جامعة أبو بكر بلقايد – تلمسان

Université Abou Bakr Belkaïd de Tlemcen

Faculté de Technologie

Département de Génie électrique et Electronique

Laboratoire de Recherche de Génie Biomédical

MEMOIRE DE PROJET DE FIN D'ETUDES

pour obtenir le Diplôme de

MASTER en GENIE BIOMEDICAL

Spécialité : **Signaux et Images en Médecine**

présenté par : GAHDOUM HAFIDA

Classification des donnée déséquilibrée médicale

Soutenu le 24 juin 2013 devant le Jury

Dr.	Abderrahim MedElamine	*MCB*	Université de Tlemcen	Président
M.	Chikh Amine	*prof*	Université de Tlemcen	Encadreur
Melle	Settouti Nesma	*MAB*	Université de Tlemcen	Examinateur
Melle	Belaroussi Sara	*Doctorant*	Université de Tlemcen	Co-encadreur

Année universitaire 2012-2013

Didicace

Je remercie tout d'abord Dieu pour l'accomplissement de ce mémoire.

Je dédie ce modeste travail à

Ceux qui ont fait de moi l'homme que je suis aujourd'hui. Mes très chers parents, ma grande mère

Que dieu les récompense et les garde.

Mes frères YOUSEF-NESRADIN-FETHI-MOHAMED

Mes amis RAZZAK-ZOHIR-BILAL-HOSSEM-MALAK-SOHIL-YASSIN.

Mes collègues de promotion.

A tous ceux qui me connaissent de près ou de loin.

Merci d'être toujours là pour moi.

Remerciments

Avant toute chose, nous tenons à remercier Dieu pour l'accomplissement de ce projet.

Je tiens à exprimer ma très profonde gratitude à mon encadrant Mr. CHIKH Mohammed Amine qui n'a ménagé aucun effort pour me prendre en charge pour la réalisation de ce travail. Je le remercie pour le temps et les connaissances et la confiance qu'il m'a dispensé.

J'adresse mes sincères remerciements, a Melle Settouti Nesma qui me fait l'honneur de présider ce jury.

Je remercie tous particulièrement Mr.ABDERRAHIM qui ont accepté de juger ce travail.

Je remercie toute l'équipe du Laboratoire de Recherche de Génie Biomédical pour leur aide et leurs conseils, tout particulièrement Mlle BOUCHIKHI SARA et ATBI AMINA et BOBENZA AMINA et Bachir SAID .

Résumé

- La médecine est une discipline scientifique mais aussi une discipline d'action qui requiert souvent une prise de décision à partir d'un ensemble de données médicales.
- La classification des données médicales déséquilibrées constitue un problème majeur dans le domaine de la santé, ce problème concerne beaucoup plus les données médicales minoritaires et qui sont d'une grande importance.
- Dans ce travail et pour résoudre cette problématique, nous avons appliqué un algorithme des moindres carrées pour équilibrer les données médicales ensuite nous avons réalisé une classification neuronale de ces données.

Mots clés : déséquilibré, donnée médicale, réseaux neurone.

Abstract

- Medicine is a scientific discipline but also a discipline of action often requires a decision from a set of medical data.
- The classification of imbalanced medical data is a major problem in the field of health, the problem is much more minority medical data that are of great importance.
- In this work and to selves this problem, we applied a square algorithm to balance the medical evidence then we realized neuronal classification of these data less.

- **Keywords**: imbalanced, medical data, neuron networks.

Table des matières

Table des matières

Table des matières

Table des matières

Table des matières

Liste des figures

Liste des figures

Liste des tableaux

Liste d'abréviations

BCD: Bladder Cancer Dataset

BMI : Body Mass Index

HGPO : Hyper Glycémie Provoquée par voie Orale

FPG : The Fasting Plasma Glucose

IMC : Indice de masse corporelle

PMC : Perceptron multicouche

RLS: Regulation least mean square

RNA : Réseau de Neurones Artificiel

RNMC : Réseau de Neurones Multi-Couches

Se : Sensibilité

Sp : Spécificité

TC : Taux de classification

Te :Taux de erreur

E :Erreur

Introduction générale

- La problématique de classification du diabète est un domaine qui a fait l'objet de plusieurs recherches.
- Certaines de ces recherches, telles quelles seront exposées dans le chapitre 2 état de l'art contribuent vers la classification des données déséquilibrées pour l'aide au diagnostic médical, par une caractérisation des données médicales.
- Le travail que nous vous présentons dans ce mémoire de Master s'inscrit dans le contexte d'aide au diagnostic. Nous donnons un intérêt plus particulier à l'apprentissage automatique structurel et paramétrique, dans le but d'augmenter les performances pour classer les malades diabétiques minoritaires par rapport au non diabétiques majoritaires a partir Des techniques telles que les réseaux de neurones , ces derniers sont les plus couramment utilisées dans les systèmes de classification et sont développées par de grand nombre d'équipes de recherches.la performance du système peut diminuer grâce aux déséquilibres des données .
- Nous avons utilisé des techniques de moindre carrée pour équilibrer les données et faire un meilleur apprentissage et assurer une bonne performance du système de classification.

1. introduction:

Le diabète est une maladie inopinée, sournoise et silencieuse qui peut survenir d'un moment à l'autre. Cette maladie est classée parmi les plus émergentes et se propage à une vitesse fulgurante. Sa gravité réside dans ses impacts néfastes sur plusieurs organes du corps. Le diagnostic de cette pathologie consiste à classer le patient suivant deux situations « diabétique ou sain » en analysant un certain nombre de paramètres qui la caractérise. Et vue le nombre important d'individus et la complexité d'interprétation des paramètres, il est utile et important de faire appel aux systèmes de classification des données. D'où la nécessité de conception d'un système d'aide au diagnostic pour seconder le médecin d'une façon complémentaire et à réduire au minimum les erreurs possibles qui peuvent survenir ; et cela en examinant des données médicales dans un temps plus court et d'une façon plus détaillée et plus précise.

2. Contexte médical :

2.1. Définition :

Le diabète est défini comme une maladie caractérisée par une hyperglycémie pathologique se déclenche, le diabète provoque des symptômes spectaculaires connus depuis la plus haute antiquité. A long terme, ce sont les complications qui font la gravité de la maladie. En pratique, on distingue les diabètes insulinodépendants (diabète de type 1) marqués par une carence absolue en insuline et les diabètes non insulinodépendants (diabète de type 2), et peut provoquer hypertension, artérite, cécité, insuffisances rénales. L'Organisation Mondiale de la Santé chiffre le nombre de diabétiques dans le monde à plus de 180 millions en 2008 et estime que ce nombre devrait plus que doubler d'ici 2030. Il existe plusieurs types de diabètes correspondant à des altérations différentes du fonctionnement normal du Métabolisme.

2.2. Les principes types de diabète :

2.2.1. Diabète de type 1 :

Dans la nouvelle classification, adoptée par les spécialistes français depuis 1999, le terme diabète insulinodépendant doit être remplacé par diabète de type 1 touche généralement les enfants et les adultes de moins de 30/40 ans (sujets jeunes). Cette nouvelle classification se fonde en effet sur la pathogénie et non plus la thérapeutique des différents diabètes. Le

diabète de type 1 correspond à la destruction de la cellule B aboutissant habituellement à une carence absolue en insuline.

2.2.2. Diabète de type 2 :

Dans l'idée d'adopter une classification plus proche des mécanismes du diabète non insulinodépendant. De façon très insidieuse chez des personnes généralement plus âgées (classiquement > 40 ans) il existe en effet un assez petit nombre de diabètes de type 2 traités par l'insuline. Depuis le début du 19^{éme} siècle, il était admis que le diabétique pouvait ne pas mourir dans la mesure ou il observait une diététique appropriée. Il fallut cependant attendre lancer eaux pour qu'en 1880 apparaisse clairement la différence entre le diabète maigre d'évolution mortelle et le diabète pléthorique. La classification reste valable de nos jours ; elle a même été confortée par la mise en évidence au moyen des dosages radio-immunologiques que la sécrétion insulinique de base dans le diabète de type 2 pouvait être normale et même élevée.

2.2.3. Diabète gestationnel :

Le diabète gestationnel est un trouble de la tolérance glucidique de gravité variable, survenant ou diagnostiqué pour la première fois pendant la grossesse, malgré le traitement appliqué et son évolution après l'accouchement.

2.3. Les symptômes:

Dans les deux types de diabète, une glycémie très élevée peut être à l'origine de

Symptômes comme :

- Soif excessive.
- Fréquente envie d'uriner (polyurie).
- Fatigue.
- Troubles de la vue.
- Perte de poids.
- Envie de manger (polyphagie).

Ces symptômes révèlent généralement le diabète de type 1. Cependant, environ 50 % des personnes atteintes de diabète de type 2 sont asymptomatiques et le diabète est découvert de façon fortuite lors d'un bilan systématique ou lorsque le patient développe des complications.

2.4. Cause du diabète :

La prévalence de cette maladie a été multipliée par cinq en moins de cinquante ans. Cette augmentation progressive est due à divers facteurs :

- le vieillissement global de la population.
- l'augmentation de l'espérance de vie du diabétique.
- l'augmentation de la fécondité des femmes diabétique.
- l'augmentation de l'obésité,

2.5. Tests pour le diagnostic du diabète :

Les tests suivants sont utilisés pour le diagnostic :

Test glycémie à jeûne (FPG) mesure la glycémie chez une personne qui n'a rien mangé pendant au moins 8 heures. Ce test est utilisé pour détecter le diabète et le (pré diabète).

Résultats du glucose plasmatique (mg/dl)	diagnostique
<=99	Normal
100-125	Le prédiabète (anomalie de la glycémie à jeun)
>=126	Diabète

Tableau 1 : Diagnostic du test FPG[MBG06]

Test oral de tolérance au glucose (HGPO) la glycémie est mesurée après le jeune d'au moins 8 heures et ensuite 2 heures après consommation d'une boisson contenant du glucose. Ce test est appliqué pour diagnostiquer le diabète ainsi que son apparition (pré-diabète).

Résultats du glucose plasmatiqu après 2heures (mg/dl)	diagnostic
<=139	Normal
140-199	Le pré-diabète (tolérance au glucose)
>=200	Diabète

Tableau 2 : Diagnostic du test HGPO [MBG06]

Test aléatoire de glucose plasmatique appelé également test de plasma glucose occasionnel, comporte des mesures de glycémie prélevées aléatoirement après le dernier repas. Ce test ainsi que l'évaluation des symptômes, permettent de déceler le diabète, et non son apparition.

La prise	Résultat de glucose plasmatique (mg/dl)
A jeun	>=95
A 1heure	>=180
A 2 heures	>=155
A 3 heures	>=140

Tableau 3 : Exemple du diabète gestationnel avec le test aléatoire de glucose évaluant les valeurs supérieures à la normale [MBG06].

2.6. Le diabète dans le Monde :

– L'OMS a estimé le nombre de diabétiques dans le monde en 2000 à plus de 177 millions de cas. La projection en 2030 serait de 350 millions de cas, que seul un remède pourra freiner [Dia02]. (Voir figures 1 et 2)

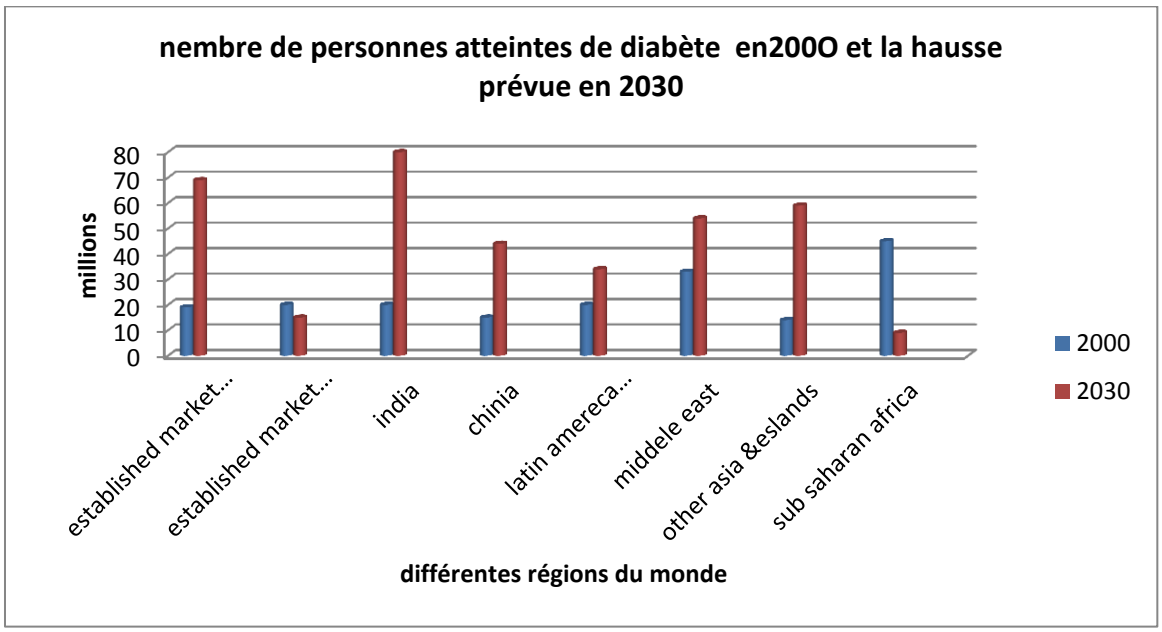

Figure 1 : Projection du nombre de personnes diabétiques dans différentes régions du monde. [Dia11].

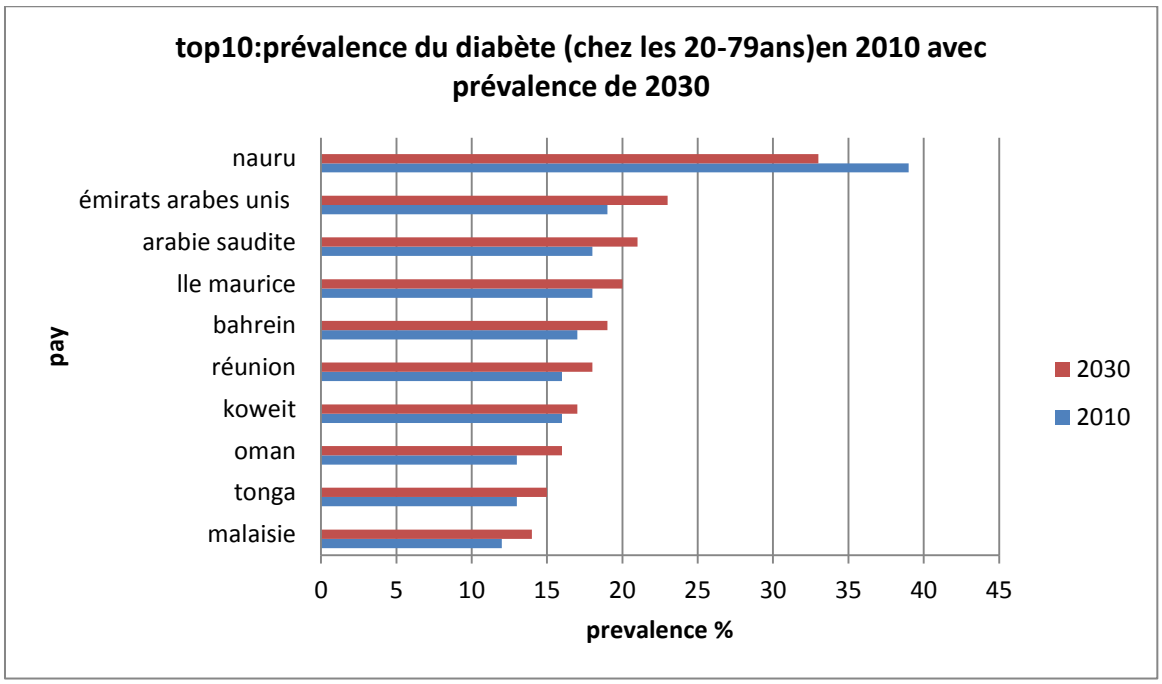

Figure 2 : Les 10 pays les plus touchés par le diabète en 2010 et leurs prévisions en 2025. [Alw11].

