les classes de la base de données.

La méthode du sous-échantillonnage proposée dans cette thèse, consiste à utiliser la technique de sélection des échantillons basée sur l'erreur d'apprentissage, pour spécifier les exemples à éliminer, et donc à *cibler* les exemples qui contribuent le moins dans la définition

de la frontière de décision. Contrairement au sous-échantillonnage aléatoire, les exemples éliminés sont ceux qui génèrent une erreur de classification inférieure à un seuil déterminé. De cette manière, le classifieur évite le problème de la perte des exemples critiques appartenant à la classe majoritaire [4].

### *V.2.2. Sous-échantillonnage ciblé appliqué au MLP*

La technique du sous-échantillonnage ciblé peut être intégrée dans les différents systèmes d'apprentissage grâce à sa simplicité en implémentation. Dans notre cas nous nous intéressons au perceptron multi-couches (MLP) vu son coût de calcul modeste. Nous résumons dans l'algorithme suivant l'usage du sous-échantillonnage ciblé dans la structure MLP.

Signalons que l'ensemble de la classe majoritaire et minoritaire sont notés respectivement par ($D_{maj}$) et ($D_{min}$). $D_{-new}$ représente le nouvel ensemble d'entraînement

**Entrées:** - *th* le seuil de l'erreur

- $N_{max}$ nombre maximal d'itérations

**Initialisation** aléatoire des poids des neurones

**Faire :**

- **Faire**
  - Entraîner MLP utilisant $D=\{ D_{min}, D_{maj}\}$
  - Calculer l'erreur de classification du classifieur
  - $D_{-th} = \{ x \in D_{maj} \mid e\_tr(x) < th\}$;
  - $D_{-new} = D_{maj} - D_{-th}$;
  - $D_{maj} = D_{-new}$;

  **jusqu'à** convergence du MLP

- validation du classifieur
- Calcul de l'erreur de validation

**Tant que** (erreur de validation décroit **ou** nombre itération $\leq N_{max}$ )

**Sortie:** Poids du MLP adaptés à la classification pour base de données non équilibrées

Comme nous pouvons le constater à partir de l'algorithme cité dessus, la sélection des échantillons se fait progressivement au cours de l'entraînement du MLP. En effet, le classifieur entraîne initialement toute la base de données, et élimine au fur et à mesure des exemples à

partir de la classe majoritaire, effectuant ainsi un sous-échantillonnage des éléments ayant une erreur de classification inferieur au seuil déterminé au début du processus.

L'usage de cette méthode nous permet de faire face au problème du déséquilibre en diminuant la taille de l'ensemble des données, permettant ainsi de diminuer les coûts de calcul et améliorer la performance de la classification.

### *V.2.3. Résultats de l'application de la méthodes proposée sur KDD-Cup'99*

Nous présentons la variation de la précision en fonction du seuil de l'erreur retenues. Nous avons tester quatre valeurs du seuil d'erreur d'apprentissage.

L'erreur de l'apprentissage varie en valeur absolue entre 0 et 2. Les valeurs retenues pour l'expérience sont $\{0.5, 1, 1.5, \text{moyenne}\}$, où "moyenne" est la moyenne du vecteur d'erreur.

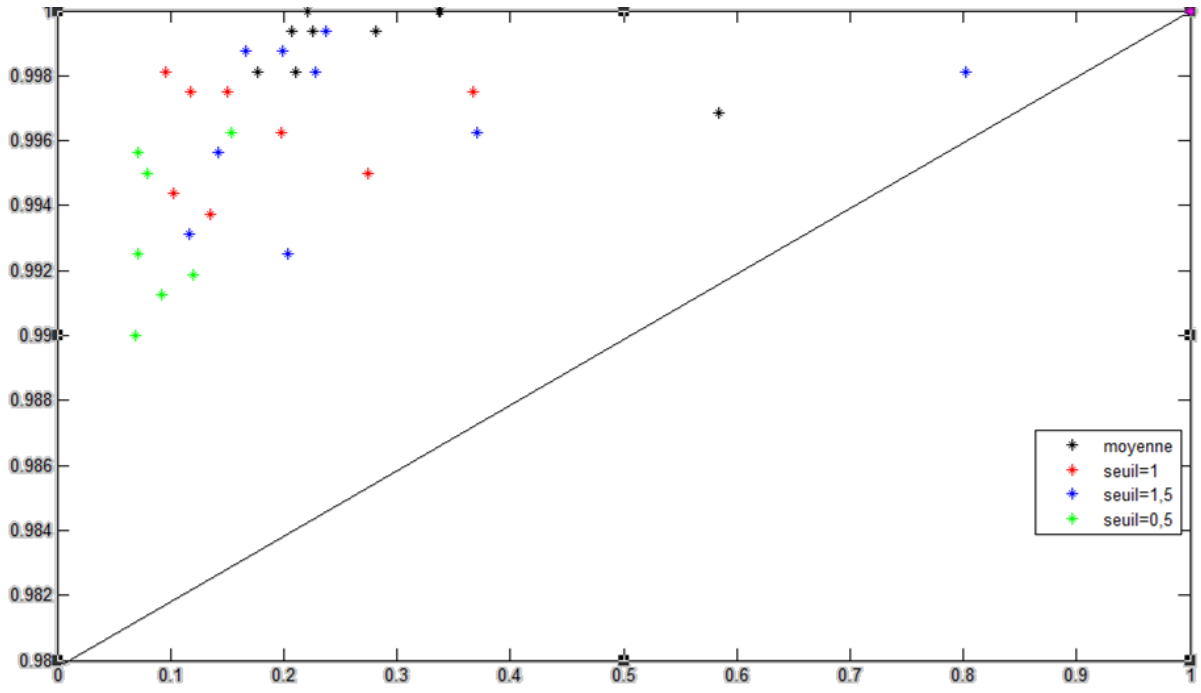



*Figure V.1Le graphe ROC de classification en fonction de variation du paramètre seuil $\delta$*

Nous avons présenté dans la figure ci-dessus, les différentes valeurs ROC du couple ( taux de faux positive, rappel ). Cette mesure nous permet de voir la performance de la classification au niveau de la classe minoritaire. Comme le montre la figure, la qualité de classification de la classe minoritaire est très bonne vu que le rappel, qui est le taux des vrais positives est généralement supérieur à 99%.

Nous pouvons constater que le seuil de l'erreur d'apprentissage influence légèrement sur la classification. Plus le seuil est élevé plus la qualité de classification est meilleure.

Cette approche nous a permis d'augmenter la qualité du classifieur MLP en augmentant le taux de la bonne classification de la classe minoritaire. La précision générale de la classification est de **95%.**

## *V.3. Méthode 2: Sous-échantillonnage à l'aide de la distance par rapport à la frontière*

Apres avoir élaboré une méthode permettant de choisir les exemples critiques à l'aide du critère de l'erreur d'apprentissage, nous nous sommes intéressés à un autre critère utilisé dans la littérature de sélection des échantillons: Il s'agit de la proximité de la frontière de décision.

### *V.3.1. Principe de la méthode*

Comme nous avons vu dans le chapitre précédent, les méthodes de sélection des échantillons sont des techniques de réduction de données qui permettent d'améliorer la performance d'un classifieur en sélectionnant les exemples les plus significatifs dans l'ensemble d'entraînement [6].

Le point commun de toutes les stratégies de SS exposées, est le critère qui nous permet de définir lesquels des échantillons sont critiques. Ce critère est la proximité à la frontière de classification.

La première approche proposée dans ce travail consiste à utiliser ce critère (la proximité) pour réaliser un sous-échantillonnage de la classe majoritaire. Pour ce faire, nous avons opté pour l'utilisation des clusters plutôt que les exemples afin de réduire le coût du calcul. Cette idée n'est pas nouvelle vu qu'elle a été utilisé par Kohonen dans [106] où il expose le classifieur supervisé LVQ

#### *Clustering à l'aide de Quantification vectorielle et application du LVQ(3)*

La méthode de quantification vectorielle (VQ) fait partie de la famille des réseaux compétitifs. Elle permet de réduire énormément la complexité du calcule de la classification en désignant des centres qui permettent de représenter la distribution générale des données.

L'entraînement du réseau se fait d'une façon adaptative à la manière suivante : étant donné un ensemble de centres $\{c_i, i = 1,2, \dots, m\}$, l'actualisation du i$^{ème}$ centre dans l'itération (k+1)$^{ième}$ est comme suit

$$\mathbf{c}_i^{(k+1)} = \begin{cases} \mathbf{c}_i^{(k)} + \alpha^{(k)} \left[\mathbf{x}^{(k)} - \mathbf{c}_i^{(k)}\right] & \text{si } \mathbf{c}_i^{(k)} \in N_k \\ \mathbf{c}_i^{(k)} & \text{si } \mathbf{c}_i^{(k)} \notin N_k \end{cases} \qquad (V.1)$$

Avec

- $x^{(k)}$ représente l'attribut présenté au réseau dans l'itération k
- $N_k$ est un domaine de $x^{(k)}$ préalablement définie
- $\alpha^{(k)}$ est le pas d'adaptation de l'algorithme

Cet algorithme présente des problèmes d'initialisation : si un centre (ou plusieurs) est initialisé vers une zone où se trouvent des entrées à haute probabilité, il existe le risque que ce poids gagne répétitivement face à ces entrées ; laissant ainsi les autres poids sans aucune activité significative. Ceci provoque un comportement du réseau inadéquat

Une solution efficace pour ces problèmes d'initialisation, consiste à réduire la capacité de victoire des centres qui gagne d'une manière répétitive. Une méthode qui a montré de bons résultats est la FSCL (Frequency Sensitive Competitive Learning) [107].

L'algorithme FSCL compte le nombre de fois que chaque poids gagne, et utilise cette information pour garantir le fait que durant le processus d'entraînement, tous les poids puissent s'actualiser approximativement le même nombre de fois. Pour cela, on introduit une nouvelle mesure de distorsion entre les données et les centres, afin de déterminer le gagnant.

$$\boldsymbol{d^*(x, c_i) = d(x, c_i).u_i} \quad (V.2)$$

Où $u_i$ est le nombre de fois que le poids $c_i$ a gagné.

Cet algorithme offre une grande équiprobabilité de victoire des centres face à une entrée déterminée. Une modification de l'algorithme de cette méthode a été proposé par Lyhyaoui *et al.* [71], permettant d'attribuer à chaque classe un ensemble initiale de centres, de manière à rendre l'algorithme supervisé. On ajoute un entraînement supervisé après l'auto-organisation.

Dans ce cas, les poids sont groupés en sous-ensembles ; chacun représente une classe. Les valeurs des poids sont définies de sorte que ces derniers déterminent directement la frontière de décision entre classes. Ces stratégies sont appelées Learning Vector Quantisation (LVQ).

LVQ3 est la dernière adaptation de l'algorithme LVQ, Elle permet que le centre cj continue à s'approcher de la distribution de sa classe, l'algorithme LVQ3 ajoute une correction dans le cas où les deux centres les plus proches de l'échantillon appartiendraient à la même classe :

$$c_j^{k+1} = c_j^k + \varepsilon . \alpha(k)[x^k - c_j^k] \quad (V.3)$$

Avec $\varepsilon$ est un paramètre dépendant de la distance entre les deux centres

### *Détermination des centres critiques: Les paires opposées les plus proches*

Apres avoir déterminé les centres correspondant à chaque classe, l'étape suivante consiste à trouver lesquels de ces centres sont signifiante dans la détermination de la frontière de décision.

Pour cela, nous utilisons la technique de détection des paires opposées les plus proches (POPP) proposée par Sklansky dans [108]. Cette méthode est caractérisée par le fait de pouvoir sélectionner les centres critiques à partir de l'ensemble des centres déterminés, sans avoir besoin d'un classifieur préalablement entraîné.

soient $c_1$ et $c_2$ deux centres de classes appartenant à deux classes différentes. Le couple ($c_1$, $c_2$) est dit " paires opposée ". Pour que cette paire opposée soit définie comme étant la plus proche, il faut qu'elle vérifie la condition suivante:

$$D(c_1 , c_2) = \min_{c_k \in C_1} D(c_1, c_k) = \min_{c_k \in C_2} D(c_2, c_k) \quad (V.4)$$

où $D(c_1 , c_2)$ est la distance euclidienne entre $c_1$ et $c_2$. Et $C_1$, $C_2$, sont les deux classes d'où proviennent les centres $c_1$ et $c_2$. Depuis la formule V.4, on dit que le couple $(c_1 , c_2)$ est une paire opposée la plus proche si $c_1$ est plus proche que tous les centres de la classes $C_1$ à $c_2$ et vice-versa [108].

