Weka_S2D : Système de Surveillance de Diabétiques

TALEB ZOUGGAR Souad(*), **ATMANI Baghdad**(*)
Souad.taleb@gmail.com, atmani.baghdad@univ-oran.dz
(*) Equipe de recherche « Simulation, Intégration et fouille de données » (SIF)
Laboratoire d'Informatique d'Oran (LIO), Université d'Oran
BP : 1524 El M'Naouer, 31000 Oran, Algérie

Mots clefs :
Veille scientifique et technologique, Apprentissage automatique, Classification, Arbres de décision, Surveillance, Diabète
Keywords:
Scientific and technical observation, Machine learning, Classification, Decision trees, Monitoring, Diabetes
Palabras clave :
Escudriñar científico y tecnológico, Aprendizaje automático, Clasificación, árboles de decisión, Seguimiento, Diabetes

Résumé

Dans la communauté apprentissage il existe un grand nombre de méthodes à base d'arbres de décision, ce sont des méthodes d'apprentissage inductif dédiées à la classification. Dans ce papier nous proposons le logiciel Weka_S2D qui est basé essentiellement sur le logiciel de datamining Weka et intègre l'algorithme ID3_improving (Induction d'Arbres de Décision par une nouvelle Mesure de Distance) qui permet de faire de l'induction d'arbres de décision. L'algorithme ID3_improving utilise le même principe de construction d'arbre que la méthode ID3, il permet de construire l'arbre par des segmentations successives jusqu'à obtenir la partition la plus fine. Son originalité réside dans la simplicité de la mesure utilisée pour le calcul de l'importance d'une variable ce qui permet de réduire la complexité de calcul.
Weka_S2D s'adresse à deux types de publics. D'un côté, il exploite l'environnement de fouille de données offert par la plateforme Weka qui le rend ainsi accessible à une utilisation de type «chargé d'études» sur des données réelles. De l'autre, du fait que les règles générées sont intelligibles, il se prête à une utilisation directe pour l'aide à la décision.

1 Introduction

Les méthodes de surveillances actuelles du diabète ne répondant pas aux besoins des patients[1] malgré que le cette surveillance se fait de différentes façons, il nécessite généralement des visites médicales semestrielles, plusieurs tests de glycémie chaque jour, et la communication insuffisante entre les docteurs et les patients ce qui peut induire du stress chez les patients avec un effet potentiellement négatif sur l'efficacité du traitement.

Le diabète de type 1 se diagnostique en général plus facilement car son début est brutal. Perte de poids soudaine, fatigue, les symptômes sont très parlants et incitent à faire des tests. Par contre le type 2 apparaît de façon sournoise, et est souvent diagnostiqué à l'occasion d'une complication du diabète lui-même. Il y a donc fréquemment un diagnostic tardif[2] !

D'après une étude récente du moi de Mars 2010 du réseau de diabète algérien DIABCARE qui ciblera tous les patients présentant un *Diabète* de *Type 1* ou de *Type 2,* en Algérie il existe 2 500 000 de Diabétiques dont 86% sont des type 1 et 14% de type 2.

Pour la conception des systèmes d'aide à la décision, la classification par apprentissage automatique, est souvent utilisée. Plusieurs méthodes ont été misent au point pour résoudre ce problème [3][4][5]; parmi lesquelles, on retrouve les méthodes statistiques, à base de réseaux de neurones, arbres de décision ou graphes d'induction [1][2][7]. Pour réaliser la tache de surveillance des diabétiques nous proposons dans cet article d'utiliser le principe des méthodes à base d'arbres de décision qui sont des méthodes symboliques de l'apprentissage inductif très utilisées dans le domaine de classification reconnue par leur propriété d'intelligibilité, l'efficacité des algorithmes qu'elles présentent et l'exactitude et la précision des résultats fournis.

Nous proposons dans ce papier le logiciel Weka_S2D et en particulier l'algorithme ID3_improving pour la génération d'arbres de décision, ce nouvel algorithme utilise le même principe que la méthode à base d'arbres de décision ID3 [3] ; le partitionnement de l'échantillon jusqu'à obtenir la partition la plus fine ou la plus homogène. La différence réside dans la mesure de qualité de données utilisée qui est de complexité réduite pour ID3_improving et fournit un modèle de prédiction avec des performances égales ou dépassant parfois celles des méthodes arborescentes existantes.