-En 2000, le diabète avait causé plus de 2,9 millions de décès dans le monde, ce qui le positionne au 5ème rang des principales causes de mortalité, selon les premières estimations de l'OMS [OMS07] ; (voir figure 3)

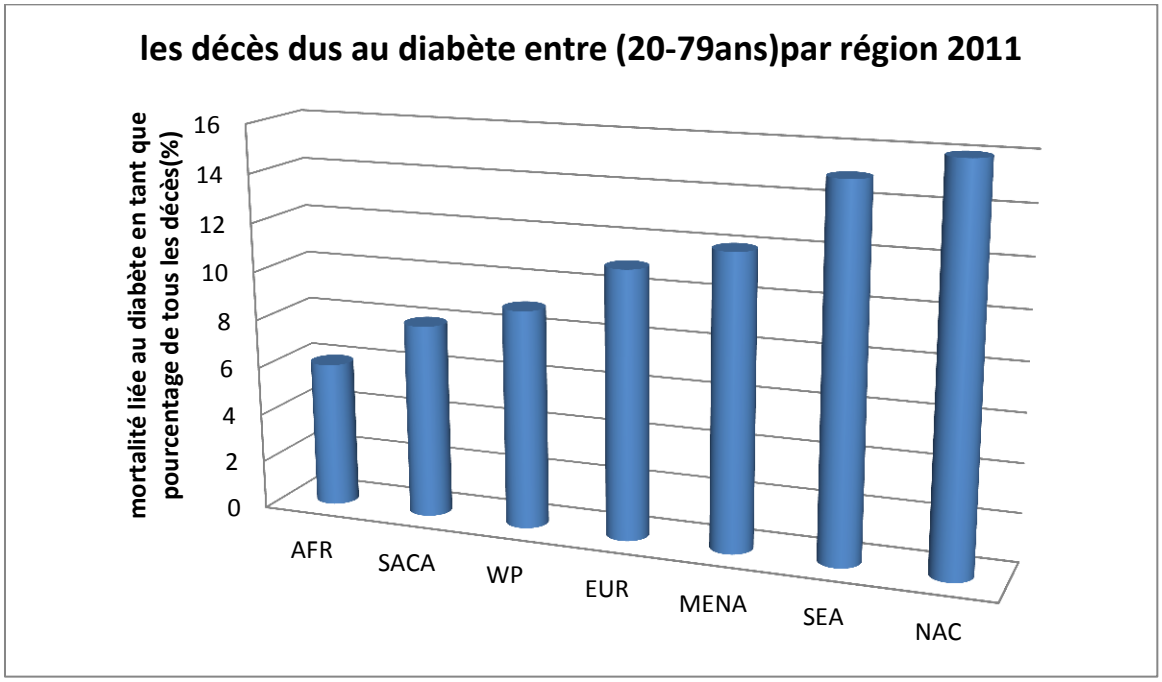

Figure 3 :% de Mort relative au diabète (20-79 ans) dans différentes régions (NAC : Amérique du Nord; SEA : Asie du sud-est; MENA : Moyen Orient; EUR : Europe; WP : Pacifique ouest; SACA : Amérique centrale et du Sud ; AFR : Afrique.)[Alw11].

– Dans le monde, la maladie rénale imputable au diabète (néphropathie diabétique) est la cause identifiable la plus répandue d'insuffisance rénale qui requiert une dialyse ou une greffe de rein. C'est aux Etats-Unis que l'incidence est la plus élevée : en effet, plus de 40% des personnes qui doivent suivre un traitement pour insuffisance rénale sont diabétiques [Dia03].

2.7. Diabète En Algérie :

Pour les praticiens algériens, le diabète est un véritable fléau. Il est considéré comme un sérieux problème de santé publique. Les diabétologues naviguent à vue car aucune étude épidémiologique n'a été lancée pour recenser les malades. Les seuls chiffres disponibles sont lancés par des experts du système de comptage de l'Organisation mondiale de la santé (OMS) qui estimaient que le nombre de diabétiques en Algérie est plus de 3 millions en 2011 dont 300 000 insulinodépendants (diabète de type 1) [Or11] et que le diabète est responsable de 4% des décès dans le pays (Figure 4).

Cette maladie est en progression continue à cause du changement rapide du mode de vie qui a tendance à s'aligner sur le modèle occidental de consommation alimentaire à base de graisse et de sucre.

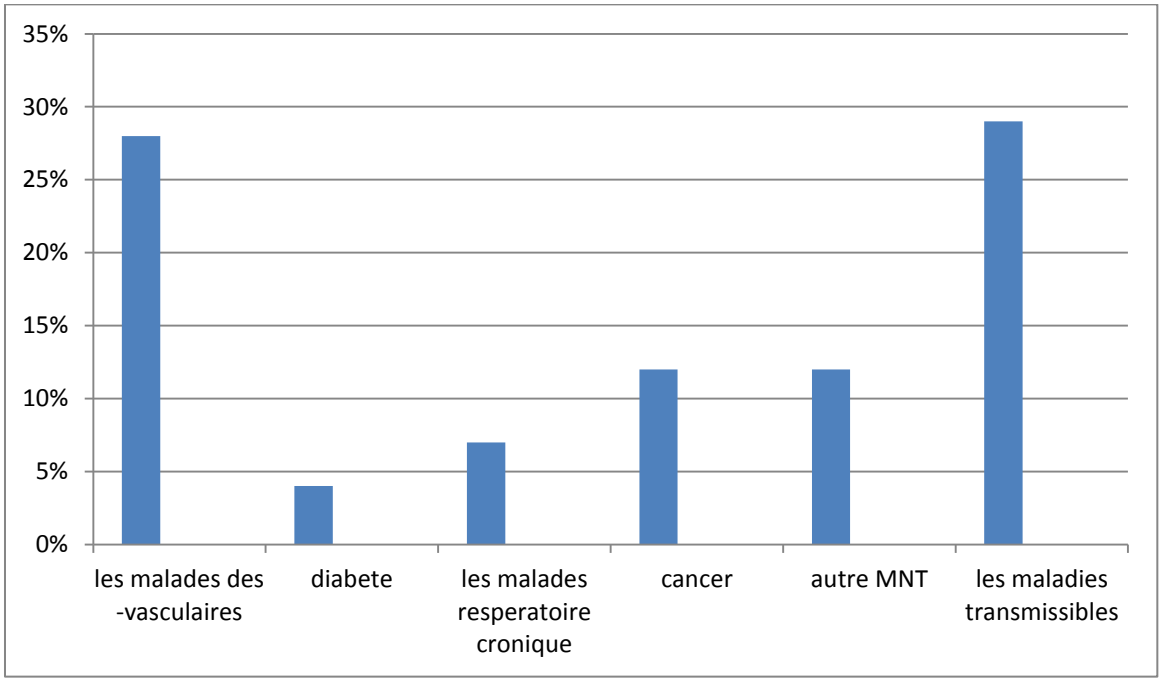

Figure 4 : Proportion de la mortalité en Algérie (% des décès).

2.8. Complications :

Le diabète sucré peut présenter des symptômes caractéristiques comme la soif, et la perte de poids. Dans ses formes les plus graves, une acidocétose ou un état hyperosmolaire non cétonique peut se développer et conduire à la stupeur, le coma et, en l'absence de traitement efficace, la mort. Des effets à long terme du diabète sucré comprennent le développement progressif des complications spécifiques de la rétinopathie et potentiellement de cécité, la néphropathie qui peut entraîner une insuffisance rénale, et / ou de neuropathie avec des risques d'ulcères du pied, amputation, articulations de Charcot, et les caractéristiques de disfonctionnement d'autonomie. Les personnes diabétiques courent un risque accru de maladie cardiovasculaires, cérébravasculaire et vasculaires périphériques.

2.9. Prévention

La prévention du diabète et ces complications implique adopter un ensemble d'actions afin d'éviter son apparition ou progression. Cette prévention peut être réalisée en trois niveaux :

2.9.1. Prévention primaire :

Elle a comme objectifs d'éviter la maladie. Dans la pratique c'est toute activité qui à lieu avant l'apparition de la maladie dans le but d'éviter son apparition, les actions suivantes sont proposées afin d'éviter la maladie :

- une éducation sanitaire principalement à travers de brochures, magazines, bulletins.
- Prévention et traitement de l'obésité en promouvant la consommation des régimes à faible teneur en matières grasses, de sucres raffinés et riche en fibres.
- Faire attention à l'indication des médicaments diabétogènes tels que les corticoïdes et enfin la stimulation de l'activité physique.

2.9.2. Prévention secondaire :

Elle permet principalement d'éviter les complications, en mettant sur la détection précoce du diabète comme stratégie de prévention. A ce niveau les objectifs sont :

- Chercher la remise, si c'est possible.
- Prévenir l'apparition de complications aiguës et chroniques.
- Ralentir la progression de la maladie.

- Les actions sont basées sur le contrôle métabolique optimal du diabète.

3. Facteurs de risque :

Plusieurs études ont tenté de découvrir les causes de cette maladie, parmi les variables qui peuvent Influencé l'apparition du diabète se trouve :

- Age: la prévalence augmente avec l'âge, pour les moins de 20 ans la possibilité d'avoir un Diabète est de 0.16 %, entre 20 et 65 ans est de 8.2 % et à partir de 65 ans le risque augmente Jusqu'à 20 %.
- Génétique: les enfants de mère diabétique ont plus de possibilité de développer un diabète.
- Nutrition: la possibilité d'avoir un diabète de type 2 augmente avec l'obésité. La graisse Abdominale est la plus dangereuse.
- Manque d'exercice: pour le diabète de type 2 le sédentarisme induit l'apparition de L'insulinoresistance. bientôt

- Infections: pour le diabète de type 1, l'incidence augmente en hiver et printemps, ce qui Conduit à penser que cette maladie pourrait être lié à certain virus.

4. Aide au diagnostic :

Le diagnostic du diabète semble s'adapter difficilement aux contraintes de l'informatique, si l'on en juge par le petit nombre de programmes existant actuellement. Cela se comprend aisément car :

- L'évocation d'un diagnostic fait intervenir non seulement des éléments objectifs mais aussi des données subjectives et interprétatives.
- il n'existe pas une seule forme d'une même maladie mais plusieurs, que nous appelons dans le jargon médical « formes cliniques ». Celles-ci se rencontrent plus souvent que la forme classique.
- Il existe de nombreuses formes d'une même affection liées à son degré de gravité et à son évolutivité. Certains signes peuvent apparaitre secondairement ou tardivement.
- Les signes principaux d'une maladie peuvent manquer sans que nous puissions rejeter la possibilité de cette affectio bien que les signes principaux de l'affection manquent, quelques signes secondaires doivent nous faire évoquer le diagnostic. Plusieurs catégories de systèmes informatiques peuvent résoudre ce type de problème tels que: les systèmes experts et les systèmes de classification.

Une première méthode possible pour résoudre ce type de problème est l'approche "systèmes experts". Dans ce contexte, la connaissance d'un expert (ou d'un groupe d'experts) est décrite sous forme de règles. Cet ensemble de règles forme un système expert qui est utilisé pour classifier de nouveaux cas. Cette procédure, largement employée dans les années 80, dépend fortement de la capacité à extraire et à formaliser les connaissances de l'expert. Nous considérons ici un autre procédé pour lequel la classification sera tirée automatiquement à partir d'un ensemble d'exemples. Un exemple consiste en la description d'un cas avec la classification correspondante. Il s'agit donc d'induire une technique de classification générale à partir de données. Le problème est alors inductif, il est question en effet d'établir sur la base d'exemples une règle générale. La procédure générée devra classifier correctement les

paramètres du patient mais surtout avoir un bon pouvoir prédictif pour classifier correctement de nouvelles descriptions.

5. Conclusion :

Nous avons vu dans ce chapitre les types de diabète, les différents traitements et testes ainsi que les complications du à cette maladie. Même s'il existe des méthodes de prévention qui permettent de réduire le risque d'avoir le diabète, parfois il impossible de l'éviter comme pour le diabète de type 1. Dans ces cas, Ce qui nécessite l'utilisation d'un système d'aide au diagnostic pour faciliter la prise de décision et combattre les complications.

1. Introduction :

Ils existent de nombreux problèmes complexes qui ne peuvent pas être résolu par un algorithme dans un temps optimal. A partir de 1980. Des méthodes nommées méta heuristiques ont commencé à apparaitre pour résoudre au mieux les problèmes complexes. Le diagnostic est parmi les activités les plus intéressantes dans la mise en œuvre de systèmes intelligents et robustes. Le principal objectif de ce chapitre est de développer un système automatisé d'aide à la décision pour la classification des données médicales déséquilibrés concernant le diabète. Le but est de déterminer un schéma de classification optimal avec une précision de diagnostic élevée pour une transparence maximale.

Le chapitre 2 présente dans la 1^{ére} partie la technique de classification et le réseau de neurone et dans la 2^{éme} partie les principales pistes de recherche Actuelles de l'apprentissage supervisé en situation d'asymétrie. Après quelques définitions de base, nous proposons d'identifier les problèmes liés à l'asymétrie, puis nous exposons les principales méthodes existantes pour essayer de régler ces problèmes.

2. Fonctionnement général des méthodes de classification :

2.1. Principe de la classification :

Une classe est un ensemble d'éléments qui sont semblables entre eux et qui sont dissemblables à ceux d'autres classes. Classifier consistera à maximiser les similarités des éléments qui sont dans la même classe et à minimiser les similarités de ces éléments avec ceux des autres classes. Inversement, on peut dire que classifier consiste à minimiser la variation intra-classe et à maximiser la variation interclasse.

2.2. Classification et techniques supervisées :

Quand on part d'un volume de données très important, on a intérêt à faire une classification préalable pour réduire l'espace de recherche des algorithmes supervisés. Comment mesurer la similarité ? Notion de distance entre les enregistrements c'est le premier problème inhérent à la classification.

2.2.1. Techniques inductives :

Elles présentent :

• une phase d'apprentissage qui permet d'élaborer un modèle. C'est la phase inductive. Remiser problème inhérent à la classification parmi la technique utilisée dans un classifiuer c'est le réseau de neurone multicouche.

• une phase de test pour vérifier le modèle obtenu (et éventuellement une phase de validation en plus).

Les phases d'apprentissage, de test et de validation sont effectuées sur des échantillons distincts de la population.

