ARQAT : plateforme exploratoire pour la qualité des règles d'association

Xuan-Hiep Huynh*, Fabrice Guillet*, Henri Briand*
*LINA CNRS FRE 2729 - Ecole polytechnique de l'université de Nantes
La Chantrerie, BP 50609, 44306 Nantes Cedex 3, France
{xuan-hiep.huynh, fabrice.guillet, henri.briand}@polytech.univ-nantes.fr

Résumé. Le choix de mesures d'intérêt pour la validation des règles d'association constitue un défi important dans le contexte de l'évaluation de la qualité en fouille de données. Mais, comme l'intérêt dépend à la fois de la structure des données et des buts de l'utilisateur (décideur, analyste), certaines mesures peuvent s'avérer pertinentes dans un contexte donné, et ne plus l'être dans un autre. Dans cet article, nous proposons un outil original ARQAT afin d'étudier le comportement spécifique de 34 mesures d'intérêt dans le contexte d'un jeu de règles, selon une approche résolument exploratoire mettant en avant l'interactivité et les représentations graphiques.

1 Introduction

L'étude et la conception de mesures d'intérêt (MI) adaptées aux règles d'association constitue un important défi pour l'évaluation de la qualité des connaissances en ECD. Les règles d'association (Agrawal et al. 1993) proposent un modèle non supervisé pour la découverte de tendances implicatives dans les données. Malheureusement, en phase de validation, l'utilisateur (expert des données, ou analyste) se trouve confronté à un problème majeur : une grande quantité de règles parmi lesquelles il doit isoler les meilleures en fonction de ses préférences. Une manière de réduire le coût cognitif de cette tâche consiste à le guider à l'aide de mesures d'intérêt adaptées à la fois à ses préférences et à la structure des données étudiées.

Les travaux précurseurs sur les règles d'association (Agrawal et al. 1993) (Agrawal et Srikant 1994) proposent l'utilisation de 2 mesures statistiques : le support et la confiance. Ce couple de mesures dispose de vertus algorithmiques accélératrices, mais n'est pas suffisant pour capter l'intérêt des règles. Afin de compenser cette limite, de nombreuses mesures complémentaires ont été proposées dans la littérature et dissociées en 2 groupes (Freitas 1999) : les mesures objectives et les mesures subjectives. Les mesures subjectives dépendent essentiellement des buts, connaissances, croyances de l'utilisateur qui doivent être préalablement recueillis. Elles sont associées à des algorithmes supervisés ad hoc (Padmanabhan et Tuzhilin 1998) (Liu et al. 1999) permettant de n'extraire que les règles conformes ou au contraire en contradiction avec les croyances de l'utilisateur, et ainsi d'orienter la notion d'intérêt vers la nouveauté (novelty) ou l'inattendu (unexpectedness). Les mesures objectives, quant à elles, sont des mesures statistiques s'appuyant sur la structure des données ou plus exactement la fréquence des combinaisons fréquentes d'attributs (itemsets). De nombreux travaux de synthèse récapitulent et comparent leurs définitions et leurs propriétés (Bayardo et Agrawal 1999) (Hilderman et Hamilton 2001) (Tan et al. 2002) (Tan et al. 2004)

(Piatetsky-Shapiro 1991) (Lenca et al. 2004) (Guillet 2004). Ces synthèses traitent deux problèmes fondamentaux et complémentaires afin d'aider l'utilisateur à repérer les meilleures règles : la caractérisation des principes sous-jacents à une "bonne" MI, et l'étude comparative de leur comportement sur des simulations et des jeux d'essai.

Dans cet article, nous présentons une nouvelle approche et une plateforme d'implémentation ARQAT (Association Rule Quality Analysis Tool) afin d'étudier le comportement spécifique des MI sur le jeu de données de l'utilisateur et selon une perspective d'analyse exploratoire.

Plus précisément, ARQAT est une boite à outil conçue pour aider graphiquement l'utilisateur analyste à repérer dans ses données les meilleures mesures et au final les meilleures règles.