Le papier est organisé de la manière suivante. Dans la section 2 on présente le logiciel Weka_S2D, son fonctionnement avec une illustration de l'environnement et l'application sous forme d'une démonstration logicielle. Pour terminer dans la section 3 nous concluons ce papier et présentons les perspectives de recherche et développement associées.

2 Le logiciel Weka_S2D

WEKA_S2D a été développé en Java (Eclipse 3.5.1) dans une architecture préétablie, plateforme de fouille de données WEKA [6] dans sa version 3.7. La plateforme Weka contient un nombre important de méthodes de fouille divisées en plusieurs familles.

Weka_S2D s'adresse à deux types de publics. D'un côté, il exploite l'environnement de fouille de données offert par la plateforme Weka qui le rend ainsi accessible à une utilisation de type «chargé d'études» sur des données réelles. De l'autre, du fait que les règles générées sont compréhensibles, il se prête à une utilisation directe par les spécialistes du domaine pour améliorer la prise de décision.

Dans la figure 1, nous présentons l'architecture fonctionnelle de WEKA_S2D: qui comportent les étapes traditionnelles de prétraitement, traitement et validation du modèle.

Comme toute méthode de classification ID3_improving permet de générer ou induire un modèle représentant des connaissances générales à partir de données élémentaires représentées dans un tableau appelé échantillon d'apprentissage.

---

[1] http://www.epinex.com/FR/the-diabetes-problem.php

[2] http://www.vivolta.com/diabete/diabete-diabetique-diagnostic-maladie-20091223448156.html

Notre échantillon d'apprentissage [2], pour lequel un extrait est présenté sur le tableau 1, est une base réelle composée d'un ensemble de 1461 patients diabétiques et de 10 descripteurs (variables exogènes), pour chaque individu (patient) de la base il s'agit de savoir s'il est classifié comme insulinodépendant ou diabétique de type 1 ou non insulinodépendant ou diabétique de type 2. Les descripteurs sur lesquels on se base pour faire cette classification sont : l'âge qui est une variable discrétisée soit entre 15 et 30 ans ou supérieure à 35 ans elle représente l'âge de découverte du diabète chez le patient, le mode de révélation qui détermine la façon dont le diabète s'est déclaré chez un patient et il peut être Cétose Diabétique Spontané, Cétose Diabétique avec Foyer Infectieux, Déséquilibre Glycémique ou Diabète Découverte Récente, nous avons aussi le poids du patient qui peut être normal, obèse, surcharge pondérale ou amaigrissement, l'attribut infection virale prend l'une des valeurs oui ou non, l'attribut statut qui détermine s'il y a un amaigrissement ou non, l'attribut association qui détermine que le diabète est en relation avec des maladies auto-immunes ou pas, l'attribut circonstance de découverte Détermine dans quelles circonstances le diabète s'est déclaré chez le sujet et pour chaque individu peut prendre l'une des valeurs : Pieds Diabétique, Découverte Fortuite, Infection Bactérienne, Rétinopathie, Comas Hyper – Osmolaire, Cétose Diabétique Inaugurale ou Comas Cétosique, l'attribut Asthénie qui détermine l'existence d'une asthénie ou pas, l'attribut antécédent qui prend les valeurs antécédent familial, personnel ou pas d'antécédents, et enfin l'attribut sexe.