- L'objectif de l'apprentissage supervisé est naturellement de pouvoir utiliser les modèles construits sur de nouveaux individus. Or, calculer les différentes mesures d'évaluation sur les individus utilisés pour l'apprentissage donne souvent des valeurs bien trop optimistes (on parle alors de tests en ("par cœur"), et empêche de connaître la capacité de généralisation du modèle (sa capacité à bien se comporter sur de nouveaux individus).

  L'idée est alors de partitionner les exemples disponibles suivant différents protocoles.

- Apprentissage / test :

  C'est le protocole le plus simple.

  Il consiste à séparer le jeu de données en deux parties: l'une sera utilisée pour construire le modèle; l'autre pour le tester.

  Il faut juste définir la taille de chacun de ces deux échantillons. Généralement on utilise 70 % des individus pour construire le modèle.

  Et 30 % pour le tester.

  On parle alors de protocole "70/30".

  Le principal inconvénient de cette méthode est qu'il oblige à se passer de beaucoup d'individus pour la construction du modèle, et dans le même temps tous les individus ne sont pas exploités pour le test.

- Validation croisée :

  La validation croisée consiste à couper le jeu de départ en k subdivisions d'effectif équivalent par tirage aléatoire sans remise.

  Un modèle est ensuite construit sur k−1 subdivisions, et tester sur la subdivision restante.

Ce procédé est répété pour que chaque subdivision se retrouve une unique fois en test. Ainsi chaque individu s'est retrouvé une fois en test sur un modèle où il n'a pas été utilisé pour la construction.

La matrice de confusion obtenue contient donc tous les individus pour le calcul des différentes mesures.

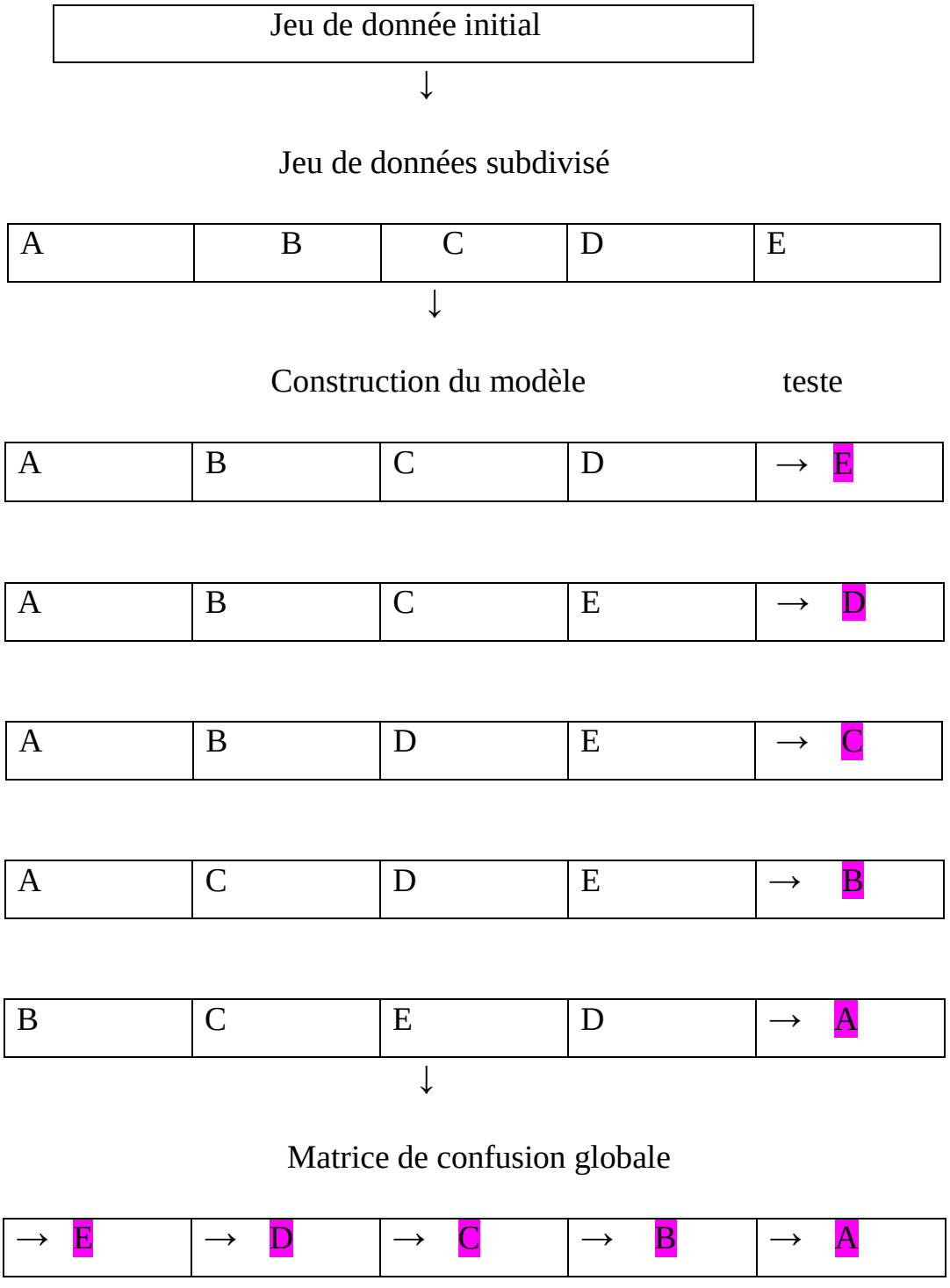

Figure 5 : principe d'une validation croisée 5 subdivisions.

Plus le nombre de subdivisions est importants, plus les déférents modèles construits utilisent d'individus simultanément, et se rapprochent donc d'un modèle construit sur la totalité des individus.

L'inconvénient devient alors le nombre de modèles à construire (égal au nombre de subdivisions), et les temps de calcul associés. Notons que la configuration extrême de la validation croisée ou chaque subdivision n'est constituée que d'un seul individu se nomme le (leave-one-out).

3. Réseaux de neurones:

Les réseaux de neurones ont d'abord été développés pour résoudre des problèmes de contrôle, de reconnaissance de formes ou de mots, de décision, de mémorisation comme une alternative à l'intelligence artificielle, et en relation plus ou moins étroite avec la modélisation de processus cognitifs (capable de connaître ou faire connaître) réels et des réseaux de neurones biologiques.

3.1. Le neurone biologique :

Le neurone biologique est une cellule vivante spécialisée dans le traitement des signaux électriques. Les neurones sont reliés entre eux par des liaisons appelées axones. Ces axones vont eux-mêmes jouer un rôle important dans le comportement logique de l'ensemble.

Ces axones Conduisent les signaux électriques de la sortie d'un neurone vers l'entrée (synapse) d'un autre neurone. Les neurones font une sommation des signaux reçus en entrée et en fonction du résultat obtenu vont fournir un courant en sortie. **(Figure 6)**

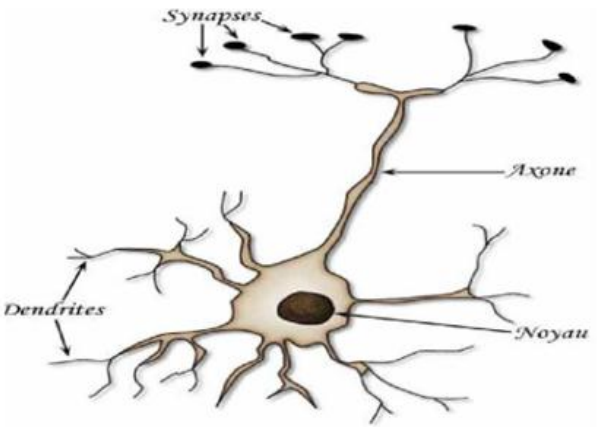

Figure 6 : le neurone biologique

3.1.1. La structure d'un neurone se compose de trois parties :

- La somma : ou cellule d'activité nerveuse, au centre du neurone.
- L'axone : attaché au somma qui est électriquement actif, ce dernier conduit L'impulsion conduite par le neurone.
- Dendrites : électriquement passives, elles reçoivent les impulsions d'autres neurones.

3.2. Le neurone formel (artificiel) (RNA) :

Le neurone artificiel (ou cellule) est un processeur élémentaire. Il reçoit un nombre variable d'entrées en provenance de neurones appartenant à un niveau situé en amont (on parlera de neurones "amont"). A chacune des entrées est associé un poids w représentatif de la force de la connexion. Chaque processeur élémentaire est doté d'une sortie unique, qui se ramifie. Ensuite pour alimenter un nombre variable de neurones appartenant à un niveau situé en aval (on parlera de neurones "avals").

A chaque connexion est associe un poids**. (Figure 7)**

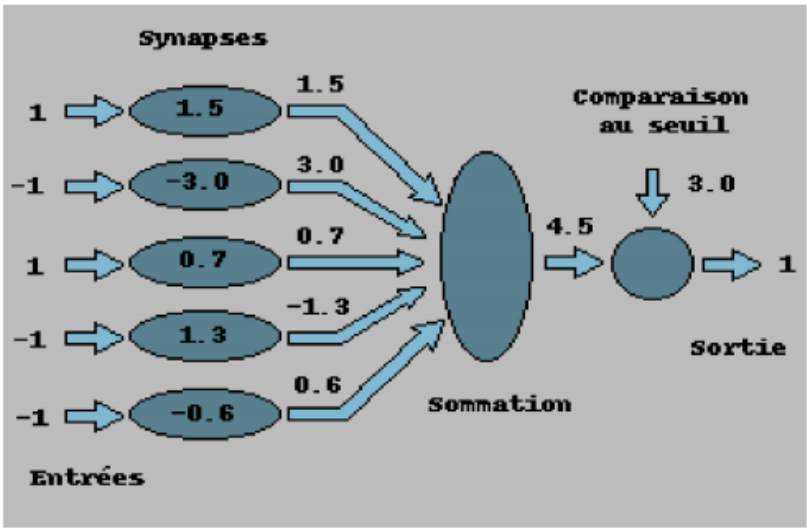

Figure 7 : Le neurone formel

3.2.1. Modélisation d'un neurone formel :

Les réseaux de neurones formels sont à l'origine d'une tentative de modélisation mathématique du cerveau humain. Les premiers travaux datent de 1943 [MP43] (et sont l'œuvre de MM. Mac Culloch et Pitts.

Ils présentent un modèle assez simple pour les neurones et explorent les possibilités de ce modèle.

La modélisation consiste à mettre en œuvre un système de réseau neuronal sous un aspect non pas biologique mais artificiel, cela suppose que d'après le principe biologique on aura une correspondance pour chaque élément composant le neurone biologique, donc une modélisation pour chacun d'entre eux.

On pourra résumer cette modélisation par **le tableau 4**, qui nous permettra de voir clairement la transition entre le neurone biologique et le neurone formel [Gur97].

Neurone biologique	**Neurone artificiel**
Synapses	Poids de connexions
Axones	Signal de sortie
Dendrite	Signal d'entrée
Somma	Fonction d'activation

Tableau 4 : la transition entre le neurone biologique et le neurone formel

3.3. Topologie en couches :

Si les unités élémentaires sont souvent très proches dans la plupart des systèmes neuronaux, c'est au niveau d'architecture de ces neurones que les systèmes se différencient. Il a été constaté que si les neurones sont placés en couches successives (les sorties d'un certain nombre de neurones sont les entrées des suivants et ainsi de suite jusqu'à la sortie), alors l'ensemble du réseau est capable de décider des problèmes plus complexes et peut aussi simuler n'importe quelle fonction booléenne. Ce type d'organisation (Figure 8) est appelé perceptron multicouche (PMC).

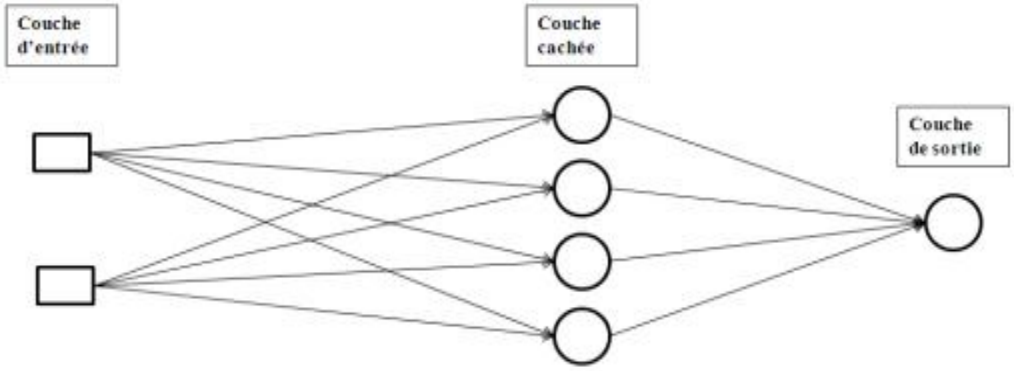

Figure 8 : Architecture d'un perceptron multicouche

Les unités de calcul ne sont plus appelées perceptrons mais plus simplement neurones ou encore nœuds. Outre la topologie en couches, la principale différence avec la version de [Fa] vient de l'utilisation de fonctions d'activation dérivables et non linéaires telles que la sigmoïde, encore appelée fonction logistique.

L'idée d'une telle topologie est ancienne et il a fallu attendre un certain nombre d'années pour voir apparaitre des algorithmes permettant de calculer les poids d'un tel réseau en particulier à cause de l'introduction des couches cachées. Proposé pour la première fois par [Pg] en 1974, l'utilisation de la rétro propagation du gradient de l'erreur dans des systèmes à plusieurs couches sera de nouveau mise au-devant de la scène en 1986 par [Pc], et simultanément, sous une appellation voisine, chez [Pcv] durant sa thèse.

Ces réseaux sont souvent totalement connectés, ce qui signifie que chaque neurone d'une couche i est connecté à tous les neurones de la couche i+1. Par contre, dans un schéma classique, les neurones d'une même couche ne sont jamais reliés entre eux.

Les PMC sont essentiellement employés à deux taches: le partitionnement d'un espace de formes pour des problèmes de classification et l'approximation de fonctions. Contrairement au perceptron de [Fa], le PMC peut représenter n'importe quelle fonction booléenne à n variables, bien que certaines puissent requérir un nombre exponentiel en n de neurones dans les couches cachées. Du fait de la non-linéarité de la sigmoïde comme fonction d'activation, les frontières de séparation s'adaptent mieux à chaque classe dans le cas d'un problème de

classification. Cette propriété se retrouve aussi dans le cas de l'approximation de fonctions qui produit des courbes continues et lisses à la fois. Les PMC possèdent des propriétés mathématiques intéressantes. Beaucoup d'entré elles sont valables pour des réseaux a seulement deux couches cachées, ce qui témoigne de la puissance potentielle des PMC. Il est à noter que ces propriétés sont rarement constructives dans le sens ou bien qu'il soit démontré qu'un certain nombre de neurones soit suffisant pour réaliser une tache, la propriété ne donne aucune information sur la topologie à choisir afin de résoudre le problème (Annexe A .Propriétés mathématiques).

La majeure partie des propriétés sont prouvées sans l'hypothèse de l'utilisation de la sigmoïde, il suffit simplement que la fonction d'activation soit bornée (majorée et minorée), croissante et continue.