2 Principes de la plateforme ARQAT

ARQAT inclut 34 mesures objectives issues des articles de synthèse précédents. Nous complétons cette liste avec 3 mesures : l'Intensité d'Implication (II) (Gras 1996) (Guillaume et al. 1998) sa version entropique (EII) (Gras et al. 2001) (Blanchard et al. 2003) et la mesure de taux informationnel modulé par la contraposée (TIC) (Blanchard et al. 2004).

ARQAT (Fig. 1) implémente 14 vues graphiques complémentaires qui sont structurées en 5 groupes selon la tâche offerte.

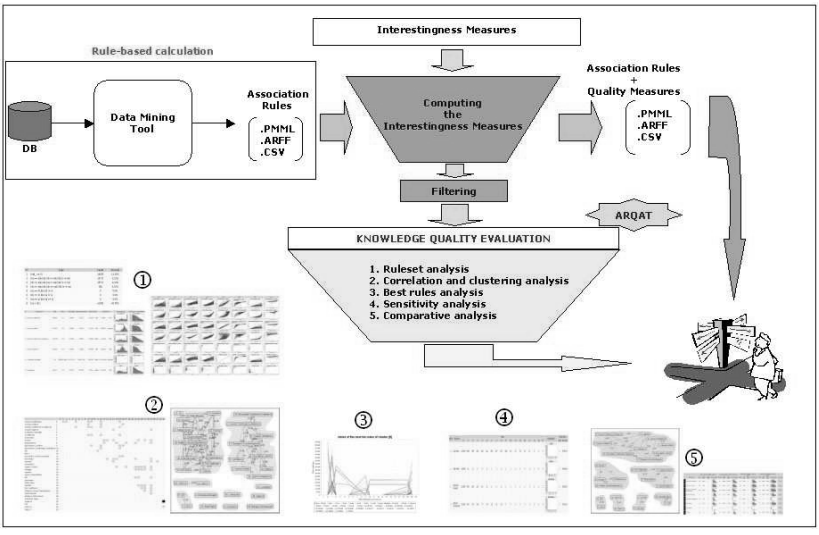

FIG. 1 – *Structure d'ARQAT.*

Les données d'entrée sont constituées d'un ensemble R de règles d'association extrait d'un jeu de données initial D, où la description de chaque règle $a \Rightarrow b$ est complétée

par ses contingences $(n, n_a, n_b, n_{a\bar{b}})$ dans D. Plus précisément, n est le nombre total d'enregistrements de D, n_a (resp. n_b) le nombre d'enregistrements de D satisfaisant a (resp. b), et $n_{a\bar{b}}$ le nombre d'enregistrements satisfaisant $a \wedge \overline{b}$ (les contre-exemples).

Dans une étape préliminaire, l'ensemble de règles R est traité afin de calculer les valeurs des MI pour chaque règle, puis les corrélations entre chaque paire de mesure. Les résultats sont stockés dans deux tables : la table des mesures (R×I) dont les lignes corespondent aux règles et les colonnes aux valeurs des mesures, et la matrice de corrélation (I×I) entre les mesures. Lors de cette étape, l'ensemble de règles R peut aussi être échantillonné afin de cibler l'étude sur un sous-ensemble de règles.

La seconde étape est ensuite interactive, l'utilisateur mène l'exploration graphique des résultats. Il s'appuie pour cela sur la structuration en 5 groupes de vues orientées tâche. Le premier groupe (1 dans Fig. 1) est dédié à la visualisation de statistiques élémentaires afin de mieux appréhender la structure de la table R×I. Le deuxième groupe (2) est orienté vers la visualisation de la table des corrélations entre mesures I×I et leur classification afin de repérer les meilleures mesures. Le troisième groupe (3) cible l'extraction des meilleures règles. Le quatrième groupe (4) permet une étude de la sélectivité des MI. Enfin, un dernier groupe offre la possibilité de mener une étude comparative des résultats obtenus sur plusieurs ensembles de règles.

Dans la suite de cet article, nous ne décrivons que la tâche d'analyse de corrélations et l'illustrons sur un même jeu de règles : 120000 règles d'association extraites par un algorithme Apriori (support 10%) de la base mushroom (Blake et Merz 1998).

