Le rôle de l'utilisateur dans un processus d'extraction de règles d'association

Cyril Nortet*, Ansaf Salleb**
Teddy Turmeaux*, Christel Vrain*

* LIFO Rue Léonard de Vinci BP 6759 45067 Orléans Cedex 02
{Cyril.Nortet, Teddy.Turmeaux, Christel.Vrain}@lifo.univ-orleans.fr

** INRIA Projet Dream, Campus Universitaire de Beaulieu, 35042 Rennes
Ansaf.Salleb@irisa.fr

Résumé. De nombreux travaux ont porté sur l'extraction de règles d'association. Cependant, cette tâche continue à intéresser les chercheurs en fouille de données car elle soulève encore plusieurs défis. En particulier, son utilisation en pratique reste difficile : d'une part, le nombre de règles apprises est souvent très grand, d'autre part, le traitement des valeurs numériques dans cette tâche est loin d'être maîtrisé. Nous nous intéressons dans cet article au rôle que peut jouer l'utilisateur pour pallier ces difficultés. Il s'agit d'impliquer l'utilisateur dans le processus de recherche de règles d'association qui est dans ce cas interactif et guidé par des schémas de règles qu'il aurait choisis. Nous illustrons notre propos avec QuantMiner qui est un outil convivial et interactif que nous avons développé. La présence de l'expert reste indispensable durant tout le processus d'extraction de règles.

Mots clé. Fouille de Données Interactive, Règle d'Association Quantitative, Optimisation.

1 Introduction

L'extraction de règles d'association est devenue aujourd'hui une tâche populaire en fouille de données. Elle a pour but de dégager des relations intelligibles entre des attributs dans une base de données. Une règle d'association (Agrawal et al. 1993) est une implication $C_1 \Rightarrow C_2$, ou C_1 et C_2 expriment des conditions sur les attributs de la base de données. La qualité d'une règle est classiquement évaluée par un couple de mesures support et confiance, définies par :

- Support(C), ou C exprime des conditions sur les attributs, est le nombre de n-uplets (lignes de la base de données) qui satisfont C.
- $\text{Support}(C_1 \Rightarrow C_2) = \text{Support}(C_1 \wedge C_2)$
- $\text{Confiance}(C_1 \Rightarrow C_2) = \text{Support}(C_1 \wedge C_2)/\text{Support}(C_1)$

Une règle d'association est dite solide, si son support et sa confiance dépassent deux seuils fixés *a priori*, MinSupp et MinConf respectivement. De nombreux travaux se sont intéressés au problème crucial de performance que pose cette tâche (par ex. (Brin et al. 1997,

Zaki 2000)) et des algorithmes de plus en plus performants sont proposés. Cependant d'autres problèmes persistent dont les deux suivants :

- **Problème 1** : le nombre de règles apprises est souvent très grand et décourageant pour l' expert qui voudrait les exploiter.
- **Problème 2** : les travaux existant ne gèrent pas bien, voire pas du tout, les valeurs numériques. Une pré-discrétisation (découpage du domaine de l'attribut numérique en intervalles) est souvent effectuée mais reste non satisfaisante. Ceci restreint sérieusement l'applicabilité de cette tâche aux données réelles.

Nous montrons dans cet article qu'il est possible de répondre à ces deux problèmes en impliquant l'utilisateur durant le processus de fouille de données. Celui-ci choisit des schémas de règles d'association qui l'intéressent (certains travaux ont montré l'intérêt de restreindre la forme des règles (Srikant et al. 1997)). Un algorithme génétique permettant de découvrir de façon dynamique les intervalles des variables numériques qui optimisent le support et la confiance de chaque schéma est ensuite employé.

2 QuantMiner

L'idée de QuantMiner (Nortet et al. 2005) est de considérer des **schémas de règles**. Un schéma de règle est une règle présentant dans ses membres gauche et droit des items catégoriques aux valeurs fixées ou non et des items numériques dont les intervalles correspondant ne sont pas encore instanciés. Puis par optimisation nous cherchons les bornes les plus adaptées pour chacun de ses intervalles, en prenant en compte la mesure du Gain proposée par (Fukuda et al. 1996) et donnée par :

$$Gain(A \Rightarrow B) = Supp(AB) - MinConf * Supp(A)$$

Seules les règles aux meilleurs Gains sont gardées. Le système prend la forme d'un assistant (wizard), limité à 5 étapes. Nous illustrons les étapes de QuantMiner à travers une application réelle portant sur la maladie de l'athérosclérose (Salleb et al. 2004, Nortet et al. 2005). L'athérosclérose est une maladie répandue et grave des artères dont les parois se durcissent provoquant une gêne considérable à la circulation du sang. Le projet STULONG[1] porte sur une étude médicale effectuée pendant 20 ans sur les facteurs de risque de cette maladie et concerne une population de patients composée de plus de 1 400 hommes adultes qui ont été classés en trois groupes : le groupe des patients normaux N, le groupe des patients à risque R et enfin le groupe P des patients présentant la pathologie. Nous nous sommes intéressés à l'extraction de règles d'association dans un but descriptif comme par exemple, décrire les patients décédés et les patients non décédés. Au total 27 attributs catégoriques et 17 numériques décrivent les patients.