La figure suivante représente ce principe

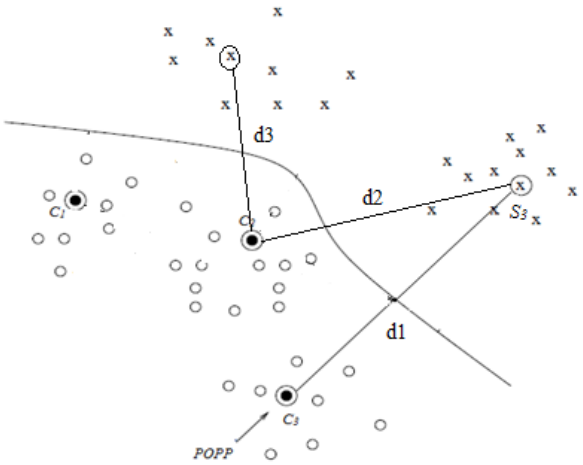



*Figure V.2: Illustration des paires opposées les plus proches*

La figure dessus montre que le couples ( $C_3$, $S_3$) constitue une POPP, même si $d_2 < d_1$, cela est parce que $C_2$ est plus proche à $S_2$ qu'à $S_3$.

### *Construction du nouvel ensemble d'entraînement*

Pour l'instant nous avons réalisé un clustering des classes et trouvé les centre critiques pour chaque classe. L'étape suivante est de former l'ensemble d'entraînement du classifieur.

Vu que notre étude traite le déséquilibre entre les classes, nous proposons donc d'appliquer un sous-échantillonnage de la classe majoritaire en se basant sur les centres de classes plutôt que les exemples.

Nous proposons de créer un nouvel ensemble d'entraînement *E*, équilibré, en éliminant les exemples qui n'appartient pas au clusters critiques de la classe majoritaire considérant ces exemples comme étant non pertinente. En effet, *E* sera la union entre l'ensemble de tous les exemples de la classe minoritaire et les exemples appartenant au clusters critique de la classe majoritaire

$$E = (CC_{maj} \cup C_{min}).$$

La méthode proposée a permis d'avoir une amélioration notable de la performance de la classification des données non-équilibrées [5]. Nous détaillerons les résultats trouvés dans le chapitre 6 de cette thèse.

### *V.3.2. Application de la méthode sur le MLP*

En appliquant cette approche sur le classifieur MLP, nous pourrons explorer les paramètres $nA$ et $nB$ , représentant le nombre des classe minoritaire et majoritaire respectivement , ainsi que $n$ le nombre des neurones.

- Exploration des centres de la classe majoritaire avec nombre de classe minoritaire fixé en 10.

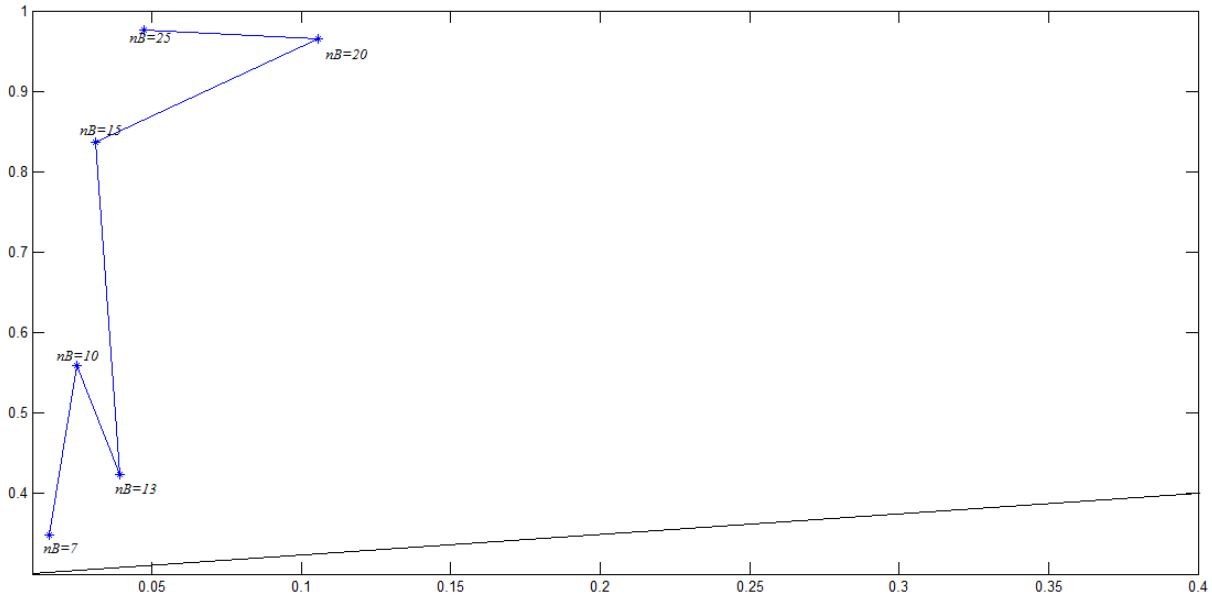



*Figure Figure V.3 Graphe ROC pour éxploration des centres de la classe majoritaire nB*

En fixant le nombre de centre de la classe minoritaire en 10, nous testons la variation de la qualité de classification en fonction de nombre des centres de la classe majoritaire. Cette variation est exposée dans la figure *V.3* présentée ci-dessus.

Ces résultats montrent que le nombre des centres *nB* a une influence significative sur la qualité de classification de la classe majoritaire. Le taux des vrais positif est de moins de 40% pour $nB = 7$, tandis qu'il est de 99% pour $nB = 25$. Ce résultat est tout à fait attendu car en augmentant le nombre des centres nous conservons la distribution initiale des données, ce qui permet d'avoir une bonne généralisation [6] [5].

- Exploration des centres de la classe minoritaire avec nombre de classe majoritaire fixé en 25.

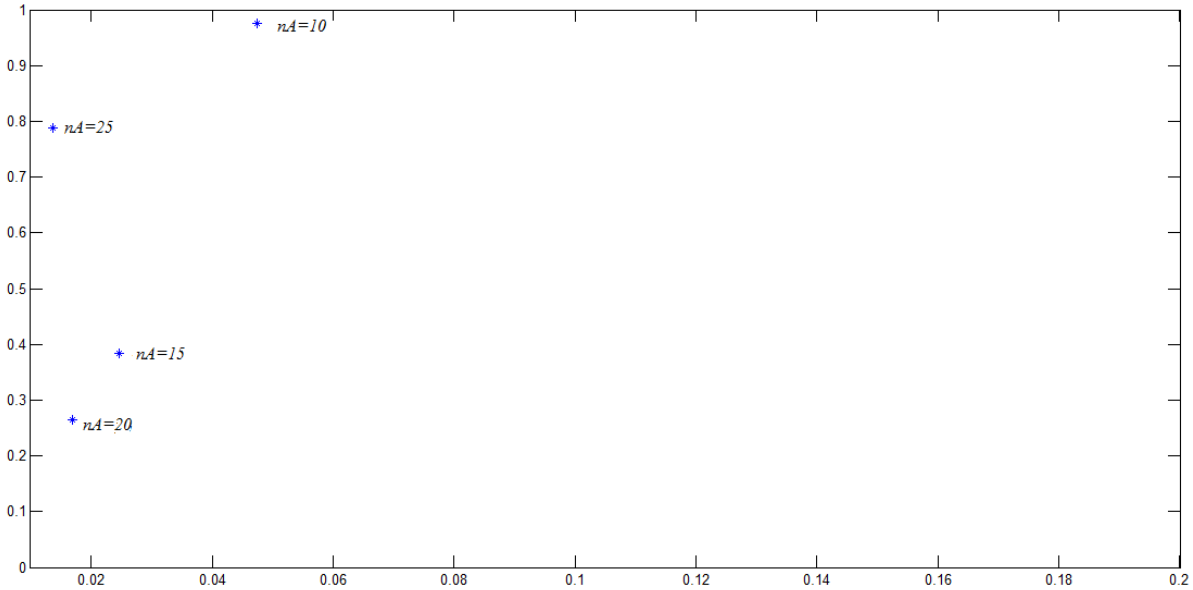



*Figure V.4: Variation du taux des FP et TP en fonction de nombre des centres de la classe minoritaire nA*

La variation des nombres des centres de la classe minoritaire influence aussi sur la qualité de la classification de la classe minoritaire. La meilleure classification est obtenue par $nA = 10$, où le taux des vrais positifs atteint 97,6%.

En général, d'après les résultats exposés ci-dessus, nous pouvons conclure que la classification en utilisant MLP, a connu une amélioration notable, puisque la précision globale atteint **96,5%**, et ce, sans ignorer la classe minoritaire. En choisissant les nombres de classes adéquats, le taux des vrais positif (rappel) est proche de 99%, c.à.d. presque toute la classe minoritaire est bien classée.

Cela nous pousse à noter que la dégradation de la précision globale est due à la dégradation du taux des vrai négative, c.à.d, la classification de la classe majoritaire n'est pas aussi bonne que celle de la classe majoritaire. Cette observation nous a permis de développer la troisième approche qui permet de se focaliser sur les exemples critiques sans pour autant perturber la distribution de la classe majoritaire.

### *V.3.3. Application sur le Active Learning*

- ***Aperçue sur la méthode Active Learning (AL)***

L'Active Learning (AL), est une méthode de classification semi-supervisée, qui réalise la classification en utilisant un nombre de données déjà étiquetés par un expert ( humain). Le AL a pour objectif de réaliser la classification des données avec le moindre nombre d'exemple étiquetés possible ; de sorte à minimiser le coût générale de l'apprentissage [109]. Le principe de cette méthode d'apprentissage est de construire itérativement l'ensemble d'entraînement, en rajoutant les exemples qui sont considérer comme critique selon une certaine mesure.

L'algorithme d'apprentissage en utilisant AL est résumé comme suit:

- Soit un ensemble S composé de $S_t$ , ensemble des donnée étiquetés et $S_u$ les données non-étiquetés
- Répéter jusqu'à critère d'arrêt définie préalablement
  - Entraîner un classifieur en utilisant $S_t$
  - Prédire les étiquettes de $S_u$
  - Calculer la mesure de sélection en utilisant l'erreur de classification de $S_u$
  - Construire l'ensemble $S_i$ des k exemples de $S_u$ qui ont la mesure de l'erreur la plus élevée.
  - Etiqueter $S_i$
  - Rajouter l'ensemble $S_i$ à l'ensemble $S_t$ ( $S_t = S_t \cup S_i \ et \ S_u = S_u - S_i$)

Le critère d'arrêt peut prendre plusieurs formes: Il peut être le nombre des itération, taille de l'ensemble $S_u$, ou tout simplement le critère de sur-apprentissage.

Le processus d'AL, ainsi défini est une sorte d'interaction entre le classifieur et l'humain afin de mettre à jour l'ensemble d'entraînement.

- ***Classification de données déséquilibrées en utilisant AL***

En général, l'application de l'algorithme Active Learning sur les donné déséquilibré permet d'augmenter la performance de la classification. Cela est dû au fait que cette technique est implicitement une méthode de sélection des échantillons [110] . En effet, le processus de la technique AL se base sur le principe de la construction de l'ensemble d'entraînement selon l'erreur l'apprentissage.

Nous appliquons deux types de classification par AL sur les données traitées. La première, nommée " Random Sampling Active learning" (RS) qui commence l'entraînement par un ensemble choisi aléatoirement, et la deuxième " Margin Sampling Active Learning" ( MS) qui utilise les exemples les plus proches de la frontière pour débuter l'apprentissage.

La figure suivante montre la comparaison entre les deux méthodes d'active Learning ( RS et MS) appliquées sur la base KDD-cup'99.

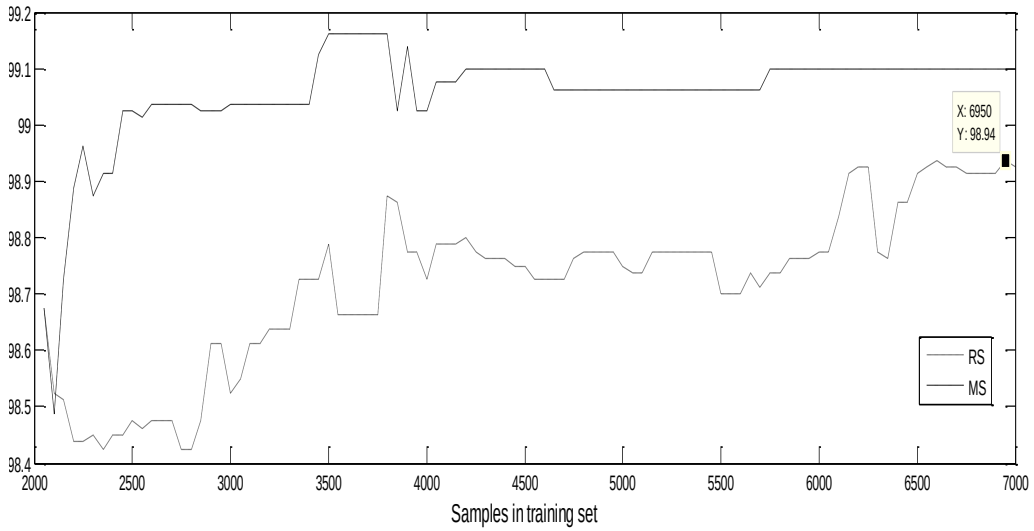



*Figure V.5: Comparaison entre les deux courbes de précision pour les méthodes RS et MS*

D'après ces résultats, nous pouvons confirmer que l'usage de la technique AL en général permet d'améliorer la précision de la classification. Bien que la méthode est beaucoup plus précise que RS, les deux techniques permettent de passer à une précision globale de **98.9%** pour la méthode RS et **99.1%** pour MS.