3 Analyse de corrélation

Cette tâche permet l'analyse des corrélations (matrice I×I) entre mesures et leur partitionnement en groupes corrélés, afin d'orienter l'utilisateur vers les mesures les mieux adaptées à ses besoins spécifiques au jeu de règles étudié. Les valeurs de corrélation sont calculées à titre provisoire selon la formule du coefficient de corrélation linéaire de Pearson. Les résultats sont présentés sous deux formes graphiques. La première vue est une visualisation élémentaire de la matrice de corrélation sous la forme d'une *matrice de niveau de gris*, où chaque valeur de corrélation est codée par un niveau de gris.

La deuxième représentation envisagée, beaucoup plus expressive, est un *graphe de corrélation* (Fig. 2). Comme les graphes constituent un excellent outil d'investigation des structures complexes, nous les utilisons afin de représenter la matrice de corrélation sous la forme d'un graphe non-orienté et valué. Chaque sommet correspond à une MI, et une arête est associée à la valeur du coefficient de corrélation entre les deux sommets reliés. Nous y ajoutons une possibilité de seuillage par une valeur d'arête minimale τ (resp. maximale θ) afin de ne retenir que le sous graphe partiel CG+ (resp. CG0) des corrélations significativement élévées (resp. significativement faibles).

Ces deux sous-graphes partiels peuvent ensuite être traités afin d'être découpés en classes de mesures, chaque classe correspondant ici à une partie connexe maximale (nous orienterons ultérieurement vers un partitionnement en cliques). Dans CG+ chaque classe rassemble des mesures significativement corrélées proposant donc un point de vue proche sur les règles, alors que dans CG0 chaque groupe est révélateur de points de vue différents.

Ainsi, chaque jeu de règle produira un couple de graphes CG0 et CG+ différent, grâce auxquels l'utilisateur pourra observer rapidement la structure des MI, et valider graphiquement son choix des meilleurs indices. Par exemple, Fig. 2, CG+ fait apparaître 11 parties connexes qui peuvent aider à choisir une base réduite de 11 mesures, parmi les 34 utilisées, composée du meilleur représentant de chaque classe, afin de simplifier la validation des règles. Autre exemple, sur CG0 on voit une partie connexe composée des deux mesures Support and Yule's Y significativement non corrélées. Un dernier exemple, sur le graphe CG+ apparaît une classe triviale associant les 2 mesures Yule's Q et Yule's Y comme fortement corrélées montrant une dépendance fonctionnelle.

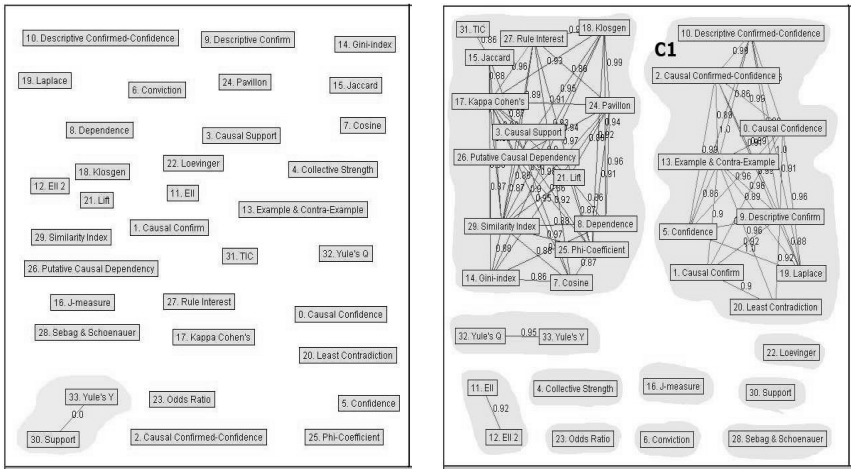

Fig. 2 – *Graphes de corrélation CG0 et CG+ sur la base mushroom (classes indiquées sur fond grisé).*

4 Conclusion

D'un point de vue technique, ARQAT est écrit en Java, offre une interface de visualisation interactive portable à travers un navigateur web et implémente pour l'instant 34 mesures. Afin de faciliter les échanges avec les logiciels externes, ARQAT supporte 3 formats de fichiers standard pour importer/exporter un jeu de règles : PMML (XML data-mining standard), CSV (Excel et SAS) and ARFF (format WEKA). ARQAT sera sera disponible sur la toile à partir de l'adresse www.polytech.univ-nantes.fr/arqat.