La difficulté d'utilisation de ce réseau réside dans le fait qu'il faille déterminer sa topologie, il s'agit de définir le nombre de neurones des différentes couches ainsi que leurs interconnexions. Si le nombre de neurones cachés est trop faible, l'algorithme d'apprentissage n'arrivera pas à construire une représentation intermédiaire du problème qui soit linéairement séparable et certains des exemples ne seront pas appris correctement. Inversement, si ce nombre est trop élevé, il y a risque d'apprentissage par cœur du problème : le réseau reconnait parfaitement les exemples d'apprentissage mais donnera des résultats médiocres sur des nouvelles données qu'il n'a pas vues durant l'apprentissage.

3.5.2. Apprentissage :

L'approche la plus connue pour l'apprentissage d'un PMC est la technique de descente de gradient. En effet, l'utilisation de fonctions d'activation différentiables permet d'utiliser cette technique à la fois simple à mettre en œuvre et surtout très efficace sur le plan calculatoire.

Le problème se résume à :

$$\text{Min } E(w) \qquad \text{(II.5)}$$

Pour résoudre ce type de problème, une technique classique d'optimisation issue de la recherche opérationnelle consiste à déterminer par itérations successives les valeurs du paramètre des poids synaptique . Au regard des objets à manipuler, la descente de gradient est une réponse adéquate à ce problème. Elle consiste à utiliser un point existant w0 et lui faire effectuer un déplacement dans la direction de l'anti gradient. Le nouveau point obtenu par la

translation $w \rightarrow w+\mu x_p \ \sigma_p$ a une plus petite valeur pour la fonction objective. Le paramètre est un pas positif appelé dans le cas présent pas d'apprentissage est le gradient de l'erreur. L'opération de translation est répétée jusqu'à l'obtention d'une solution satisfaisante. En utilisant la rétro propagation du gradient de l'erreur, le résumé du déroulement de la méthode est donné par (Algorithme 1).

Algorithme 1 : Apprentissage d'un PMC par rétro-propagation du gradient

1: Initialisation aléatoire des poids du réseau
2: répéter pour chaque échantillon de la base d'apprentissage faire
- Propager l'échantillon dans le réseau
- Calcule de l'erreur sur la couche de sortie
- Propagation de l'erreur sur les couches inférieures
- Ajustement des poids
Fin
Mise à jour de l'erreur totale
Jusqu'à Critère d'arrêt

Bien que l'erreur soit minimisée localement, la technique permet de converger vers un minimum et donne de bons résultats pratiques. Dans la plupart des cas, peu de problèmes dus aux minima locaux sont rencontrés. Il persiste cependant deux problèmes que l'on rencontre dans une application réelle qui sont d'une part la lenteur de la convergence si est mal choisi et d'autre part le possible risque de converger vers un minimum local et non global de la surface d'erreur.

Le principal défaut de cette méthode est un temps de convergence restant assez long qui dépend de différents paramètres comme l'initialisation à l'instant t= 0 des poids synaptiques ou de la valeur initiale du paramètre. Il n'en reste pas moins qu'elle donne de bons résultats expérimentaux.

Dans une implémentation de l'algorithme de rétro propagation de l'erreur, il est aussi difficile de déterminer quand l'ajustement des poids du PMC doit s'achever. Plusieurs critères d'arrêt sont employés : les itérations cessent quand la norme du gradient est proche de zéro (les poids ne varient alors que très peu), ou bien alors dés que l'erreur en sortie est en dessous d'un certain seuil.

Le premier critère est plus intéressant mathématiquement car il correspond à la stabilisation de la solution dans un minimum, le second est plus proche de critères réels (interprétables) de bonne corrélation entre solution calculée et solution attendue. Dans ce dernier cas, si le problème étudié concerne une tâche de classification, on peut considérer que l'apprentissage s'achève quand toutes les formes sont classifiées, ce qui permet de s'affranchir de la détermination du taux d'erreur à ne pas dépasser. En pratique, nous allions ce dernier critère d'arrêt à un deuxième qui tient compte d'un nombre maximum d'itérations à ne pas franchir. En effet, il n'est pas garanti que le réseau puisse classifier toutes les formes, même avec un nombre infini d'itérations. La combinaison des deux conditions permet d'obtenir une solution correcte dans un temps raisonnable.

La validation croisée est une technique s'assurant principalement de la bonne généralisation du réseau, c'est-à-dire de son bon fonctionnement sur de nouveaux échantillons. Elle consiste à utiliser deux bases : l'une pour l'apprentissage et l'autre pour le test d'arrêt.

La première, comme son nom l'indique, sert uniquement à l'algorithme de rétro-propagation du gradient, la seconde permet de tester, à la fin de chaque itération, la qualité du réseau. Tant que l'erreur globale du réseau sur la base de test diminue, les itérations continuent. Dès que l'erreur augmente, l'apprentissage est stoppé même s'il aurait été possible de diminuer encore l'erreur sur la base d'apprentissage. Cette solution, quand on dispose d'un grand nombre d'échantillons permet d'éviter le phénomène de sur-apprentissage sur les données d'apprentissage ayant comme conséquence une mauvaise généralisation [Rp]. Les résultats en termes de taux de lecture ou de taux d'erreur sont alors donnés pour une troisième base, indépendante des deux autres, nommée base de validation. Reste le choix des échantillons lors de l'apprentissage qui est aussi un problème crucial. Il existe dans ce domaine peu de résultats théoriques concernant la création d'une «bonne» base d'apprentissage. Il est évident que dans un cas réel, afin d'avoir une bonne fiabilité et un grand pouvoir de généralisation, les exemples doivent être d'autant plus nombreux que le problème est complexe et sa topologie peu structurée. Et pour éviter des phénomènes de sur apprentissage de certaines classes, il est recommandé de fournir au réseau un nombre d'exemples similaire pour chacune des classes et de les présenter de manière aléatoire lors de l'apprentissage.

4. Problèmes de déséquilibre de classes :

4.1. Problématique :

Dans beaucoup de problèmes de classification réelle, la base de données est déséquilibrée: les différentes classes ne sont pas représentées de manière équitable dans l'ensemble d'apprentissage. Un déséquilibre trop important affecte généralement négativement la précision des algorithmes d'apprentissage (qui ont tendance à favoriser la classe majoritaire) et différentes méthodes ont été proposées dans la littérature pour pallier à ce problème.

4.1.1. Apprentissage supervisé et asymétrie :

Comme nous l'avons souligné en introduction , la prise en compte de l'asymétrie des classes en apprentissage est un problème relativement récent apparu dès lors que le data mining est devenu une technologie amplement utilisée dans l'industrie ,dans des exemples réels comme le diagnostic des maladies de la thyroïde [MA94], la gestion des défauts des boites de vitesses des hélicoptères[JMG95], la détection de fraudes téléphoniques [FP97], ou encore la recherche de gisements de pétrole sur des images satellites [KHM98]. L'asymétrie est devenu un défi majeur de l'apprentissage supervisé, le déséquilibre des jeux de données pouvant atteindre 1 pour 100, 1 pour 1000, 1 pour 10000 et souvent encore plus [CJK04]. Comme le notent Florian Ver hein et Sanjay Chawla [VC07] "dans des applications comme le diagnostic médical ou la détection de fraudes, les jeux de données déséquilibrés sont la norme et non l'exception". La communauté scientifique du data mining s'est attelée à ces problèmes, et les ateliers qui y sont consacrés dans les principales conférences [WCS00, WLI00, WLI03, ID-03, SI-04] témoignent de l'ampleur des définitions qui sont posés.

- Une catégorie de l'apprentissage supervisé a donc émergé, regroupant plusieurs types de problématiques, que l'on peut résumer à deux thèmes principaux : l'apprentissage sur données déséquilibrées d'une part ; et l'asymétrie des coûts d'autre part.

- Mais nous verrons que ces deux problèmes sont étroitement liés. Avant d'aborder les approches que nous avons considéré pour traiter le problème de l'asymétrie dans la classification diabétique, il est donc indispensable de présenter les principaux problèmes, méthodes et résultats du domaine de l'apprentissage supervisé en situation d'asymétrie.

4.1.2. Problèmes d'asymétrie :

Dans un article très complet sur le sujet ; Gary Weiss [Wei04] propose de les différents problèmes de l'asymétrie, en associant à chacun les méthodes adaptées. Ses conclusions sont les suivantes :

- le problème de la rareté absolue doit être résolu par du sur-échantillonnage, tandis que la rareté relative peut être abordée avec l'ensemble des méthodes que nous avons présenté.
- Le problème des métriques inadaptées doit être réglé en le remplaçant par des mesures sensibles aux coûts.
- Le problème des données bruitées peut être considéré sous l'angle de la marge d'induction à adapter, ou par des méthodes d'échantillonnage avancées.
- Enfin la fragmentation des données peut être gérée par des méthodes d'apprentissage ciblées sur une classe, ou par des techniques d'échantillonnage.

Des études ont été menées pour déterminer quelles sont les méthodes les plus adaptées selon les caractéristiques du problème. Concernant l'échantillonnage, Houles [HKN07] préconise le sous-échantillonnage tant que le déséquilibre n'est pas trop fort (jusqu'à environ 10%), préférant le sur-échantillonnage lorsqu'il est au-delà (conclusion partagée par [BSGR03]). Japkowicz [Jap00b] confirme également Houles lorsqu'il préconise les méthodes d'échantillonnage aléatoires simples, ne constatant pas d'améliorations significatives avec les méthodes plus fines mais plus couteuses. Ce même auteur remarque par ailleurs que lorsque les classes sont facilement séparables, le déséquilibre affecte peu les modèles [Jap00a, VR05]. De plus un important résultat de Weiss et Prouvost montre que l'équilibre des classes n'est pas forcement la distribution qui permet d'avoir systématiquement les meilleurs résultats [WP03] ; la question que se posent à présent de nombreux auteurs est «quelle est la meilleure distribution ?"[VR05]. Enfin concernant les arbres de décision, Elkan [Elk01] préfère modifier uniquement le seuil de décision plutôt que modifier la composition du jeu de données.

Weiss propose une étude comparative entre approches sensibles aux coûts, sur-échantillonnage et sous-échantillonnage [WMZ07]. Constatant qu'aucune méthode ne domine les autres systématiquement, il évalue ces différentes approches en fonction des caractéristiques du problème. Il conclue que sur les grands jeux de données (plus de 10 000 individus), l'apprentissage sensible aux coûts fournit de meilleurs résultats que

l'échantillonnage. Sur les petits jeux de données par contre, c'est le sur-échantillonnage qui l'emporte. Enfin d'une manière générale, les auteurs notent qu'on ne peut pas départager le sur-échantillonnage du sous-échantillonnage : les résultats varient beaucoup d'un jeu de données à l'autre. On retrouvé aussi que la méthode de moindre carré utiliser pour résoudre le problème de déséquilibre dans le processus d'apprentissage parce que cette méthode utilisée dans le chapitre 3 comme un algorithme de régulation de la base de donnés.

4.2. Notation et concepts :

Nous positionnerons enfin, Notre travail vis-à-vis du domaine de notations et concepts. On travaille à partir d'un ensemble d'apprentissage composé de individus noté $\Omega = \{w_1, .., w_n\}$, décris par un ensemble de variables descriptives $X = \{x_1, \ldots, x_p\}$ numériques ou catégorielles. A chaque individu est associé une classe c_i, appartenant à un ensemble de k classes $C=\{c_1, \ldots, c_k\}$.

L'objectif de l'apprentissage supervisé est de construire un modèle de prédiction associant une classe (ou variable endogène) c_i à tout nouvel individu dont on ne connaît que les descripteurs. Une distribution de probabilités des classes de la variable endogène est notée : $(p_1, \ldots, p_k)$, tel que $P_i \geq 0$ et $\sum_{i=1}^{k} P_i = 1$ Nous noterons $I= (\frac{1}{k}, \ldots, \frac{1}{k})$ la distribution uniforme des classes.

4.3. Apprentissage supervisé et classification d'asymétries :

Une grande partie des travaux sur l'apprentissage sur données déséquilibrées considère des problèmes à deux classes, où l'une minoritaire, est également la classe la plus importante (on parle également de classe d'intérêt). Les individus de la classe minoritaire sont appelés les individus positifs (et la classe positive est notée c+), et ceux de la classe majoritaire sont les individus négatifs (et la classe négative.est notée c−).

4.4. Problèmes de l'asymétrie en apprentissage supervisé

Différentes études [CJK04, Wei04] résument les problématiques et les avancées de l'apprentissage en situation d'asymétrie. Dans cette section nous allons présenter les différents problèmes que peut recouvrir le terme d'asymétrie.

4.4.1. Apprentissage sur données déséquilibrés :

Foster Prouvost rappelle dans l'éditorial de l'atelier sur les jeux de données déséquilibrés de la conférence AIII [Pro00] les fondements de la prise en compte du déséquilibre des classes en apprentissage. La plupart des algorithmes sont basés sur deux hypothèses : (1) le critère à minimiser est le nombre d'erreurs et (2) le jeu de données d'apprentissage est un échantillon représentatif de la population sur laquelle le modèle sera appliqué. Ce sont ces deux hypothèses qui font que les modèles ne sont pas satisfaisants quand ils sont construits à partir de données déséquilibrées.

On peut l'illustrer par un exemple simple : si 99 % des donnée s'appartiennent à une seule classe, il sera difficile de faire mieux que le 1% d'erreur obtenu en classant tous les individus dans cette classe : selon les hypothèses que nous venons de citer c'est même la meilleure chose à faire. Il convient donc de vérifier dans quelle mesure il est possible de se passer de ces hypothèses sans remettre en cause les fondements des algorithmes. Propose de distinguer plus précisément les différents problèmes des données déséquilibrées, et de l'apprentissage des classes rares. Nous allons les exposer dans points suivants.

a) Métriques inappropriées:

Que ce soit pour guider l'apprentissage, ou pour en évaluer les résultats, les mesures utilisées au cours du processus d'apprentissage ne sont pas adaptées aux classes déséquilibrées. En remplaçant le critère à optimiser, par exemple le taux d'erreur, par un critère plus pertinent, on doit pouvoir adapter simplement les algorithmes. Le rappel et la précision sont par exemple plus adéquats. Manque "absolu" de données: Il s'agit du problème principal du déséquilibre : les données disponibles ne sont pas suffisantes pour définir clairement le concept.