- ***Intégration de l'approche proposée dans la technique AL***

Nous avons appliqué la méthode proposée dans l'étape du choix de l'ensemble d'entraînement avec lequel le classifieur commencera l'apprentissage. Nous présentons dans la figure suivante les résultats trouvés.

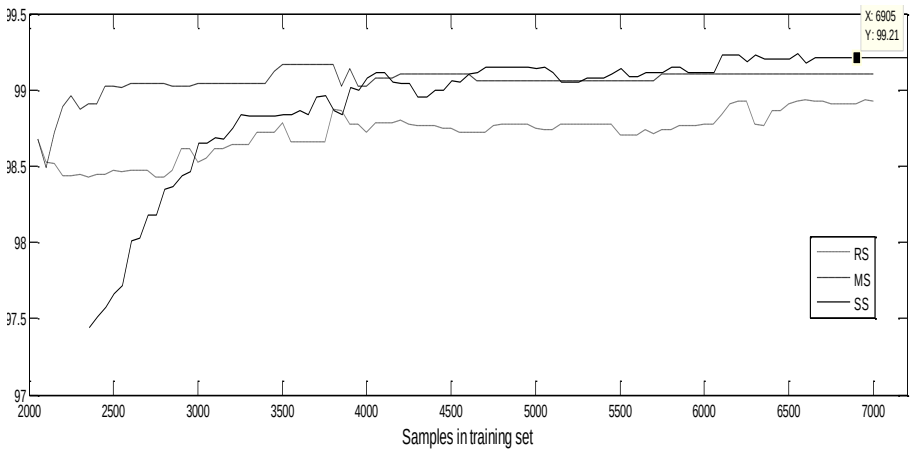



*Figure V.6: Comparaison entre méthode proposée et RS et MS*

Dans cette présentation nous pouvons trouver les trois courbes de précision des trois méthodes d'AL. Le RS, MS et enfin SS qui est l'application de la méthode proposée sur le processus classique de AL. Nous constatons d'une part qu'en général la performance globale est bonne, d'autre part, l'application de la deuxième approche a permis de faire augmenter la qualité de la classification pour atteindre et même dépasser celle de la méthode MS.

L'application de la méthode proposée au sein du processus d'AL, a permis d'atteindre une précision globale égale à **99.21%**.

## *V.4. Méthode 3: Sous- échantillonnage à l'aide de la fonction indicatrice*

Les méthodes de sélection des échantillons ont connu beaucoup de développement depuis leurs apparition. Des travaux sur le critère de sélection ont montré que la proximité et l'erreur d'apprentissage ne sont pas toujours les meilleurs moyens pour décider sur la criticité d'un exemple [111], [71].

Dans ce sens, la deuxième approche que nous proposons apporte une nouvelle notion de criticité et utilise une fonction indicatrice pour choisir de nouveaux centres critiques à partir des centres proche de la frontière. De cette manière on assurera que la distribution initiale de la classe majoritaire est maintenue.

### *V.4.1. Présentation de la méthode*

#### *Notion de criticité*

Le nouveau concept de criticité adopté par les techniques récentes de sélection des échantillons, est celui de choisir non seulement les exemples qui sont proche de la frontière, mais plutôt ceux qui ont une influence sur la définition de celle-ci.

La figure suivante illustre ce concept, qui permet de mesurer la criticité d'un exemple en se basant sur le comportement du classifieur.

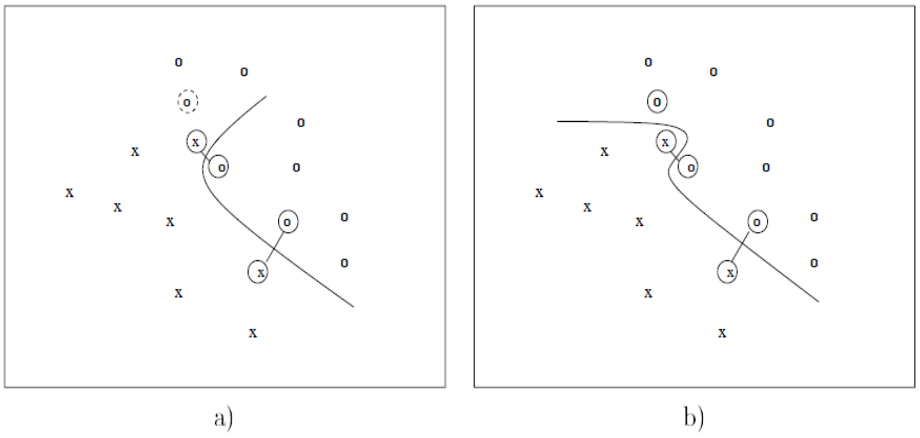



*Figure V.7: Illustration du nouveau concept de criticité*

Comme le montre cette figure, le fait de ne pas considérer les échantillons qui participent à la définition de la frontière comme étant critiques, peut mener au changement de l'allure de celle-ci et conduit à une détérioration de la performance de la classification.

Notre concept de criticité, proposé dans cette session, permet de maintenir la distribution de l'ensemble initial autour de la frontière de manière à ce que l'ensemble sélectionné soit plus représentatif. Nous utiliserons pour ce fait, la fonction indicatrice qui permettra de tenir en compte la distribution des centres par rapport aux paires opposées les plus proches déjà sélectionnées.

#### *Fonction indicatrice*

Comme nous avons cité précédemment, le choix des exemples critiques doit tenir en compte la représentativité et la distribution de l'ensemble initial. Dans ce sens un nouveau critère nommé "Typicité" a été défini par Lyhyaoui *et al.* afin de compléter le critère de la proximité [112].

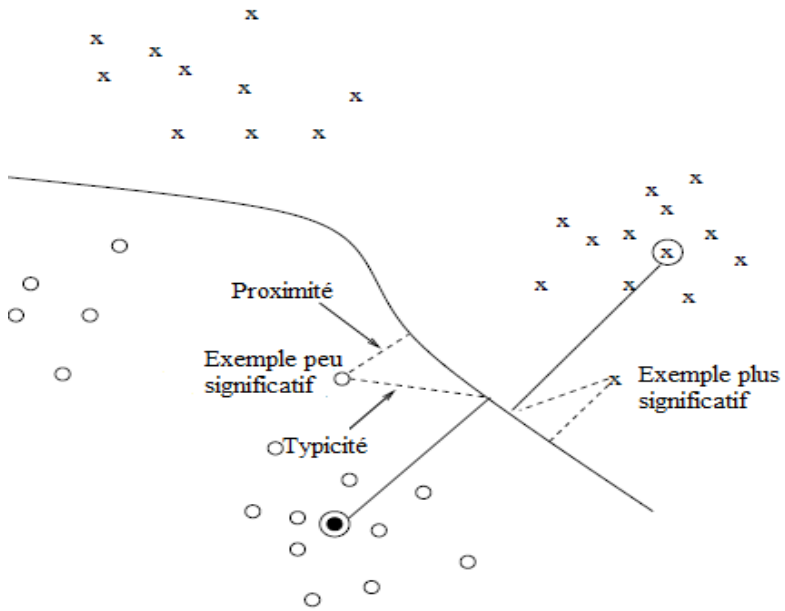



*Figure V.8: Illustration de la notion de typicité et proximité*

La fonction indicatrice comme défie dans [112] est la mesure qui permet d'indiquer si un exemple est considérer comme critique ou pas. Elle rassemble les deux critère de sélection : La proximité et la typicité.

La fonction indicatrice d'un exemple x est définie comme suit:

$$\boldsymbol{I(x,\delta) = \| w^T x + b\| + \delta \left\| x - x_0^p \right\|} \quad (V.5)$$

Où, $\delta$ est une pondération du paramètre de la typicité de x.

La fonction $I(x,\delta)$ainsi définie est composée de deux termes: $\| w^T x + b\|$ qui représente la distance de $x$ par rapport à la frontière, et $\left\| x - x_0^p \right\|$ , qui est la distance entre $x$ et la projection du centre $x_0$ sur la frontière. A partir de cette définition, les exemples considérés critiques sont ceux qui minimise le plus la fonction indicatrice $I$.

### *Fonction indicatrice modifiée*

Vu que l'objectif de notre approche est de garder l'allure de la distribution initiale des données autour de la frontière, une modification de la fonction indicatrice, définie précédemment, s'avère nécessaire.

La sélection de nouveaux centres critiques se réalisera en fonction des paires opposées les plus proches déjà trouvée par la première approche.

Dans ce cadre, la notion de typicité, sera attribuée aux centres les plus loin des POPP. La figure suivante illustre la notion de typicité proposée.

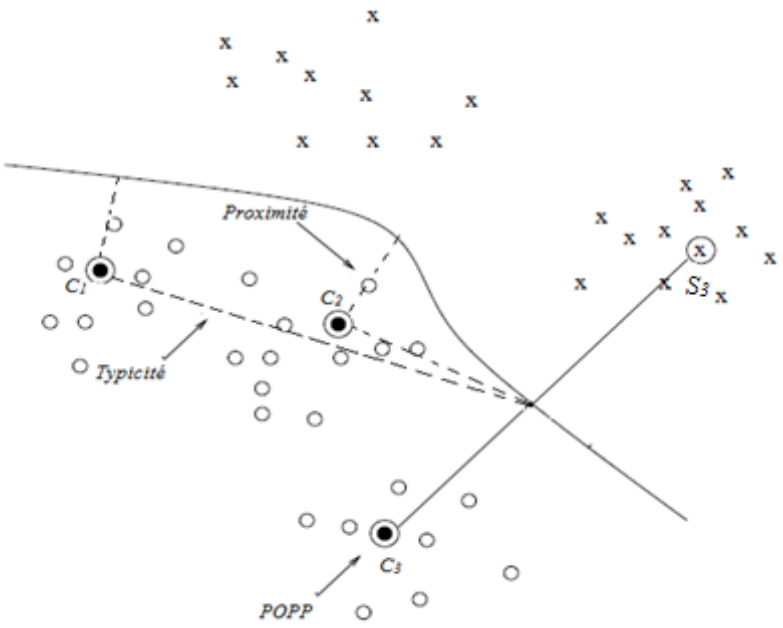



*Figure V.9: Concept de proximité et de typicité dans un espace de données*

Dans la figure ci-dessus, considérons le couple *($C_3$, $S_3$)* une des paires opposées les proches trouvée précédemment. Et *$C_1$, $C_2$* sont deux centres. Notre approche consiste à choisir entre *$C_1$* et $C_2$ le centre qui représentera le mieux la distribution des données autour de la frontière.

Dans la figure *V.9*, le centre *$C_1$* vérifie mieux cette condition que le centre *$C_2$* vu qu'il permet de rajouter un ensemble d'exemples se situant autour de la frontière et qui n'ont pas été sélectionner à cause d'absence d'un centre opposés qui vérifie la condition du POPP.

Pour tenir en compte ce concept de typicité, la nouvelle fonction indicatrice est donc une minimisation de la distance entre le centre et la frontière et une maximisation de la distance entre le centre et la projection du centre critique sur la frontière. Soit, une minimisation de $\| w^T x + b\|$ et maximisation de $\left\|x - x_0^p\right\|$.

Dans ce sens, la fonction indicatrice modifié est

$$FI(x,\delta) = \| w^T x + b\| + \delta . exp\ (-\left\|x - x_0^p\right\|)\ V.6)$$

Pour chaque entre critique, le calcule de la fonction indicatrice se fait localement au sein de l'ensemble des centres qui sont les plus proches du centre critique en question.

