Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

Une plateforme exploratoire pour la qualité des règles d'association: apports pour l'analyse implicative

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. De nombreuses mesures d'intérêt sont disponibles dans la littérature, et de nombreux auteurs ont discuté et comparé leurs propriétés dans ce but. 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. Par conséquent, il est nécessaire de concevoir de nouvelles approches contextuelles pour guider l'utilisateur dans son choix. Dans cet article, nous proposons un outil original ARQAT afin d'étudier le comportement spécifique de 35 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. L'outil ARQAT implémente 14 vues graphiques complémentaires structurées en 5 tâches d'analyses. Une partie de ces vues est décrite et illustrée sur un même jeu de 120000 règles issues de la base MUSHROOMS (MLRepository), afin de montrer l'intérêt de cet outil exploratoire et de la complémentarité de ses vues.

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 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. Étant donné que l'intérêt dépend à la fois des préférences de l'utilisateur et des données, les MI peuvent être 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 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 (voir (Bayardo 1999), (Hilderman 2001), (Tan et al. 2002), (Tan et al. 2004), (Piatetsky 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 cette optique, (Vaillant et al. 2003) proposent un premier outil d'expérimentation : HERBS.

Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

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.
Dans une première partie nous présentons la structure et les principes de la plateforme ARQAT. Puis dans les parties suivantes nous ciblons la présentation sur 3 tâches complémentaires munies de représentations graphiques : les statistiques élémentaires sur le jeu de règles, l'analyse de corrélation, et enfin l'aide au choix des meilleures règles. Chacune de ces fonctionnalités est illustrée sur le même jeu de 120000 règles (issu de la base MUSHROOMS - MLRepository), afin de montrer l'intérêt de l'approche exploratoire s'appuyant sur un ensemble de vues complémentaires que nous proposons.

2 Principes de la plateforme ARQAT

ARQAT inclut 35 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) (cf Annexe 1 pour une liste récapitulative).
ARQAT (Fig. 1) implémente 14 vues graphiques complémentaires qui sont structurées en 5 groupes selon la tâche offerte.
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 $a \wedge \bar{b}$ (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.

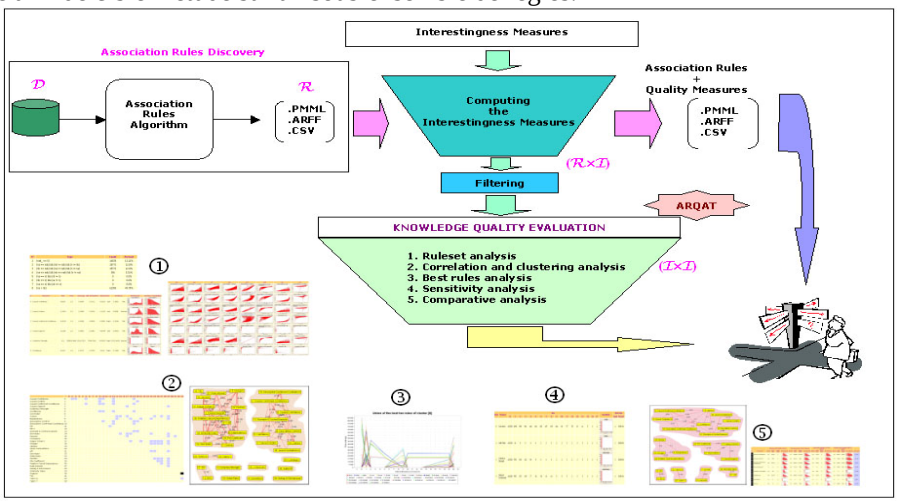


FIG. 1 - *Structure d'ARQAT.*

Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

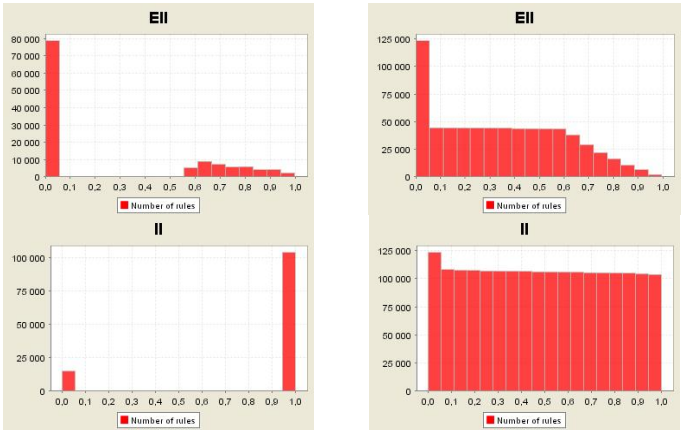
La seconde étape est ensuite interactive, l'utilisateur mène l'exploration graphique des résultats. Il s'appuie pour cela sur la structuration en 5 groupes de vues orientées tâche. Le premier groupe (1 dans Fig. 1) est dédié à la visualisation de statistiques élémentaires afin de mieux appréhender la 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 décrivons les trois premiers groupes et les illustrons sur un même jeu de règles : 120000 règles d'association extraites par un algorithme Apriori (support 10%) de la base MUSHROOMS (Blake 1998).