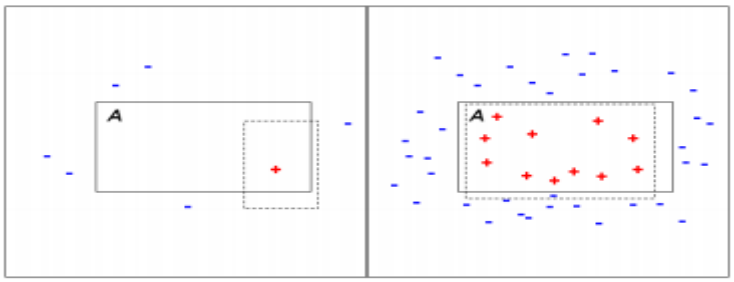

Figure 9 : L'effet d'un manque "absolu" de données

La figure 9 illustre ce principe : dans l'échantillon de la partie gauche les données disponibles concernant le concept A ne sont pas suffisantes pour définir ses frontières, estimées par les rectangles en pointillés. Le concept est beaucoup mieux défini sur l'échantillon de droite, pour lequel plus de données sont disponibles. Ce problème est à mettre en relation avec celui des "cas rares"(Small disjuncts) [Wei03] : différents sous-concepts d'une classe d'intérêt sont représentés par quelques individus dans des régions spécifiques de l'espace. On a alors un problème de manque absolu de données, mais uniquement pour une sous-partie de la classe.

b) Manque relatif de données:

Les objets d'une classe ne sont pas rares au sens absolu, mais beaucoup moins représentés que ceux des autres classes. Le problème est donc le ratio classe majoritaire/classe minoritaire plus que le nombre d'individus disponible pour apprendre le concept de la classe minoritaire : apprendre sur un jeu de données où la répartition est 5 : 95 (manque absolu) est un problème très différent d'un autre où la répartition est 500 : 9500(manque relatif). On a donc un problème de déséquilibre que Weiss [Wei04] illustre par la phrase populaire "trouver une aiguille dans une botte de foin". De nombreuses méthodes peuvent être utilisées pour gérer ce problème.

c) Fragmentation des données:

Ce problème est lié aux algorithmes ayant une approche "divise et conquiert" (Divide and conquer), comme les arbres de décision, qui partent de l'espace de tous les individus et le partitionnent récursivement en sous-espaces de plus en plus petit. Les invariants sont à chercher dans de petites partitions contenant de moins en moins de données. Si la fragmentation des données est toujours un problème, elle l'est encore plus dans le cadre de l'apprentissage des cas rares et est à mettre en relation avec le problème de "manque de données".

d) Marge d'induction inappropriée:

Il s'agit de la marge appliquée à la règle apprise sur les données d'apprentissage pour pouvoir généraliser. On peut considérer la marge de généralité maximum : une fois qu'on a sélectionné un groupe d'individu comme appartenant à un même concept, on définit ce dernier grâce au nombre minimum de conditions qui mènent à ces individus. A l'inverse la marge de spécificité maximum va quant à elle conserver toutes les règles possibles satisfaites par ces individus. Or la marge de généralité maximum est appropriée aux cas

fréquents mais pas aux cas rares. Pourtant de nombreux systèmes d'induction préfèrent la généralité à la spécialisation, favorisant la classe la plus présente en cas d'incertitude. Cette question est un point fondamental de l'apprentissage en situation d'asymétrie.

4.4.2. Asymétrie des coûts :

L'apprentissage sensible aux coûts répond à une problématique un peu différente que l'apprentissage sur données déséquilibrées, mais les deux sont très liés. L'objectif est de prendre en compte l'asymétrie des classes en termes d'importance, ou de coûts des erreurs. Ainsi dans notre exemple d'aide au diagnostic médical, si faire une erreur sur la classe majoritaire (classer comme malade un individu sain) est coûteux en termes d'examens inutiles et de stress pour le patient, faire une erreur sur la classe minoritaire (ne pas détecter la maladie chez un patient) est bien plus grave : elle peut entraîner des complications, voir le décès de la personne. Cette asymétrie des coûts n'est pas prise en compte par les systèmes d'apprentissage basiques. De plus les coûts ne sont généralement pas précisément connus [ZE01a].

Ce problème est lié à celui du déséquilibre, car bien souvent les classes rares sont les plus importantes, et les erreurs sur ces dernières sont plus coûteuses. Nous verrons que les problèmes comme les méthodes de l'apprentissage sensible aux coûts (Cost-sensitive learning) peuvent être adaptées dans le cadre de l'apprentissage sur données déséquilibrées, comme le préconise par exemple Maloof [Mal03] en observant que l'échantillonnage, l'ajustement d'une matrice de coûts et le déplacement du seuil de décision ont des effets similaires, il existe différents types d'asymétries des coûts en apprentissage. Le type de coûts le plus traité dans la littérature concerne les coûts de mauvaises classifications, dont nous parlerons plus bas ; mais des méthodes existent pour tenir compte des coûts de tests (coût d'acquérir la valeur d'une variable pour un individu [Tan93]), ainsi que du couplage des deux [Tur95,DHR+06].

4.5. Apprentissage supervisé sensible à l'asymétrie :

Différentes méthodes ont été proposées pour traiter les problèmes de l'asymétrie en apprentissage supervisé, que nous pouvons regrouper en deux catégories principales [KKP06, Wei04] :

1. Au niveau des données, les stratégies d'échantillonnage permettent de redresser les jeux de données déséquilibrés, ou de constituer des échantillons de manière dirigée pour encourager les algorithmes d'apprentissage à se diriger vers un type de modèle spécifique.

2. Au niveau algorithmique, on retrouve des méthodes qui tiennent intrinsèquement compte de l'asymétrie via une matrice de coûts, une distribution de référence, ou des objectifs spécifiques spécifiés par l'utilisateur.

Des approches ensemble permettent de rendre n'importe quel type d'algorithme sensible à l'asymétrie, notamment par des méthodes de boosting ou dubagging.

4.5.1 Stratégies d'échantillonnage :

Un premier type d'approches pour gérer le déséquilibre des jeux de données est de les redresser par des méthodes d'échantillonnage. Deux catégories ont été considérées : le sous-échantillonnage de la classe majoritaire, ou le sur-échantillonnage de la classe minoritaire.

a) Sous-échantillonnage:

La méthode la plus évidente et la plus simple consiste à supprimer aléatoirement du jeu d'apprentissage des individus appartenant à la classe majoritaire, de manière à rééquilibrer le jeu de données. Cette méthode a l'avantage d'être très simple à mettre en œuvre, mais elle risque de supprimer du jeu de données d'apprentissage des individus importants pour le concept de la classe majoritaire.

Pour éviter les inconvénients de cette première approche, on peut guider l'échantillonnage de la classe majoritaire pour le rendre moins aveugle. Pour cela il est intéressant de prendre en compte la notion de frontières entre les classes. Considérons deux individus ω_1 et ω_2 appartenant respectivement à la classe i et à la classe j, et d (ω_1,ω_2) la distance entre ces deux individus. La paire (ω_1,ω_2) est un lien de Tomek [Tom76] s'il n'existe aucun individu ω_3 tel que d (ω_3,ω_1) < d (ω_1,ω_2) ou d(ω_3,ω_2) < d(ω_1,ω_2). Si ces deux individus forment un lien de Tomek, c'est que l'un des deux est du bruit, ou que les deux sont des points frontières. Kubat [KM97] utilise le lien de Tomek comme méthode de sous-échantillonnage en supprimant les individus de la classe

majoritaire qui forment un lien de Tomek, les individus éloignés de la frontière étant plus sûrs pour l'apprentissage, et moins sensibles au bruit. Ces mêmes auteurs proposent une seconde méthode pour sous-échantillonner les individus de la classe majoritaire proches de la frontière [KHM97]. Il s'agit de tirer aléatoirement un individu dans la classe majoritaire et tous les individus de la classe minoritaire pour former un nouvel échantillon Ω. Puis ils classent tous les individus de Ω avec la classe de leur plus proche voisin (1-PPV) dans Ω. Les individus mal classés sont ensuite déplacés vers Ω.

L'objectif est donc également de ne conserver parmi les exemples de la classe majoritaire uniquement ceux qui sont éloignés de la frontière de décision, car ils sont considérés comme plus pertinents pour l'apprentissage (Figures 10 et 11).

Le choix de supprimer ou de conserver des individus frontière peut être vu comme un compromis entre la sensibilité du modèle et sa précision.

Lors d'un sous-échantillonnage des individus de la classe majoritaire, choisir ceux qui sont proches d'individus de la classe minoritaire diminuera le rappel sur cette dernière, mais augmentera la précision.

A l'inverse, sélectionner des individus négatifs qui sont éloignés d'individus positifs élargira la marge autour des individus positifs, et ainsi augmentera le rappel sur cette classe.

Sur cette idée [TJP08] proposent la méthode d'échantillonnage Funss (Fitting User Needs Sampling Strategy) effectuée à chaque étape de la construction [Bre01].

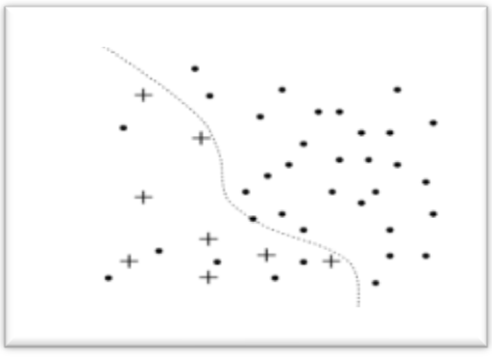

Figure 10 : Echantillon d'apprentissage d'origine

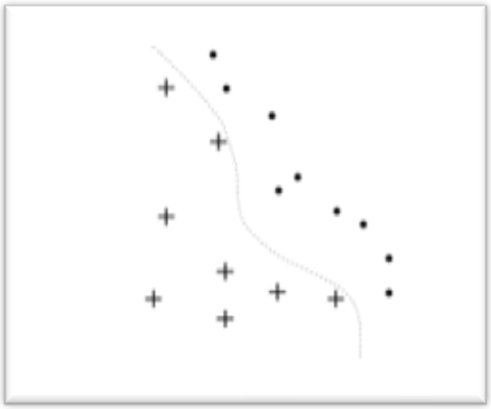

Figure 11 : Echantillon d'apprentissage après le retrait des individus négatifs redondants ou trop proches d'individus positifs

b) Sur-échantillonnage :

A l'inverse du sous-échantillonnage, un moyen pour rééquilibrera les jeux de données est l'augmentation du nombre d'individus appartenant à la classe minoritaire. La première solution est de répliquer aléatoirement des individus. Le risque de cette approche simpliste est de ralentir les algorithmes en ajoutant des individus, tout en fournissant des modèles incapables de généraliser (risque de sur-apprentissage): une règle ayant un bon support dans le jeu d'apprentissage peut en fait porter sur un seul individu.

Pour éviter ces inconvénients, la méthode Smote (Synthetic Minority Over sampling Technique) [CBHK02] permet de générer des individus artificiels dans la classe minoritaire. Pour chaque individu de la classe minoritaire, ses k plus proches voisins appartenant à la même classe sont calculés, puis un certain nombre d'entre eux (selon le taux de sur-échantillonnage voulu) sont sélectionnés.

Des individus artificiels sont ensuite disséminés aléatoirement le long de la ligne entre l'individu de la classe minoritaire et ses voisins sélectionnés. Ainsi le problème du sur-apprentissage est évité et la frontière de la classe minoritaire tend à se rapprocher de l'espace de la classe majoritaire.

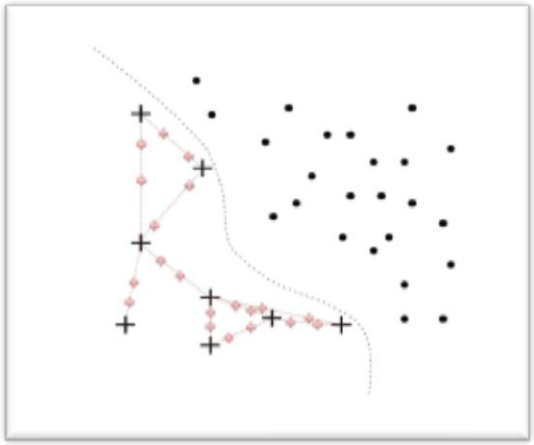

Figure 12 : Illustration du principe de Smote

4.5.2. Stratégies algorithmiques :

Les méthodes d'échantillonnage permettent essentiellement de traiter le problème des jeux de données déséquilibrés.

Une autre approche pour tenir compte de l'asymétrie est d'introduire un biais dans les algorithmes d'apprentissage. [BSGR03] proposent d'utiliser une mesure de distance pondérée dans l'algorithme des k plus proches voisins (k PPV). L'objectif de cette distance est de compenser le déséquilibre du jeu de données sans altérer la distribution des classes: des poids sont assignés non aux individus prototypes mais aux classes. Ainsi les distances aux prototypes de la classe minoritaire deviennent plus faibles qu'à ceux de la classe majoritaire.

a) Modification des seuils de décision :

Certains algorithmes fournissent une probabilité pour chaque individu d'appartenir à telle ou telle classe. La décision est donc prise en fixant un seuil sur cette probabilité (généralement 50%). C'est le cas du baesine naïf, ou de certains réseaux de neurones. Il est donc possible de tenir compte du déséquilibre des données en diminuant ce seuil pour la classe minoritaire (et à l'inverse d'augmenter ce seuil pour les individus de la classe majoritaire), ce qui améliorerait mécaniquement la sensibilité du modèle sur la classe minoritaire, le risque étant de dégrader la précision.

b) Apprentissage centré sur une classe :

L'apprentissage centré d'une seule classe (One class Learning) est une solution intéressante, face aux méthodes discriminantes comme les réseaux de neurones ou les arbres de décision

[Jap01]. Dans cette catégorie entrent des algorithmes de recherche supervisée de règles d'association, On peut également citer Ripper [Coh95], système d'induction qui construit des règles itérativement pour couvrir les individus qui n'ont pas été couverts auparavant. Des règles sont générées de la manière habituelle, mais de la classe la plus rare à la plus fréquente.

De part son architecture il lui est ainsi simple d'apprendre des règles uniquement sur la règle minoritaire.

Raskutti et Kowalczyk [RK04] montrent que l'apprentissage d'une seule classe est particulièrement approprié lorsque les données sont très déséquilibrées, et que l'espace est hautement dimensionnel, ou bruité.

5. Discussion :

5.1. Synthèse

Nous avons donc résumé les différents problèmes que nous regroupons sous le terme d'asymétrie, qui sont principalement le déséquilibre des classes et l'asymétrie du coût des erreurs. Puis nous avons exposé les méthodes permettant de traiter ces problèmes, également regroupées en deux catégories : les techniques d'échantillonnage, et les algorithmes sensibles à l'asymétrie. Enfin nous avons évoqué quelques études récentes qui comparent ces approches, en essayant de dégager la méthode la plus adaptée selon le type de problème. Il apparait que la méthode optimale dépend d'une part du jeu de données considéré, et d'autre part de l'objectif attendu des modèles. En effet, l'apprentissage sur données déséquilibrées revient souvent à effectuer un arbitrage entre la sensibilité du modèle qui est sa capacité à détecter les individus de la classe minoritaire, et sa précision qui correspond à la proportion d'individus réellement positifs parmi ceux classés comme positifs par le modèle asymétrie.

6. Conclusion :

Nous avons donc présenté un ensemble de problèmes en apprentissage supervisé que nous regroupons sous le terme d'asymétrie, ainsi que plusieurs méthodes permettant de prendre ces problèmes en compte. Dans le chapitre trois nous présentons et nous discutons les résultats obtenus.

