Université Abou Bekr Belkaid جامعة ابي بكر بلقايد
تلمسان الجزائر

République Algérienne Démocratique et Populaire
Université Abou Bakr Belkaid– Tlemcen
Faculté des Sciences
Département d'Informatique

Mémoire de fin d'études

pour l'obtention du diplôme de Master en Informatique

Option: *Modèle Intelligent et Décision(M.I.D)*

Thème

L'utilisation de l'approche Boosting pour le diagnostic du diabète

Réalisé par :

- **HAMIDI Yacine Nacer Eddine**

Présenté le 01 Juillet 2012 devant le jury composé de MM.

- M.BENAZZOUZ *M (MAA)* *(Président)*
- *M.CHIKH M.A (Professeur)* *(Encadreur)*
- *M^{elle} SAIDI M (Doctorante)* *(Co-Encadreur)*
- *M.HADJILA F (MAA)* *(Examinateur)*
- *M.BENMOUNA Y (MAB)* *(Examinateur)*

Année universitaire : 2011-2012

Résumé

L'utilisation des méthodes ensemblistes pour améliorer les performances des classifieurs faibles est une nouvelle voie dans le domaine de l'apprentissage artificiel. Parmi ces méthodes nous citons l'algorithme d'Adaboost, l'un des algorithmes du Boosting. Ce mémoire présente une approche basée sur l'amélioration de classifieur Kppvp (K plus proche voisins pondéré) par l'algorithme Adaboost afin de classifier la base de données Pima Indians diabetes. Les performances obtenues seront comparées en utilisant des critères comme le taux de classification, la sensibilité et la spécificité.

Mots-clés : Classification, méthodes ensemblistes, Diagnostic, Boosting, Adaboost, K plus proches voisins pondéré, Base de données Pima Indians diabetes.

Abstract

The use of ensemblist methods to improve the performance of weak classifiers is a new way in the field of machine learning. Among these methods Adaboost is one of the Boosting algorithms. This master project presents an approach based on improving K-NNW (weighted K nearest neighbors) classifier by the Adaboost algorithm to classify the database Pima Indians Diabetes (PID). The performances obtained are compared against the criteria of classification rate, sensitivity and specificity.

Keywords: Classification, Ensemblist methods, Diagnosis, Boosting, Adaboost, k- Nearest Neighbors, Pima Indians Diabetes database.

Je dédie ce travail à :

Mes parents,

Toute ma famille,

Mes amis,

Qu'ils trouvent ici l'expression de toute ma reconnaissance.

Remerciements

Je commence à adresser mes vifs remerciements à M.CHIKH M.A, professeur à l'université de Tlemcen, en tant que Directeur de mémoire, qui s'est toujours montré à l'écoute et disponible tout au long de la réalisation de ce travail malgré ses charges administratives et pédagogiques, et surtout pour sa patience et ses judicieux conseils.

Ensuite, je désire adresser mes sincères remerciements Mlle SAIDI Meryem, doctorante à l'Université de Tlemcen et co-encadreur de ce mémoire. Son savoir, sa méthodologie et son expérience m'ont beaucoup aidé.

J'adresse mes sincères remerciements, au Monsieur M. BEN AZOUZ qui me fait l'honneur de présider ce jury.

Je remercie tous particulièrement, M F. HADJILA et M . Y.BENMOUNA Maîtres de conférences à l'université Abou Bekr Belkaid-Tlemcen, qui ont accepté de juger ce travail.

Je tiens à remercier Mlle Settouti Nesma, doctorante à l'université de Tlemcen, pour l'intérêt porté à ce travail.

Table des matières

Liste des tableaux

Liste des figures

Introduction générale

Introduction générale

L'apprentissage artificiel s'intéresse à l'écriture de programmes d'ordinateur capables de s'améliorer automatiquement au fil du temps, soit sur la base de leur propre expérience, soit à partir de données antérieures fournies par d'autres programmes. Dans le domaine scientifique relativement jeune de l'informatique, l'apprentissage artificiel joue un rôle de plus en plus essentiel.

La discipline de l'apprentissage artificiel a revendiqué des succès dans un grand nombre de domaines d'application. Des logiciels de fouille de données sont utilisés à grande échelle pour découvrir quelle prescription est la plus efficace pour quel patient, à partir de l'analyse de fichiers médicaux antérieurs. Il semble certain que le rôle de l'apprentissage artificiel ne cessera de croitre au centre de cette science.

Le diagnostic médical est un processus de classification. L'utilisation de l'informatique pour la réalisation de cette classification devient de plus en plus fréquente. Même si la décision de l'expert est le facteur le plus important lors du diagnostic, les systèmes de classification fournissent une aide substantielle, car elles réduisent les erreurs dues à la fatigue et le temps nécessaire pour le diagnostic. Actuellement, la plupart des hôpitaux modernes sont bien équipés avec des dispositifs de collecte de données. Ces données seront partagées en inter- et intra-systèmes d'information hospitaliers. Ce qui était avant une base de données isolée ou un système d'information de laboratoire est maintenant intégrée dans un système d'information médicale à plus grande échelle (ministères, hôpitaux, ou à base communautaire). L'augmentation du volume de données entraîne des difficultés à extraire des informations utiles pour l'aide à la décision. Les méthodes traditionnelles d'analyse de données sont devenues insuffisantes, et les méthodes dites intelligentes sont indispensables.

Les méthodes du datamining (ou apprentissage) ensemblistes présentent actuellement des techniques très sollicitées dans la communié scientifique, grâce à leur apport significatif en terme de précision. L'idée est basée sur la combinaison des résultats de plusieurs algorithmes du datamining, appliqué chacun sur un ensemble diversifié de données. Le boosting est une technique d'apprentissage qui vise à rendre plus performant un système d'apprentissage (faible). Pour ce faire, le système d'apprentissage est entraîné successivement sur des échantillons d'apprentissage surpondérant les exemples difficiles à apprendre. A chaque fois, une hypothèse h_t est

produite, et l'hypothèse finale est une combinaison linéaire de ces hypothèses pondérées par des coefficients liés à leur performance. Le boosting est d'un emploi très large et fait l'objet de nombreux travaux et applications.

Dans ce mémoire de master nous nous intéresserons au diagnostic du diabète qui est un dysfonctionnement du système de régulation de la glycémie. Notre problématique réside dans l'amélioration du diagnostic de diabète par l'approche de boosting (l'algorithme Adaboost), utilisant l'algorithme de K plus proche voisin pondéré (K-PPVP) comme algorithme de base, appliqué sur la base des diabétique Pima indian.

Beaucoup de travaux ont été mené afin d'effectuer la classification ou le diagnostic du diabète diabète. Dans [KS07], utilisent l'analyse de la composante principale et l'inférence neuro flou adaptative pour la classification de la base de données Pima Indian diabetes. Ils ont obtenu une précision de 89.47%. Purnami & al [SAJ09], ont obtenu une précision de 93.2% en utilisant un nouveau « smooth support vector machine » et ses applications dans le diagnostic du diabète. Dans [SKH05],Sahan & al. ont utilisé un "Attribute Weighted Artificial Immune System" avec 10-fold cross validation et ont obtenu un taux de classification de 75.87%. Salami & al [MAA10] ont obtenu 80.00% et 80.65% en utilisant "CVNN-based CAR" et "RVNNbased AR". Dans [TDM08], Exarchos & al., les auteurs ont utilisé "Automated creation of transparent fuzzy models based on decision trees", ils ont obtenu un taux de classification de 75.91%. Ganji & al [MM10] ont utilisé un "fuzzy Ant Colony Optimization" ; et ils ont obtenu un taux de 79.48%. Dans [TA10], Jayalakshmi & al utilisent un réseau de neurones pour le diagnostic du diabète avec la base Pima Indian diabetes dataset sans données manquantes et ont obtenu un taux de 68.56%.

Ce travail de Master se situe dans le contexte général de l'Aide au Diagnostic médical, qui a pour but de réaliser un classifieur de diabète. Le plan de mémoire est composé de :

- Le 1er chapitre : présente un aperçu général sur la maladie du diabète en citant les causes, le diagnostic, les traitements et la prévention.
- Le 2eme chapitre : présente les principes des méthodes ensemblistes notamment la méthode de boosting et l'algorithme Adaboost utilisé dans ce travail.
- Dans le 3ème chapitre, nous présentons la base de données PID utilisée pour évaluer et comparer les approches. Finalement les résultats sont présentés,

comparés et interprétés. Le manuscrit est clôturé par une conclusion générale résumant les idées fondamentales que nous a apportées ce travail tout en discutant les pistes de recherche futures ouvertes pour la suite.

Chapitre I

Présentation du diabète

1. Introduction

Le diabète est une maladie lourde de conséquences par ses complications. C'est pourquoi il constitue un problème de santé publique au niveau national et international dont le poids humain et économique augmente graduellement. En effet, ses complications en font une maladie dont la morbidité et la mortalité sont fortement accrues par rapport à la population générale. Actuellement, il y a 200 millions de diabétiques dans le monde. Plus de 90% des diabétiques ont le diabète de type II, et seulement 10% présentent le diabète de type I. En 2030, il y aura 330 millions de diabétiques dans le monde. Cette épidémie du diabète est surtout due aux modifications du mode de vie, l'occidentalisation, l'alimentation très calorique, la sédentarité, le dépistage plus actif du diabète, le vieillissement de la population [DEC12].

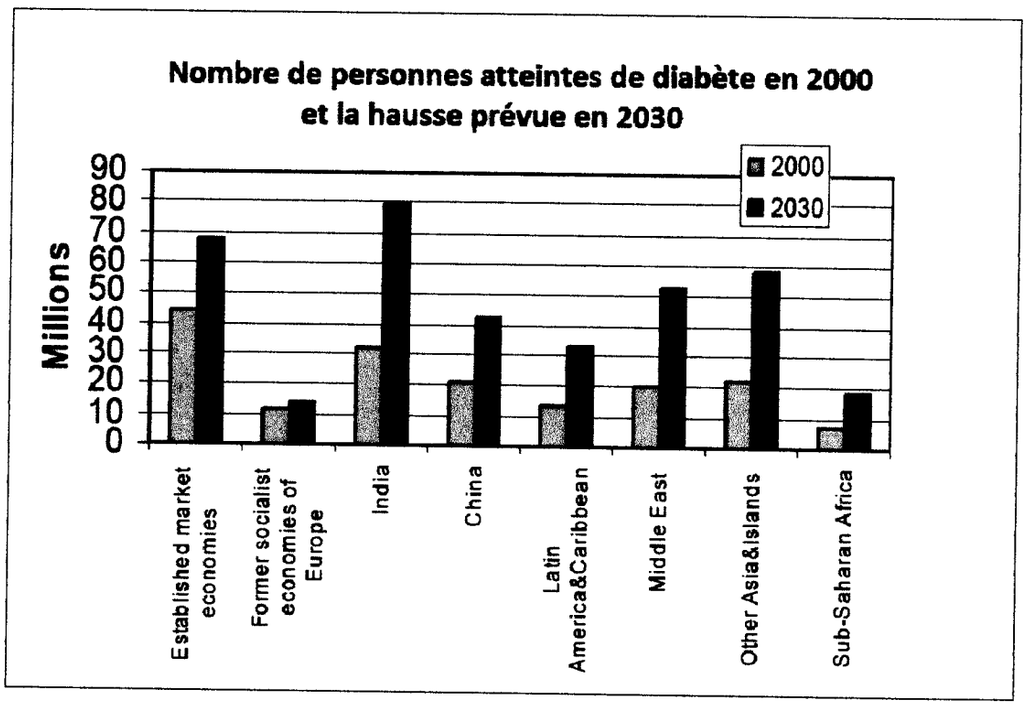


FIGURE 1.1: Projection du nombre de personnes diabétiques dans différentes régions du monde [OMS11].