Soit $C = \{C_1, C_2, \dots, C_n\}$ et $S = \{S_1, S_2, \dots, S_m\}$ l'ensembles des centres des classes majoritaire et minoritaire respectivement. et soit $PP = \{(C_i, S_i), tq\ C_i \in C\ et\ S_i \in S\}$ l'ensemble des POPP. Pour chaque paire $(C_i, S_i)$, nous déterminons une nouveau centre critique $C_j$ à partir de la classe majoritaire. La détermination du centre critique selon la nouvelle fonction indicatrice se fait comme suit:

- Nous déterminons à partir de l'ensemble *C*, l'ensemble $\pi$ des centres les plus proches de $C_i$.
- Nous déterminons la frontière locale formée par la paire $(C_i, S_i)$. La frontière est une ligne d'équation $w^T z + b = 0$ qui passe par le milieu du segment $[C_i, S_i]$.
- Pour chaque centre $C_k$ de $\pi$ , nous calculons la distance entre le centre et la frontière utilisant la formule suivante

$$Proximité_k = \frac{\left|wx_{C_k} - y_{C_k} + b\right|}{\sqrt{w^2+1}}\ \ (V.7)$$

- Pour chaque centre $C_k$ de $\pi$, nous calculons la distance entre le centre est le centre du segment $[C_i, S_i]$ , noté $c_{i,}$ en utilisant la norme: $Typicité_k = \|x - c\|$
- La valeur de la fonction indicatrice du centre $C_k$ est $FI_k = Proximité_k + Typicité_k$
- Le centre qui a la plus petite valeur de $FI$ est celui qui est choisit comme centre critique.

Apres la sélection des nouveaux centres critiques de la classe majoritaire, l'ensemble d'entraînement est formé par toute la classe minoritaire et tous les exemples de la classe majoritaires qui appartiennent aux centres critiques.

### *V.4.2. Résultats de l'application de la méthode proposée*

#### *V.4.2.1. Simulation d'usage de la fonction indicatrice sur l'ensemble Ripley*

- ***Les données Ripley***

Les données Ripley sont considérées comme une référence dans le domaine de l'apprentissage automatique vu qu'elles permettent de tester la prestation des algorithmes appliqués dans différents domaines. Ces données sont disponibles dans http://marcov.States.ox.ac.uk/pub/neural/papers.

Leurs popularités est du à deux caractéristiques importantes de ces données:

- Les deux classes présentent un haut degré de chevauchement aux données
- Le nombre des données d'entraînement est très réduit par rapport à celles du test

Afin de mieux adapter ces données à notre étude, nous avons créé un déséquilibre en rajoutant des données synthétiques en gardant le même pourcentage que la base KDD-Cup'99 traitée (80% des données correspond à la classe 1 et 20% pour la classe 0).

- ***Application de l'approche proposée sur Ripley déséquilibré***

Afin de visionner la méthode de sélection des centres critiques à partir des paires opposées les plus proches, nous présentons les étapes de l'application de cette procédure.

**Etape1:** En appliquant un clustering comme il est appliqué dans la méthode 2, nous trouvons les centres présentés dans la figure suivante

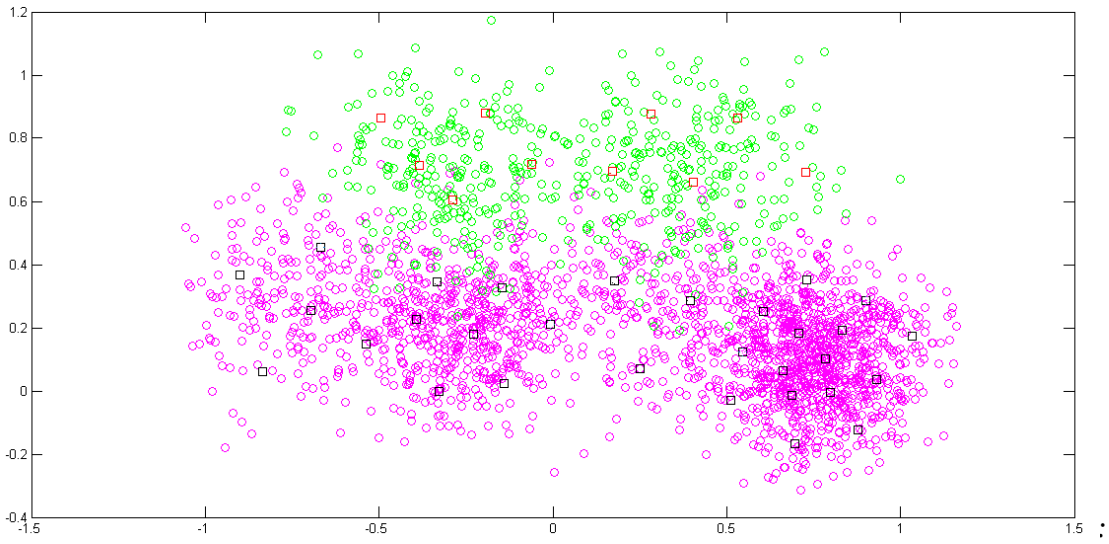


*Figure V.10Clustering des données de la base Ripley déséquilibrée*

La classe minoritaire (présentée en vert) contient 10 centres tandis que la classe majoritaire (présenté en rose) est segmentée en 30 centres.

**Etape 2:** Nous appliquons la procédure de sélection des paires opposées les plus proches

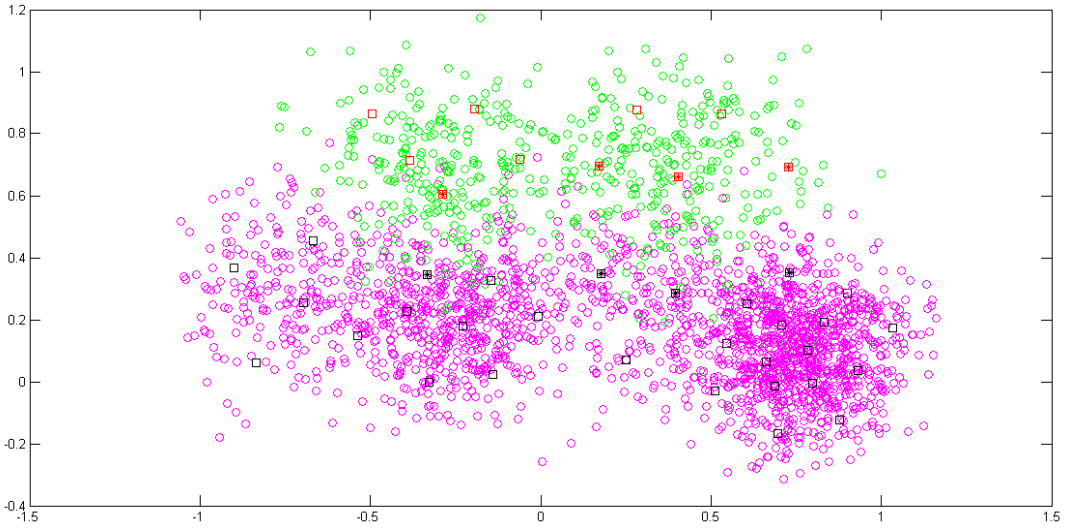


*Figure V.11: Détermination des POPP pour Ripley*

La figure ci-dessus montre la présence de quatre paires opposées les plus proches (marquées sous forme d'étoiles)

**Etape 3:** Nous rajoutons à la classe majoritaire un nombre de centres sélectionné à l'aide de la fonction indicatrice.

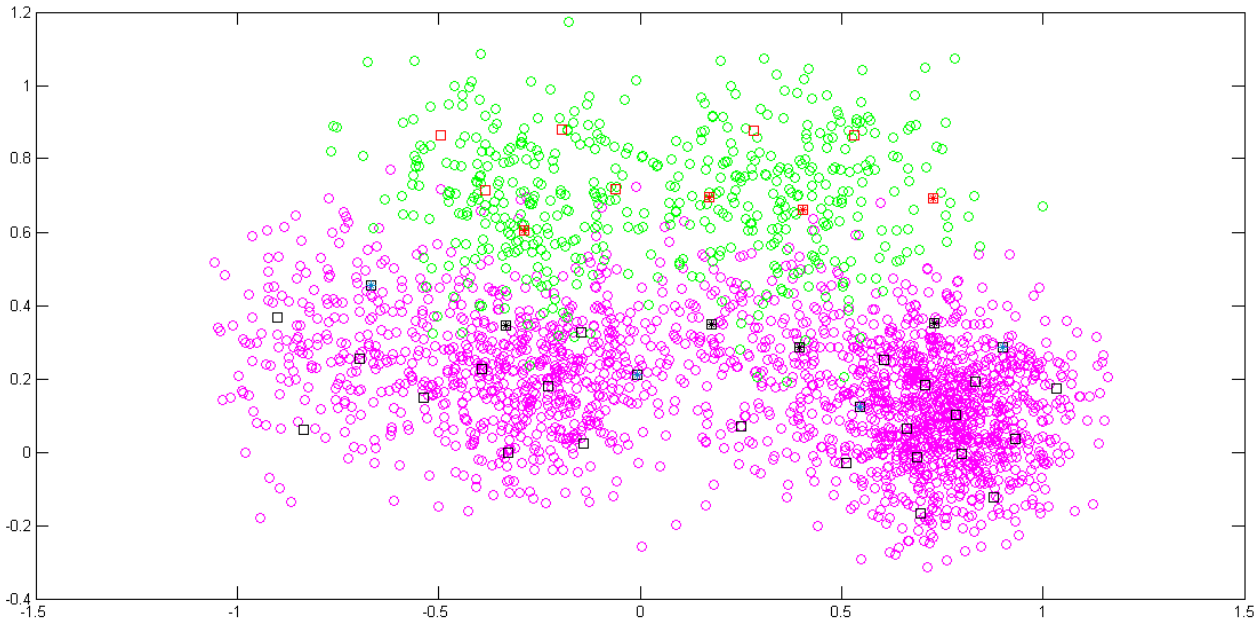



*Figure V.12: Ajout des centres critiques à l'aide de la fonction indicatrice*

La figure montre l'ajout de quatre nouveaux centres dans la classe majoritaire (marqués en étoiles bleu) qui sont les plus proches possibles à la frontière et éloignés des centres critiques initialement trouvés.

La méthode proposée pour sélectionner l'ensemble des centres critiques, permet de parcourir toute la frontière. De cette manière nous pouvons assurer que la distribution de la classe majoritaire est sauvegardée autour de la frontière.

### *V.4.2.2. Application de l'approche 3 sur KDD-Cup'99*

Dans cette expérience nous avons fixé les paramètres du MLP et des nombre des centres selon les résultats des anciennes approches. Pour les paramètres du MLP nous avons opté pour $n = 8$ et $lr = 10^{-5}$. Et pour les nombre de classe, nous avons choisit, 10 centres pour la classe minoritaire $nA = 10$ et 30 centres pour la classe majoritaire $nB = 30$.

Cette approche nous a donné une précision globale de **97%** avec un rappel de **98.7%.**

Enfin nous pouvons dire que cette méthode a permis d'augmenter la précision globale de l'apprentissage par rapport à celle de la deuxième approche, cela en gardant une bonne précision de la classification de la classe minoritaire.

## *V.5. Discussion et synthèse*

Dans ce chapitre nous avons exposé les résultats trouvés pour les trois approches proposées. Nous pouvons conclure que la performance de la classification dépend de quelques paramètres (comme le cas des seuils d'apprentissage et le nombre de classes) ainsi que la qualité des exemples sélectionnés pour la réalisation de l'entrainement.

Grace à ces approche, la classification a pu atteindre un pourcentage de 99% de précision globale dont une précision de classification de la classe minoritaire de 99% .

Le tableau suivant récapitule les méthodes exposées dans ce chapitre, présentant une comparaison de leurs précisions et rappel ( taux des vrai positifs).

| **Méthode de classification** | **Précision globale** | **Taux des vrai positive ( Rappel)** |
|---|---|---|
| MLP classique | 93.4% | 60% |
| SS avec erreur d'apprentissage | 95% | 99% |
| SS avec les POPP sur MLP | 96.5% | 99% |
| SS avec les POPP sur AL | 99.2% | 99% |
| SS avec fonction indicatrice | 97% | 98.7% |

*Tableau V-1: Comparaison des performances des méthodes de classification proposées*

D'après ce tableau, nous pouvons conclure que l'usage des techniques de sélection des échantillons pour le sous-échantillonnage permettent en générale d'augmenter la précision de l'apprentissage, ainsi que de prêter attention à la classe minoritaire. La meilleure précision est obtenue par l'usage des technique de SS combinés à l'Active Learning.

# *Chapitre VI: Conclusion et perspectives*

Durant nos travaux de recherche, nous avons traité un problème majeur rencontré par les méthodes d'apprentissage supervisé face aux données du monde réel. Ce problème est le déséquilibre des classes.

Le double objectif que nous nous sommes fixés dans cette thèse est d'une part améliorer la performance de la classification, en utilisant des techniques connues par leurs apport comme la sélection des échantillons, et d'autre part, à équilibrer les données, afin d'éviter la marginalisation de la classe minoritaire puisqu'elle a tendance à être ignorer par les algorithmes de classification classiques.

Notre apport durant cette thèse se résume en trois approches basées sur la combinaison des techniques de sélection des échantillons et le sous-échantillonnage de la classe majoritaire.