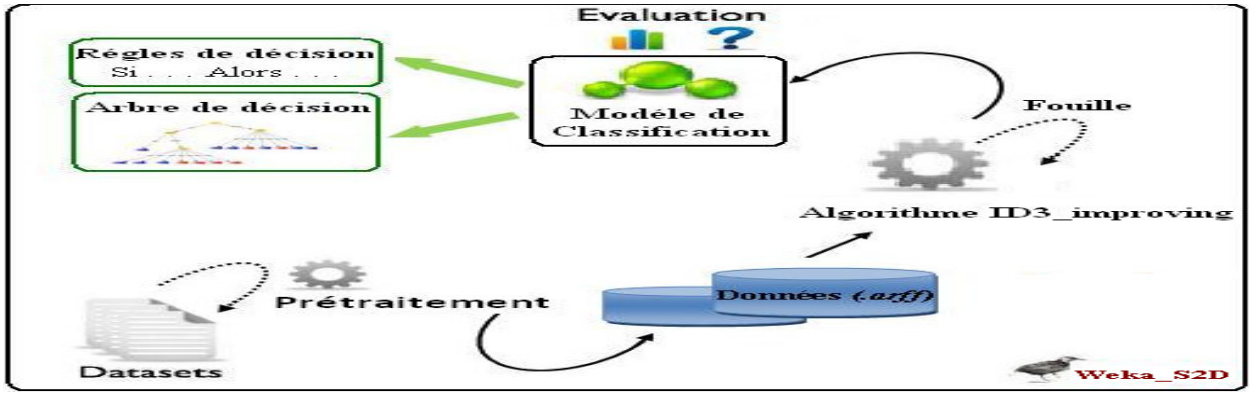

Figure 1 : Architecture fonctionnelle de Weka_S2D.

Tableau 1 : Extrait de la base Diabète

Class Nominal	Age Nominal	Revelation_mode Nominal	Poinds Nominal	Infection_Virale Nominal
Type_2	>35	G	N	Y
Type_2	other	G	N	Y
Type_2	>35	G	N	Y
Type_1	15-30	I	Ov	N
Type_1	>35	I	Ov	Y
Type_1	>35	G	Ov	N
Type_2	>35	G	N	N

Les données sont chargées sous forme d'un tableau Variables/Classe, il est exigé d'utiliser des données au format .arff (Attribute Relation File Format), les étapes de chargement sont les suivantes illustrées sur la figure 2:

1- Cliquer sur le bouton « Open file »,

2- Choisir la base d'apprentissage,

3- Cliquer sur le bouton ouvrir du jfilechooser.

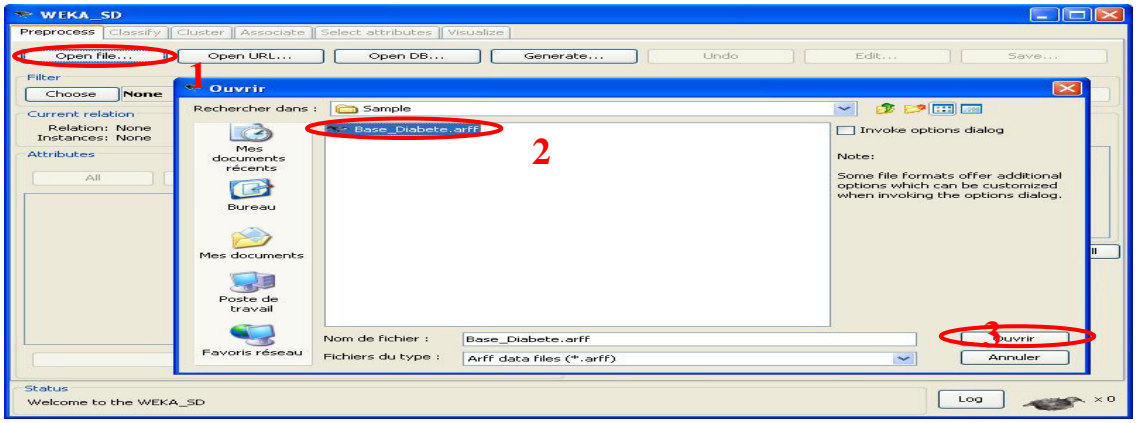

Figure 2 : Ouverture du fichier Diabète.arff.

Dans la figure 3 nous présentons les étapes à suivre pour construire le modèle de classification des individus, on commence d'abord par à l'onglet Classify, pour pouvoir choisir la méthode de classification parmi plusieurs familles de méthodes d'apprentissage supervisé ou non-supervisé, dans notre cas la méthode ID3_improving, que nous avons intégré dans Weka, se trouve dans le sous package trees.