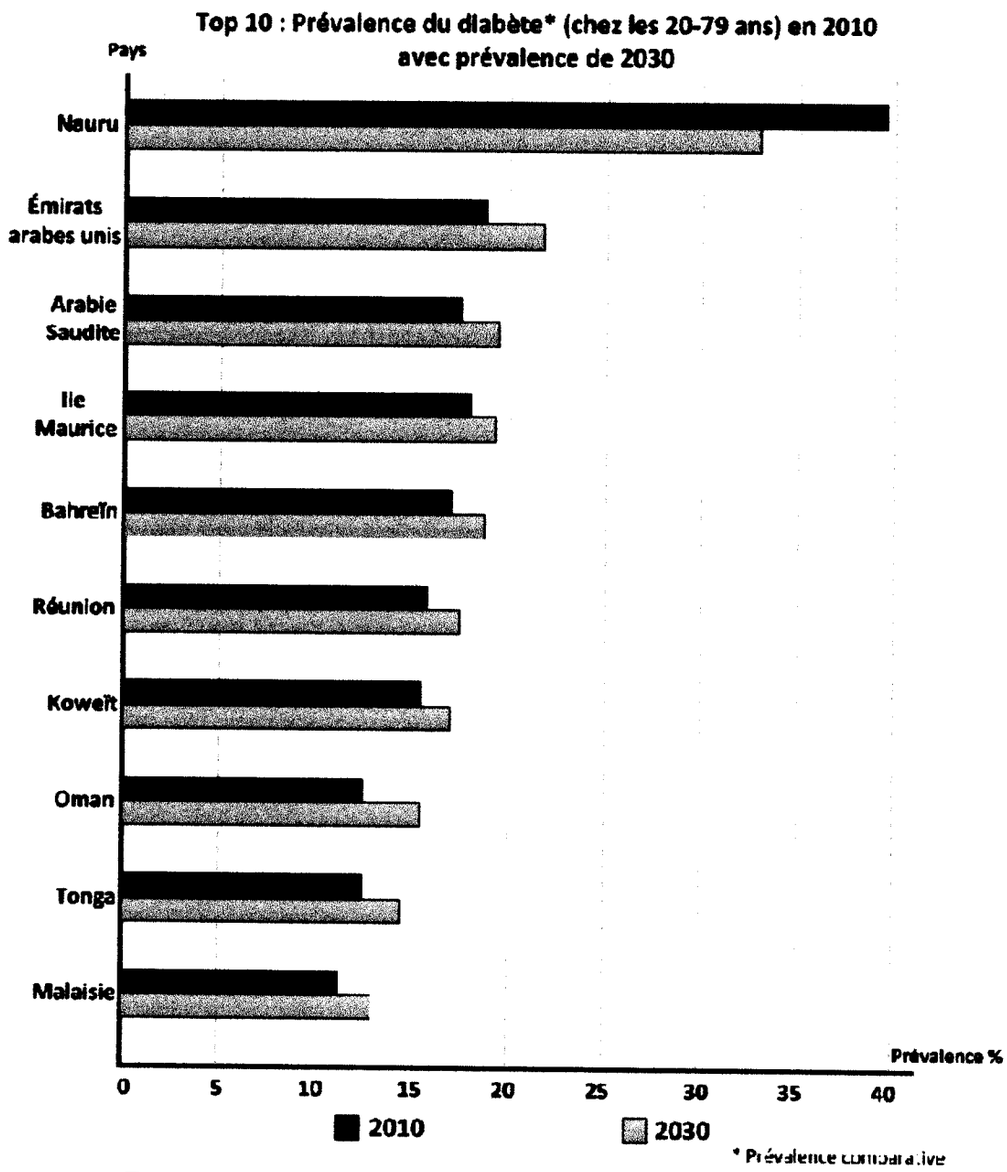


FIGURE 1.2: Les 10 pays les plus touchés par le diabète en 2010 et leurs prévisions en 2030[ALW11].

2. Définition

Le diabète est une maladie chronique incurable causée par une carence ou un défaut d'utilisation de l'insuline entraînant un excès de sucre dans le sang. Produite par le pancréas, l'insuline est une hormone qui permet au glucose (sucre) contenu dans les aliments d'être utilisés par les cellules du corps humain. Les cellules disposent de toute cette énergie dont elles ont besoin pour fonctionner.

Si l'insuline est insuffisante où si elle ne remplit pas son rôle adéquatement, comme c'est le cas dans le diabète, le glucose (sucre) ne peut pas servir de carburant

aux cellules. Il s'accumule dans le sang et est ensuite déversé dans l'urine. À la longue, l'hyperglycémie provoquée par la présence excessive de glucose dans le sang entraîne certaines complications, notamment au niveau des yeux, des reins, des nerfs, du cœur et des vaisseaux sanguins [ACD01].

2.1 Le diabète dans le monde

- Ils existent 356 millions personnes diabétiques dans le monde.
- Le diabète a tué environ 3,4 million de personnes en 2004.
- Le diabète est l'un des causes de décès, de plus de 80% dans des pays à revenu faible ou intermédiaire.
- Selon les données de l'OMS, le nombre de décès par diabète va doubler entre 2005 et 2030.
- Un régime alimentaire sain, une activité physique régulière, le maintien d'un poids normal et l'arrêt du tabac permettent de prévenir ou de retarder l'apparition du diabète de type 2 [OMS12].

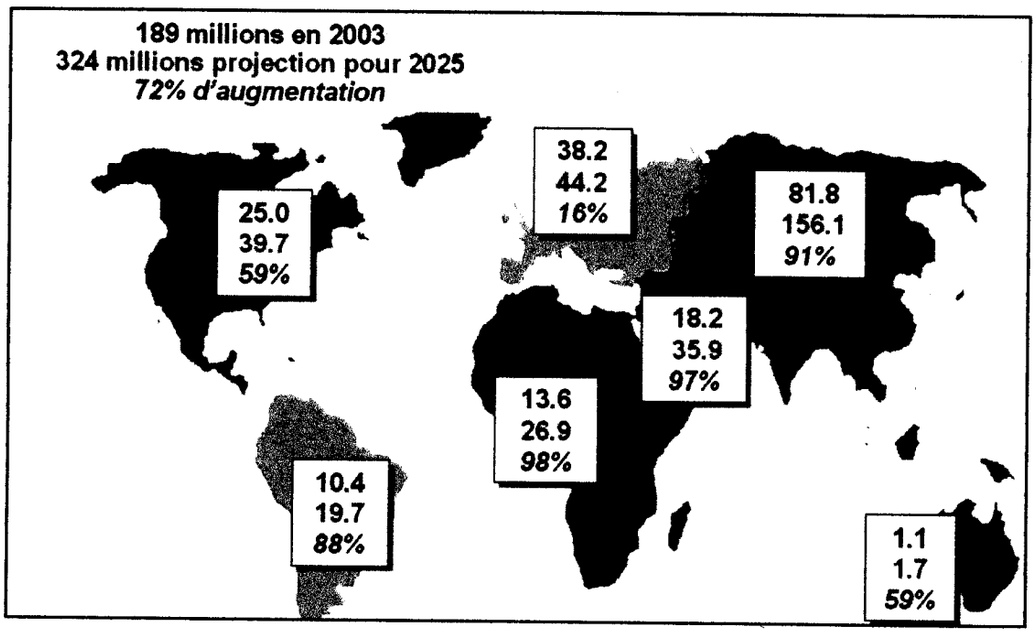


FIGURE 1.3 : Augmentation de nombre des diabétiques dans le monde 2003-2025 [DEC12].

2.2 Types de diabète

Il existe différents types de diabète, le diabète insulinodépendant ou diabète de type I, le diabète non insulinodépendant ou diabète de type II (plus fréquent que le diabète de type I), le diabète gestationnel qui concerne 6% des grossesses.

2.2.1 Diabète de type 1

Le diabète de type 1 se manifeste soit dès l'enfance, à l'adolescence ou chez les jeunes adultes. Il se caractérise par l'absence totale de la production d'insuline. Les personnes diabétiques de type 1 dépendent d'injections quotidiennes d'insuline pour vivre. Il est présentement impossible de prévenir ce type de diabète. Les recherches s'effectuent principalement vers la compréhension des mécanismes détruisant les cellules responsables de la production d'insuline.

2.2.2 Diabète de type 2

Dans ce type de diabète l'altération métabolique n'est pas aussi intense que pour DM1 et l'évolution de la maladie est progressive. Ce diabète est caractérisé par la résistance ou la faible sensibilité du corps à l'insuline, c'est-à-dire que le taux d'insuline endogène peut se trouver dans les paramètres normaux, mais les tissus sont incapables de l'assimiler et par conséquent le taux de glucose dans le sang augmente [MNT12].

L'insuline agit au niveau cellulaire à travers de certains récepteurs de membrane (Figure 1.4). La liaison insuline-récepteur active un deuxième messager qui induit la synthèse des protéines et l'activation et inhibition des enzymes intracellulaire. Les malades souffrant de diabète de type 2 ont des altérations dans les mécanismes post-récepteurs, ce qui oblige l'organisme à augmenter la sécrétion d'insuline pour compenser et ceci peut conduire à l'épuisement des cellules Béta. Pour des personnes avec une certaines prédisposition les cellules ne seront pas capable de maintenir un taux de glucose normale, ce qui conduit à l'apparition du diabète. Même si les diabétiques de type 2 ne nécessitent pas les injections d'insuline pour survivre, prés du 40 % des malades finissent par en avoir besoin pour contrôler la glycémie. L'hyperinsulinisme de plus de 80 % de diabétique de type 2 est la conséquence de leur obésité (la graisse abdominal est la plus dangereuse) [MNT12].

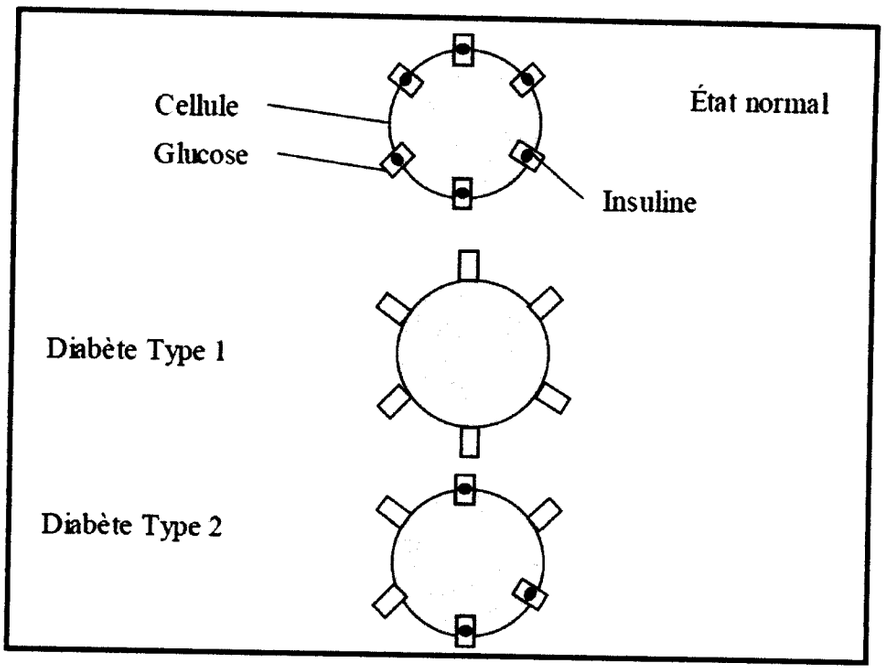


FIGURE 1.4 : Fonctionnement de l'insuline.

2.2.3 Le diabète de grossesse

Le diabète de grossesse se manifeste pendant la grossesse, généralement vers la fin du 2ème trimestre et au cours du 3ème. Il est aussi connu sous le nom de diabète gestationnel. Dans 90% des cas, il disparaîtra après l'accouchement. Le diabète gestationnel (4 à 6% des grossesses) affecte à la fois le bébé et la mère. L'enfant risque d'être plus gros que la normale et risque de faire un diabète plus tard. Chez la mère, la présence du diabète accroît les risques d'infections, augmente le niveau de fatigue et peut causer des complications lors de l'accouchement.

Le diabète de grossesse se traite et se contrôle par une saine alimentation, et l'adoption d'une bonne hygiène de vie. Si, malgré ces changements, le diabète n'est pas bien contrôlé, l'utilisation d'insuline deviendra nécessaire car l'emploi d'antidiabétiques oraux est contre-indiqué lors d'une grossesse.