[1] « The study (STULONG) was realized at the 2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital, Unemocnice 2, Prague 2 (head. Prof. M. Aschermann, MD, SDr, FESC), under the supervision of Prof. F. Boudik, MD, ScD, with collaboration of M. Tomeckova, MD, PhD and Ass. Prof. J. Bultas, MD, PhD. The data were transferred to the electronic form by the European Centre of Medical Informatics, Statistics and Epidemiology of Charles University and Academy of Sciences (head. Prof. RNDr. J. Zvarova, DrSc). The data resource is on the web pages {http://euromise.vse.cz/STULONG}. At present time the data analysis is supported by the grant of the Ministry of Education CR Nr LN 00B 107. »

2.1 Étapes 1 et 2 : Choix de attributs puis choix des schémas

Après une première étape consistant à choisir les attributs, la seconde étape permet une répartition fine des attributs ($A_i = v_i$ ou $A_i \in [l_i,u_i]$, les valeurs l_i, u_i non instanciées, v_i instanciée ou non) aux places où on souhaite les voir apparaître dans les règles. Chacun peut être indépendamment placé à gauche de la règle (condition), à droite (objectif), ne pas y apparaître, ou apparaître impérativement dans la règle. Par exemple, si l'on souhaite travailler sur des règles portant sur l'influence de la consommation de tabac sur le décès de patients, il suffira d'indiquer que tous les attributs n'apparaissent nulle part, sauf ``DEATH ?" à gauche et ``TOBA_CONSO" et "TOBA_DURA" à droite, comme indiqué dans la figure ci-dessous.

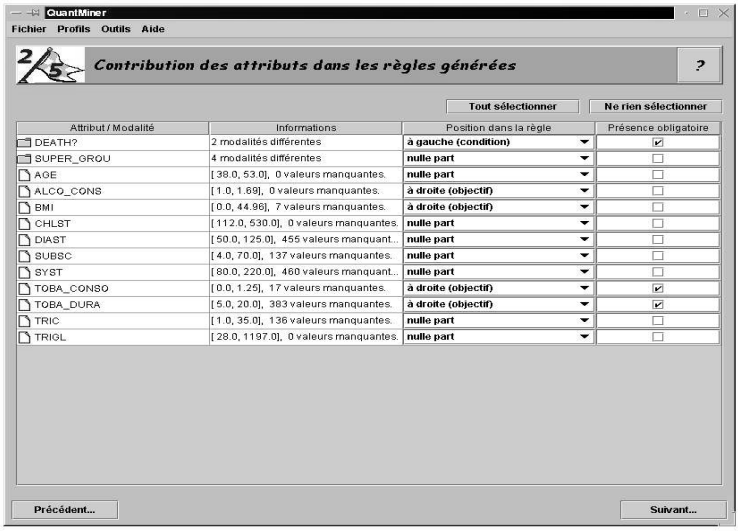

FIG. 1 – *Étape 2 de QuantMiner*.

2.2 Étapes 3 : Choix de la méthode

Il s'agit de choisir une technique d'optimisation et régler ses paramètres. En plus du support minimum et de la confiance minimum, des paramètres spécifiques à la technique d'optimisation sont fixés par l'utilisateur. Par exemple, dans le cas de l'algorithme génétique employé ici, se sont la taille de la population, le nombre de générations, les taux de mutations et de croisements.

2.3 Étapes 4 et 5 : Exécution puis visualisation des règles

L'utilisateur lance l'exécution de l'algorithme d'optimisation choisi. Les règles apprises sont affichées au fur et à mesure du calcul. Dans la dernière étape, un affichage détaillé des

règles est présenté à l'utilisateur. Pour rendre les règles plus exploitables, de nombreux paramètres statistiques les accompagnent dans l'affichage. Pour les attributs numériques, la proportion de l'intervalle dans la règle par rapport au domaine de l'attribut est présentée pour montrer la pertinence de l'intervalle optimisé.

