# Exploitation des données de la cohorte STANISLAS par des techniques de fouille de données numériques et symboliques utilisées seules ou en combinaison

Sandy Maumus[1-2], Amedeo Napoli[2], Laszlo Szathmary[2], Sophie Visvikis-Siest[1]
[1] INSERM U525, 30 rue Lionnois, 54000 Nancy, France, [2] LORIA – UMR 7503, B.P. 239, 54506 Vandoeuvre-Lès-Nancy, France

**Résumé**

La cohorte STANISLAS est une population de familles d'origine française supposées saines, recrutées au Centre de Médecine Préventive de Vandoeuvre-lès-Nancy et suivies tous les cinq ans sur une période de dix ans. Les données de la cohorte, de types numérique et textuel, représentent une richesse et un volume considérables, exploitées jusque là par des méthodes statistiques classiques. Nous proposons d'étudier ces données par des techniques de fouille de données numériques et symboliques, en nous intéressant plus exactement à l'étude du syndrome métabolique dans la cohorte STANISLAS, une affection correspondant à la présence simultanée chez un individu de plusieurs facteurs de risques cardiovasculaires. Nous présentons ici nos recherches en cours sur ce domaine, qui font intervenir l'utilisation de la boîte à outils Weka et de J-Close, une implémentation en Java d'un algorithme d'extraction de motifs fréquents et de règles d'association. Ultérieurement nous projetons le couplage d'un module de classification de Weka avec J-Close.

## 1. Contexte

La cohorte STANISLAS (Suivi Temporaire Annuel Non Invasif de la Santé des Lorrains Assurés Sociaux) est une étude familiale lancée en 1993 au centre de Médecine Préventive de Vandoeuvre-lès-Nancy, dont objectif principal est d'étudier le rôle et la contribution de facteurs génétiques et environnementaux sur les facteurs de risque cardiovasculaire [1]. C'est une étude longitudinale sur dix ans, où des familles de la Meurthe-et-Moselle et des Vosges ont été invitées par la Caisse primaire d'assurance maladie à venir passer un examen de santé tous les cinq ans. Lors du recrutement initial (1993-1995, $t_0$), les critères d'inclusion étaient les suivants : familles d'origine française, supposées saines, exemptes de maladies aiguës et/ou chroniques, composées de deux parents et de deux enfants de plus de six ans. 1006 familles (4295 sujets) ont ainsi pu être recrutées. Lors de la deuxième visite (1998-2000, $t_{+5}$), 75% des familles sont revenues. La troisième visite (2003-2005, $t_{+10}$) est en cours de réalisation.

Comme nous allons le voir ci-dessous, les données de la cohorte STANISLAS sont véritablement hétérogènes, de natures très différentes, et par conséquent méritent d'être qualifiées de données complexes. Les données recueillies dans la cohorte peuvent se diviser en trois catégories : cliniques et environnementales, biologiques, et génétiques. Les données cliniques se divisent en examens cliniques systématiques (mesures morphologiques telles que taille ou poids, mesure de la pression artérielle, …) et en examens cliniques sur projets spécifiques. Par ailleurs, des questionnaires sont remplis pour chaque individu et concernent entre autres les antécédents familiaux cardiovasculaires, les habitudes de vie et santé (consommation d'alcool et de tabac…). Concernant les données biologiques, des dosages

systématiques sont réalisés : (i) biochimie du sérum; (ii) numération de formule sanguine ; (iii) analyse d'urine. Par ailleurs des dosages spécifiques sur projets sont faits, entre autres : apolipoprotéine (apo) apo CIII, apo E, fibrinogène, insuline, vitamines. Enfin, pour chaque sujet, nous disposons de 116 SNPs (Single Nucleotide Polymorphisms ou polymorphismes génétiques) correspondant à tous les processus métaboliques impliqués dans les maladies cardiovasculaires : métabolisme lipidique, pression artérielle, coagulation, adhésion cellulaire, et l'inflammation.

Les données de la cohorte représentent une richesse et un volume important. Les techniques utilisées couramment pour exploiter ces données de la cohorte STANISLAS appartiennent aux méthodes statistiques (études d'association, méthodes de régression, ressemblance familiale). Cependant, face à la complexité de ces données, nous avons pensé que l'utilisation de techniques de fouille de données serait judicieuse, parce qu'elles apportent des points de vue nouveaux sur les données, et permettent peut-être plus facilement d'interpréter les éléments extraits des données en termes de connaissances.

Différents problèmes se posent. Ainsi, en ce qui concerne les données numériques, nous sommes confrontés à la question de la discrétisation des données, obligatoire pour l'utilisation de certains outils de fouille. Par ailleurs, certains individus présentent des valeurs manquantes pour les attributs de la base de données. Enfin, nous disposons pour chaque individu de données cliniques, environnementales et biologiques pour chaque visite, ce qui pose le problème de la variation des données dans le temps (les données génétiques ne varient pas avec le temps).