	Diabète de type 1	Diabète de type 2
Fréquence relative	10-15 %	85-90 %
ATCD familiaux	+	+++
Age de début	Avant 30 ans	Après 40 ans
Mode de début	Brutal	Progressif
Surpoids	Absent	Présent
Symptômes	+++	/
Insulinosécrétion	Néant	Persistante
Cétose	Fréquente	Absente
MAI associées*	Oui	Non
Auto-anticorps	Présents	Absents
Groupe HLA	Oui	Non
Traitement	Insuline	Régime, exercice, ADO**

Tableau 1.1 : Caractéristiques des diabètes de type 1 et 2.

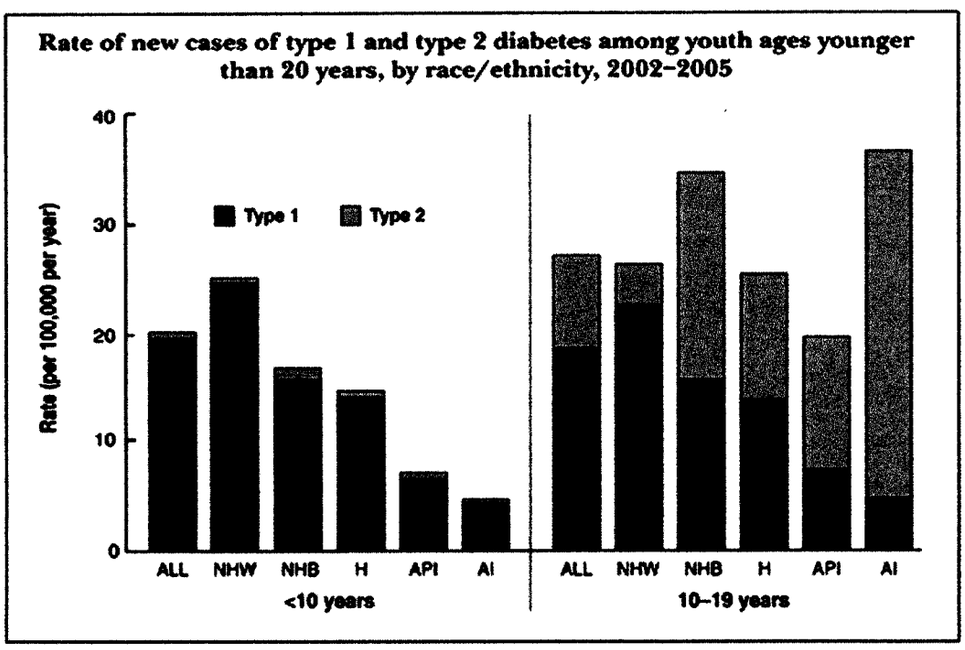


FIGURE 1.5 : Taux de nouveaux cas de diabète de type1 et de type 2 parmi les jeunes âgés de moins de 20 ans par la race, 2002-2005.

3. Causes de diabète

Le diabète de type 1 est une affection auto-immune. On a lieu de penser qu'une prédisposition génétique associée à d'autres facteurs (pas encore identifiés) incite le système immunitaire à attaquer les cellules productrices d'insuline dans le pancréas et à les détruire.

4. Symptômes

Les symptômes du diabète ne se présentent pas tous de la même manière ni avec la même intensité. Qu'il s'agisse du type 1, du type 2 ou du diabète de grossesse, une consultation avec le médecin s'impose. Les symptômes sont :

- Fatigue, somnolence.
- Augmentation du volume des urines.
- Soif intense.
- Faim exagérée
- Amaigrissement.
- Vision embrouillée.
- Cicatrisation lente.
- Infection des organes génitaux.
- Picotements aux doigts ou aux pieds.
- Changement de caractère.

5. Complication du diabète

Les conséquences du diabète peuvent être lourdes pour la santé. Le diabète est un facteur de risque important de maladies cardiovasculaires, infarctus, insuffisance cardiaque, artérite, accident vasculaire cérébral, de neuropathie, ou encore de troubles micro-angiopathiques pouvant conduire à la cécité (rétinopathie), à une insuffisance rénale chronique (néphropathie). Il a été aussi clairement défini comme un facteur de risque majeur prédisposant à la maladie parodontale (Figure 1.6).

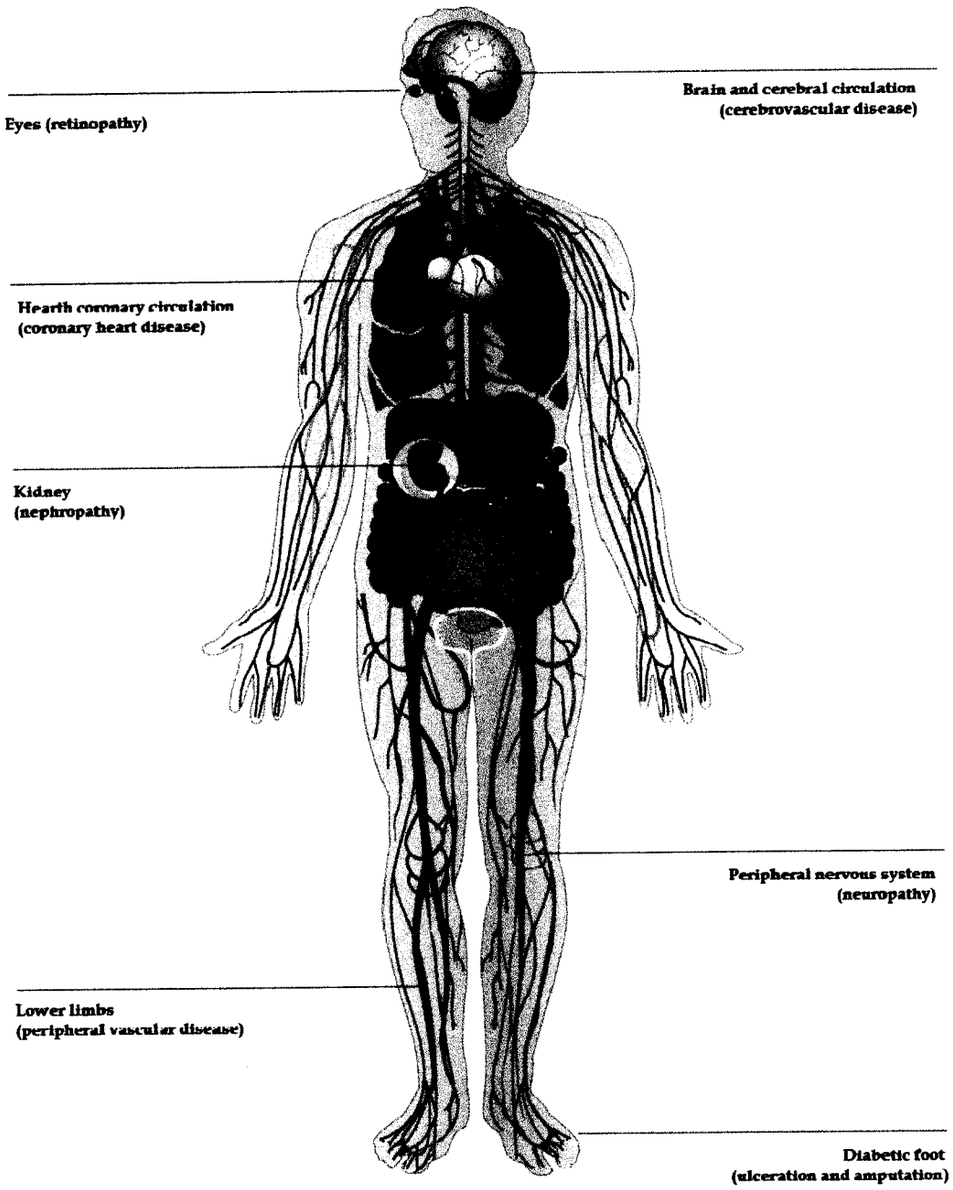


FIGURE 1.6 : Les majeures complications du diabète [ALW11].

- Le diabète augmente le risque de cardiopathie et d'accident vasculaire cérébral. 50% des diabétiques meurent d'une maladie cardio-vasculaire (principalement cardiopathie et accident vasculaire cérébral) (Figure 1.7).

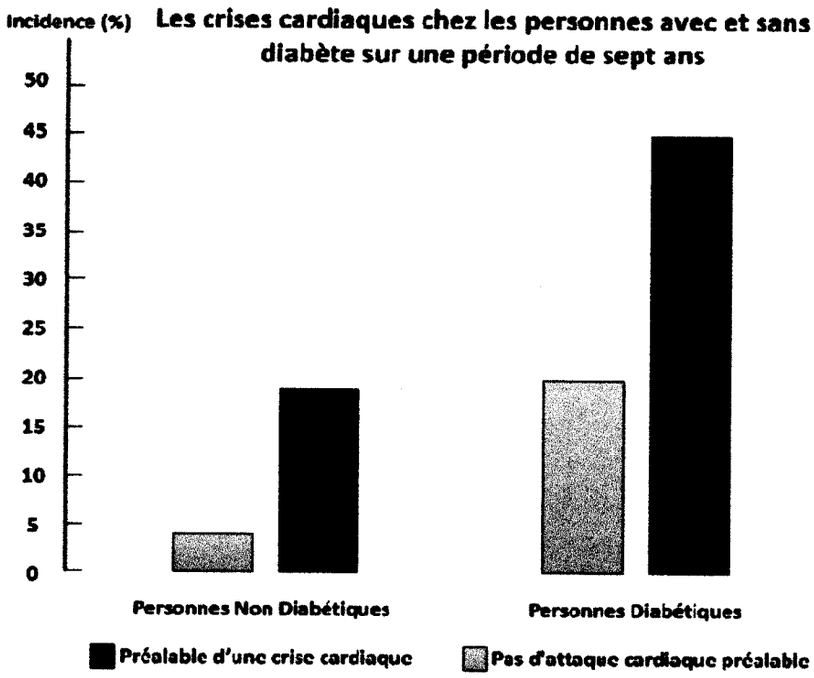


FIGURE 1.7 : Incidence des crises cardiaques pour les personnes diabètiques et non diabèques durant une periode de 7 ans[ALW11].

- Associée à une diminution du débit sanguin, la neuropathie qui touche les pieds augmente la probabilité d'apparition d'ulcères des pieds et impliquant l'amputation des membres.
- La rétinopathie diabétique est une cause importante de cécité et survient par suite des lésions des petits vaisseaux sanguins de la rétine qui s'accumulent avec le temps. Au bout de 15 ans de diabète, près de 2% des sujets deviennent aveugles et environ 10% présentent des atteintes visuelles graves [OMS11].
- Le diabète figure parmi les principales causes d'insuffisance rénale. 10 à 20% des diabétiques meurent d'une insuffisance rénale [OMS11] (Figure 1.8).

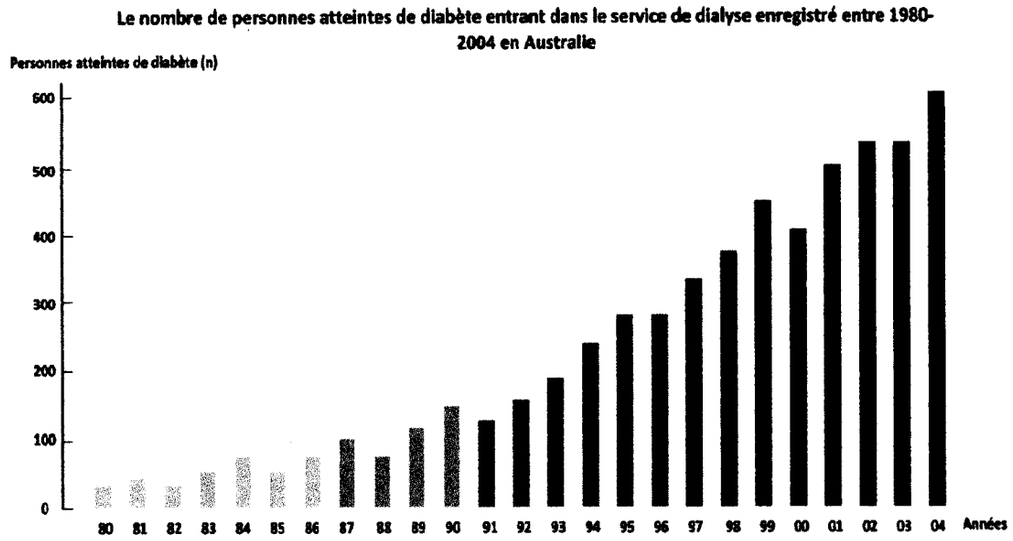


FIGURE 1.8 : Evolution des malades diabétiques souffrant d'insuffisance rénale signalés en Australie [ALW11].