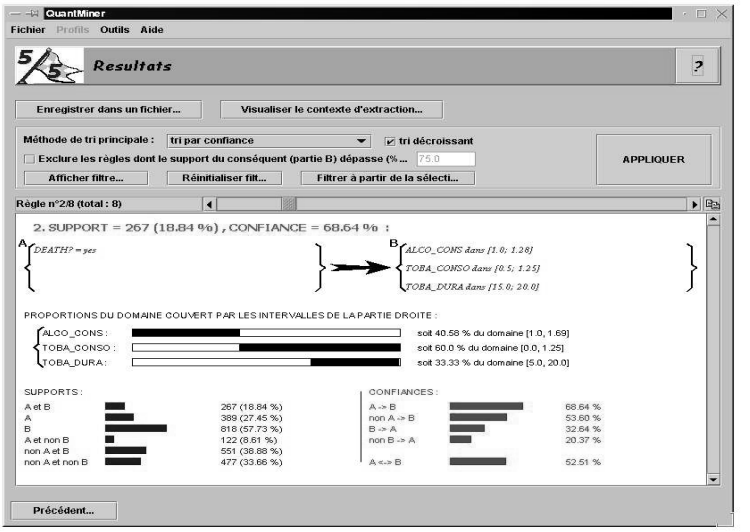

FIG. 2 – *Étape 5 de QuantMiner.*

3 Discussion et Conclusion

Quelles leçons pouvons nous tirer de l'expérience de QuantMiner ?

- Un processus d'extraction doit être aussi simple que possible avec un nombre d'étapes d'extraction et un nombre de paramètres limités.
- Une visualisation conviviale des résultats d'extraction est primordiale pour permettre à l'utilisateur de manipuler et identifier les règles intéressantes.
- Tel que nous l'avons vu, QuantMiner ne se contente pas de donner la confiance d'une règle de la forme A⇒B, mais il donne aussi la confiance de ¬A⇒B, B⇒A, ¬B⇒ A. Cependant, lors de l'application aux données médicales, il nous a semblé que ces critères devaient être affinés dans le cas de la caractérisation de plusieurs classes. Dans le cas où A représente des caractéristiques des patients à risques (exemple ci-dessous), ¬A regroupe les patients bien portants et malades; il serait plus intéressant de séparer ces deux classes.

GROOUP=R ⇒ ALCO_CONS ∈[1.0,1.29] BMI∈[22.28,30.72]
TOBA_CONSO ∈ 0.5,1.25]
supp(A⇒B)=39%, conf(A⇒B)=64%, conf(¬A⇒B)=38%

- Souvent l'expert aimerait bien découvrir dans ses données une connaissance surprenante, *exceptionnelle*, alors que les règles d'association fréquentes risquent d'être déjà connues.

Par conséquent, la recherche de règles surprenantes s'oriente vers des associations ayant des supports relativement faibles mais des confiances fortes. Baisser le support n'est pas une solution car survient alors le problème du grand nombre de règles générées. Nous nous intéressons à cette problématique (Duval et al. 2004) qui nous semble prometteuse dans la voie d'extraction de connaissances de qualité.

– On peut enfin se demander jusqu'à quel point impliquer l'expert dans un processus de fouille de données et l'impact que cela entraîne sur l'intérêt des connaissances apprises. Un compromis reste sans doute à trouver en fonction des applications et du besoin de l'utilisateur final.

Références

Agrawal R., Imielinski T. et Swami A. N. (1993). Mining association rules between sets of items in large databases. Dans Buneman P. et Jajodia S., éditeurs, Proceedings of the 1993 ACM SIGMOD, pp 207-216,Washington, D.C.

Brin S., Motwani R. et Silverstein C. (1997). Beyond market baskets: generalizing association rules to correlations. Dans Proc. of ACM SIGMOD, pp 265-276.

Duval B., Salleb A. et Vrain C. (2004). Méthodes et mesures d'intérêt pour l'extraction de règles d'exception. Revue des Nouvelles Technologies de l'Information - Mesures de Qualité pour la Fouille de Données RNTI-E-1, pp 119-140.

Fukuda T., Morimoto Y., Morishita S. et Tokuyama T. (1996). Data mining using two dimensional optimized'association rules: Scheme, algorithms and visualization. Dans Proc. of the Int'l Conf. ACM SIGMOD, pp 12-23.

Nortet C., Salleb A., Turmeaux T. et Vrain C. (2005). Extraction de Règles d'Association Quantitatives - Application à des Données Médicales. Dans EGC 2005, volume II, pp 495-506. RNTI-Cépaduès éditions.

Salleb A., Turmeaux T., Vrain C. et Nortet C. (2004). Mining quantitative association rules in a atherosclerosis dataset. Dans Proceedings of the PKDD Discovery Challenge 2004 (co-located with the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases), pp 98-103, Pisa, Italy.

Srikant R., Vu Q. et Agrawal R. (1997).Mining association rules with item constraints. Dans 3rd KDD, pp 67-73. AAAI Press.

Zaki M. J. (2000). Scalable algorithms for association mining. IEEE Transactions on Knowledge and Data Engineering, 12(3):372-390.

Summary

We propose QuantMiner, a genetic-based system for mining quantitative association rules. Our algorithm starts with a set of rule templates chosen by the user and then looks dynamically for the ``best'' intervals for the numeric attributes present in these templates. An optimization criterion based on both support and confidence is used to keep only high quality and interesting rules. In QuantMiner, the user is highly solicited in order to guide the mining process thus avoiding the discovery of hundreds of rules, as it is usually the case in association rule mining.