3 Statistiques sur le jeu de règles

Ce premier groupe offre trois vues graphiques résumant les caractéristiques statistiques du jeu de règles étudié. La première vue (Tab. 1) récapitule la *distribution des contingences* sous-jacente aux règles, et facilite la détection des cas limites. Par exemple, la première ligne décrit le nombre de règles "logiques" (i.e. sans contre-exemple $n_{a\bar{b}} = 0$, ou encore avec une confiance à 100%). On peut noter que le nombre de règles logiques est très élevé (13%).

N°	TYPE	NOMBRE DE CAS	POURCENTAGE
1	$n_{a\bar{b}} = 0$	16158	13,11%
2	$n_{a\bar{b}} = 0 \,\&\, n_a < n_b$	15772	12,80%
3	$n_{a\bar{b}} = 0 \,\&\, n_a < n_b \,\&\, n_b = n$	0	0%
4	$n_a > n_b$	61355	49,79%
5	$n_b = n$	0	0%

TAB 1 – *Quelques caractéristiques des règles issues de la base MUSHROOMS.*



Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

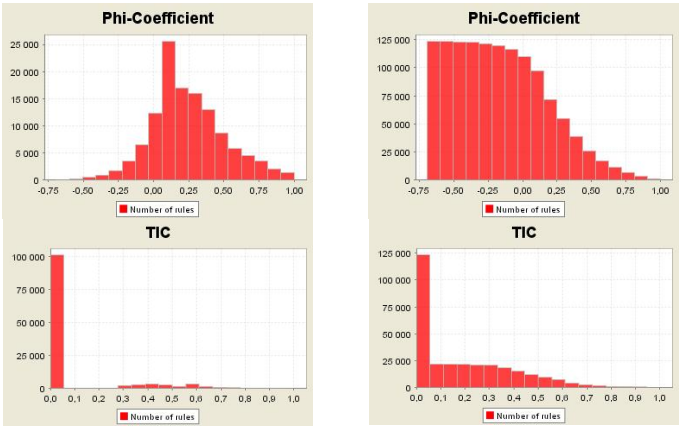


FIG. 2 -*Distributions (fréquence, inverse-cumulative) de quelques mesures sur la base MUSHROOMS.*

La deuxième vue, *distribution des mesures* (Fig. 2), présente l'histogramme de chaque mesure, en le complétant de divers indicateurs (minimum, maximum, écart-type, ...). On peut par exemple y observer que la mesure II possède une distribution très irrégulière; alors que la mesure Phi-Coefficient montre une distribution très différente.

En complément, la troisième vue (Fig. 3, Fig. 4) montre les *distributions croisées* des couples de mesures, récapitulées dans une représentation graphique matricielle très utile pour visualiser la forme de la liaison existant entre deux mesures. Les deux axes sont les valeurs des mesures. Par exemple, la Fig. 3 permet d'observer 6 différentes formes de non liaison : EII vs II (établissement de valeurs élevées, 1), EII vs Phi-Coefficient (2), EII vs TIC (3), EII 2 vs Phi-Coefficient (4), EII 2 vs TIC (5), et II vs Support (6), cette dernière révélant une forte indépendance. A l'opposé (la Fig. 4), les cellules EII vs EII 2 (7), Phi-Coefficient vs Rule Interest (8), Phi-Coefficient vs Similarity Index (9), et Yule's Q vs Yule's Y (10), montrent des formes de liaison forte dont la dernière révèle une dépendance fonctionnelle.