Dans cet article, nous avons souhaité montrer sur des illustrations l'intérêt de notre approche exploratoire, où l'organisation en tâches, l'usage intensif de représentations graphiques, et de leur complémentarité, améliore et facilite l'analyse des mesures d'intérêt par la communauté scientifique.

Références

Agrawal R. et Srikant R. (1994), Fast algorithms for mining association rules, Proc. of the 20th VLDB Conference, pp 487-499, 1994.

Agrawal R., Imielinski T. et Swami A. (1993), Mining association rules between sets of items in large databases, Proc. of 1993 ACM-SIGMOD Inter. Conf. on Management of Data, pp 207-216, 1993.

Bayardo Jr.R.J. et Agrawal R. (1999), Mining the most interestingness rules, Proc. of KDD'99, pp 145-154, 1999.

Blake C.L. et Merz C.J. (1998), UCI Repository of machine learning databases, http ://www.ics.uci.edu/∼mlearn/MLRepository.html, University of California, Irvine, Dept. of Information and Computer Sciences, 1998.

Blanchard J., Guillet F., Gras R. et Briand H. (2004), Mesurer la qualité des règles et de leurs contraposés avec le taux informationnel TIC, EGC'04, pp 287-298, 2004.

Blanchard J., Kuntz P., Guillet F. et Gras R. (2003), Implication Intensity : from the basic statistical definition to the entropic version (Chapter 8), Statistical Data Mining and Knowledge Discovery, Chapman & Hall CRC Press, pp 475-493, 2003.

Freitas A.A. (1999), On rule interestingness measures, Knowledge-Based Systems, 12(5-6), pp 309-315, 1999.

Gras R., Kuntz P., Couturier R. et Guillet F. (2001), Une version entropique de l'intensité d'implication pour les corpus volumineux, ECA 1(1-2), pp 69-80, 2001.

Gras R. (1996), L'implication statistique - Nouvelle méthode exploratoire de données, La pensée sauvage édition, 1996.

Guillaume S., Guillet F. et Philippé J. (1998), Improving the discovery of association rules with intensity of implication, Proc. of PKDD'98, Springer, pp 318-327, 1998.

Guillet F. (2004), Mesures de la qualité des connaissances en ECD, Actes des tutoriels, EGC'04, http ://www.isima.fr/∼egc2004/, pp 1-60, 2004.

Hilderman R.J. et Hamilton H.J. (2001), Knowledge Discovery and Measures of Interestingness, Kluwer Academic Publishers, 2001.

Lenca P., Meyer P., Picouet P., Vaillant B. et Lallich S. (2004), Evaluation et analyse multi-critères des mesures de qualité des règles d'association, Mesures de Qualité pour la Fouille de Données, RNTI-E-1, pp 219-246, 2004.

Liu B., Hsu W., Mun L. et Lee H. (1999), Finding interestingness patterns using user expectations, IEEE Trans. on Knowl. and Data Mining (11), pp 817-832, 1999.

Padmanabhan B. et Tuzhilin A. (1998), A belief-driven method for discovering unexpected patterns, Proc. of KDD'98, pp 94–100, 1998.

Piatetsky-Shapiro G. (1991), Discovery, analysis and presentation of strong rules, Knowledge Discovery in Databases, MIT Press, pp 229–248, 1991.

Tan P.N., Kumar V. et Srivastava J. (2004), Selecting the right objective measure for association analysis, Information Systems 29(4), pp 293-313, 2004.

Tan P.N., Kumar V. et Srivastava J. (2002), Selecting the Right Interestingness Measure for Association Patterns, Proc. of KDD'02, pp 32–41, 2002.