- La neuropathie diabétique fait suite aux lésions nerveuses dûes au diabète et touche jusqu'à 50% des diabétiques. Bien que de nombreux problèmes différents puissent résulter d'une neuropathie diabétique, les symptômes courants sont les suivants: fourmillement, douleur, engourdissement ou faiblesse au niveau des pieds et des mains.

6. 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.
- 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.
- Race: le DM1 est plus répandu entre des personnes de race blanche. Pour le DM2 le type de régime alimentaire pourrait exacerber certains hyperinsulinisme génétiquement conditionné.
- Niveau socioéconomique: plus le niveau est bas, plus le risque d'avoir un DM2 et de souffrir des complications augmente, puisque il y'a une tendance à suivre un régime déséquilibré et consommer des aliments hautement énergétiques, ce qui favorise les altérations métaboliques.

7. Diagnostic du diabète

Le diabète peut être diagnostiqué grâce à de simples analyses de sang. Un diagnostic de diabète peut être posé lorsque le taux de glucose dans le sang après 8 heures de jeûne est égal ou supérieur à 7,0 mmol/L. Si le taux de sucre sanguin (la glycémie) à jeûne se situe entre 6,1 et 6,9 mmol/L, cela signifie que vous avez peut-être un trouble appelé une hyperglycémie modérée à jeûne (ou prédiabète) qui pourrait se transformer en diabète un jour.

Le diagnostic de diabète est également confirmé si le taux de sucre sanguin mesuré à n'importe quel moment de la journée, sans tenir compte des repas, est égal ou supérieur à 11,1 mmol/L et si des symptômes caractéristiques du diabète sont présents (par exemple: une soif accrue, un besoin d'uriner plus souvent, une perte de poids inexpliquée). Il est important de préciser qu'une seule mesure élevée du taux de sucre ne signifie pas qu'une personne fait du diabète. Dans la plupart des cas, il faut au moins 2 lectures élevées avant que le médecin pose le diagnostic de diabète, sauf s'il soupçonne la présence du diabète de type 1.

Une autre méthode peut être employée pour diagnostiquer le diabète, soit l'épreuve d'hyperglycémie provoquée par voie orale ou HPO. Ce test consiste à ingérer une boisson contenant 75 g de glucide à la suite d'une période de jeûne. Le taux de sucre sanguin est mesuré à jeun puis 2 heures après l'ingestion de la boisson.

Un taux de sucre sanguin supérieur à 11,1 mmol/L dans les 2 heures confirme le diagnostic de diabète [CCD12].

8. Prévention et traitement

Outre un dépistage permettant un traitement plus précoce, un régime alimentaire adapté, une augmentation de l'activité physique (baisse de poids), avec une sensibilisation et un programme d'éducation continu peuvent fortement diminuer la prévalence du diabète. C'est ce qu'a notamment montré, selon l'OMS, une expérience chinoise conduite sur six ans au sein d'une population sensible, qui a réduit de près des deux tiers l'apparition de cas de diabète.

De telles mesures sont lourdes mais très rentables à long et moyen termes si appliquées à toute une population. Des conséquences secondaires positives concerneront de plus l'obésité, les maladies cardio-vasculaires et certains cancers d'origine socio-environnementale.

Chez les patients ayant déjà développé un diabète, divers moyens existent afin de diminuer l'impact :

- Le traitement précoce de l'hypertension artérielle et de l'hyperlipémie, le contrôle de la glycémie (antidiabétiques oraux pour le diabète de type II et insuline pour le diabète de type I) réduisent les complications et freinent l'évolution vers les formes graves de diabète. La détection et le traitement précoces de la protéinurie limitent ou freinent l'évolution vers l'insuffisance rénale.
- La prévention de l'ulcération des pieds par une éducation et des soins appropriés divise par deux l'incidence des amputations (source OMS).
- Le dépistage et le traitement précoces des rétinopathies évitent nombre de cécités et diminuent les coûts globaux (dont indirects et immatériels) du diabète.
- Une lutte plus efficace contre le tabagisme et l'alcoolisme, facteurs d'aggravation du diabète (hypertension et cardiopathie) est également recommandée par l'OMS.

9. Aide au diagnostic :

Le diagnostic médical est un processus de classification. L'utilisation de l'informatique pour la réalisation de cette classification devient de plus en plus fréquente. Même si la décision de l'expert est le facteur le plus important lors du diagnostic, les systèmes de classification fournissent une aide substantielle, car elles réduisent les erreurs dues à la fatigue et le temps nécessaire pour le diagnostic. Actuellement, la plupart des hôpitaux modernes sont bien équipés avec des dispositifs de collecte de données. Ces données seront partagées en inter- et intra-systèmes d'information hospitaliers. Ce qui était avant une base de données isolées ou un système d'information de laboratoire est maintenant intégrée dans un système d'information médicale à plus grande échelle (ministères, hôpitaux, ou à base communautaire). L'augmentation du volume de données entraîne des difficultés à extraire des informations utiles pour l'aide à la décision. Les méthodes traditionnelles d'analyse de données sont devenues insuffisantes, et les méthodes dites intelligentes sont indispensables.

10. Conclusion

Nous avons vu dans ce chapitre les différents types de diabète, les différents traitements et testes ainsi que les complications dû à cette maladie. Même s'il existe des méthodes de prévention qui permettent de réduire le risque d'avoir le diabète, parfois il est impossible de l'éviter comme pour le diabète de type 1. Dans ces cas là, la seule solution est de pouvoir le diagnostiquer très tôt et faire tout son possible pour combattre les complications.

Dans ce travail nous présentons un classifieur basé sur l'approche ensembliste «boosting» où nous utilisons le classifieur faible K-PPVP (k-plus proche voisins pondéré), pour la reconnaissance du diabète.

Chapitre II

Principes de la méthode Boosting

1 Introduction

Depuis les années 90, la combinaison de classifieurs a été une des directions de recherche les plus soutenues dans le domaine de la classification. L'amélioration des performances des systèmes de classification est finalement le principal enjeu des recherches menées ces dernières années sur les systèmes de combinaison.

De manière étonnante, des recherches en apprentissage artificiel montrent qu'il est possible d'atteindre une décision aussi précise que souhaitée par une combinaison judicieuse d'experts imparfaits mais correctement entraînés. Plusieurs algorithmes d'apprentissage ont été développés à la suite de ces travaux.

Le mot boosting s'applique à des méthodes générales capables de produire des décisions très précises (au sens d'une fonction de perte) à partir d'un ensemble de règles de décision (faibles), c'est-à-dire dont la seule garantie est qu'elles soient un peu meilleures que le hasard. Ces méthodes s'appliquent aussi bien à l'estimation de densité qu'à la régression ou à la classification. Pour simplifier, nous nous concentrons ici sur la tâche de classification binaire.

Dans sa version (par sous-ensembles), cette technique fait produire à l'algorithme trois résultats selon la partie de l'ensemble d'apprentissage sur laquelle il apprend, puis combine les trois apprentissages réalisés pour fournir une règle de classification plus efficace. Examinons d'abord cette technique avant de voir comment la généraliser à l'aide de distributions de probabilité sur les exemples.

2 Les méthodes ensemblistes

Les méthodes du datamining (ou apprentissage) ensemblistes présentent actuellement des techniques très sollicitées dans la communié scientifique, grâce à leur apport significatif en terme de précision. L'idée est basée sur la combinaison des résultats de plusieurs algorithmes du datamining, appliqué chacun sur un ensemble diversifié de données.

2-1 La classification par les méthodes ensemblistes

La classification est une des tâches du datamining qui permet de prédire si une instance de donnée est membre d'une classe prédéfinie. Elle utilise un ensemble S de données appelées ensemble d'apprentissage. Chaque donnée est typiquement représentée sous forme d'un vecteur d'attributs $x = <x_1, x_2, \ldots., x_m, y>$ avec y un attribut de classe. L'objectif de la classification est d'entraîner un algorithme de

classification *A* sur l'ensemble *S*, pour trouver une bonne approximation d'une certaine fonction *f(x)= y*. La fonction approximative *Cl* calculée est appelée Pclassificateur. L'évaluation de la précision de *Cl* est faite sur un ensemble de données *T* indépendant de *S*, appelé ensemble de test. Le classificateur sera par la suite capable de prédire la valeur de classe *y* pour de nouvelles données *d*, en calculant *Cl(d)*.

Dans le cas des méthodes ensemblistes, *N* classificateurs de base *Cli* sont construits, à partir de *N* ensembles de données *Si*. Le classement d'une nouvelle donnée se fait par la combinaison des prédictions des *N* classificateurs de base, par un vote majoritaire par exemple. Malgré la simplicité de cette idée intuitive « l'union fait la force », elle repose sur une théorie statistique [LBR96] renforcée par plusieurs études empiriques.

Ces études ont montré dans différents travaux de recherche [[LBB96], [LBR01], [YFE96], [YK02], [JRQ96]] que la précision d'un algorithme d'apprentissage peut être améliorée d'une façon significative en appliquant le principe de perturbation et combinaison [LBR96]. Les algorithmes les plus appropriés à l'application de cette approche sont ceux considérés comme non stable, ca-d que des petites modifications dans les données d'apprentissage pourraient induire à un grand changement dans la fonction *Cl* estimée. Les arbres de décision par exemple sont considérés comme de bons candidats [LBR96].

Cette perturbation permet de générer plusieurs ensembles d'apprentissage, à partir d'un ensemble de base, comme dans les techniques de boosting et bagging. Elle peut aussi être appliquée sur les algorithmes de construction des classificateurs, en utilisant plusieurs algorithmes différents, ou en modifiant certains paramètres [RPW02]. Les résultats expérimentaux montrent que 50 répliques sont en générale suffisantes [LBR96], mais le temps de calcul est encore un champ d'investigation. Nous présentons dans ce qui suit, des approches parmi les plus répandues dans la génération des ensembles d'apprentissage, ainsi que les techniques de combinaison.

3 Génération des ensembles d'apprentissage

Afin d'aboutir à divers ensembles d'apprentissage qui seront utilisés pour construire les classificateurs de base, plusieurs techniques peuvent être appliquées, parmi les plus utilisées, nous présentons les techniques: *bagging*, *boosting*, *comité de validation croisée*.

3.1 Bagging (Boostrap AGGregatING)

Le bagging est une méthode qui, comme le boosting, combine des hypothèses pour obtenir une hypothèse finale. Cependant la méthode est plus simple et généralement moins performante [LBB96]. L'idée de base est d'entraîner un algorithme d'apprentissage (arbre de décision, réseau connexionniste, etc.) sur plusieurs bases d'apprentissage obtenues par tirage avec remise de m` (avec m`<m) exemples d'apprentissage dans l'échantillon d'apprentissage S. Pour chaque tirage b (pour *bag*), une hypothèse h_b est obtenue. L'hypothèse finale est simplement la moyenne des hypothèses obtenues sur B tirages au total :

$$H(x) = \frac{1}{B} \sum_{b=1}^{B} h_b(x)$$

L'une des justifications de cette méthode est que si les hypothèses h_b calculées pour chaque tirage b ont une variance importante (donc sont sensibles à l'échantillon des m` exemples d'apprentissage), alors leur moyenne H aura une variance réduite.