1. Introduction :

Dans la classification de diagnostic médical, nous sommes souvent confrontés au nombre des échantillons de données déséquilibrées entre les classes dans lesquelles il n'y a pas suffisamment d'échantillons dans les classes rares. Malheureusement, la plupart des méthodes classiques d'apprentissage concurrentiels ne sont pas adaptés à cette situation parce que les classes secondaires peuvent être ignorés et on a tendance à réclamer toutes les données soient les classes majoritaires. Dans ce chapitre nous élaborons nos différentes contributions proposées en parcourant l'état de l'art (voir chapitre 2), Pour cela nous divisons ce chapitre en 2 parties où chacune d'elles traite des résultats d'une contribution appliquée. La procédure de classification du diabète est représentée dans le schéma suivant :

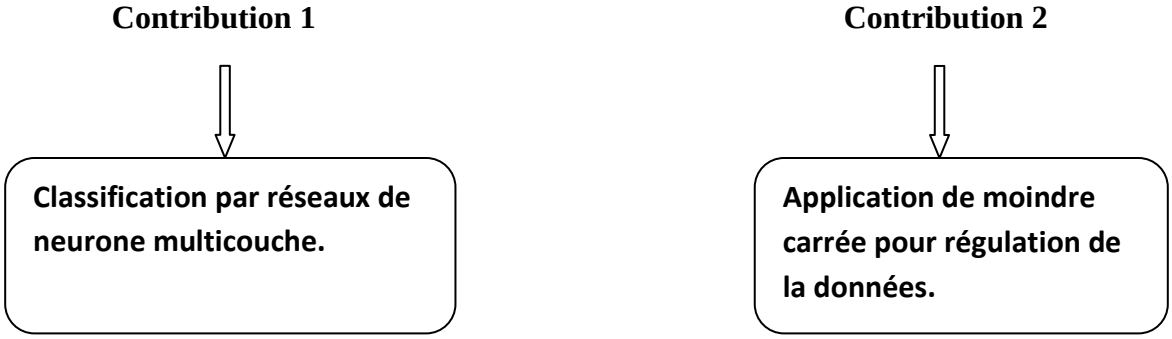

Figure 13 : Schéma représentatif de la procédure de classification du diabète

Dans la contribution 1 la classification du diabète est mise en œuvre avec L'algorithme supervisé réseau de neurone multicouche. La contribution 2 présente l'application de moindre carré en amont du réseau de neurones multicouches afin d'augmenter son efficacité. Pour l'implémentation de ces approches, nous avons fait recourt au logiciel Matlab.

2. Base de données PIMA

Les tests de la méthode proposée sont effectués sur la base de données Pima diabète. L'ensemble de données a été choisi du l'hôpital de maghnia qui réalise une étude sur 280 femmes (138 non diabétique 62Diabétiques).

Le diagnostic est une valeur binaire variable « classe » qui permet de savoir si le patient montre des signes de diabète selon les critères de l'organisation Mondiale de la santé. Les huit descripteurs cliniques sont :

- Npreg : nombre de grossesses.
- Glu : concentration du glucose plasmatique.
- BP : tension artérielle diastolique,(mm Hg).
- SKIN : épaisseur de pli de peau du triceps,(mm).
- Insuline : dose d'insuline,(mu U/ml).
- BMI : index de masse corporelle,(poids en kg/(taillye m)
- PED : fonction de pedigree de diabète (l'hérédité).
- Age : âge (Année).

N° Attribut	Description attribut	Moyenne
1	Nombre de grossesses (Ngross)	2.8
2	Concentration du glucose plasmatique (mg/dl)	109
3	Pression artérielle d iastolique (mm Hg) (PAD)	67
4	Epaisseur de la peau au niveau du triceps (mm) (Epa i)	19.77
5	Taux d'insuline au bout de 2 heures (mU¼ ml) (INS)	82
6	Indice de masse corporelle (poids en kg/ m^2) (IM C)	29.75
7	Fonction pédigrée du diabète (Ped)	0.55
8	Age	36

Tableau 5 : Description des attributs de la base

3. Contribution 1 : classification par réseaux de neurone perceptron multicouche :

3.1. **Principe**

Perceptron multicouche a été utilisé pour la classification des données.

Cette classification passe par six étapes :

- Créer Modèle :

Une méthode standard utilisée pour la création d'un modèle du classifieur en fixant son architecture.

- Apprentissage du classifieur :

une méthode standard utilisée pour l'apprentissage du modèle crée . Elle a comme paramètres le pas d'apprentissage, le maximum d'itération, la base d'apprentissage et la base de validation.

- Test du Modèle :

Une méthode standard utilisée pour le test du modèle après l'apprentissage. Elle a comme critère la base de test.

- Initialisation du Modèle :

Une méthode standard pour initialiser aléatoirement le modèle créé.

- Fonction d'activation :

Une méthode utilisée au niveau locale du classifieur comme fonction d'activation des neurones. Cette fonction est de type sigmoïde.

- Test Erreur Validation :

Une méthode utilisée pour tester le changement de l'erreur générée par la base de validation entre deux itérations a but de simuler le principe de l'arrêt par validation croisée.

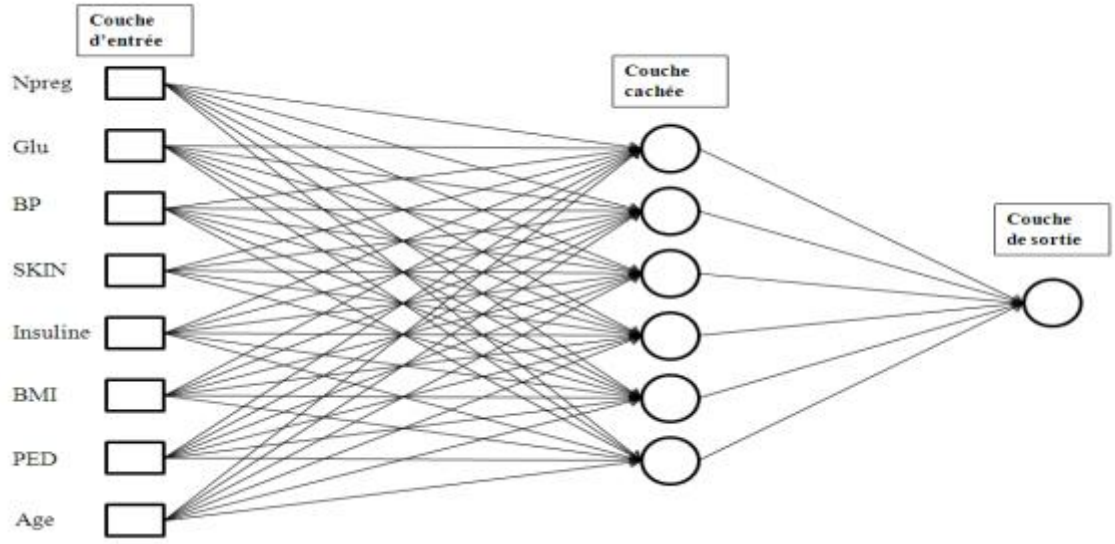

Figure 14 : Architecture utilisée dans l'expérimentation

Algorithme 3Principe de l'algorithme de rétro propagation

1: Initialisation: Affecter à tous les poids des valeurs aléatoires réelles.

2: Présentation des entrées et la sortie désirée :

Présenter le vecteur d'entrée x(1), x(2),……, x(N) et leurs correspondantes

sorties désirées d(1), d(2),……, d(N) une paire à la fois, où N est le nombre d'exemple d'apprentissage.

3: Calcul des sorties réelles : calcule les sorties $\mathbf{y_1, y_2, \dots \dots, y_{NM}}$

$$\mathbf{y_i}=Ø\left(\sum_{\mathbf{j=1}}^{\mathbf{N_{M-1}}} \mathbf{w_{ij}^{(M-1)}}\, \mathbf{x_i^{(M-1)}} \mathbf{b_j^{(M-1)}}\right) \qquad I = 1,\dots\dots, \mathbf{N_{M-1}}$$

4: Adaptation des poids ($\mathbf{w_{ij}}$) et les biais ($\mathbf{b_i}$)

$$\mathbf{\Delta w_{ij}^{l-1}}(n) = \boldsymbol{\mu} \mathbf{x_j}(n)\, \boldsymbol{\delta}_\mathbf{i}^{\mathbf{l-1}}(n)$$

$$\mathbf{\Delta b_i^{l-1}}(n) = \boldsymbol{\mu}(n)\, \boldsymbol{\delta}_\mathbf{i}^{\mathbf{l-1}}(n)$$

Avec :

– $\mathbf{x_j}(n)$ représente la sortie du nœud j à l'itération n,

– l est la couche,

– K est le nombre de nœuds de sortie du réseau de neurones,

– M est la couche de sortie,

– Ø est la fonction d'activation.

– Le pas d'apprentissage est représenté par μ.

3.2.Repartitionnement de la base :

Dans cette expérimentation, nous avons repartitionné les échantillons de la base sur trois ensembles:

– Un ensemble d'apprentissage.

– Un ensemble de validation.

– Et un ensemble de test.

3.3. Les critères d'évaluation :

Les performances de classification des données ont été évaluées par le calcul des vrais positifs (VP), vrais négatifs (VN), faux positifs (FP) et faux négatifs (FN), le pourcentage de sensibilité (SE), la spécificité (SP) et le taux de classification (TC), leurs définitions respectives sont les suivantes :

– Sensibilité (Se%) : [Se = 100 * VP / (VP + FN)] on appelle sensibilité (Se) du test sa capacité de donner un résultat positif quand la maladie est présente.

– Spécificité (Sp %) : [Sp = 100 *VN / (VN + FP)] on appelle spécificité du test cette capacité de donner un résultat négatif quand la maladie est absente. Elle est représentée pour détecter les patients non diabétiques.

– Taux de classification (CC %) : [CC = 100 * (VP + VN) / (VN + VP + FN +FP)] est le taux de reconnaissance.

– VP : diabétique classé diabétique

– FP : non diabétique classé diabétique

– VN : non diabétique classé non diabétique

– FN : diabétique classé non diabétique.

Dans les 5 expérimentations nous avons utilisé des perceptrons multicouche comme entrée les huit paramètres de la base PIMA.

2.4. Expérimentation et discussion :

Expérimentation 1 :

Dans cette étape nous avons utilisé une base de données équilibrée avec 50 cas diabétiques et 50 cas non diabétiques. Et on a obtenu les performances indiquées dans le tableau 6.

Base de données	La sensibilité	La spécificité	Le taux d'erreur	L'erreur
50 diabétique/50 non diabétique	66.67	46.15	60.00	40 .00

Tableau 6 : Résultat de test a l'équilibre

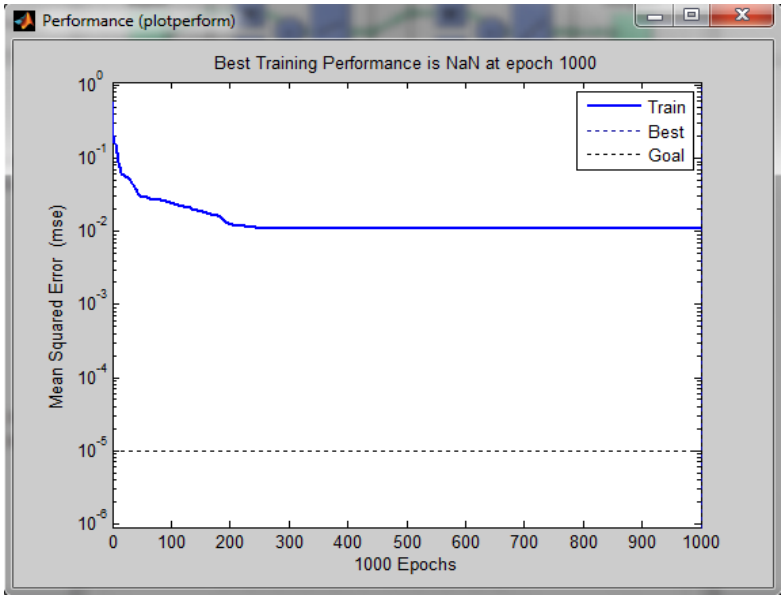

Figure 15 : Résultat de test a l'équilibre (50_50)

Expérimentation 2 :

Dans cette étape nous avons utilisé une base de données non équilibrée avec 40 cas diabétiques et 60 cas non diabétiques. Et on a obtenu les performances indiqué dans le tableau 7.

Base de données	La sensibilité	La spécificité	Le taux d'erreur	L'erreur
40 diabétique /**60** non diabétique	66.67	53.85	62.50	**37 .50**

Tableau 7 : Résultat de test non équilibrée (40_60)

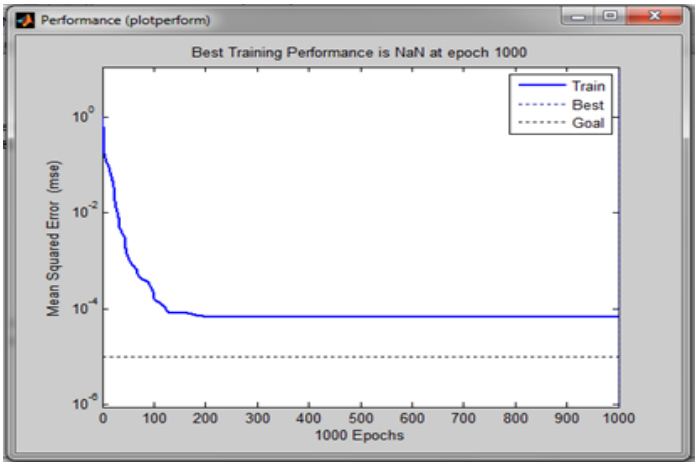

Figure 16 : Résultat de test non équilibrée (40_60)

Expérimentations 3 :

Dans cette étape nous avons utilisé une base de données non équilibrée avec 30 cas diabétiques et 70 cas non diabétiques. Et on a obtenue les performances indiqué dans le tableau 8.

Base de données	La sensibilité	La spécificité	Le taux d'erreur	L'erreur
30diabétique /**70** non diabétique	64.81	**57.69**	**63.75**	**37.50**

Tableau 8 : Résultat de test non équilibrée (30_70)

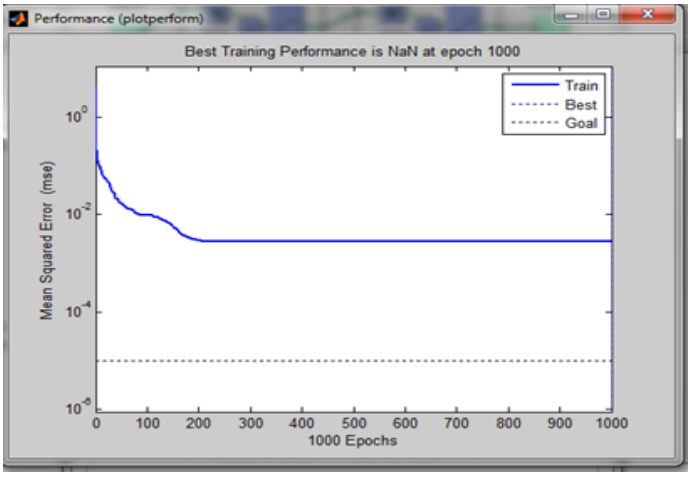

Figure 17 : Résultat de test non équilibrée (30_70)

Expérimentations 4 :

Dans cette étape nous avons utilisé une base de données non équilibrée avec 20 cas diabétiques et 80 cas non diabétiques. Et on a obtenue les performances indiqué dans le tableau 9.