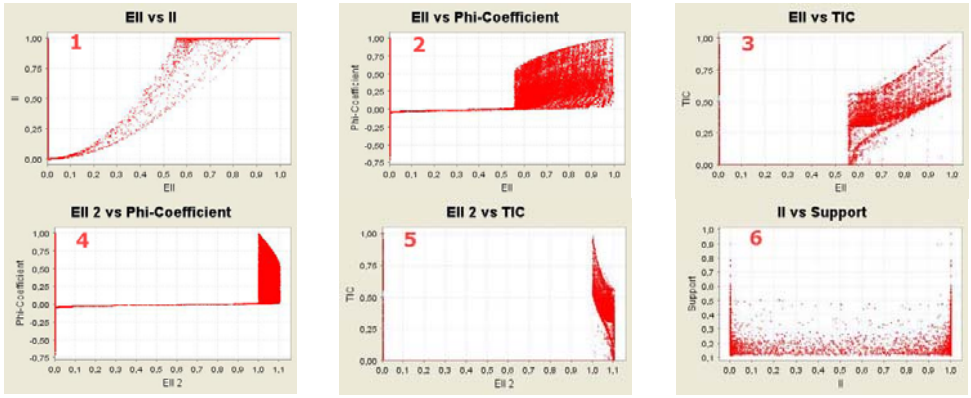


FIG. 3 - *Matrice de quelques distributions croisées sur la base MUSHROOMS (non liaison).*

Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

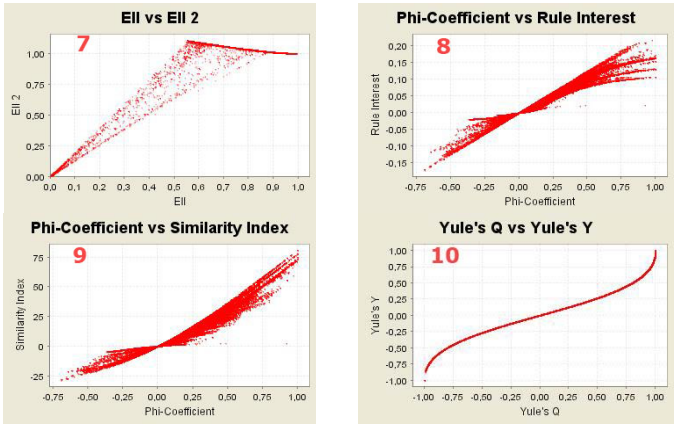


FIG. 4 - *Matrice de quelques distributions croisées sur la base MUSHROOMS (liaison forte).*

4 Analyse de corrélation

Ce deuxième groupe de vues est orienté vers la tâche d'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 (Fig. 5) 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. Par exemple (Fig. 5), les deux cellules noires mettent en évidence une non corrélation significative pour les couples de mesures: II vs Support et Yule's Y vs Support. Les 75 cellules grises correspondent à des corrélations significatives.

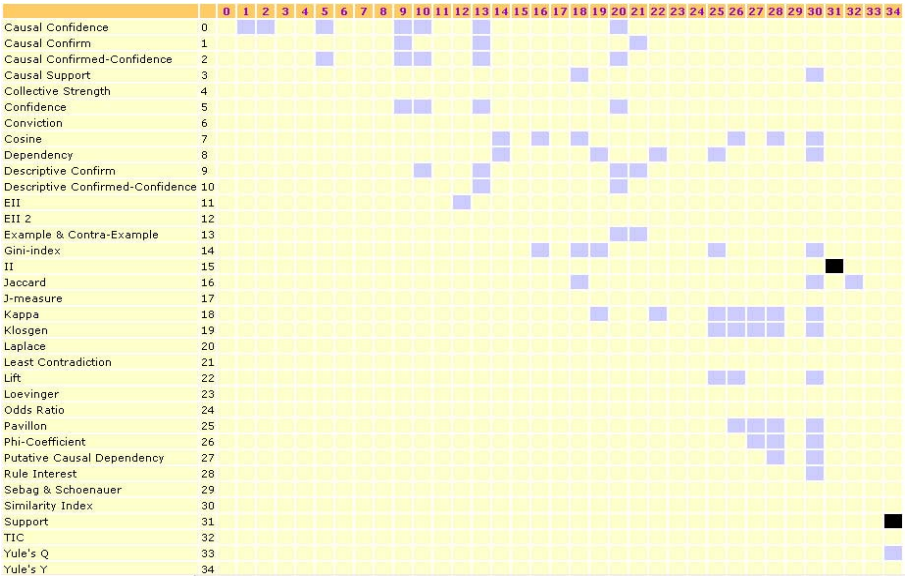
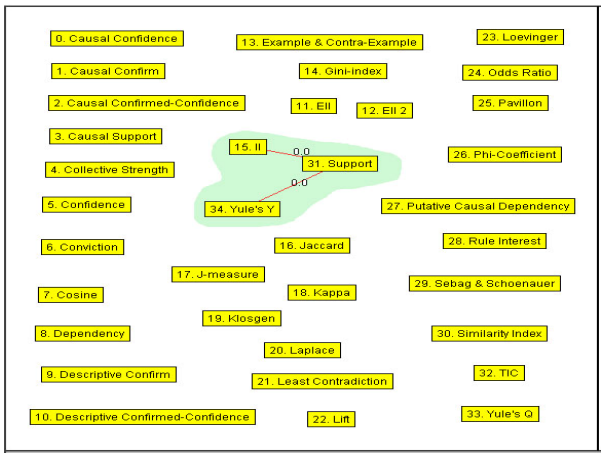