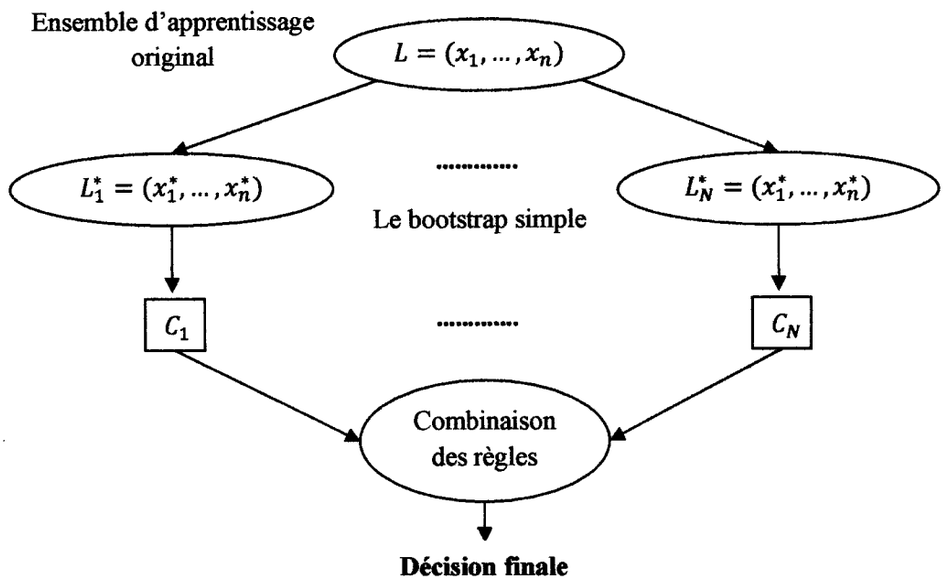


FIGURE 2.1: Manipuler les Données d'Apprentissage: Bagging.

3.2 Boosting

L'idée de base est de construire un nouveau classificateur, selon la performance d'une série de classificateurs précédents, dans un processus séquentiel.

L'ensemble d'apprentissage d'origine est renforcé par des poids qui seront ajustés à chaque étape, dans l'objectif 'd'amplifier' (boost) les exemples mal classés. Les poids des exemples bien classés -par le dernier modèle construit- seront alors décrémentés, et les poids des exemples mal classés seront incrémentés, en permettant ainsi au système de 'prêter plus d'attention' aux exemples mal classés [RES90]. Les modèles sont combinés par vote à majorité pondéré où la pondération est déterminée par la précision de prédiction de chaque classificateur. Nous présentons Adaboost [YFR95] comme un algorithme utilisant cette technique.

3.2.1 Le premier algorithme de Boosting

Schapire [RSC90] développa le premier algorithme de boosting pour répondre à une question de Kearns : est-il possible de rendre aussi bon que l'on veut un algorithme d'apprentissage (faible), c'est-à-dire un peu meilleur que le hasard? Shapire montra qu'un algorithme faible peut toujours améliorer sa performance en étant entraîné sur trois échantillons d'apprentissage bien choisis. Nous ne nous intéressons ici qu'à des problèmes de classification binaire.

L'idée est d'utiliser un algorithme d'apprentissage qui peut être de natures très diverses (un arbre de décision, une règle bayésienne de classification, une décision dépendant d'un hyperplan, etc.) sur trois sous-ensembles d'apprentissage.

1- On obtient d'abord une première hypothèse h1 sur un sous-échantillon S1 d'apprentissage de taille m1 <m (m étant la taille de S l'échantillon d'apprentissage disponible).
2- On apprend alors une deuxième hypothèse h2 sur un échantillon S2 de taille m2 choisi dans S - S1 dont la moitié des exemples sont mal classés par h1.
3- On apprend finalement une troisième hypothèse h3 sur m3 exemples tirés dans S - S1 - S2 pour lesquels h1 et h2 sont en désaccord.
4- L'hypothèse finale est obtenue par un vote majoritaire des trois hypothèses apprises :

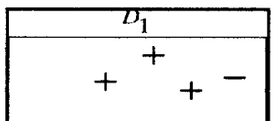
H = vote majoritaire (*h1,h2, h3*)

Le théorème de Schapire sur la (force de l'apprentissage faible) prouve que H a une performance supérieure à celle de l'hypothèse qui aurait été apprise directement sur l'échantillon S.

Une illustration du boosting selon cette technique de base est donnée dans les figures: (2.2, 2.3, 2.4, 2.5, 2.6).

3.2.2 Les étapes d'apprentissage par le boosting

Étape 1 : l'ensemble initial



Étape 4 : en augmentant les poids des exemples mal classés et en appliquant le troisième classifieur faible sur l'ensemble d'apprentissage.

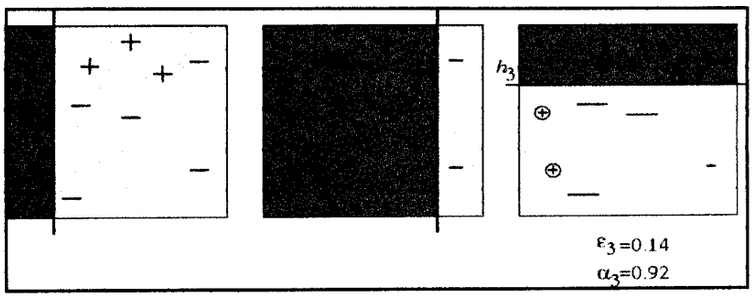


FIGURE 2.5: le résultat du troisième

Étape 5 : le résultat est une combinaison de décision entre les 3 classifieurs.

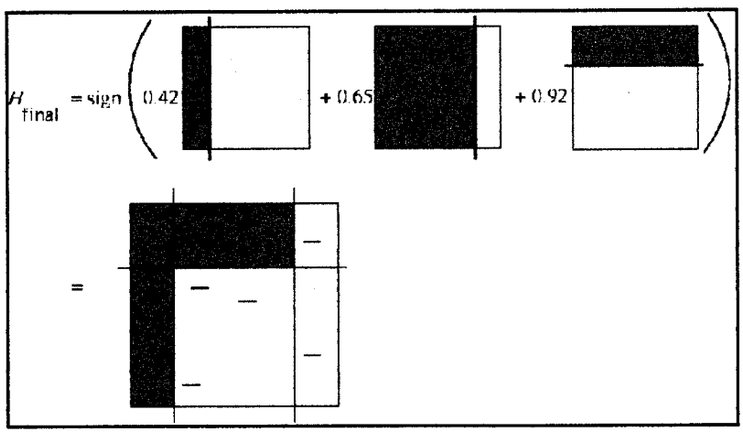


FIGURE 2.6: le résultat final du boosting

3.2.3 L'utilisation du Boosting

Le boosting, et particulièrement l'algorithme AdaBoost, a été employé avec succès avec de nombreux algorithmes d'apprentissage (faibles) (par exemple C4.5 : un système d'apprentissage d'arbre de décision [JQU90] ou Ripper: un système d'apprentissage de règles) et sur des domaines d'application variés. En général, l'utilisation du boosting a pour résultat d'améliorer souvent sensiblement les performances en apprentissage.

Les avantages du boosting et de l'AdaBoost en particulier sont qu'il s'agit d'une méthode facile à programmer et aisée d'emploi. Elle ne nécessite pas de connaissance a priori sur l'algorithme d'apprentissage (faible) utilisé, et elle peut s'appliquer de fait à n'importe quel algorithme d'apprentissage faible. Les seuls paramètres à régler sont la taille de l'ensemble d'apprentissage m et le nombre total

d'étapes T, qui peuvent être fixés par l'utilisation d'un ensemble de validation. De plus, des garanties théoriques sur l'erreur en généralisation permettent de contrôler l'apprentissage. Une autre propriété intéressante du boosting est qu'il tend à détecter les exemples aberrants (outliers) puisqu'il leur donne un poids exponentiellement grand en cours d'apprentissage. Cependant, la contrepartie de ce phénomène est que le boosting est sensible au bruit et ses performances peuvent être grandement affectées lorsque de nombreux exemples sont bruités. Récemment des algorithmes ont été proposés pour traiter ce problème (comme Gentle AdaBoost [TRJ01] ou BrownBoost [YFR99]).

Il est à noter que l'adaptation aux problèmes multiclasses n'est pas immédiate, mais elle a cependant fait l'objet d'études menant aussi à des algorithmes efficaces. De même qu'il existe des extensions à la régression.

4 L'algorithme Adaboost

Trois idées fondamentales sont à la base des méthodes de boosting probabiliste :

1- L'utilisation d'un comité d'experts spécialisés que l'on fait voter pour atteindre une décision.

2- La pondération adaptative des votes par une technique de mise à jour multiplicative.

3- La modification de la distribution des exemples disponibles pour entraîner chaque expert, en surpondérant au fur et à mesure les exemples mal classés aux étapes précédentes.

L'algorithme le plus pratiqué s'appelle AdaBoost (pour adaptive boosting). L'une des idées principales est de définir à chacune de ses étapes $1 <= t <= T$, une nouvelle distribution de probabilité a priori sur les exemples d'apprentissages en fonction des résultats de l'algorithme à l'étape précédente. Le poids à l'étape t d'un exemple $(x_i; y_i)$ d'indice i est noté $Dt(i)$. Initialement, tous les exemples ont un poids identique, puis à chaque étape, les poids des exemples mal classés par l'apprenant sont augmentés, forçant ainsi l'apprenant à se concentrer sur les exemples difficiles de l'échantillon d'apprentissage.

A chaque étape t, l'apprenant cherche une hypothèse ht X : {-1; +1} bonne pour la distribution Dt sur X. La performance de l'apprenant est mesurée par l'erreur :

$$\varepsilon_t = PD_t [h_t (x_i) \neq y_i] = \sum_{i:h_t (x_i) \neq y_i} D_t(i)$$

On note que l'erreur est mesurée en fonction de la distribution D_t sur laquelle l'apprenant est entraîné. En pratique, soit les poids des exemples sont effectivement modifies, soit c'est la probabilité de tirage des exemples qui est modifié et l'on utilise un tirage avec remise (bootstrap). Chaque hypothèse h*t* apprise est affectée d'un poids α_t mesurant l'importance qui sera donnée à cette hypothèse dans la combinaison finale. Ce poids est positif si $\varepsilon_t < 1/2$ (on suppose ici que les classes '+' et '-' sont équiprobables, et donc que l'erreur d'une décision aléatoire est de 1/2). Plus l'erreur associée à l'hypothèse h*t* est faible, plus celle-ci est dotée d'un coefficient α_t important.

L'examen des formules de mise à jour des poids des hypothèses dans l'algorithme (Adaboost) suggère que vers la fin de l'apprentissage, le poids des exemples difficiles à apprendre devient largement dominant. Si une hypothèse peut être trouvée qui soit performante sur ces exemples (c'est-à dire avec $\varepsilon_t \approx 0$), elle sera alors dotée d'un coefficient α_t considérable. L'une des conséquences possibles est que les exemples bruités, sur lesquels finit par se concentrer l'algorithme, perturbent gravement l'apprentissage par boosting. C'est en effet ce qui est fréquemment observé.

Algorithme 1 Adaboost

Entrée : Base d'exemples : $A=\{(x_1, y_1), \ldots, (x_m, y_m)\}$, avec $y_i \in \{+1, -1\}$, $i=1, \ldots, m$, Algorithme de classification "faible » : L, Nombres d'itérations : T

- Initialisation : Poids initiaux : $D_1(x_i)= 1/m$ pour tout $i=1, \ldots, m$
- Pour chaque étape t de 1 à T, faire

1- Apprendre par l'algorithme L une hypothèse de classification h_t sur A suivant la distribution D_t.

2- Calculer l'erreur pondérée ε_t de h_t sur A en considérant la distribution D_t. (somme des poids des ex. mal classés)

$$\varepsilon_t = \sum_{i}^{n} (x_i) \times | h_t(x_i) - y_i|$$

3- Si $\varepsilon t \geq 0.5$ alors $T= t-1$; exit.