Base de données	La sensibilité	La spécificité	Le taux d'erreur	L'erreur
20 diabétique **/80** non diabétique	**64.81**	**61.54**	**63.75**	**36.25**

Tableau 9 : Résultat de test non équilibrée (20_80)

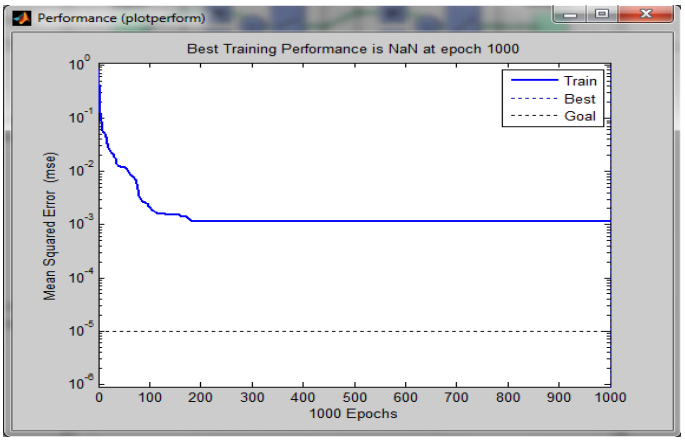

Figure 18 : Résultat de test non équilibrée (20_80)

Expérimentations 5 :

Dans cette étape nous avons utilisé une base de données non équilibrée dont il ya 10 cas diabétiques et 90 cas non diabétiques. Et on a obtenue les performances indiqué dans le tableau 10.

Base de données	La sensibilité	La spécificité	Le taux d'erreur	L'erreur
10 diabétique **/90** non diabétique	62.96	**61.54**	62.50	**37.50**

Tableau 10 : Résultat de test non équilibrée (10_90)

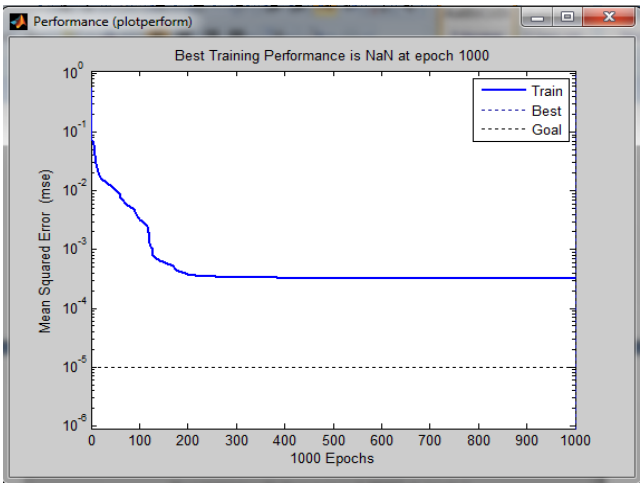

Figure 19 : Résultat de test non équilibrée (10_90)

Résultats et interprétation :

La répartition de l'échantillon entre les deux ensembles se fait en général sur la base des proportions 1/2, 1/2 pour chacun des deux ensembles et 2/3 pour l'ensemble d'apprentissage et 1/3 pour l'ensemble de test.

L'évaluation des performances de cette approche est estimée par le taux de classification, la sensibilité, la spécificité. Nous avons réalisé une étude sur la base de données diabétique suivant déférente considérations.

Dans la première considération, nous avons utilisé le perceptron multicouche.

Dans cette étape nous avons étudié 5 cas différents, où la base de données a été partitionnée comme indiqué au tableau 11 :

- La première [50 diabétique - 50 non diabétique] : la spécificité obtenus 46.15 ; la sensibilité 66.67 ; le taux d'erreur 60.00
- la deuxième [40 diabétique - 60 non diabétique] : la spécificité 0.5385 peu augmenté ; la sensibilité 66.67 ; le taux d'erreur 62.50
- la troisième [30 diabétique - 70 non diabétique] : la spécificité toujours augmenté 57.69 ; la sensibilité 64.81diminue ; le taux d'erreur **63.75** augmenté
- la quatrième [20 diabétique – 80 nom diabétique] : la spécificité **61.54** augmente presque égale a la sensibilité **64.81** ; le taux d'erreur bon **63.75**
- la cinquième [10 diabétique – 90 non diabétique] : la spécificité 61.54; la sensibilité 62.96 ; le taux d'erreur 62.50 sont résumés les performances dans le tableau 11 :

Base de données	La sensibilité	La spécificité	Le taux d'erreur	L'erreur
50 diabétiques /50 non diabétiques	66.67	46.15	60.00	40 .00
40 diabétiques / 60 non diabétiques	66.67	53.85	62.50	**37 .50**
30 diabétiques /70 non diabétique	64.81	57.69	**63.75**	**37.50**
20 diabétiques / 80 non diabétiques	64.81	**61.54**	**63.75**	**36.25**
10 diabétiques / 90 non diabétiques	62.96	**61.54**	62.50	**37.50**

Tableau 11 : Les performances du déférent résultat de perceptron multicouche

Les résultats de ces expérimentation montrent que le taux d'erreur va augmenter a partir de l'augmentation des données de la classe minoritaire.

Donc lorsqu'un patient est non diabétique notre modèle le détecte avec beaucoup de succès. Par contre la sensibilité du système est très faible ce qui veut dire que le système a fait une mauvaise reconnaissance des données positives. Ce qui peut générer un risque majeur pour la santé du patient. Avec ces performances, nous pouvons dire que le modèle a donné une bonne spécificité. Par contre il a donné une faible sensibilité un taux de classification moyen. Ce qui reste un inconvénient à étudier pour équilibrée ses données.

4. Considération 2 : méthode de moindre carrée :

Moindre carré est un algorithme de classification récemment développé sur la théorie de base de la réglementation [CE07].

4.1. Principe :

Dans l'algorithme RLS, les carrés des erreurs sont minimisés par la résolution d'un système d'équations linéaires. En raison de leur simplicité, leur faible computationnelle. En l'algorithme RLS, les erreurs de tous les échantillons de données sont le même coût. Dans ce document, pour surmonter le problème de l'apprentissage de l'ensemble de données asymétriques, nous avons utilisé une extension sensible aux coûts de l'algorithme RLS qui pénalise des erreurs de différents échantillons.

Algorithme de moindre carré régularisé de classification (RLS)

résolvons contraint problème de minimisation définie comme

- $\mathbf{min_w \frac{1}{n} \sum_{i=1}^{n} (Y_i - w.x_i)^2}$

- **Ajoutons un vecteur de paramètres β,appelé coût sensible du poids.**
- $\mathbf{min_w \frac{1}{n} \sum_{i=1}^{n} \beta (Y_i - w.x_i)^2}$

$$= \mathbf{min_w \frac{1}{n} \sum_{i=1}^{n} \left(\beta_i^{\frac{1}{2}} Y_i - w.\beta_i^{\frac{1}{2}} x_i \right)^2}$$

Sujet a $\|w\|^2$<=α, $\beta = [\beta_1, \beta_2, \ldots \beta_n]^T$,0<= β

- **V=diag ($\beta_1, \beta_2, \ldots \beta_n$)**

$\mathbf{V^{\frac{1}{2}} = (V^{\frac{1}{2}})^T = diag\ (\beta_1^{\frac{1}{2}}, \beta_2^{\frac{1}{2}}, \ldots, \beta_n^{\frac{1}{2}})}$

et

- $x_i^* = \beta_i^{\frac{1}{2}} x_i$ & $x^* = V^{\frac{1}{2}} x$
- $Y_i^* = \beta_i^{\frac{1}{2}} Y_i$ & $Y^* = V^{\frac{1}{2}} Y$

En substituant (4.3) à (4.2)minimisé normale contrainte de problème

$$\min_w \frac{1}{n} \sum_{i=1}^{n} \left(\beta_i^{\frac{1}{2}} Y_i - w . \beta_i^{\frac{1}{2}} x_i \right)^2$$

- $\min_w \frac{1}{n} \sum_{i=1}^{n} \beta_i (Y_i^* - w . x_i^*)^2$

Le système linéaire correspondant est :

$((x^*)^T x^* + \lambda n I_d) w = (x^*)^T y^*$

4. Expérimentation et discussion:

Expérimentations 1 : Dans cette étape nous avons utilisé un algorithme moindre carrée sur base de données équilibrée dont il ya 50 cas diabétiques et 50 cas non diabétiques. Et on a obtenue les performances indiqué dans le tableau 12.

Base de données	La sensibilité	La spécificité	Le taux d'erreur	L'erreur
50 diabétique /50 non diabétique	100	**92.31**	**97.50**	02.50

Tableau 12 : Résultat de test a l'équilibre (50_50)

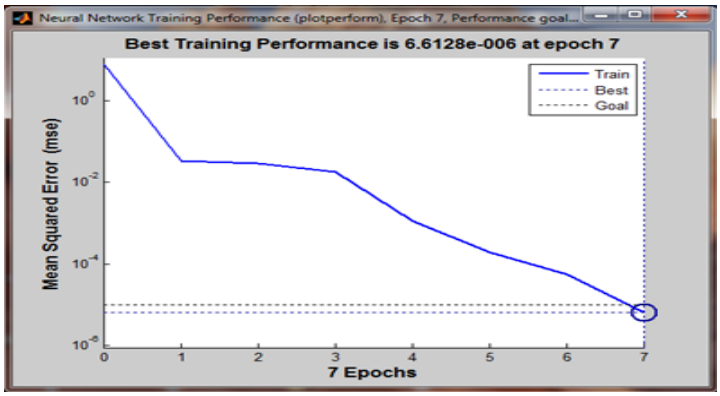

Figure 20 : Résultat de test équilibré (50_50)

Expérimentations 2 :

Dans cette étape nous avons utilisé un algorithme moindre carrée sur base de données non équilibrée dont il ya 40 cas diabétiques et 60 cas non diabétiques. Et on a obtenue les performances indiqué dans le tableau 13.

Base de données	La sensibilité	La spécificité	Le taux d'erreur	L'erreur
40 diabétique **/60** non diabétique	100	**96.15**	**98.75**	01.25

Tableau 13 : Résultat de test non équilibrée (40_60)

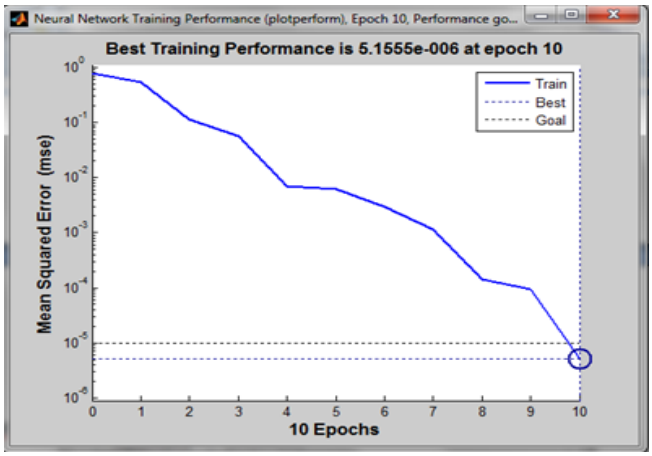

Figure 21 : Résultat de test non équilibrée (40_60)

Expérimentations 3 :

Dans cette étape nous avons utilisé un algorithme moindre carrée sur base de données non équilibrée dont il ya 30 cas diabétiques et 70 cas non diabétiques. Et on a obtenue les performances indiqué dans le tableau 14.

Base de données	La sensibilité	La spécificité	Le taux d'erreur	L'erreur
30diabétique **/70** non diabétique	100	**96.15**	**98.75**	01.25

Tableau 14 : Résultat de test non équilibrée (30_70)

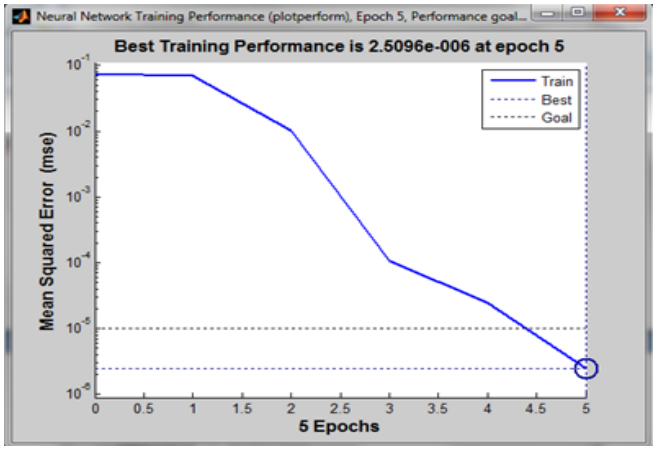

Figure 22 : Résultat de test non équilibrée (30_70)

Expérimentations 4 :

Dans cette étape nous avons utilisé un algorithme moindre carrée sur base de données non équilibrés dont il ya 20 cas diabétiques et 80 cas non diabétiques. Et on a obtenue les performances indiqué dans le tableau 15.

Base de données	La sensibilité	La spécificité	Le taux d'erreur	L'erreur
20 diabétique **/80** non diabétique	100	**92.31**	**97.50**	02.50

Tableau 15 : Résultat de test non équilibrée (20_80)

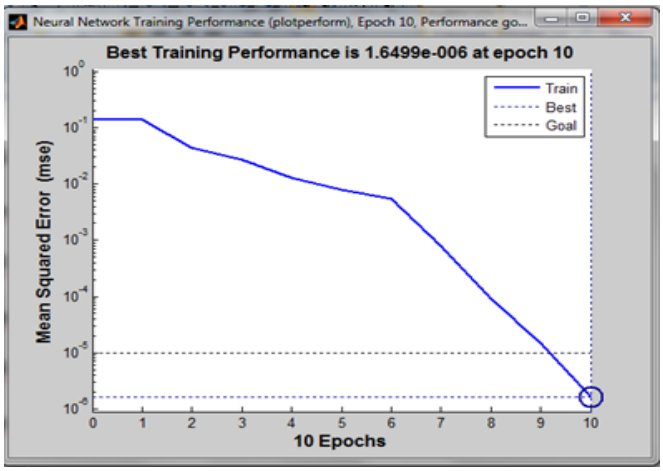

Figure 23 : Résultat de test non équilibrée (20_80)

Expérimentations 5 :

Dans cette étape nous avons utilisé un algorithme moindre carrée sur base de données non équilibrée dont il ya 10 cas diabétiques et 90 cas non diabétiques. Et on a obtenue les performances indiqué dans le tableau 16.