- La première approche se base sur le critère de l'erreur d'apprentissage pour sélectionner les exemples pertinents à partir de la classe majoritaire. C'est une méthode qui réalise un sous-échantillonnage ciblé au fur et à mesure de l'entraînement du classifieur.
- La deuxième approche permet de sélectionner les clusters les plus proches à la frontière sans avoir recours au classifieur. Elle se base seulement sur le critère de la distance et des paires opposées les plus proches.
- La dernière approche, utilise une mesure se basant sur la distance (fonction indicatrice) permettant de sélectionner les centres étendus tout au long de la frontière, nous permettant ainsi, de garder la distribution initiale des données autour de la frontière.

Notre choix de domaine d'application, est tombé sur les systèmes de détection d'intrusion, vu que ses données sont connues par leurs difficultés d'apprentissage. En effet, l'analyse qui a été réalisé à la base de données a montré plusieurs anomalies comme les redondances, les bruits et aberrance et bien entendu, le déséquilibre.

Un prétraitement de données s'est donc imposé: Apres la réalisation d'une analyse descriptive des données, nous avons traité les anomalies de bruit et la dispersion par des méthodes statistiques connues comme la normalisation. D'autre part, nous avons réalisé une réduction de la dimension des données à l'aide de l'analyse bi-variée et l'analyse des composantes principales (ACP).

Enfin, les résultats trouvés semblent être satisfaisantes en comparaison avec d'autre méthodes.

Cette étude nous a permis de s'ouvrir à d'autres domaines d'apprentissage automatique permettant de traiter les données dans un temps réel. Nous nous sommes posés comme perspective d'intégrer les méthodes de sélection des échantillons avec la détection de nouveauté, qui est un domaine de recherche très actif de nos jours, vu son application dans plusieurs variétés de données. Aussi nous comptons nous diriger vers la prévention des intrusions et le développement des systèmes de détection d'intrusion.

Du côté algorithmique, nous voudrons développer une méthode qui permet de combiner le critère de la proximité et celui de l'erreur de l'apprentissage au cours de la technique de la sélection des échantillons. D'autre part, nous comptons améliorer la fonction indicatrice par l'intégration d'une pondération convexe ou par l'entraînement du paramètre de pondération.

# *Annexe*

***Annexe 1: Statistique descriptives des données quantitatives de la base KDD-Cup'99***

| Variables | Minimum | Maximum | Moyenne | Ecart type |
|---|---|---|---|---|
| V1 | 0 | 58329 | 47,98 | 707,747 |
| V5 | 0 | 693375640 | 3025,62 | 988219,101 |
| V6 | 0 | 5155468 | 868,53 | 33040,035 |
| V8 | 0 | 3 | ,01 | ,135 |
| V9 | 0 | 3 | ,00 | ,006 |
| V10 | 0 | 30 | ,03 | ,782 |
| V11 | 0 | 5 | ,00 | ,016 |
| V13 | 0 | 884 | ,01 | 1,798 |
| V16 | 0 | 993 | ,01 | 2,013 |
| V17 | 0 | 28 | ,00 | ,096 |
| V18 | 0 | 2 | ,00 | ,011 |
| V19 | 0 | 8 | ,00 | ,036 |
| V20 | 0 | 0 | ,00 | ,000 |
| V23 | 0 | 511 | 332,29 | 213,147 |
| V24 | 0 | 511 | 292,91 | 246,323 |
| V25 | ,0000 | 7,0000 | ,176997 | ,3957733 |
| V26 | ,0000 | 7,0000 | ,177175 | ,3893141 |
| V27 | ,0000 | 7,0000 | ,054767 | ,2295711 |
| V28 | ,0000 | 7,0000 | ,057423 | ,2391362 |
| V29 | ,0000 | 7,0000 | ,915497 | ,9643196 |
| V30 | ,0000 | 7,0000 | ,412825 | 1,6403414 |
| V31 | ,0000 | 7,0000 | ,028551 | ,3173025 |
| V32 | 0 | 255 | 232,47 | 64,745 |
| V33 | 0 | 255 | 188,67 | 106,040 |
| V34 | ,0000 | 7,0000 | ,918957 | 1,1678004 |
| V35 | ,0000 | 7,0000 | ,654765 | 2,0322888 |
| V36 | ,0000 | 7,0000 | ,604767 | ,5992231 |
| V37 | ,0000 | 7,0000 | ,022608 | ,3917251 |
| V38 | ,0000 | 7,0000 | ,177475 | ,3963772 |
| V39 | ,0000 | 7,0000 | ,176299 | ,3826328 |
| V40 | ,0000 | 7,0000 | ,054085 | ,2437343 |
| V41 | ,0000 | 7,0000 | ,053344 | ,2418473 |

*Annexe 2: Répartition des modalités de la variable V2*

| | V2 | | |
|---|---|---|---|
| | ICMP | TCP | UDP |
| Effectif | 283 602 | 190 065 | 20 354 |
| N % ligne | 57,4% | 38,5% | 4,1% |

*Annexe 3: Répartition des modalités de la variable 3*

| | Effectif | Proportion | Proportion cumulée |
|---|---|---|---|
| **ecr_i** | 281 400 | 56,96% | 57,0% |
| **private** | 110 893 | 22,45% | 79,4% |
| **http** | 64 292 | 13,01% | 92,4% |
| **Smtp** | 9 723 | 1,97% | 94,4% |
| **Other** | 7 237 | 1,46% | 95,9% |
| **domain_u** | 5 863 | 1,19% | 97,0% |
| **ftp_data** | 4 721 | ,96% | 98,0% |
| **eco_i** | 1 642 | ,33% | 98,3% |
| **ftp** | 798 | ,16% | 98,5% |
| **finger** | 670 | ,14% | 98,6% |
| **urp_i** | 538 | ,11% | 98,7% |
| **telnet** | 513 | ,10% | 98,8% |
| **ntp_u** | 380 | ,08% | 98,9% |
| **Auth** | 328 | ,07% | 99,0% |
| **pop_3** | 202 | ,04% | 99,0% |
| **Time** | 157 | ,03% | 99,1% |
| **csnet_ns** | 126 | ,03% | 99,1% |
| **remote_job** | 120 | ,02% | 99,1% |
| **gopher** | 117 | ,02% | 99,1% |
| **imap4** | 117 | ,02% | 99,2% |
| **discard** | 116 | ,02% | 99,2% |
| **domain** | 116 | ,02% | 99,2% |
| **iso_tsap** | 115 | ,02% | 99,2% |
| **systat** | 115 | ,02% | 99,2% |
| **Echo** | 112 | ,02% | 99,3% |
| **Shell** | 112 | ,02% | 99,3% |
| **Rje** | 111 | ,02% | 99,3% |
| **sql_net** | 110 | ,02% | 99,3% |
| **whois** | 110 | ,02% | 99,4% |
| **printer** | 109 | ,02% | 99,4% |

| | | | |
|---|---|---|---|
| **courier** | 108 | ,02% | 99,4% |
| **nntp** | 108 | ,02% | 99,4% |
| **Mtp** | 107 | ,02% | 99,4% |
| **netbios_ssn** | 107 | ,02% | 99,5% |
| **sunrpc** | 107 | ,02% | 99,5% |
| **Bgp** | 106 | ,02% | 99,5% |
| **klogin** | 106 | ,02% | 99,5% |
| **Uucp** | 106 | ,02% | 99,6% |
| **uucp_path** | 106 | ,02% | 99,6% |
| **vmnet** | 106 | ,02% | 99,6% |
| **Nnsp** | 105 | ,02% | 99,6% |
| **Ssh** | 105 | ,02% | 99,6% |
| **supdup** | 105 | ,02% | 99,7% |
| **hostnames** | 104 | ,02% | 99,7% |
| **Login** | 104 | ,02% | 99,7% |
| **daytime** | 103 | ,02% | 99,7% |
| **Efs** | 103 | ,02% | 99,7% |
| **Link** | 102 | ,02% | 99,8% |
| **netbios_ns** | 102 | ,02% | 99,8% |
| **Ldap** | 101 | ,02% | 99,8% |
| **pop_2** | 101 | ,02% | 99,8% |
| **Exec** | 99 | ,02% | 99,8% |
| **http_443** | 99 | ,02% | 99,9% |
| **netbios_dgm** | 99 | ,02% | 99,9% |
| **kshell** | 98 | ,02% | 99,9% |
| **Name** | 98 | ,02% | 99,9% |
| **Ctf** | 97 | ,02% | 99,9% |
| **netstat** | 95 | ,02% | 100,0% |
| **Z39_50** | 92 | ,02% | 100,0% |
| **IRC** | 43 | ,01% | 100,0% |
| **urh_i** | 14 | ,00% | 100,0% |
| **X11** | 11 | ,00% | 100,0% |
| **tim_i** | 7 | ,00% | 100,0% |
| **pm_dump** | 1 | ,00% | 100,0% |
| **red_i** | 1 | ,00% | 100,0% |
| **tftp_u** | 1 | ,00% | 100,0% |

***Annexe 4: Etudes descriptive des variables qualitatives***

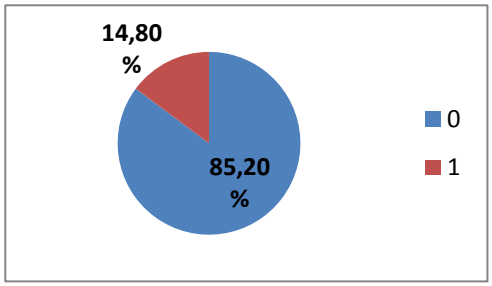
Variable V7

| | Effectif | N % colonne |
|---|---|---|
| 0 | 493 998 | 100,0% |
| 1 | 22 | ,0% |

Presque la totalité de la base traitée prend la même valeur pour la variable V7, ce qui montre que cette variable n'est pas signifiante dans notre étude

Variable V12

| | Effectif | N % colonne |
|---|---|---|
| 0 | 420 784 | 85,2% |
| 1 | 73 236 | 14,8% |




Variable V14

| | Effectif | N % colonne |
|---|---|---|
| 0 | 493 965 | 100,0% |
| 1 | 55 | ,0% |

Presque la totalité de la base traitée prend la même valeur pour la variable V14, ce qui montre que cette variable n'est pas signifiante dans notre étude.

Variable V15

| | Effectif | N % colonne |
|---|---|---|
| 0 | 494008 | 100,0% |
| 1 | 6 | ,0% |
| 2 | 6 | ,0% |

Presque la totalité de la base traitée prend la même valeur pour la variable V15, ce qui montre que cette variable n'est pas signifiante dans notre étude

Variable V21

| | | Effectif | N % colonne |
|---|---|---|---|
| 1 | V2 0 | 494020 | 100,0% |

Cette variable prend une seule valeur dans toute la base, elle est donc considérée constante.

Variable V22

| | | Effectif | N % colonne |
|---|---|---|---|
| 2 | V2 0 | 493335 | 99,9% |
| | 1 | 685 | ,1% |

Presque la totalité de la base traitée prend la même valeur pour la variable V22, ce qui montre que cette variable n'est pas signifiante dans notre étude