4- En déduire la «capacité prédictive» de h_t : $\alpha_t = {}^1/_2 \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) > 0$

$$\begin{cases} \alpha_t > 0 \; si \; \varepsilon_t < 0.5 \\ \alpha_t \to +\infty \; si \; \varepsilon_t \to 0 \end{cases}$$

5- Mettre à jour les poids, i.e. pour i de 1 à m

$$\frac{D_t(x_i)}{z_t} \; D_{t+1} = \begin{cases} e^{-\alpha_t} \; si \;\; y_i = h_t(x_i) & x \; bien \; classé \\ e^{\alpha_t} \; si \;\; y_i \neq h_t(x_i) & x \; mal \; classé \end{cases} \quad z_t : facteur \; denormalisation$$

Fournir l'hypothèse finale $H_{final}(x) = Sng(\sum_t \alpha_t \; h_t(x))$

5 Comité de validation croisée [TGD00]

Cette technique consiste à diviser l'ensemble S en k parties disjointes $\{S_1, S_2, \ldots S_k\}$. Le processus suivant est répété k fois : construction d'un classificateur C_i avec l'ensemble S privé de S_i $(S-S_i)$, ensuite évaluer la précision de C_i testé sur S_i. La précision globale est obtenue par la moyenne. Les ensembles construits de cette manière sont appelés 'cross validated commitees'.

6 Méthodes de combinaison

Une fois les classificateurs de base construits, différentes techniques peuvent être utilisées pour combiner les résultats de chaque classificateur. On présente quelques unes parmi les plus citées dans la littérature : le vote majoritaire, le vote pondéré, stacking.

6.1 Le vote à majorité

C'est une technique simple et intuitive, qui consiste à classer la nouvelle instance selon la prédiction majoritaire des classificateurs de base. L'inconvénient de cette méthode est dans le cas ou plus de la moitié des classificateurs de base obtiennent de faux résultats.

6.2 Le vote à majorité pondéré

C'est un vote basé sur des poids associés aux classificateurs de base. Ces poids peuvent être diminués ou augmentés au fur et à mesure que les classificateurs s'entraînent, suivant qu'ils produisent respectivement une bonne ou une mauvaise prédiction.

6.3 Stacking

C'est une méthode qui permet de combiner plusieurs classificateurs de base. La première phase consiste à induire N classificateurs Cl_i, à partir de N ensembles de données $\{S_1, S_2, ..., S_N\}$. Le test est ensuite fait sur un ensemble d'évaluation $T=\{t_1, t_2, ..., t_L\}$, indépendant des ensembles d'apprentissage Si. Dans la deuxième phase, un nouvel ensemble de données M est formé par les valeurs calculées $Cl_i(t_j)$ et la vraie classe de l'instance tj, classe(t_j). Chaque instance de M sera de la forme $<Cl_1(t_j), Cl_2(t_j), ..., Cl_N(t_j), classe(t_j)>$. Dans la dernière étape, un classificateur global est construit à partir de M. Les classificateurs de base peuvent être construits avec des algorithmes différents (arbres de décision, réseaux de neurone..) selon les contraintes du problème [DHW92].

7 Le classificateur faible proposé

7.1 L'algorithme KNN

L'algorithme KNN figure parmi les plus simples algorithmes d'apprentissage artificiel. Dans un contexte de classification d'une nouvelle observation $\boldsymbol{x}$, l'idée fondatrice simple est de faire voter les plus proches voisins de cette observation. La classe de $\boldsymbol{x}$ est déterminée en fonction de la classe majoritaire parmi les k plus proches voisins de l'observation $\boldsymbol{x}$. La méthode KNN est donc une méthode à base de voisinage, non-paramétrique ; Ceci signifiant que l'algorithme permet de faire une classification sans faire d'hypothèse sur la fonction $y=f(x1,x2,...xp)$ qui relie la variable dépendante aux variables indépendantes [EMD10].

7.2 Algorithme 1-NN

La méthode du plus proche voisin est une méthode non paramétrique où une nouvelle observation est classée dans la classe d'appartenance de l'observation de l'échantillon d'apprentissage qui lui est la plus proche, au regard des covariables utilisées. La détermination de leur similarité est basée sur des mesures de distance.

Formellement, soit L l'ensemble de données à disposition ou échantillon d'apprentissage :

$$L = \{(y_i, x_i), 1 = 1, .. n_L\}$$

où $y\ i \in \{1,.. c\}$ dénote la classe de l'individu i et le vecteur $x = (x_{i1},\ldots,x_{ip})$ représente les variables prédictrices de l'individu i. La détermination du plus proche voisin est basée sur un fonction distance arbitraire $d(.,.)$. La distance euclidienne ou dissimilarité entre deux individus caractérisés par p covariables est définie par:

$$d\left((x_1, x_2, .., x_p), (u_1, u_2, .., u_p)\right) = \sqrt{(x_1 - u_1)^2 + (x_2 - u_2)^2 + \cdots + (x_p - u_p)^2}$$

Ainsi, pour une nouvelle observation (y,x) le plus proche voisin $(y_{(1)}, x_{(1)})$ dans l'échantillon d'apprentissage est déterminé par :

$$d(x, x_{(1)} = \min_i(d(x, x_i))$$

et $\hat{y} = y_{(1)}$, la classe du plus proche voisin, est sélectionnée pour la prédiction de y. Les notations $x_{(j)}$ et $y_{(j)}$ représentent respectivement le $j^{ème}$ plus proche voisin de x et sa classe d'appartenance. Parmi les fonctions distance types, la distance euclidienne est définie comme suit :

$$d(x_i, x_j) = \left(\sum_{s=1}^{p} |x_{is} - x_{js}|^2\right)^{1/2}$$

et plus généralement la distance de Minkowski :

$$d(x_i, x_j) = \left(\sum_{s=1}^{p} |x_{is} - x_{js}|^q\right)^{1/q}$$

La méthode est justifiée par l'occurrence aléatoire de l'échantillon d'apprentissage. La classe $Y_{(1)}$ du voisin le plus proche $x_{(1)}$ d'un nouveau cas x est une variable aléatoire. Ainsi la probabilité de classification de x dans la classe $y_{(1)}$ est $P[Y_{(1)} / x_{(1)}]$. Pour de grands échantillons d'apprentissage, les individus x et $x_{(1)}$ coincident de très près, si bien que $P[y_{(1)} / \mathbf{x}_{(1)}] \approx P[y / \mathbf{x}]$. Ainsi, la nouvelle observation (individu) x

est prédite comme appartenant à la vraie classe y avec une probabilité égale approximativement à P[y/x] [EMD10].

7.3 Algorithme K-NN

Une première extension de cette idée, qui est largement et communément utilisée en pratique, est la méthode des k plus proches voisins. La plus proche observation n'est plus la seule observation utilisée pour la classification. Nous utilisons désormais les k plus proches observations. Ainsi la décision est en faveur de la classe majoritairement représentée par les k voisins. Soit k_r le nombre d'observations issues du groupe des plus proches voisins appartenant à la classe r

$$\sum_{r=1}^{c} k_r = k$$

Ainsi une nouvelle observation est prédite dans la classe l avec :

$$1 = max_r(k_r)$$

Ceci évite que la classe prédite ne soit déterminée seulement à partir d'une seule observation. Le degré de localité de cette technique est déterminé par le paramètre k : pour k=1, on utilise la méthode du seul plus proche voisin comme technique locale maximale, pour $k \to n_1$ on utilise la classe majoritaire sur l'ensemble intégral des observations (ceci impliquant une prédiction constante pour chaque nouvelle observation à classifier).

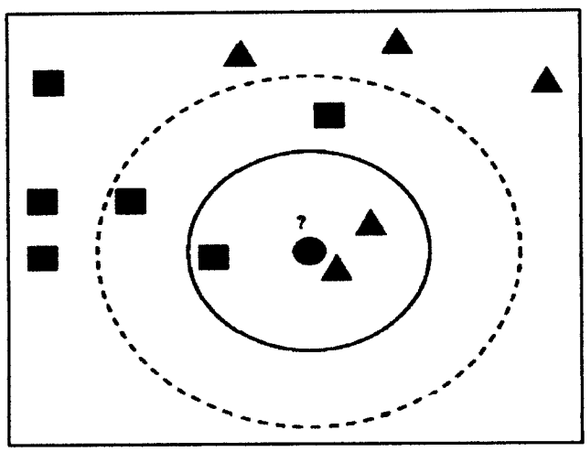


FIGURE 2.7: la sélection des k plus proche voisins

7.4 Quelques règles sur le choix de k

Le paramètre k doit être déterminé par l'utilisateur : $k \in N$. En classification binaire, il est utile de choisir k impair pour éviter les votes égalitaires. Le meilleur choix de k dépend du jeu de donnée. En général, les grandes valeurs de k réduisent l'effet du bruit sur la classification et donc le risque de sur-apprentissage, mais rendent les frontières entre classes moins distinctes. Il convient donc de faire un choix de compromis entre la variabilité associée à une faible valeur de k contre un 'oversmoothing' ou surlissage (i.e gommage des détails) pour une forte valeur de k. Un bon k peut être sélectionné par diverses techniques heuristiques, par exemple, de validation-croisée. Nous choisirons la valeur de k qui minime l'erreur de classification.

8 Méthode des k plus proches voisins pondérés

Cette technique étend la méthode des k plus proches voisins selon deux voies :

1) Tout d'abord, un schéma de pondération des plus proches voisins est introduit en fonction de leur similarité avec la nouvelle observation à classer.
2) Basé sur le fait que le vote des plus proches voisins est équivalent au mode de la distribution de la classe, la seconde extension utilise la médiane ou la moyenne de cette distribution, si la variable cible est relative à une échelle ordinale ou de niveau plus élevé.

Cette extension est fondée sur l'idée que les observations de l'échantillon d'apprentissage, qui sont particulièrement proches de la nouvelle observation (y,x), doivent avoir un poids plus élevé dans la décision que les voisins qui sont plus éloignés du couple (y,x). Ce n'est pas le cas avec la méthode KNN : en effet seuls les k plus proches voisins influencent la prédiction, mais l'influence est identique pour chacun des voisins, indépendamment de leur degré de similarité avec (y,x). Pour atteindre ce but, les distances, sur lesquelles la recherche des voisins est fondée dans une première étape, sont transformées en mesures de similarité, qui peuvent être utilisées comme poids [EMD10].

Algorithme 2 K-NN pondéré

1- Affecter des poids aléatoires w_i à tous les éléments de la base d'apprentissage .
2- Pour chaque exemple x_i de la base de test :
   - Trouver les k plus proche voisins, on utilisant la distance euclidienne.
   - Calculer la classe :
     $\sum w_k * x_{j\,k}$ j : la classe de l'exemple
   - Si la classe actuelle ≠ la classe calculé alors
     Appliquer la descente de gradient
     - Erreur = la classe actuelle – la classe calculé
     - Pour chaque w_k : $w_k = w_k + \alpha * erreur$
3- Calculer le taux de classification :
   Taux = nombre d'exemple bien classé / nombre d'exemple de la base d'apprentissage

9 Conclusion

Le boosting est une technique d'apprentissage qui vise à rendre plus performant un système d'apprentissage (faible). Pour ce faire, le système d'apprentissage est entraîné successivement sur des échantillons d'apprentissage surpondérant les exemples difficiles à apprendre. A chaque fois, une hypothèse ht est produite, et l'hypothèse finale est une combinaison linéaire de ces hypothèses pondérées par des coefficients liés à leur performance. Le boosting est d'un emploi très large et fait l'objet de nombreux travaux et applications.

Ce chapitre était destiné à présenter les techniques du boosting, les différentes versions de l'algorithme proposé k-nn ainsi que l'algorithme de l'Adoboost. Le chapitre suivant sera consacré à l'implémentation de cette technique pour la reconnaissance du diabète.

Chapitre III

Implémentation et discussion des résultats

1. Introduction

Ce chapitre présente l'approche boosting utilisée pour effectuer la classification du diabète. Cette approche est l'une des extensions des méthodes ensemblistes, le principe est basée sur la combinaison des résultats de plusieurs algorithmes du datamining, appliqué chacun sur un ensemble diversifié de données.

Dans le but d'améliorer la classification des diabétiques via cette approche de boosting, nous utilisons l'algorithme K-NNW (K plus proche voisin pondéré) comme un algorithme de base pour chaque itération de l'algorithme Adaboost.

2. Description de la base de données utilisée

Dans ce mémoire nous avons utilisé la base Pima Indian Diabetes (PID), disponible sur le site officiel de l'UCI (machine learning repository), offerte par Vincent Sigillito. Cette base est une collection de rapports de diagnostic médicaux de 768 femmes âgé de plus de 21 ans, ces patientes sont des indiennes Pima, une population vivant près de Phoenix, Arizona, USA.