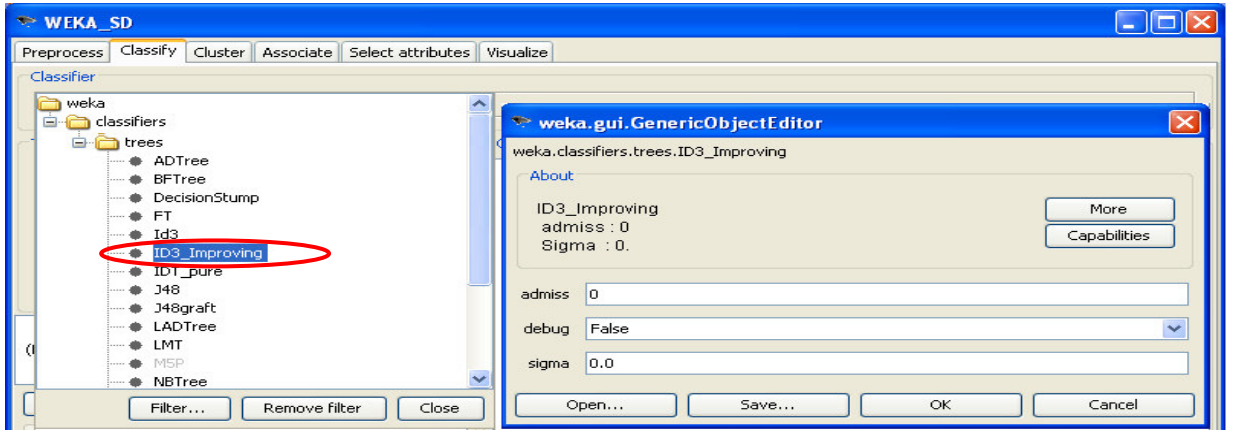

Figure 3 : Choix de la méthode de classement et lancement de l'apprentissage.

Après le choix de notre méthode de classification, il reste maintenant à lancer la classification en cliquant sur le bouton START qui se trouve dans le même onglet comme illustré sur la figure 4. Les résultats de classification apparaissent dans l'éditeur à droite et à la fin de l'apprentissage, on peut visualiser arbre et règles de décision comme présenté dans la figure 5.

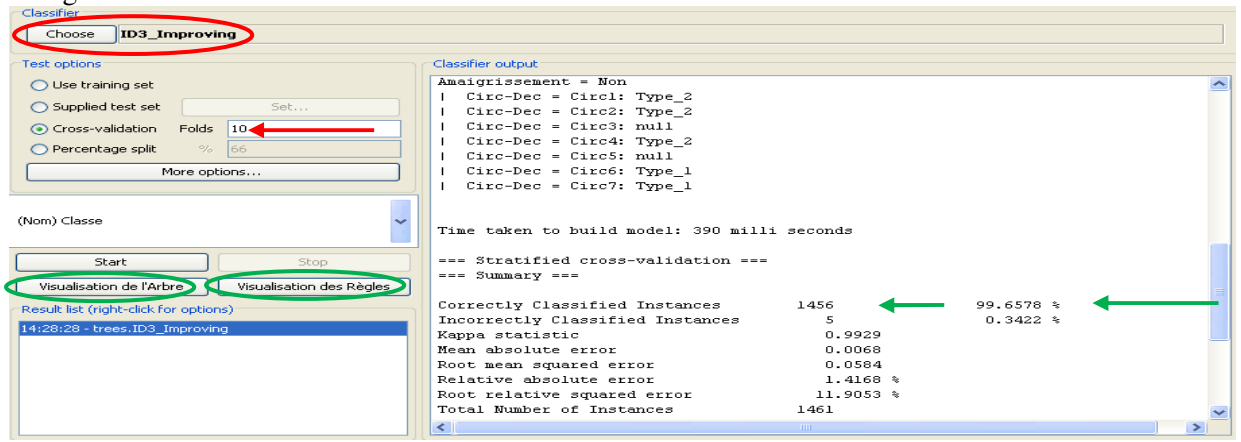

Figure 4 : Résultats de validation du modèle construit.