Dans ce papier, après une présentation brève des techniques de fouille de données choisies, nous présentons ce qu'est le syndrome métabolique, qui est la caractéristique cible de notre étude, puis les premiers résultats obtenus avec Weka. Enfin, nous concluons et donnons nos perspectives de travail.

## 2. Le logiciel Weka

Weka est une boîte à outils développée par l'Université de Waikato en Nouvelle-Zélande [2], téléchargeable gratuitement (http://www.cs.waikato.ac.nz/ml/weka) et dont les sources sont en accès libre et modifiables. C'est une collection de librairies de classes Java implémentant un ensemble d'algorithmes d'apprentissage. Weka possède des outils pour le pré-traitement des données, la classification, la régression, le clustering, l'extraction de règles d'association et la visualisation des données. Nous avons utilisé des outils de Weka pour le pré-traitement des données (filtres de discrétisation supervisée et non supervisée) et la classification (algorithmes OneR et J48 qui est la version Weka de C4.5).

## 3. Le syndrome métabolique (SM) dans la cohorte STANISLAS

Dans la cohorte STANISLAS, nous nous intéressons plus particulièrement à l'étude du syndrome métabolique (SM). Ce trouble prédispose au diabète de type 2 et aux maladies cardiovasculaires. C'est un ensemble de facteurs de risque cardiovasculaire regroupant l'insulinorésistance, la dyslipidémie, l'hypertension et l'obésité. Le SM, qui atteint 20 à 25% des individus aux Etats-Unis, touche aussi la France. Dans la cohorte STANISLAS, 8,4% des hommes et 6,4% des femmes en sont atteints [3]. Ces éléments, ajoutés à l'augmentation inquiétante du nombre de personnes touchés par l'obésité, y compris les enfants, font comprendre que le SM est devenu dans nos sociétés industrialisées un enjeu majeur de santé publique. Le but de nos expérimentations est de voir comment les techniques de fouille de données numériques et symboliques peuvent nous aider à mieux comprendre les mécanismes physiopathologiques du SM, en déterminant les facteurs influençant le SM, et avec quelle force.

## 4. Premiers résultats obtenus avec Weka pour l'étude du SM dans la cohorte STANISLAS

Les premières expérimentations ont été menées sur deux sous-populations de la cohorte STANISLAS : (i) la table « Donnéesbio », composée de 249 individus et contenant des données biologiques cliniques et environnementales décrites précédemment, nous a permis de faire une pré-étude afin de mieux appréhender les données de la cohorte avec les outils de Weka ; (ii) la table « Donnéescomplètes » composée de 1255 individus et qui contient les mêmes données biologiques, cliniques et environnementales que dans Donnéesbio, et aussi tous les médicaments et tous les polymorphismes génétiques déterminés sur les individus.

En plus des données qui sont disponibles, nous avons créé une nouvelle variable nominale dans la cohorte STANISLAS, « SM », qui décrit pour chaque individu s'il est atteint ou non par le syndrome métabolique. Pour cela, nous avons utilisé la définition du NCEP-ATP III qui déclare qu'un individu présente un SM si il possède au moins trois des cinq critères suivants : une hyperglycémie (glucose plasmatique à jeun ≥ 110 mg/dl), une hypertriglycéridémie (triglycérides ≥150 mg/dl), un HDL-cholestérol bas (c'est le "bon" cholestérol, <40 mg/dl chez les hommes ; <50 mg/dl chez les femmes), une hypertension (≥130/85 mmHg) et une obésité centrale (tour de taille>102 cm chez les hommes ; > 88 cm chez les femmes). Ainsi, dans les expérimentations que nous avons menées avec Weka, SM est la classe cible sur laquelle nous cherchons à classifier les individus.

Nous avons utilisé l'outil « Classify » de Weka. Dans un premier temps, nous avons testé sur « Donnéesbio » et sur « Donnéescomplètes » des filtres de discrétisation de données. Weka propose un mode de discrétisation supervisée (qui utilise la méthode MDL de Fayyad et Irani [4]) et un mode de discrétisation non supervisée (qui divise la variable en intervalles de mêmes fréquences) avec diverses options. L'une d'elles nous a particulièrement intéressée : Discretize –O, où l'utilisateur choisit un nombre d'intervalles maximum en fonction duquel les attributs seront discrétisés. Ensuite, nous avons utilisé l'arbre de décision J48 sur nos données, avec ou sans discrétisation préalable. J48 est l'implémentation de Weka de l'algorithme C4.5 [5]. Nous avons comparé les résultats obtenus selon que l'on avait choisi ou pas de discrétiser les données. Comme on pouvait s'y attendre, les arbres de décision générés diffèrent. Par exemple, si on utilise au préalable sur « Donnéesbio » un filtre de discrétisation non supervisée optimisée avec un nombre maximal d'intervalles égal à 4, alors l'arbre de décision généré classifiera les individus sur « SM » avec les attributs triglycérides et BMI (c'est l'indice de masse corporelle). Si ces mêmes données ne sont pas discrétisées, alors les individus seront classifiés avec les attributs BMI, triglycérides, cholestérol, pression artérielle systolique et âge (cf Figure 1).