## *Annexe 5: Matrice de corrélation de toutes les variables*

| | 1 | 5 | 4 | 8 | 9 | 10 | 11 | 13 | 16 | 17 | 18 | 19 | 20 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | ,004** | ,005** | ,003* | ,004** | ,013** | ,005** | ,058** | ,057** | ,075** | ,000 | ,026** | a | ,105** | ,080** | ,030** | ,031** | ,013** | ,012** | ,004** | ,017** | ,006** | ,010** | ,118** | ,052** | ,021** | ,016** | ,003* | ,030** | ,031** | ,006** | ,009** |
| 5 | ,004** | | ,000 | ,000 | ,000 | ,004** | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 | a | ,003* | ,003 | ,001 | ,001 | ,001 | ,001 | ,001 | ,001 | ,000 | ,002 | ,003* | ,001 | ,001 | ,001 | ,000 | ,001 | ,001 | ,001 | ,001 |
| 4 | ,005** | ,000 | | ,001 | ,016** | ,004** | ,049** | ,023** | ,021** | ,005** | ,000 | ,009** | a | ,040* | ,031* | ,011* | ,011* | ,006* | ,006* | ,002 | ,007* | ,003* | ,049* | ,006* | ,002 | ,008* | ,014* | ,006** | ,011* | ,012* | ,005* | ,005* |
| 8 | ,003* | ,000 | ,001 | | ,000 | ,002 | ,000 | ,000 | ,000 | ,001 | ,000 | ,001 | a | ,062* | ,048* | ,021* | ,022* | ,011* | ,011* | ,005* | ,010* | ,004* | ,005* | ,059* | ,029* | ,015* | ,031* | ,001 | ,021* | ,022* | ,011* | ,011* |
| 9 | ,004** | ,000 | ,016** | ,000 | | ,000 | ,142** | ,014** | ,009** | ,015** | ,000 | ,020** | a | ,004* | ,003* | ,001 | ,001 | ,001 | ,001 | ,000 | ,001 | ,000 | ,007* | ,005* | ,002 | ,000 | ,003 | ,000 | ,001 | ,001 | ,001 | ,001 |
| 10 | ,013** | ,004** | ,004** | ,002 | ,000 | | ,009** | ,007** | ,001 | ,025** | ,006** | ,002 | a | ,068* | ,052* | ,020* | ,020* | ,010* | ,010* | ,003* | ,011* | ,003 | ,026* | ,039* | ,022* | ,012* | ,041* | ,002 | ,016* | ,020* | ,028** | ,008* |
| 11 | ,005** | ,000 | ,049** | ,000 | ,142** | ,009** | | ,007** | ,003* | ,004** | ,000 | ,003* | a | ,015* | ,012* | ,004* | ,004* | ,025** | ,024** | ,000 | ,002 | ,001 | ,025* | ,015* | ,001 | ,003* | ,006* | ,001 | ,010** | ,010** | ,002 | ,002 |
| 13 | ,058** | ,000 | ,023** | ,000 | ,014** | ,007** | ,007** | | ,994** | ,011** | ,009** | ,412** | a | ,009* | ,007* | ,003 | ,003 | ,001 | ,001 | ,000 | ,001 | ,000 | ,008* | ,005* | ,002 | ,001 | ,005* | ,000 | ,002 | ,003 | ,001 | ,001 |
| 16 | ,057** | ,000 | ,021** | ,000 | ,009** | ,001 | ,003* | ,994** | | ,010** | ,013** | ,414** | a | ,009* | ,007* | ,003 | ,003 | ,001 | ,001 | ,000 | ,001 | ,000 | ,012* | ,008* | ,004* | ,002 | ,005* | ,000 | ,003 | ,003 | ,001 | ,001 |
| 17 | ,075** | ,000 | ,005** | ,001 | ,015** | ,025** | ,004** | ,011** | ,010** | | ,067** | ,080** | a | ,017* | ,013* | ,005* | ,005* | ,002 | ,002 | ,001 | ,003 | ,003 | ,019* | ,014* | ,007* | ,002 | ,009* | ,001 | ,005* | ,005* | ,010** | ,002 |
| 18 | ,000 | ,000 | ,000 | ,000 | ,000 | ,006** | ,000 | ,009** | ,013** | ,067** | | ,025** | a | ,015* | ,012* | ,004* | ,005* | ,002 | ,002 | ,000 | ,002 | ,004** | ,017* | ,012* | ,007* | ,002 | ,008* | ,001 | ,004* | ,005* | ,002 | ,002 |
| 19 | ,026** | ,000 | ,009** | ,001 | ,020** | ,002 | ,003* | ,412** | ,414** | ,080** | ,025** | | a | ,043* | ,032* | ,012* | ,013* | ,007* | ,007* | ,002 | ,007* | ,008** | ,021* | ,001 | ,018* | ,006* | ,026* | ,001 | ,012* | ,013* | ,005* | ,006* |
| 20 | a | a | a | a | a | a | a | a | a | a | a | a | a | a | a | a | a | a | a | a | a | a | a | a | a | a | a | a | a | a | a | a |
| 23 | ,105* | ,003* | ,040* | ,062* | ,004* | ,068* | ,015* | ,009* | ,009* | ,017* | ,015* | ,043* | a | | 944** | ,297* | ,306* | ,215* | ,200* | 079** | ,186* | ,136* | 533** | 515** | 107** | ,221* | 684** | ,088* | ,298* | ,308* | ,196* | ,172* |
| 24 | ,080* | ,003 | ,031* | ,048* | ,003* | ,052* | ,012* | ,007* | ,007* | ,013* | ,012* | ,032* | a | 944** | | ,512* | ,521* | ,276* | ,278* | 103** | ,291* | ,101* | 402** | 718** | 084** | ,370* | 753** | ,067* | ,512* | ,527* | ,256* | ,255* |
| 25 | ,030* | ,001 | ,011* | ,021* | ,001 | ,020* | ,004* | ,003 | ,003 | ,005* | ,004* | ,012* | a | ,297* | ,512* | | 950** | ,107* | ,097* | ,117* | 321** | ,039* | 149** | ,745* | 013** | 456** | ,448* | ,025* | 923** | 950** | ,099* | ,088* |
| 26 | ,031* | ,001 | ,011* | ,022* | ,001 | ,020* | ,004* | ,003 | ,003 | ,005* | ,005* | ,013* | a | ,306* | ,521* | 950** | | ,109* | ,107* | ,119* | 328** | ,038* | 152** | ,756* | 015** | 465** | ,457* | ,026* | 937** | 972** | ,101* | ,100* |

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 27 | ,013** | ,001 | ,006* | ,011* | ,001 | ,010* | ,025** | ,001 | ,001 | ,002 | ,002 | ,007* | a | ,215* | ,276* | ,107* | ,109* | | 945** | ,052* | 273** | 010** | ,092* | ,313* | 010** | 260** | ,201* | 020** | ,105* | ,109* | 886** | 840** |
| 28 | ,012** | ,001 | ,006* | ,011* | ,001 | ,010* | 024** | ,001 | ,001 | ,002 | ,002 | ,007* | a | ,200* | ,278* | ,097* | ,107* | 945** | | ,059* | 266** | 008** | ,085* | ,318* | 001 | 251** | ,204* | 021** | ,096* | ,110* | 849** | 845** |
| 29 | ,004* | ,001 | 002 | ,005* | 000 | 003* | 000 | 000 | 000 | 001 | 000 | 002 | a | 079** | 103** | ,117* | ,119* | ,052* | ,059* | | ,112* | 007** | ,027* | 149** | 155** | ,076* | 084** | 005** | ,118* | ,121* | ,049* | ,060* |
| 30 | ,017* | ,001 | ,007* | ,010* | ,001 | ,011* | ,002 | ,001 | ,001 | ,003 | ,002 | ,007* | a | ,186* | ,291* | 321** | 328** | 273** | 266** | ,112* | | ,022* | 086** | ,428* | ,084* | 370** | ,253* | ,014* | 324** | 335** | 258** | 267** |
| 31 | ,006* | 000 | 003* | ,004* | 000 | ,003 | ,001 | 000 | 000 | 003 | 004** | 008** | a | ,136* | ,101* | ,039* | ,038* | 010** | 008** | 007** | ,022* | | ,153* | 009** | ,019* | ,025* | ,062* | 022** | ,039* | ,041* | 008** | ,002 |
| 32 | 010** | ,002 | ,049* | ,005* | ,007* | ,026* | ,025* | ,008* | ,012* | ,019* | ,017* | ,021* | a | 533** | 402** | 149** | 152** | ,092* | ,085* | ,027* | 086** | ,153* | | ,027* | 034** | 096** | 181** | ,180* | 148** | 157** | ,081* | ,027* |
| 33 | ,118* | ,003* | ,006* | ,059* | ,005* | ,039* | ,015* | ,005* | ,008* | ,014* | ,012* | ,001 | a | 515** | 718** | ,745* | ,756* | ,313* | ,318* | 149** | ,428* | 009** | ,027* | | 061** | ,538* | 568** | 017** | ,747* | ,771* | ,294* | ,322* |
| 34 | ,052* | ,001 | ,002 | ,029* | ,002 | ,022* | ,001 | ,002 | ,004* | ,007* | ,007* | ,018* | a | 107** | 084** | 013** | 015** | 010** | 001 | 155** | ,084* | ,019* | 034** | 061** | | 012** | 075** | 002 | 013** | 015** | 010** | 000 |
| 35 | ,021* | ,001 | ,008* | ,015* | 000 | ,012* | ,003* | 001 | 002 | ,002 | ,002 | ,006* | a | ,221* | ,370* | 456** | 465** | 260** | 251** | ,076* | 370** | ,025* | 096** | ,538* | 012** | | ,323* | ,018* | 457** | 474** | 246** | 254** |
| 36 | ,016* | ,001 | ,014* | ,031* | ,003 | ,041* | ,006* | ,005* | ,005* | ,009* | ,008* | ,026* | a | 684** | 753** | ,448* | ,457* | ,201* | ,204* | 084** | ,253* | ,062* | 181** | 568** | 075** | ,323* | | ,022* | ,437* | ,463* | ,178* | ,194* |
| 37 | ,003* | 000 | 006** | ,001 | 000 | ,002 | 001 | 000 | 000 | ,001 | ,001 | ,001 | a | ,088* | ,067* | ,025* | ,026* | 020** | 021** | 005** | ,014* | 022** | ,180* | 017** | 002 | ,018* | ,022* | | ,024* | ,026* | 015** | 000 |
| 38 | ,030* | ,001 | ,011* | ,021* | ,001 | ,016* | 010** | ,002 | ,003 | ,005* | ,004* | ,012* | a | ,298* | ,512* | 923** | 937** | ,105* | ,096* | ,118* | 324** | ,039* | 148** | ,747* | 013** | 457** | ,437* | ,024* | | 956** | ,098* | ,088* |
| 39 | ,031* | ,001 | ,012* | ,022* | ,001 | ,020* | 010** | ,003 | ,003 | ,005* | ,005* | ,013* | a | ,308* | ,527* | 950** | 972** | ,109* | ,110* | ,121* | 335** | ,041* | 157** | ,771* | 015** | 474** | ,463* | ,026* | 956** | | ,102* | ,101* |
| 40 | 006** | ,001 | ,005* | ,011* | ,001 | 028** | ,002 | ,001 | ,001 | 010** | ,002 | ,005* | a | ,196* | ,256* | ,099* | ,101* | 886** | 849** | ,049* | 258** | 008** | ,081* | ,294* | 010** | 246** | ,178* | 015** | ,098* | ,102* | | 813** |
| 41 | 009** | ,001 | ,005* | ,011* | ,001 | ,008* | ,002 | ,001 | ,001 | ,002 | ,002 | ,006* | a | ,172* | ,255* | ,088* | ,100* | 840** | 845** | ,060* | 267** | ,002 | ,027* | ,322* | 000 | 254** | ,194* | 000 | ,088* | ,101* | 813** | |

La variable 20 est constante et donc peut être éliminée.

***Annexe 6: Test de chi-deux réalisé sur les variables qualitatives***

Variable 3 :

| | | V42 |
|---|---|---|
| V3 | Khi-Chi-deux | 414458,457 |
| | ddl | 65 |
| | Sig. | ,000*,a |

Variable 6 :

| | | V42 |
|---|---|---|
| V6 | Khi-Chi-deux | 26607,544 |
| | ddl | 10 |
| | Sig. | ,000* |

Variable V7 :

| | | V42 |
|---|---|---|
| V7 | Khi-Chi-deux | 3,191 |
| | ddl | 1 |
| | Sig. | ,074a |

Variable 12

| | | V42 |
|---|---|---|
| V12 | Khi-Chi-deux | 312453,188 |
| | ddl | 1 |
| | Sig. | ,000* |

Variable 14 :

| | | V42 |
|---|---|---|
| V14 | Khi-Chi-deux | 17,031 |
| | ddl | 1 |
| | Sig. | ,000* |

Variable 15 :

| | | V15 |
|---|---|---|
| V42 | Khi-Chi-deux | 39,840 |
| | ddl | 2 |
| | Sig. | ,000*,a |

Variable 21 :

Le test ne peut pas être réalisé vu que V21 est constante

Variable 22 :

| | | V22 |
|---|---|---|
| V42 | Khi-Chi-deux | 515,392 |
| | ddl | 1 |
| | Sig. | ,000* |

## *Annexe 7: Analyse des composantes principales*

**Qualité de représentation**

| | Initial | Extraction |
|---|---|---|
| var00005 | 1,000 | ,982 |
| var00006 | 1,000 | ,956 |
| var00012 | 1,000 | ,983 |
| var00023 | 1,000 | ,699 |
| var00025 | 1,000 | ,999 |
| var00026 | 1,000 | ,999 |
| var00029 | 1,000 | ,973 |
| var00032 | 1,000 | ,606 |
| var00033 | 1,000 | ,957 |
| var00034 | 1,000 | ,982 |
| var00038 | 1,000 | ,999 |
| var00039 | 1,000 | ,999 |

Méthode d'extraction : Analyse en composantes principales.

**Variance totale expliquée**

| | Valeurs propres initiales | | | Extraction Sommes des carrés des facteurs retenus | | |
|---|---|---|---|---|---|---|
| Composante | Total | % de la variance | % cumulés | Total | % de la variance | % cumulés |
| 1 | 9,791 | 81,590 | 81,590 | 9,791 | 81,590 | 81,590 |
| 2 | 1,345 | 11,206 | 92,796 | 1,345 | 11,206 | 92,796 |
| 3 | ,443 | 3,691 | 96,487 | | | |
| 4 | ,334 | 2,786 | 99,274 | | | |
| 5 | ,039 | ,328 | 99,602 | | | |
| 6 | ,026 | ,217 | 99,819 | | | |
| 7 | ,017 | ,141 | 99,960 | | | |
| 8 | ,002 | ,020 | 99,980 | | | |
| 9 | ,001 | ,012 | 99,992 | | | |
| 10 | ,001 | ,004 | 99,997 | | | |
| 11 | ,000 | ,002 | 99,999 | | | |
| 12 | ,000 | ,001 | 100,000 | | | |

Méthode d'extraction : Analyse en composantes principales.