Base de données	La sensibilité	La spécificité	Le taux d'erreur	L'erreur
10 diabétique **/90** non diabétique	100	**88.46**	**96.25**	03.75

Tableau 16 : Résultat de test non équilibrée (10_90)

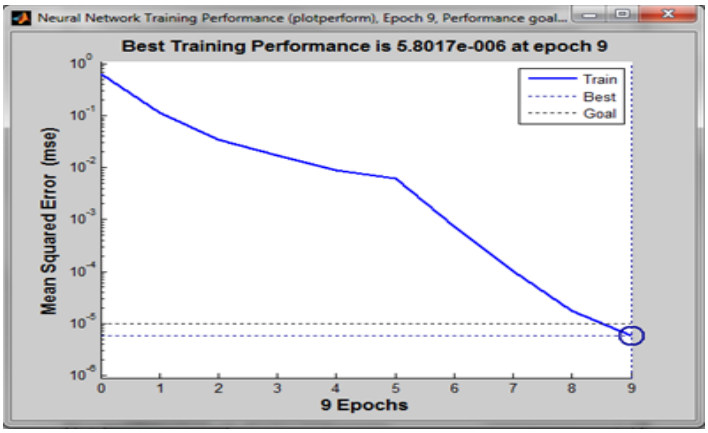

Figure 24 : Résultat de test non équilibrée (10_90)

5. Comparaison entre les deux considérations:

Afin de situer les performances de l'approche proposée nous avons réalisé une étude comparative entre les résultats des deux considérations avec la base PIMA.

	méthode	La sensibilité	La spécificité	Le taux d'erreur	L'erreur
50 diabétique / 50 non diabétique	RNMC	66.67	46.15	60.00	40.00
	LSC	100	**92.31**	**97.50**	02.50
40 diabétique / 60 non diabétique	RNMC	66.67	53.85	62.50	37.50
	LSC	100	**96.15**	**98.75**	01.25
30	RNMC	64.81	57.69	62.50	37.50

diabétique / 70 non didiabétique	LSC	100	**96.15**	**98.75**	01.25
20 diabétique / 80 non diabétique	RNMC	64.81	57.69	62.50	37.50
	LSC	100	**92.31**	**97.50**	02.50
10 diabétique / 90 non	RNMC	62.96	61.54	62.50	37.50
	LSC	100	**88.46**	**96.25**	03.75

Tableau 17 : Déférentes résultat entre la RNMC et RLS

Exemple : cas 30 diabétique et 70 non diabétique

	méthode	La sensibilité	La spécificité	Le taux d'erreur	L'erreur
30 diabétique / 70 non diabétique	RNMC	64.81	57.69	62.50	37.50
	LSC	100	**96.15**	**98.75**	01.25

Tableau 18 : La performance entre le RNMC et RLS

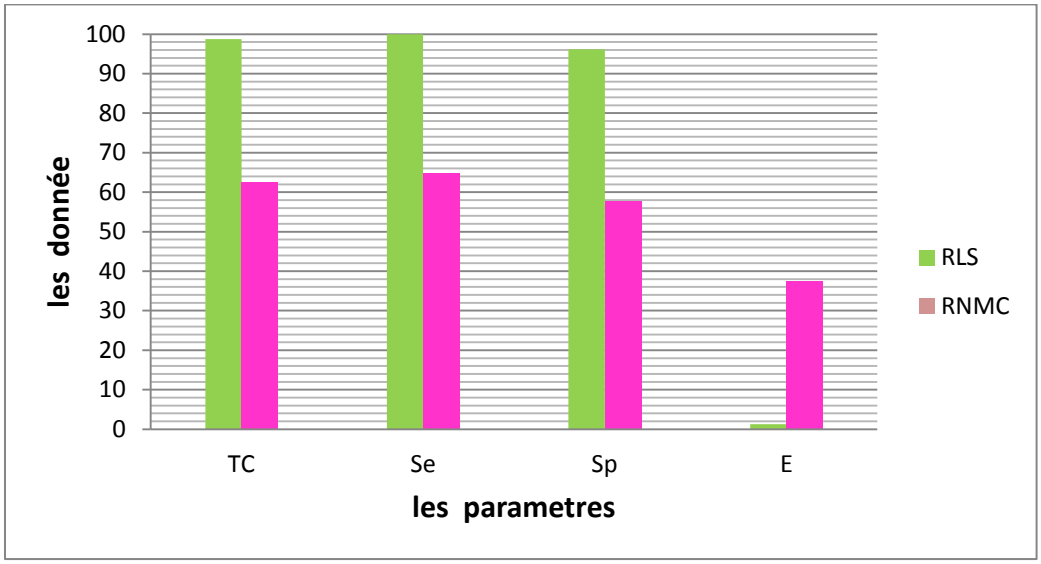

Figure 25 : Le résultat de performance de RLS a RNMC

- Nous remarquons que l'utilisation de la méthode des moindres carrées pour la régularisation des données a données de meilleurs performances de classification (SE, SP, TC).

6. Conclusion :

- l'ensemble de données asymétriques, est un problème trouvé souvent en aide de diagnostic, et qui peut causer des effets négatifs sérieux sur les performances de classification de l'apprentissage.
- Ce problème peut être résolu en utilisant les méthodes d'apprentissage sensibles aux coûts, ce qui pénalise inégalement différentes erreurs de classification.
- Dans ce chapitre, nous avons proposé d'utiliser l'algorithme RLS pour résoudre le problème des données asymétrique et augmenter les performances du classifieur.
- Les résultats expérimentaux montrent que la précision de l'algorithme RLS après équilibrage est relié à l'ensemble de données du nombre d'échantillons
- L'étude faite nous a montré que l'utilisation de la méthode de moindre carrée est effectivement trés pertinente pour la régulation de la base de données du diabète.

Conclusion Générale

- Afin de créer une application performante utilisée pour la reconnaissance du diabète, nous avons implémenté une méthode de moindre carrée comme but de minimisée l'erreur pour effectuer une meilleure classification.
- Les résultats obtenus après l'utilisation de cette méthode sont très prometteurs et sont bien situés parmi les travaux déjà réalisés dans ce domaine ce qui confirme la rigueur de la contribution proposée pour la résolution de notre problématique.
- Dans les perspectives d'avenir, nous prévoyons d'assurer l'interprétabilité des résultats du modèle en intégrant la notion du concept avec le flou. Nous voulons aussi généraliser cette modeste application sur tous les types de maladies afin de l'intégrer dans l'avenir dans un système d'aide au diagnostic applicable dans un hôpital ou dans un cabinet médical.

Bibliographie :

[Alw11]: Ala Alwan. Aperçu régional. Technical report, Fédération internationale du diabète, http : //www.diabetesatlas.org/, Access Mars 2011.

[Dia02] : OMS Diabète. Le coût du diabète. Technical report, Aide-mémoire No.236, 2002.

[Dia03] : OMS Diabète. Diabète et maladies rénales : il est temps d'agir. Technical report, Fédération Internationale du Diabète, 2003.

[Dia11] : OMS Diabète. Diabète. Technical report, Aide-mémoire No.312, Janvier 2011.

[MBG06]: Boyd E. Metzger, Susan A. Biastre, and Beverly Gardner. National diabetes information clearinghouse. Technical report, NIH Publication No.06–5129, April 2006.

[Or11]: W. H. Organization, "Ned country profiles," Algeria, 2011.

[Bre01] Leo Breiman. Random forests. Machine Learning, 45(1) :5–32, 2001.

[BSGR03] Ricardo Barandela, José Salvador Sánchez , Vicente García , and E. Rangel. Strategies for learning in class imbalance probleme. Pattern Recognition, 36(3) :849–851, 2003.

[CBHK02] Nitesh V. Chawla , Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote : Synthetic minority over-sampling technique.

Journal of Artificial Intelligence and Research, 16 : 321–357, 2002.

[CJK04] Nitesh V. Chawla , Nathalie Japkowicz, and Aleksander Kotcz. Editorial: special issue on learning from imbalanced data sets. SIGKDD Explorations, 6(1) :1–6, 2004.

[Coh95] William W. Cohen. Fast effective rule induction. In Armand Prieditis and Stuart Russell, editors,Proc. of the 12th International Conference on Machine Learning, pages 115–123, Tahoe City, CA, July 9–12, 1995.Morgan Kaufmann.

[DHR+06] Jason V. Davis, Jungwoo Ha, Christopher J. Rossbach, Hany E. Ramadan, and Emmett Witchel. Cost-sensitive decision tree learning for forensic classification. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, ECML, volume 4212 of Lecture Notes in Computer

Science, pages 622–629. Springer, 2006.

[FP97] Tom Fawcett and Foster J. Provost. Adaptive fraud detection. Data Min. Knowl. Discov., 1(3) :291–316, 1997.

[Gur97] :Kevin Gurney.An Introduction to Neural Networks. Taylor & Fran-cis, Inc., Bristol, PA, USA, 1997.

[HKN07] Jason Van Hulse, Taghi M. Khoshgoftaar, and Amri Napolitano. Experimental perspectives on learning from imbalanced data. In Zoubin Ghahramani, editor,ICML, volume 227 of ACM International Conference Proceeding Series, pages 935–942. ACM, 2007.

[ID-03] Imbalanced Data in Midwest Artificial Intelligence and Cognitive Science Conference (MAICS), 2003.

[Jap00b] Nathalie Japkowicz. The class imbalance problem :Significance and strategies. InProceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000), volume 1, pages 111–117, 2000.

[Jap01] Nathalie Japkowicz. Concept-learning in the presence of between-class and within-class imbalances. InAI '01 : Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence, pages 67–77, London, UK, 2001. Springer-Verlag.

[JMG95] Nathalie Japkowicz, Catherine Myers, and Mark A. Gluck. A novelty

[KHM97] Miroslav Kubat, Robert C. Holte, and Stan Matwin. Learning when negative examples abound. In Maarten van Someren and Gerhard Widmer, editors, ECML, volume 1224 of Lecture Notes in Computer Science, pages 146–153. Springer, 1997. Detection approach to classification. InIJCAI, pages 518–523, 1995.

[KHM98] Miroslav Kubat, Robert C. Holte, and Stan Matwin. Machine learning for the detection of oil spills in satellite radar images. Machine Learning,30(2-3) :195–215, 1998.

[KKP06] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas. Handling imbalanced datasets : A review. GESTS International Transactions on Computer Science and Engineering, 30(1) :25–36, 2006.

[KM97] Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets : One-sided selection. In Douglas H. Fisher, editor,ICML, pages 179–186. Morgan Kaufmann, 1997.

[LHM98] Bing Liu, Wynne Hsu, and Yiming Ma. Integrating classification and association rule mining. InKnowledge Discovery and Data Mining, pages 80–86, 1998.

[MA94] P. M. Murphy and D. W Aha. Uci repository of machine learning data bases, 1994.

[Mal03] M. Maloof. Learning when data sets are imbalanced and when costs are unequal and unknown. InICML Workshop on Learning from Imbalanced Data Sets II, 2003.

[MP43] :W. S. Mcculloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysic, 5 :115–133, 1943.

[Pro00] Foster Provost. Machine learning from imbalanced data sets 101. In Invited paper for the AAAI'2000 Workshop on Imbalanced Data Sets,2000.

[RK04] Bhavani Raskutti and Adam Kowalczyk. Extreme re-balancing for svms : a case study.SIGKDD Explor. Newsl., 6(1) :60–69, 2004.

[RT03]Rifkin, R., G. Yeo, and T . Poggio, Regularized least squares classification.2003.

[SI-04] SIGKDD Explorations Special Issue on Learning from Imbalanced Data sets, volume 6, 2004.

[Tan93] Ming Tan. Cost-sensitive learning of classification knowledge and its applications in robotics. Mach. Learn., 13(1) :7–33, 1993.

[TJP08] Julien Thomas, Pierre-Emmanuel Jouve, and Elie Prudhomme. Échantillonnage adaptatif de jeux de données déséquilibrés pourles forêts aléatoires. In Fabrice Guillet and Brigitte Trousse, editors, EGC, volume RNTI-E-11 ofRevue des Nouvelles Technologies de l'Information, pages 213–214. CépaduèsÉ ditions, 2008.

[Tom76] I. Tomek. Two modifications of cnn. IEEE Transactions on Systems, Man and Cybernetics, 6 :769–772, 1976.

[Tur95] Peter D. Turney. Cost-sensitive classification : Empirical evaluation of a hybrid genetic decision tree induction algorithm.J. Artif. Intell. Res. (JAIR), 2 :369–409, 1995.

[Tur02] Peter D. Turney. Types of cost in inductive concept learning. In Workshop on Cost-Sensitive Learning at ICML2000, pages 15–21, Stanford University, California, 2002.

[VC07] Florian Verhein and Sanjay Chawla. Using significant, positively associated and relatively class correlated rules for associative classification of imbalanced datasets. In ICDM, pages 679–684. IEEE Computer Society,2007.

[VCC99] K. Veropoulos, C. Campbell, and N. Cristianini. Controlling the sensitivity of support vector machines. InSixteen International Joint Conference on Artificial Intelligence, 1999.

[VCC99] K. Veropoulos, C. Campbell, and N. Cristianini. Controlling the sensitivity of support vector machines. InSixteen International Joint Conference on Artificial Intelligence, 1999.

[VR05] Sofia Visa and Anca Ralescu. Issues in mining imbalanced data sets a review paper. InSixteen Midwest Artificial Intelligence and Cognitive Science Conference, MAICS-2005, pages 67–73, Dayton, April 2005.

[W94] P. J.Werbos, The Roots of Back propagation : from Ordered Derivatives to Neural Networks and Political Forecasting. Wiley-Interscience, 1994.

[WC03] G. Wu and E. Y. Chang. Class-boundary alignment for imbalanced dataset learning. InICML'03 Workshop on Learning from Imbalanced Data Sets, 2003.

[WCS00] Workshop on Cost-Sensitive Learning, The Seventeenth International Conference on Machine Learning (ICML-2000), 2000.

[WC03] G. Wu and E. Y. Chang. Class-boundary alignment for imbalanced data set learning. InICML'03 Workshop on Learning from Imbalanced Data Sets, 2003.

[Wei03] Gary Mitchell Weiss. The effect of small disjuncts and class distribution on decision tree learning. PhD thesis, Rutgers University, New Brunswick, NJ, USA, 2003. Director-Haym Hirsh.

[Wei04] Gary M. Weiss. Mining with rarity : a unifying framework. SIGKDD Explorations, 6(1) :7–19, 2004.

[WLI00] 1st Workshop on Learning from Imbalanced Data Set (held withAAAI2000), 2000.

[WLI03] 2nd Workshop on Learning from Imbalanced Data Set (held withICML2003), 2003.

[WMZ07] Gary Weiss, Kate McCarthy, and Bibi Zabar. Cost-sensitive learning vs.sampling : Which is best for handling unbalanced classes with unequal error costs ? In Robert Stahlbock, Sven F. Crone, and Stefan Lessmann, editors,DMIN, pages 35–41. CSREA Press, 2007.

[WP03] Gary M. Weiss and Foster J. Provost. Learning when training data are costly : The effect of class distribution on tree induction.J. Artif. Intell. Res. (JAIR), 19 :315–354, 2003.

[ZE01a] Bianca Zadrozny and Charles Elkan. Learning and making decisions when costs and probabilities are both unknown. InKDD, pages 204–213, 2001.

[Ro58] F. Rosenblatt, “The perceptron : a probabilistic model for information storage and organization in the brain,”Psychological Review, vol. 06, pp. 386–408, 1958.

[GW86] G. E. H. e. R. J. W. D. E. Rumelhart, “Learning internal representations by error propagation,”Parallel Data Processing : Explorations in the Microstructure of Cognition, vol. 01, pp. 318–362, 1986.

[LC85] Y. L. Cun, “Une procédure d’apprentissage pour réseau à seuil asymmetrique (a learning scheme for asymmetric threshold networks),” Proceedings of Cognitiva, pp. 599–604, 1985.

[TA95] D. J. L. I. V. Tetko and A. I. Luik, “Neural network studies. 1. comparison of overfitting and overtraining,”Journal of Chemical Information and Computer Sciences, vol. 35, No 5, pp. 826–833, 1995.

[CE07] P. Cornillon et É. Matzner-Løber, *Régression*, Paris, Springer, 2007.