2.1 L'intérêt de la base de données

Des études ont montré que les Indiens Pima d'Arizona, dont le régime alimentaire et le mode de vie est semblable à la plupart des Américains, ont un taux de diabète de type 2 beaucoup plus élevé que la moyenne nationale, ce qui les rend le groupe le plus prédisposé au diabète dans le monde, elle atteint 70 % pour les personnes âgées entre 55 et 64 ans. Les modifications du mode de vie au cours du siècle dernier sont à l'origine de l'apparition du diabète chez les Indiens Pimas mais elles ne peuvent être tenues pour responsables de son caractère épidémique. Cette population présente une prédisposition génétique particulière au diabète de type 2. Avec un degré d'obésité et un niveau glycémique similaire, les Indiens Pimas, comparés aux autres populations, présentent une résistance à l'action de l'insuline. Cette caractéristique est certainement en partie génétiquement déterminée [PIH].

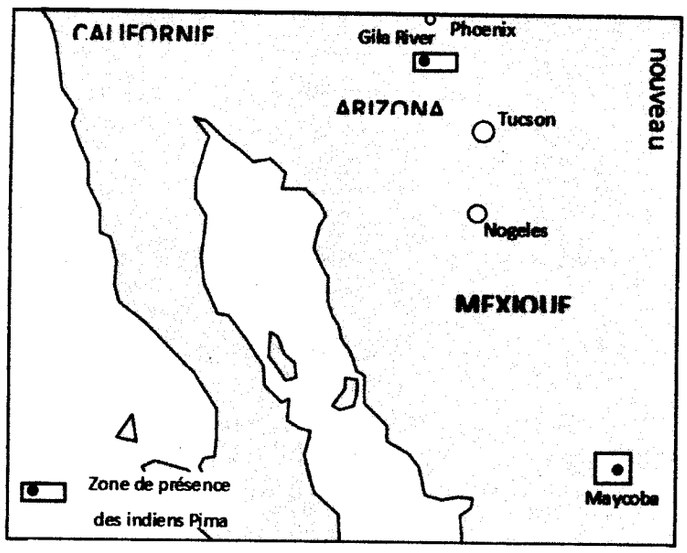


FIGURE 3.1 : Zones de présence des indiens Pima

2.2 Base de données utilisées

La base Pima Indian Diabetes est constitué de 768 cas dont 268 sont diabétique et 500 non diabétique. Chaque cas est formé de 9 attributs, dont 8 représentent des facteurs de risque et le 9eme représente la classe du patient, le tableau 6 présente une description de ces attributs:

N° attribut	*Description d'attribut*	*Moyenne*	*Déviation standard*
1	Nombre de grossesses (Ngross)	3.8	3.4
2	Concentration du glucose plasmatique (mg/dl)	120.9	32.0
3	Pression artérielle diastolique (mm Hg) (PAD)	69.1	19.4
4	Epaisseur de la peau au niveau du triceps (mm) (Epai)	20.5	16.0
5	Taux d'insuline au bout de 2 heures (*m*U¼ml) (INS)	79.8	115.2
6	Indice de masse corporelle (poids en kg/ m^2) (IMC)	32.0	7.9
7	Fonction pédigrée du diabète (Ped)	0.50	0.3
8	Age (années)	33.20	11.8

Tableau 3.1 : Description des attributs de la base

Malheureusement, il existe quelques cas avec des données manquantes qui ont été remplacé par des zéros, ce qui donne des valeurs biologiquement impossible tel qu'une pression artérielle égal à 0. Après élimination de ces cas, nous obtenons une base de 392 patientes dont 262 non-diabétiques et 130 diabétiques.

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 :

1. Npreg : Nombre de grossesses,
2. Glu : Concentration du glucose plasmatique,
3. BP : Tension artérielle diastolique,
4. SKIN : Epaisseur de pli de peau du triceps,
5. Insuline : Dose d'insuline,
6. BMI : Index de masse corporelle,
7. PED : Fonction de pedigree de diabète (l'héridité),
8. Age : Age.

La figure (3.2) montre respectivement les histogrammes et la répartition des diabétiques et non diabétiques des paramètres de la base de données :

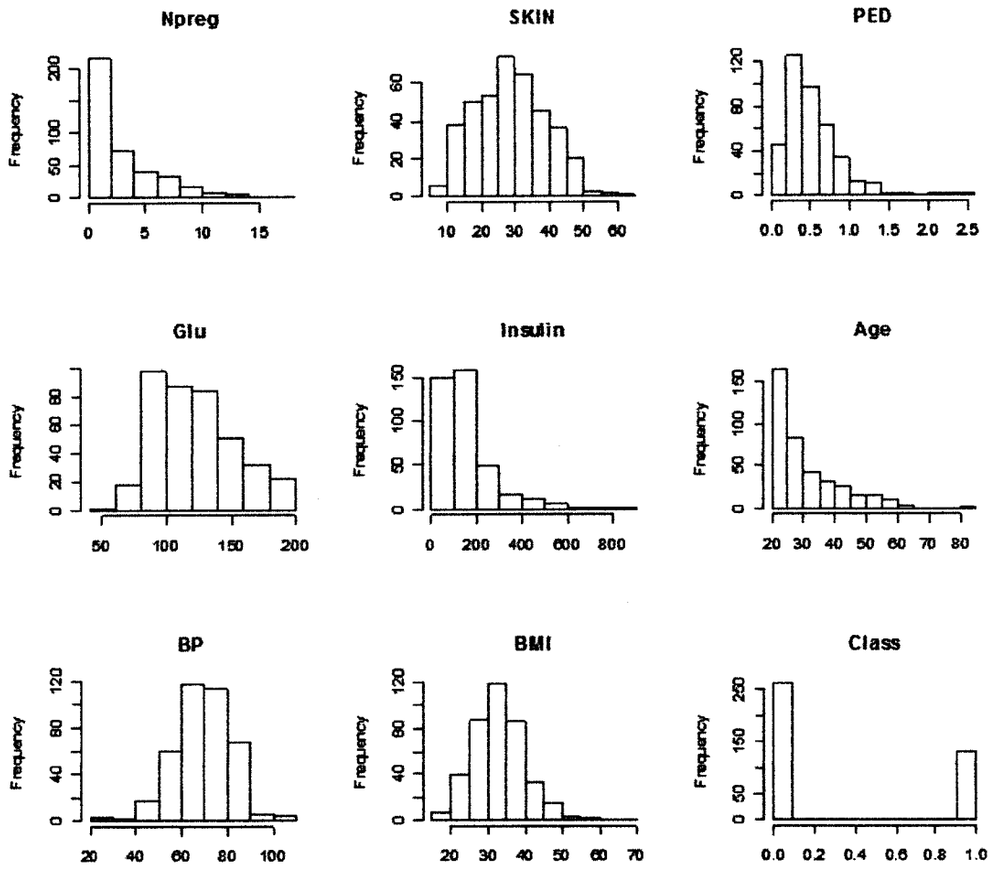


FIGURE 3.2: Histogrammes des 9 paramètres de la base de données.

3. Approches utilisées

3.1 L'approche boosting

L'idée de base de cette approche est de construire un nouveau classifieur, selon la performance (l'erreur de classification) d'une série de classifieurs précédents, l'objectif étant de booster les exemples mal classés. A la fin, les modèles sont combinés par vote à majorité pour classer les nouveaux exemples [AZ02].L'algorithme de boosting utilisé est Adaboost présenté par Yoav Freund et Robert Schapire en 1995 [YRS96], c'est un méta-algorithme qui utilise le principe du boosting pour améliorer les performances des classifieurs. L'idée est d'attribuer un poids à chaque élément de l'ensemble d'apprentissage. Au début, ils ont tous le même poids mais à chaque itération, les poids des éléments mal classés seront augmentés tandis que ceux des éléments bien classés seront décrémentés. Ainsi, le classifieur suivant sera forcé à se focaliser sur les cas difficiles de l'ensemble d'apprentissage. Par conséquence, les classifieurs seront complémentaires [[YRS96], [YFS99]].

Ensemble de test $S=\{x_1, \dots, x_m\}$
Initialisation $D(i)=1/m$, $i=1..m$
Pour $t=1..N$

- *Apprendre le classifieur C1 sur St*
- *Calculer l'erreur de C1 sur S*
- *Calculer* $\alpha_t = {}^1/_2 \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) > 0$
- *Mise à jour :*

$$D_{t+1} = \frac{D_t(x_i)}{z_t} \begin{cases} e^{-\alpha_t} \; si \;\; y_i = h_t(x_i) & x \; bien \; classé \\ e^{\alpha_t} \; si \;\; y_i \neq h_t(x_i) & x \; mal \; classé \end{cases}$$

z_t: facteur denormalisation

Combiner les N classifieurs par vote majoritaire, utilisant les poids α_t.

Algorithme 3: Pseudo code de l'algorithme Adaboost

3.2 Algorithme de base proposé

Nous proposons l'algorithme K-NNW (K plus proche voisin pondéré) comme un algorithme de base pour chaque itération de boosting, c'est un algorithme d'apprentissage simple, dans un contexte de classification d'une nouvelle observation x, l'idée fondatrice est de faire voter les plus proches voisins de cette observation. La classe de x est déterminée en fonction de la classe majoritaire parmi les k plus proches voisins de l'observation x. (voir chapitre 2)

Pour chaque exemple de la base de test x_i :

- *Trouver les k plus proches voisins de* x_i *utilisant la distance eucludienne.*
- *Calculer la valeur de la classe :*

$$\sum w_k * x_{j\,k}$$

- *Calculer la précision :*

Précision = (Nombre des exemples bien classés / nombre des exemples de test) x 100.

Algorithme 3: Pseudo code de l'algorithme K-NNW

4. Résultats obtenus et discussion

Dans ce travail, nous utilisons une approche ensembliste pour la reconnaissance automatique du diabète :

- L'Adaboost avec l'algorithme des K-plus proches voisins pondéré.

4.1 Paramètres d'évaluation

Ces deux classifieurs seront évalués en fonction de leur sensibilité, spécificité, taux de classification et matrice de confusion.

- **Sensibilité (Se)** : représente la probabilité que le test soit positif si la patiente est diabétique.

Sensibilité (%) = VP/ (VP+FN)*100

- **Spécificité (Sp)** : représente la probabilité que le test soit négatif si la patiente n'est pas diabétique.

Spécificité (%) = VN/ (VN+FP)*100

- **Taux de classification :**

$$TC = (VN+VP)/ (VN+FN+VP+FP)*100$$

Avec VP, VN, FP et FN représente respectivement :

- ✓ Vrai positif : un Diabétique classé Diabétique.
- ✓ Vrai négatif : un No Diabétique classé No Diabétique.
- ✓ Faux positif : un No Diabétique classé Diabétique.
- ✓ Faux négatif : un Diabétique classé No Diabétique.

- **Matrice de confusion** : elle contient des informations sur les classifications réelles et prédite par un système de classification. La matrice de confusion permet d'identifier les erreurs de classification.

	Prédite	
Réelle	**Négatif**	**Positif**
Négatif	**A**	**b**
Positif	**C**	**d**

Tableau 3.2: Matrice de confusion

4.2 Les performances obtenues du classifieur

4.2.1 Implémentation du classificateur

Afin d'évaluer les performances du nouveau classifieur (Adaboost avec k-NNW), nous avons implémenté l'algorithme de classification K-NNW. Nous avons divisé la base Pima en deux, une première partie (3/4) pour l'apprentissage, et la deuxième (1/4) pour le test.