Matrice des composantes

| | Composante | |
|---|---|---|
| | 1 | 2 |
| var00005 | ,965 | ,225 |
| var00006 | ,950 | ,230 |
| var00012 | ,965 | ,226 |
| var00023 | -,803 | -,233 |
| var00025 | -,877 | ,479 |
| var00026 | -,877 | ,479 |
| var00029 | ,959 | ,232 |
| var00032 | -,738 | -,247 |
| var00033 | ,952 | ,226 |
| var00034 | ,964 | ,228 |
| var00038 | -,877 | ,479 |
| var00039 | -,878 | ,479 |

Méthode d'extraction : Analyse en composantes principales.

## *Annexe 8: Les variables de la base KDD-Cup 1999*

| Numéro de variable | Nom d'attributs | Description | type |
|---|---|---|---|
| **Attributs des connexions TCP individuelles** | | | |
| 1 | Durée | longueur (nombre de secondes) de la connexion | continu |
| 2 | Protocol_type | type du protocole, par exemple TCP, UDP ,.. | discret |
| 3 | Service | service de réseau pour la destination, par exemple, HTTP, Telnet, etc.. | discret |
| 4 | Src_bytes | nombre de bytes de données de la source à la destination | continu |
| 5 | Dst_bytes | nombre de bytes de données de la destination à la source | continu |
| 6 | Flag | statut normal ou erreur de la connexion | discret |
| 7 | Land | 1 si la connexion est *from/to* même *host/port* ; 0 autrement | discret |
| 8 | Wrong_fragment | nombre de "faux" fragments | continu |
| 9 | Urgent | nombre de paquets urgent | continu |
| | **Attributs de contenu** | | |
| 10 | Hot | Nombre de *hot* indicateurs | continu |
| 11 | Num_failed_logins | nombre de tentatives d'ouverture échouées | continu |
| 12 | Logged_in | 1 si entré avec succès ; 0 autrement | discret |
| 13 | Num_compromised | le nombre de conditions compromises | continu |
| 14 | Root_shell | 1 si *root shell* est obtenue ; 0 autrement | discret |

| | | | |
|---|---|---|---|
| 15 | Su_attempted | 1 si la commande su root racine a été essayée ; 0 autrement | discret |
| 16 | Num_root | nombre d'accès root | continu |
| 17 | Num_file_créations | nombre d'opérations de création de dossier | continu |
| 18 | Num_shells | Nombre de *shell* sollicités | continu |
| 19 | Num_access_files | nombre d'opérations sur des dossiers de contrôle d'accès | continu |
| 20 | Num_outbound_cmds | nombre de commandes venant d'une session FTP | continu |
| 21 | is_hot_login | 1 si l'ouverture appartient à la *hot* liste ; 0 autrement 122 | discret |
| 22 | is_guest_login | 1 si l'ouverture est un *guest login* ; 0 autrement | discret |
| | **Attributs des connexions de même machine** | | |
| 23 | Count | nombre de connexion à la même machine que la connexion courante durant les deux dernières | continu |
| 24 | Serror_rate | nombre de connexions qui ont des erreurs de SYN | continu |
| 25 | Rerror_rate | nombre de connexions qui ont des erreurs de REJ | continu |
| 26 | Same_srv_rate | nombre de connexions au même service | continu |
| 27 | diff _srv _rate | nombre de connexions au service différents | continu |
| 28 | Srv_count | nombre de connexions au même service que le raccordement courant dans les dernières deux secondes | continu |
| | **Attributs des connexions de même-service** | | |
| 29 | Srv_serror_rate | nombre de connexions qui ont des erreurs de SYN | continu |
| 30 | Srv_rerror_rate | nombre de connexions qui ont des erreurs de REJ | Ccontinu |
| 31 | Srv_diff_host_rate | nombre de connexions aux différents *hosts machines* | continu |
| | **Relatifs à un hôte de destination particulier pendant les 100 dernières connexions** | | |
| 32 | dst_host_count | Nombre de connexion | continu |
| 33 | dst_host_srv_count | nombre de connexions relatif à service et un hôte | continu |
| 34 | dst_host_same_srv_rate | nombre de connexions au même service | continu |
| 35 | dst_host_diff_srv_rate | nombre de connexions au service différents | continu |
| 36 | dst_host_same_src_port rate | taux de connexions au même service | continu |
| 37 | dst_host_srv_diff_host_ rate | taux de connexions au service différents | continu |
| 38 | dst_host_serror_rate | nombre de connexions qui ont des erreurs de | continu |
| 39 | dst_host_srv_serror_rat e | nombre de connexions qui ont des erreurs de REJ et SYN | continu |
| 40 | dst_host_rerror_rate | nombre de connexions qui ont des erreurs de REJ | continu |
| 41 | dst_host_srv_rerror_rate | nombre de connexions qui ont des erreurs de REJ | continu |

# *Bibliographie*


[1] S. Tufféry, Data Mining et Scoring, Dunod, 2002.

[2] F. Moutarde, Apprentissage Artificiel, MINES ParisTech, 2008.

[3] H. Larochelle, Étude de techniques d'apprentissage non-supervisé pour l'amélioration de l'entraînement supervisé de modéles connexionistes, Thése de doctorat. Université de Monréal, 2008.

[4] I. Chairi, S. Alaoui and A. Lyhyaoui, "Learning from imbalanced data using methods of sample selection," *IEEE Explore. International Conference on Multimedia Computing and Systems,* pp. 254 - 257, 2012.

[5] I. Chaïri, S. Alaoui and A. Lyhyaoui, "Balancing Distribution of Intrusion Detection Data Using Sample Selection," *Journal of Information Security Research,* vol. 3, pp. 153-163, 2012.

[6] I. Chairi, S. Alaoui and A. Lyhyaoui, "Intrusion Detection based Sample Selection for imbalanced data distribution," *IEEE Xplore. Second International Conference on Innovative Computing Technology,* pp. 259 - 264, 2012 .

[7] C. Lemnaru, Strategies for dealing with real world classification problems, Universitatea Technica, Din Cluj-Napoca: (Unpublished PhD thesis) Faculty of Computer Science and Automation, 2012.

[8] P. Pendharkar, S. Nanda, J. Rodger and R. Bhaskar, "An Evolutionary Misclassification Cost Minimization Approach for Medical Diagnosis," *P. Pendharkar (Ed.), Managing Data Mining Technologies in Organizations: Techniques and Applications,* pp. 32-44, 2003.

[9] H. Haibo and A. E. Garcia, "Learning from Imbalanced Data," *IEEE Transactions On Knowledge and Data Engineering,* pp. Vol.2, N° 9, Septembre 2009.

[10] V. Nitesh, N. Chawla and A. Kolcz, "Editorial: Special Issue on Learning from Imbalanced Data Sets," *SIGKDD Explorations,* vol. 6, no. 1, June 2004.

[11] G. Wu and E. Y. Chang, "Class-boundary alignment for imbalanced dataset learning," *ICML'03 Workshop on Learning from Imbalanced Data Sets,* 2003..

[12] X. Chai, L. Deng, Q. Yang and C. Ling, "Test-Cost Sensitive Naive Bayes Classification," *Proceedings of the 4th IEEE International Conference on Data Mining,* pp. 51-58., 2004.

[13] A. Aliamiri, Statistical Methods for Unexploded Ordnance Discrimination, Northeastern University. Boston: PhD Thesis. Department of Electrical and Computer Engineering., 2006.

[14] D. Williams, V. Myers and M. Silvious, "Mine classification with imbalanced data," *IEEE Geoscience and Remote Sensing Letters,* vol. 3, no. 6, p. 528–532, 2009.

[15] N. Chawla, "C4.5 and imbalanced datasets ànvestigating the effect of sampling method, probabilistic estimate, and décision tree structure," *ICML'Workshop on Learning from Imbalanced Data Sets,* 2003.

[16] S. S. Japkowicz N., "The class imbalance problem: A systematic study," *Intelligent Data Analysis,* vol. VI, pp. 429 - 449, 2002.

[17] D. A. Z. a. G. R. Simon Marcellin, "Évaluation des critères asymétriques pour les arbres de décision," *Fabrice Guillet, and Brigitte Trousse (Eds.), EGC, (RNTI-E-11),* pp. 209-210, 2008.

[18] F. Weiss and Provost, "The Effect of Class Distribution on Classifier Learning: An Empirical Study," MLTR, 43, Dept. of Computer Science, Rutgers Univ, 2001.

[19] A. Estabrooks, T. Jo and N. Jakowicz, "A Multiple Resampling Method for Learning from Imbalanced Data Sets," *Computational Intelligence,* vol. 20, pp. 18-36, 2004.

[20] D. Mease, A. Wyner and A. Buja, "Boosted Classification Trees and Class Probability/Quantile Estimation," *J. Machine Learning Research,* vol. 8, pp. 409-439, 2007.

[21] L. A. a. B. P. R.C. Holte, "Concept Learning and the Problem of Small Disjuncts," *Proc. Int'l J. Conf. Artificial Intelligence,* pp. 813-818, 1989.

[22] C. Drummond and R.C. Holte, "C4.5, Class Imbalance, and Cost Sensitivity: Why Under Sampling Beats Over-Sampling," in *Workshop Learning from Imbalanced Data Sets II*, 2003.

[23] X.-Y. Liu, J. Wu and Z.-H. Zhou, "Exploratory Undersampling for Class-Imbalance Learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B ,* vol. 39, no. 2, pp. 539-550, April 2009.

[24] G. I. Webb, "MultiBoosting: A technique for combining boosting and wagging," *Machine Learning,* vol. 40, pp. 159-196, 2000.

[25] H. G. a. H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach," *ACM SIGKDD Explorations,* vol. 6, no. 1, pp. 30-39, 2004.

[26] J. Z. a. I. Mani, "KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction," *Proc. Int'l Conf Machine Learning (ICML'2003), Workshop Learning from Imbalanced Data Sets,* 2003.

[27] N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," *J.Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002.

[28] B. W. a. N. Japkowicz, "Imbalanced Data Set Learning with Synthetic Samples," *Proc. IRIS Machine Learning Workshop,* 2004.

[29] H. Han, W. Wang and a. B. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," *Proc. Int'l Conf. Intelligent Computing,* pp. 878-887, 2005.

[30] H. He, Y. Bai, E. Garcia and a. S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," *Proc. Int'l J. Conf. Neural Networks,* pp. 1322-132, 2008.

[31] M. Lebbah and Y. Bennani, " Sous-échantillonnage topographique par apprentissage semi-supervisé," *Extraction et Gestion des Connaissances (EGC),* vol. 19, pp. 121-126, 2010.

[32] I. Tomek, "Two Modifications of CNN," *IEEE Trans. System, Man, Cybernetics,* vol. 6, no. 11, pp. 769-772, 1976.

[33] M. K. a. S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One-Sided Selection," *Proc. Int'l Conf. Machine Learning,* pp. 79-186, 1997.

[34] R. P. a. M. M. G.E.A.P.A. Batista, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *ACM Sigkdd Explorations Newsletter,* vol. 6, no. 1, pp. 20-29, 2004.

[35] J. Laurikkala, "Improving Identification of Difficult Small Classes by Balancing Class Distribution," *Proc. Conf. AI in Medicine in Europe: Artificial Intelligence Medicine,* pp. 63-66, 2001.

[36] T. J. a. N. Japkowicz, "Class Imbalances versus Small Disjuncts," *ACM SIGKDD Explorations Newsletter,* vol. 6, no. 1, pp. 40-49, 2004.

[37] C. Elkan, "The Foundations of Cost-Sensitive Learning," *Proc. Int'l Joint Conf. Artificial Intelligence,* pp. 973-978, 2001.

[38] P. Domingos, "Metacost : a gênerai method for making classifiers cost-sensitive," *Proceedings ofthe fifth ACMSIGKDD international conférence on Knowledge discovery and data mining,* pp. 155-164, 1999.

[39] C. Ling, Q.Y.J.W and S. Zhang, "Decision trees with minimal costs," *ICML '04 Proceedings of the twenty-first international conference on Machine learning,* p. 69, 2004.

[40] Z.-H. Zhou and X.-Y. Liu, "On multi-class cost-sensitive learning," *Nat. Conf. on Artificial Intelligence,* pp. 567-572, 2006.

[41] G. M. Weiss, K. McCarthy and B. Zabar, "Cost-sensitive learning vs. sampling : Which is best for handling unbalanced classes with unequal error costs ?," *DMIN,* pp. 35-41, 2007.