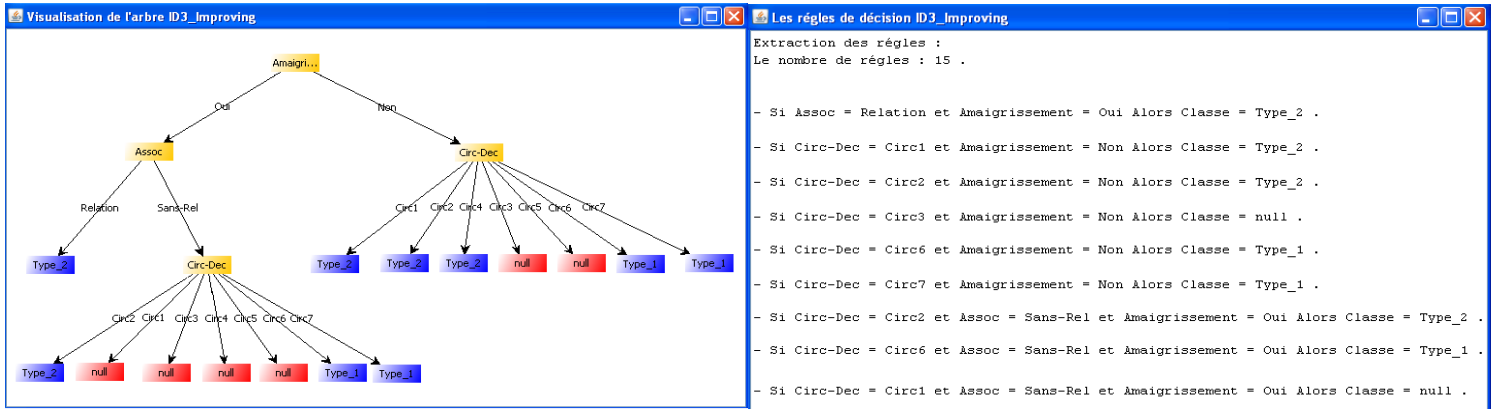

Figure 5 : Visualisation de l'arbre et des règles de décision ID3_improving.

Les règles sont validées par l'algorithme ID3_Improving mais cette base de règles sera mise à la disposition de l'expert médecin pour la validation finale. On remarque que le diabète de type 1 « diabète insulino-dépendant » est en relation directe avec les circonstances de découverte 6 et 7, par contre circonstance 2 donne le diabète de type2 « diabète non insulinodépendant ».
L'existence de l'amaigrissement chez un patient et association du diabète avec d'autres maladies auto-immunes impliquent un diabète de type2.

3 Conclusion

Une surveillance régulière des diabétiques optimise le traitement et permet de réagir efficacement en cas de déséquilibre. Très impliquant au quotidien pour le malade et sa famille, le diabète peut néanmoins s'avérer gratifiant car chaque point marqué contre la maladie est une garantie supplémentaire pour l'avenir.
Nous avons présenté dans ce papier un outil de surveillance des diabétiques qui est à base d'arbres de décision ID3_Improving qui utilise le même principe de partitionnement de la méthode ID3 mais avec une mesure de qualité de partition plus simple et plus économique en temps de calcul.
Comme perspective nous proposons d'utiliser les graphes d'inductions [7] qui sont une généralisation des arbres de décision où la prémisse des règles de décisions est sous forme de disjonctions de conjonction ce qui va nous aider à couvrir plusieurs cas spéciaux de diabète qui nécessitent plusieurs examens pour pouvoir identifier le type de diabète.

4 Bibliographie

[1] ATMANI B. et BELDJILALI B., *Neuro-IG: A Hybrid System for Selection and Elimination of Predictor Variables and non Relevant Individuals*. Informatica, Journal International, Vol. 18, No 2 163–186 (2007).

[2] ATMANI B. et BELDJILALI B., *Knowledge Discovery in Database: Induction Graph and Cellular Automaton*. Computing and Informatics Journal, Vol.26, No 2 171-197 (2007).

[3] QUINLAN J.R., Induction of Decision Trees, Machine Learning 1 81-106 (1986).

[4] RABASEDA S., RAKOTOMALALA R., Sebban M, *Génération automatique de connaissances par induction*. Actes des 3èmes rencontres de la société francophone de classification (1995) 45-46.

[5] RABASEDA S., RAKOTOMALALA R., ZIGHED D.A., *Rules extracted automatically by induction*. Proceeding of the 6th conference on information processing and management of uncertainty (1996) 551-556.

[6] WITTEN I.H. et FRANK E., *Data Mining: Practical Machine Learning Tools and Techniques (2nd edition).* Morgan Kaufmann (2005).

[7] ZIGHED D.A., AURAY J.P., DURU G., *SIPINA: Méthode et Logiciel.* Lacassagne (1992).