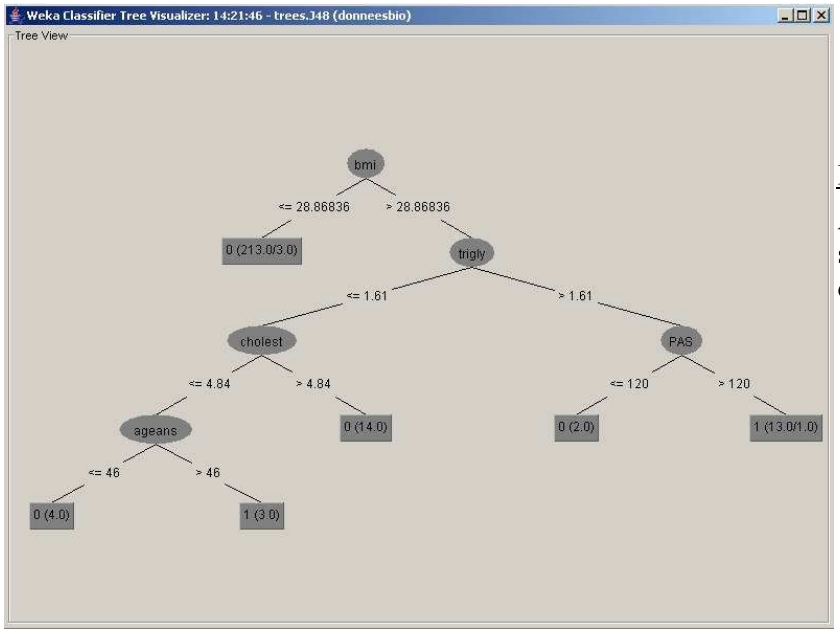



Figure 1 :

Arbre de décision généré sur « Donnéesbio » sans discrétisation préalable

Ces premières analyses sont encourageantes, car les résultats sont en accord avec les connaissances de l'expert. Ils ont été validés et interprétés par l'expert comme étant des points de vue intéressants des données.

## 5. Conclusions – Perspectives

Les premières analyses menées avec les méthodes numériques de Weka se sont avérées positives. J48 paraît être un bon outil de visualisation pour nos études. Pour la suite des expérimentations, nous projetons de répéter ces analyses en utilisant, pour définir l'attribut « SM » chez un individu d'autres définitions que celle du NCEP-ATPIII. Il sera alors intéressant de comparer les différents arbres obtenus. Nous espérons que ces différentes analyses contribueront à l'établissement d'une définition du SM mieux adaptée aux valeurs des attributs possédés par les individus de la cohorte STANISLAS.

Par ailleurs, J-Close est un logiciel développé par l'équipe Orpailleur du LORIA, qui inclut une implémentation en Java de l'algorithme Close [6], permettant l'extraction de motifs fermés fréquents et de règles d'association informatives. Ces règles d'association ont la particularité de posséder une prémisse qui correspond à un générateur minimal. Les premiers résultats qui ont été générés avec la méthode de fouille de données symboliques J-Close sont actuellement en cours de validation et d'interprétation par l'expert et sont prometteurs.

Une fois que les deux études avec Weka et J-Close auront été finalisées, une perspective intéressante de travail est d'évaluer le résultat de la combinaison de ces méthodes (outils de classification et d'extraction de règles d'association) pour faire émerger des profils d'intérêt avec les facteurs impliqués dans la physiopathologie cardiovasculaire et pour permettre une meilleure compréhension des mécanismes mis en jeu dans le SM.

## Références


[1] MANSOUR-CHEMALY, M., HADDY, N., SIEST, G. and VISVIKIS, S. (2002) : *Family studies: their role in the evaluation of genetic cardiovascular risk factors.* CCLM, 40, pp. 1085-96.

[2] WITTEN, I.H. and FRANK, E. (2000) : *Data Mining: Practical machine learning tools with Java implementations.* Morgan Kaufmann, San Francisco.

[3] MAUMUS, S., MARIE, B., SIEST, G.VISVIKIS-SIEST, S. *A Prospective Study on the Prevalence of Metabolic Syndrome (MS) among Healthy French Families. Two Cardiovascular Risk Factors (HDL-C and TNF-α) are revealed in MS Offspring.* (Diabetes Care, article accepté).

[4] FAYYAD, U.M. and IRANI, K.B. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proc. Thirteenth International Joint Conference on Artificial Intelligence*, Chambery, France, Morgan Kaufmann, San Francisco, pp. 1022-27.

[5] QUINLAN, J.R. (1993): C4.5: Programs for Machine Learning Morgan Kaufmann, San Mateo, CA.

[6] BASTIDE, Y., TAOUIL, R., PASQUIER, N., STUMME, G. and LAKHAL, L. (2002) : *Pascal: un algorithme d'extraction des motifs fréquents.* Technique et Science Informatiques, 21, pp. 65-96.