[42] S. Z. D.-A. a. R. G. Marcellin, " An asymmetric entropy measure for décision trees," in *Ilth Information Processing and Management, of Uncertainty in knowledge-based Systems*, Paris, 2006, pp. 1292-1299.

[43] S. L. P. a. V. B. Lallich, "Construction of an off-centered entropy for supervised learning," in *Xllth International Symposium on Applied Stochastic Models and Data Analysis*, Chania, Crète, Greece, 2007.

[44] C. Chen and L. Breiman, "Using random forest to learn imbalanced data," Technical report, Berkeley, Department of Statistics, University of California, 2004.

[45] J. Du, Z. Cai and C. X. Ling, "Cost-sensitive décision trees with pre-pruning," *Canadian Conférence on AI,* pp. 171-179, 2007.

[46] M. Maloof, "Learning When Data Sets Are Imbalanced and When Costs Are Unequal and Unknown," *Proc. Int'l Conf. Machine Learning, Workshop Learning from Imbalanced Data Sets II,* 2003.

[47] T. F. a. R. K. F.J. Provost, "The Case against Accuracy Estimation for Comparing Induction Algorithms," *Proc. Int'l Conf. Machine Learning,* pp. 445-453, 1998.

[48] C. Metz, "Basic principles of roc analysis," *Seminars in Nuclear Medecine,* vol. 3, 1978.

[49] T. Fawcett, "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers," Technical Report HPL-2003-4, HP Labs, 2003.

[50] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters,* vol. 27, no. 8, pp. 861-874, 2006.

[51] J. D. a. M. Goadrich, "The Relationship between Precision- Recall and ROC Curves," *Proc. Int'l Conf. Machine Learning,* pp. 233-240, 2006.

[52] R. Drummond and C. Holte, "Cost Curves: An Improved Method for Visualizing Classifier Performance," *Machine Learning,* vol. 65, no. 1, pp. 95-130, 2006.

[53] R. Drummond and Holte, "Explicitly Representing Expected Cost: An Alternative to ROC Representation," *Proc. Int'l Conf. Knowledge Discovery and Data Mining,* pp. 198-207, 2000.

[54] R. Gurley, Intrusion détection, MacMillan Technical Publishing, 2000.

[55] L. Mé, Z. Marrakchi, C. Michel, H. Debar and F. Cuppens, " La détection d'intrusions : les outils doivent coopérer," *Revue de l'Electricité et de l'Electronique,* pp. 50-55, 2001.

[56] F. Meunier, " Detection d'intrusions : notions avancées de NIDS axées sur le logiciel ManHunt (Recourse Technologies)," *Watch4net,* 2002.

[57] C. Bidan, G. Hiet, L. Mé, B. Morin and J. Zimmermann, "Vers une détection d'intrusions à fiabilité et pertinence prouvables," *Revue de l'Electricité et de l'Electronique (REE),* vol. 9, 2006.

[58] D. E. Denning, "An intrusion détection model," *IEEE Transaction on Software Engineering,* pp.

222-232, 1987.

[59] S. Smaha and Haystack, "An intrusion détection System," *Proc. of the IEEE Fourth Aerospace Computer Security Applications Conférence,* pp. 37-44, 1988.

[60] S. Forrest, S. A. Hofmeyr, A. Somayaji and T. A. Longstaff, "A sensé of self for unix processes," *IEEE Symposium on Security and Privacy,* pp. 120-128, 1996.

[61] A. Valdes and K. Skinner, "Adaptive, model-based monitoring for cyber attack détection," *Récent Advances in Intrusion Détection (RAID 2000),* vol. 1970, pp. 80-92, 2000.

[62] C. Kruegel, D. Mutz, W. Robertson and F. Valeur, "Bayesian event classification for intrusion détection," *ACSAC '03 : Proceedings of the 19th Annual Computer Security Applications Conférence,* p. 14, 2003.

[63] S. J. Stolfo, W. Fan, W. Lee, A. Prodromidis and P. K. Chan, " Cost-based modeling and évaluation for data mining with application to fraud and intrusion détection.," Results frorn the JAM Project., 1999.

[64] M. Tavallaee, E. Bagheri, W. Lu and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," *IEEE Symposium on Computational Intelligence for Security and Defense Applications,* 2009.

[65] P. E. Hart, "The condensed nearest neighbor rule," *IEEE Trans. Information Theory,* p. 515–516, 1968.

[66] J. Sklansky and L. Michelotti, "Locally trained piecewise linear classifiers," *IEEE Trans. Pattern Anal. Machine Intelligence,* no. 2, p. 101–111, 1980.

[67] P. Munro, "Repeat until bored: A pattern selection strategy," *Adv. in Neural Inf. Proc. Sys.,* vol. 4, p. 1001–1008, 1992.

[68] C. Cachin, "Pedagogical pattern selection strategies," *Neural Networks,* vol. 7, p. 171–181, 1994.

[69] E. M. Strand and W. T. Jones, "An Active Pattern Set Strategy for Enhancing Generalization while Improving Back-Propagation Training Efficiency," *Proceedings of the International joint Conference on Neural Networks,* vol. 1, pp. 830-834, 1992.

[70] M. Plutowski and H. White, "Selecting concise training sets from clean data," *IEEE Trans. Neural Networks,* vol. 4, p. 305–318, 1993.

[71] A. Lyhyaoui, M. Martinez-Ramon, I. Mora-Jimenez, M. Vazquez-Castro, J. L. Sancho-Goméz and A. R. Figueiras-Vidal, "Sample Selection via Clustering to Construct Support Vector-like Classifiers," *IEEE Transactions on Neural Networks,* vol. 10, pp. 1474 - 1481, 1999.

[72] S. Choi and P. Rockett, "The training of neural classifiers with condensed datasets," *IEEE Trans. Sys., Man, and Cybernetics,* vol. 32, p. 202–206, 2002.

[73] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine Learning,* vol. 37, p. 297–336, 1999.

[74] V. Gomez-Verdejo, M. Ortega-Moral, J. Arenas- Garcia and A. R. Figueiras-Vidal, "Boosting by weighting critical and erroneous samples," *Neurocomputing,* vol. 69, p. 679–685, 2006.

[75] V. Gomez-Verdejo, J. Arenas-Garcia and A. R. Figueiras-Vidal, "A dynamically adjusted mixed emphasis method for building boosting ensembles," *IEEE Trans. Neural Networks,* vol. 19, p. 3–17, 2008.

[76] K. Huyser and A. Hororwitz, "Generalization in Connectionist Networks that realize boolean Functions," *Connectionnist Models Summer School,* pp. 191-200, 1988.

[77] J. Denker, D. Schwartz, B. Wittner, S. Solla, R. Howard and L. Jackel, "Large Automatic Learning, Rule Extraction, and Generalization," *Complex Systems1,* pp. 877-922, 1987.

[78] S. Haykin, Neural Network: A Comprehensive Foundation, New York: Macmillan, 1999.

[79] C. Bishop, Neural Networks for Pattern Recognition, New York: Oxford University Press Inc., 1995.

[80] M. Wann, T. Hidegir and N. Greenbaun, "The Influence of Training Sets on Generalization in Feed-Forward Neural Networks," *Proceedings of the International Joint Conference on Neural Networks,* vol. 3, pp. 137-142, 1990.

[81] N. Ohnishi, A. Okamoto and N. Sugi, "Selective Presentation of Learning Samples for Efficient Learning in Multilayer Perceptron," *Proceedings of the IEEE International Conference on Neural Networks,* vol. 1, pp. 688-690, 1991.

[82] R. K. M. Cheung, I. Lusting and A. L. Kornhauser, "Relative Effectiveness of Training Set Patterns for Back Propagation," *Proceedings of the IEEE International Conference on Neural Networks,* vol. 1, pp. 673-678, 1992.

[83] E. M. Strand and W. T. Jones, "An Active Pattern Set Strategy for Enhancing Generalization while Improving Back-Propagation Training Efficiency," *Proceedings of the International joint Conference on Neural Networks,* vol. 1, pp. 830-834, 1992.

[84] B. T. Zhang and G. Veenker, "Neural Networks That Teach Themselves Through Genetic Discovery of Novel Examples," *Proceedings of the International Joint Conference on Neural Networks,* vol. 1, pp. 690-685, 1991.

[85] B. T. Zhang and H. Mühlenbein, "Genetic Programming of Minimal Neural Nets Using Occam's Razor," *Proceedings of the International Conference Genetic Algorithms,* pp. 342-349, 1993.

[86] B. T. Zhang, "Accelerated Learning by Active Example Selection," *International Journal of Neural Networks.,* vol. 5, no. 1, pp. 67-75, 1994.

[87] B. T. Zhang, "An Incremental Learning Algorithm That Optimizes Network Size and Sample Size in One Trial," *Proceedings of the IEEE International Conference on Neural Networks,* pp. 215-220, 1994.

[88] E. I. Chang and R. P. Lippmann, "A Boundary Hunting Radial Basis Function Classifier which Allocates Centers Constructively," *Advances in Neural Information Processing Systems,* vol. 5, pp. 139-146, 1993.

[89] C. Cortes and V. Vapnik, "Support Vector Networks," *Machine Learning,* vol. 20, pp. 273-297, 1995.

[90] M. B. Almeida, A. Braga and J. P. Braga, "SVM-KM: Speeding SVMs Learning with a Priori Cluster Selection and k-Means," *Proceedings of the 6th Brazilian Symposium on Neural Networks,* pp. 162-167, 2000.

[91] H. J. Shin and S. Cho, "Pattern Selection for Support Vector Classifiers," *Proceedings of the 3rd International Conference on Intelligent Data Engineering and Automated Learning, Lecture Notes in Computer Science (LNCS 2412),* pp. 469-474, 2002.

[92] H. J. Shin and S. Cho, "Fast Pattern Selection for Support Vector Classifiers," *Proceedings of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Lecture Notes in Artificial Intelligence (LNAI 2637),* pp. 376-387, 2003.

[93] R. Shapire, "The strength of weak learnability," *Machine Learning,* vol. 5, pp. 197-227, 1990.

[94] Y. Freund and R. Schapire, "Experiments with a new Boosting algorithm," *Machine Learning Proceedings of the Thirteenth National Conférence,* pp. 148-156, 1996.

[95] N. Garcia-Pedrajas, "Constructing Ensembles of Classifiers by Means of Weighted Instance Selection," *Neural Networks, IEEE Transactions ,* vol. 20, no. 2, pp. 258-277, 2009.

[96] M. Dash and H. Liu, " Feature Selection for Classification," *Intelligent Data Analysis. INSTICC Press,* vol. 1, p. 131–156, 1997.

[97] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research,* no. 3, pp. 1157-1182, 2003.

[98] J. Gennari, P. Langley and D. Fisher, "Models of incremental concept formation," *Artificial Intelligence,* no. 40, pp. 11-61, 1989.

[99] R. Kohavi and J. John, "Wrappers for feature subset selection," *Artificial Intelligence,* vol. 7, no. 1-2, 1997.

[100] L. Molina, L. Belanche and A. Nebot, "Feature Selection Algorithms: A Survey and Experimental Evaluation," *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02),* p. 306, 2002.

[101] M. Hall, "Correlation-based feature selection for machine learning," PhD Thesis, Department of Computer Science, Waikato University, New Zealand, 2000.

[102] P. Langley and S. Sage, "Induction of selective Bayesian classifiers," *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence,* pp. 399-406, 1994.

[103] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol,* 1933.

[104] S. Nedevschi, I. Peter, I.-A. Dobos and C. Prodan, "An improved PCA type algorithm applied in face recognition," *Intelligent Computer Communication and Processing ,* pp. 259-262, 2010.

[105] J. Benzecri, L'Analyse des donées, vol. 1, Dunod, Ed., 1973.

[106] T. Kohonen, "The Self-Organizing Map," *Proc. IEEE,* vol. 78, pp. 1464-1480, 1990.

[107] S. Ahalt, C. Stanley, K. Krishnamurthy, P. Chen, Douglas and E. Melton, " Competitive learning algorithms for vector quantization," *Neural Networks,* vol. 3, pp. 277-290, 1990.

[108] J. Sklansky and L. Michelotti, "Locally trained piecewise linear classifiers," *IEEE Trans. Pattern Anal. Machine Intelligence,* vol. 2, p. 101–111, 1980.

[109] Bonwell and J. Eison, Active Learning: Creating Excitement in the Classroom, Jossey-Bass, 1991.

[110] S. Burr, "Active Learning Literature Survey.," Computer Sciences Technical Report 1648, University of Wisconsin, Madison, 2009.

[111] L. Yuhua and L. Maguire, "Selecting Critical Patterns Based on Local Geometrical and Statistical Information," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 33, pp. 1189 -1201, 2011.

[112] A. Lyhyaoui, Tesis doctoral : Classificadores RBF via técnicas de agrupamiento y selecton de muestras, Madrid: Universidad Carlos III de Madrid, 1999.