4.2.2 Paramètres de l'algorithme

a. l'algorithme K-NNW

Paramètre	Valeur
Valeur de K	3
Pas d'apprentissage	0.3

Tableau 3.3: Valeurs des paramètres de K-NNW

b. l'algorithme Adaboost

Le nombre des itérations T de boosting est fixé à 17 après plusieurs expérimentations, le graphe suivant représente la variation de taux de classification en fonction du nombre d'itérations :

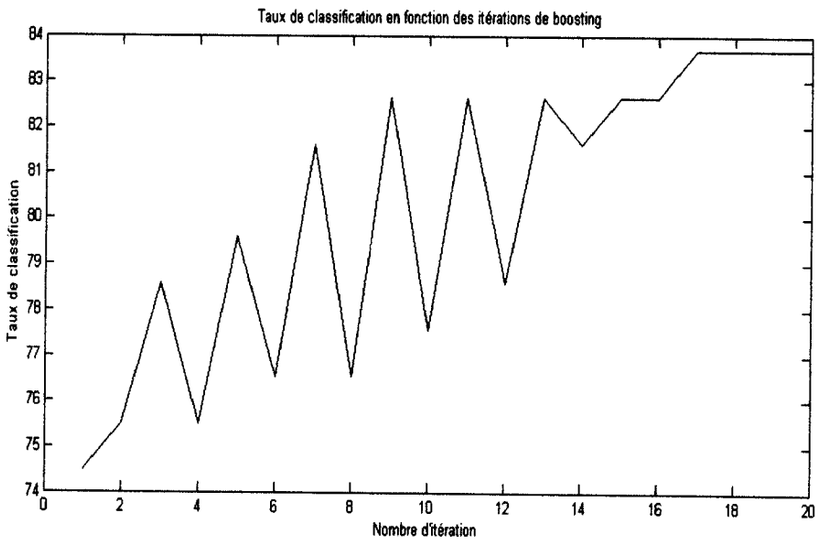


FIGURE 3.3: Le taux de classification en fonction des itérations de boosting

5. Résultats obtenus

Le tableau suivant (Tableau 5) représente la distribution des patients (diabétiques / non-diabétiques) selon les deux classifieurs K-NNW et Adaboost :

Réelle / Prédite	Diabétiques	Non-diabétiques	Méthode
diabétique	18	13	K-NNW
Non-diabétique	13	54	
Diabétique	**20**	**5**	**Adaboost(avec k-nnw)**
Non-diabétique	**11**	**62**	

Tableau 3.4 : Distribution des patients selon les deux classifieurs

5.1 Comparaison entre K-NNW et Adaboost

Le tableau suivant présente les performances de chaqu'un des classifieurs K-NNW et Adaboost selon les différents critères d'évaluations :

Critère\ Méthode	K-NNW	Adaboost	Différence
Taux de classification	73.46 %	83.67 %	+ 10.21 %
Exemples bien classés	72	82	+ 10
Exemples mal classés	26	16	+ 10
Sensibilité	58.06 %	80.00 %	+ 21.94
Spécificité	80.59 %	84.93 %	+ 4.34 %

Tableau 3.5: Performances de K-NNW et Adaboost

5.2 Comparaison des résultats avec les autres travaux

Le tableau suivant présente les taux de classification obtenu par différents classifieurs implémentés dans la littérature utilisant la base Pima Indian Diabetes.

Auteur	Méthode	Taux de classification
Purnami & al.(2009)	smooth SVM	93.20 %
Saidi & al.(2011)	MAIRS2	89.10 %
Polat & al.(2007)	AIRS with fuzzy resource allocation mechanism	84.42 %
Notre travail	**Adaboost (K-NNW)**	**83.67 %**
Aibinu & al.(2010)	CVNN-based CAR	81.00 %
Aibinu & al.(2010)	RVNN- based AR	80.65 %

Tableau 3.6: Comparaison des résultats obtenus avec l'état de l'art

Nous remarquons que le taux de classification obtenu par notre méthode est parmi les meilleurs résultats obtenus jusqu'à maintenant pour la classification du diabète avec un taux de classification de 83.67 %. Notre objectif n'est pas de présenter une méthode qui surpasse les algorithmes de classification déjà existants et qui ont démontré leur puissances tels que ceux cités dans l'état de l'art mais de présenter une méthode qui a pour but d'améliorer les classifieurs faibles.

6. Conclusion

Dans ce chapitre, nous avons évalué les performances de l'Adaboost pour la classification du diabète sur la base Pima Indian Diabetes. Nous avons mené nos expérimentations pour le classifieur k-PPVP qui est un algorithme stable, il a présenté une amélioration de 10.21% après l'application du boosting, nous pouvons conclure que l'utilisation de la méthode ensembliste Adaboost pour la classification du diabète est effective quoique le choix du classifieur de base peut fortement influencer les résultats obtenus.

Conclusion et perspectives

Conclusion et perspectives

Le diabète est considéré actuellement comme la maladie du siècle vu le nombre de diabétiques qui ne cesse d'augmenter. Selon une enquête menée par l'institut national de santé publique, le diabète se situe dans la quatrième place des maladies chroniques non transmissibles en Algerie. La prévalence du diabète de type 2 dans l'est et l'ouest du pays varie entre 6.4 % et 8.2 % chez des patients allant de 30 à 64 ans. Le diabète est l'une des maladies les plus répondu au monde avec plus de 220 millions de personnes diabétiques. L'augmentation du nombre de diabétiques est tellement rapide que l'organisation mondiale de la santé (OMS) l'a identifié comme étant une épidémie.

Beaucoup de travaux ont été mené afin d'effectuer la classification ou le diagnostic du diabète. Dans ce mémoire de Master, nous avons présenté la méthode de boosting pour le diagnostic du diabète basée sur la combinaison des classifieurs. L'algorithme de K-PPVP est utilisé comme un algorithme de base dans chaque itération de boosting.

Nous avons testé notre méthode sur la base Pima Indian Diabetes qui est une collection de rapports de diagnostic médicaux de 392 femmes âgées de plus de 21 ans. Les patientes de cette base sont des indiennes Pima, une population située près de Phoenix, Arizona, qui ont la plus haute prévalence du diabète dans le monde.

Le taux de classification obtenu avec notre méthode est compétitif par rapport aux autres travaux pour la classification du diabète, mais nous rappelons que notre objectif n'est pas de présenter une méthode qui surpasse les algorithmes de classification déjà existants et qui ont démontré leur puissances tels que ceux cités dans l'état de l'art mais de présenter une méthode qui a pour but d'améliorer les classifieurs faibles.

Les principales perspectives de recherche qui apparaissent à l'issue de ce travail sont de faire appel aux différents algorithmes faibles avec le boosting, l'adaptation de l'algorithme Adaboost aux problèmes multi-classes. Aussi nous souhaitons trouver une solution efficace pour traiter les données manquantes de la base de données médicales d'une manière générale.

Glossaire

Glossaire

ADO : Anti-Diabétiques Oraux.

DM1: Diabète de type 1.

DM2: Diabète de type 2.

Epai : Epaisseur de la peau au niveau du triceps.

HPO : Hyperglycémie Provoquée par voie Orale.

IMC : Indice de Masse Corporelle.

K-NN: K Nearest Neighbor.

K-NNW: K Weighted Nearest Neighbor.

K-PPV : K Plus Porches Voisins.

MAI : Maladies Auto-Immunes.

OMS: Organisation Mondiale de Santé.

PAD : Pression Artérielle Diastolique.

PED : Fonction Pédigrée du Diabète.

PID : Pima Indian Diabetes.

UCI : University California Irvine.

Bibliographie

Bibliographie

[ACD01] Diabète Québec, Association canadienne du diabète, Association médicale Canadienne, American Diabetes Association. Mars 2001.

[ALW11] Ala Alwan. Aperçu régional. Technical report, Fédération internationale du diabète, http ://www.diabetesatlas.org/, Access Mars 2011.

[AZ02] A. Lazarevic and Z. Obradovic, "Boosting algorithms for parallel and distributed learning," Distributed and Parallel Databases : An International Journal, Special Issue on Parallel and Distributed Data Mining, vol. 2, pp. 203–229, 2002.

[CCD12] Canoe cause diabète http://sante.canoe.ca/condition_info_details.asp ?disease_id=244

[DEC12] le diabète en chiffres. http ://www.lediabète.net

[DHW92] David H. Wolpert: Stacked generalization. Neural Networks, 5(2) :241-259 (1992).

[EMD10] Eve MATHIEU-DUPAS. Algorithme des K plus proches voisins pondérés (WKNN) et Application en diagnostic. Publié dans "42èmes Journées de Statistique 2010.

[JQU90] J. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[JRQ96] J. Ross Quinlan: Bagging, boosting, and C4.5. In Proceedings of the Thirteenth National Conference on Artificial Intelligence, AAAI 96, AAAI Press, pp 725 730, Portland, Oregon (August 1996).

[KS07] K. Polat and S. Gunes, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," Digital Signal Processing, pp. 702–710, 2007.

[LBB96] Leo Breiman: Bagging predictors. Machine Learning, 24:123–140 (1996).

[LBR01] Leo Breiman : Random forest. Machine Learning, 45:5–32 (2001).

[LBR96] Leo Breiman: Arcing Classifiers. technical report. Dept. of Statistics, University of California, Berkeley (1996).

[MAA10] M. Salami, A. Shafie, and A. Aibinu, "Application of modeling techniques to diabetes diagnosis," in IEEE EMBS Conference on Biomedical Engineering & Sciences, 2010.

[MM10] M. Ganji and M. Abadeh, "Using fuzzy ant colony optimization for diagnosis of diabetes disease," IEEE, 2010.

[MNT12] Medical News Today. [Online] http:www.medicalnewstoday.com/info/diabetes/

[OMS11] OMS Diabète. Diabète. Technical report, Aide-mémoire No.312, Janvier 2011.

[OMS12] Organisation mondiale de la santé (OMS) http://www.who.int/fr/

[PIH] The Pima Indian pathfinders for Health. The Pima Indian pathfinders for Health. [Online]. http://diabetes.niddk.nih.gov/dm/pubs/pima/index.htm

[RES90] Robert E. Shapire: The strength of weak learnability. Machine Learning, 5(2) : 197- 227. (1990).

[RPW02] Robert P.W. Duin: The Combining Classifier: to Train or Not to Train?. Proceedings 16th International Conference on Pattern Recognition, 2: 765- 770 (2002).

[RSC90] R. Schapire. The strength of weak learnability. Machine Learning journal, 5(2):197{227,1990.

[SAJ09] S. Purnami, A. Embong, J. Zain, and S. Rahayu, "A new smooth support vector machine and its applications in diabetes disease diagnosis," Journal of Computer Science, pp. 1003–1008, 2009.

[SKH05] S. Sahan, K. Polat, H. Kodaz, and S. Gunes, "The medical applications of attribute weighted artificial immune system (awais) : Diagnosis of heart and diabetes diseases," in ICARIS, 2005, pp. 456 – 468.

[TRJ01] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning. Data mining, inference, and prediction. Springer Verlag, 2001.

[TA10] T. Jayalakshmi and A. Santhakumaran, "A novel classification method for diagnosis of diabetes mellitus using artificial neural networks," in International Conference on Data Storage and Data Engineering, 2010.

[TDM08] T. Exarchos, D. Fotiadis, and M. Tsipouras, "Automated creation of transparent fuzzy models based on decision trees - application to diabetes diagnosis," in 30th Annual International IEEE EMBS Conference, 2008.

[TGD00] T. G. Dietterich: Ensemble methods in machine learning. In J. Kittler and F. Roli, editors, First International Workshop on Multiple Classifier Systems, pp 1-15 (2000).

[YFE96] Yoav Freund and Robert E. Schapire: Experiments with a new boosting algorithm. In Lorenza Saitta, editor, Machine Learning: Proceedings of the

Thirteenth International Conference, pp 148–156, Morgan Kaufmann, Bari, Italy (July, 1996).

[YFR95] Yoav Freud: Boosting a weak learning algorithm by majority. Information and Computation, 121(2) : 256-258 (Septembre 1995).

[YFS99] Y. Freund and R. Schapire, "A short introduction to boosting," Journal of Japanese Society for Artificial Intelligence, vol. 14(5), pp. 771–780, 1999.

[YK02] Yongdai Kim: Convex hull ensemble machine. In Proceedings of 2002 IEEE International Conference on Data Mining (ICDM 2002), IEEE Computer Society, pp 243 – 249, Maebashi City, Japan (2002).

[YRS96] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in the Thirteenth International Conference In Machine Learning, 1996.

[YFR99] Y. Freund. An adaptive version of the boost by majority algorithm. In Proc.of the 12th Annual Conf. on Computational Learning Theory, COLT'99, pages 102{113. Morgan Kaufmann, 1999.