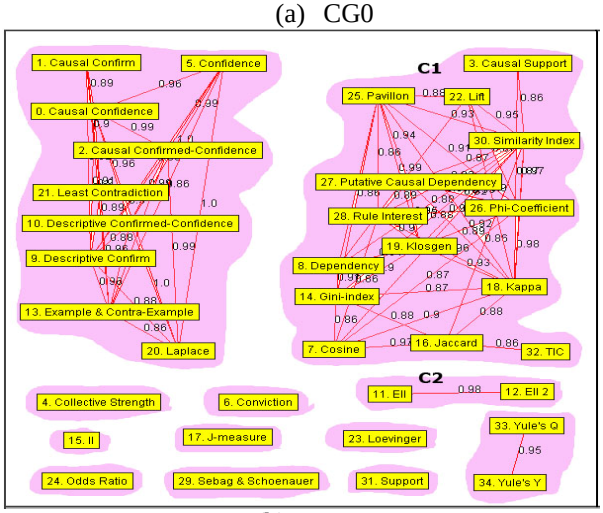
FIG. 5 - *Matrice de corrélation codée par niveau de gris sur la base MUSHROOMS.*

Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

La deuxième représentation envisagée, beaucoup plus expressive, est un *graphe de corrélation* (Fig. 6). 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).



(a) CG0



(b) CG+

FIG. 6 - *Graphes de corrélation CG0 et CG+ sur la base MUSHROOMS (classes indiquées sur fond grisé).*

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. 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

Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

exemple, Fig. 6, CG+ fait apparaître 12 parties connexes qui peuvent aider à choisir une base réduite de 12 mesures, parmi les 35 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 trois mesures II, Support et Yule's Y significativement non corrélées. Ce phénomène avait déjà été révélé par la matrice des niveau de gris (Fig. 5), et peut aussi être recoupé avec les distributions croisées de la Fig. 3 cellule (8). Un dernier exemple, sur le graphe CG+ apparaît une classe triviale associant les deux mesures Yule's Q et Yule's Y comme fortement corrélées; ce qui se retrouve sur la figure (Fig. 4 cellule (10)) montrant une dépendance fonctionnelle.
On peut observer (Fig. 6) que EII et EII 2 forment un cluster fortement corrélé (C2), Phi-Coefficient participe fortement à un cluster (C1) basé principalement sur les mesures de similarité (Cosine, Similarity Index, Jaccard, ...) , la relation entre TIC et le cluster C1 n'est pas très forte car il n'est en liaison qu'avec Jaccard.
Ces deux exemples illustrent l'intérêt d'utiliser conjointement les différentes vues complémentaires d'ARQAT. Ici, la matrice des distributions croisées (Fig. 3, Fig. 4) permet d'évaluer la nature des liens corrélatifs portés dans les graphes CG0 et CG+, afin de contourner les limites du coefficient de corrélation linéaire.

5 Analyse des meilleures règles

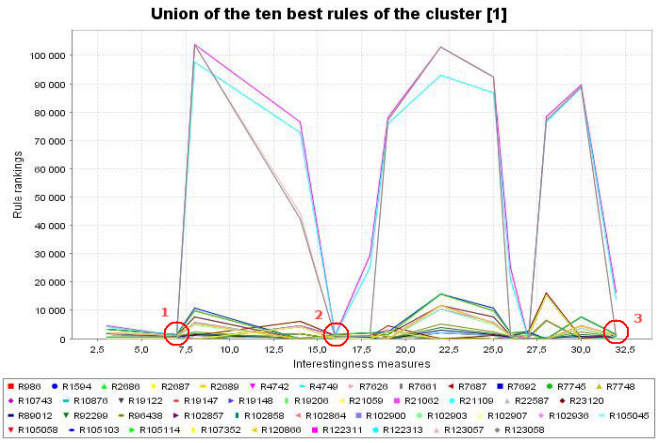


FIG. 7 - *Coordonnées parallèles des rangs des 10 meilleures règles pour chaque mesure de la classe C1 sur la base MUSHROOMS (les cercles numérotés sont les points de concentration).*

Deux vues ont été spécifiquement implémentées pour guider l'utilisateur vers l'interprétation des meilleures règles. La première vue rassemble les *n meilleures règles* d'une classe (partie connexe du graphe de corrélation), celles qui sont bien évaluées par au moins une mesure de cette classe. Les règles choisies y sont visualisées par une *représentation en coordonnées parallèles* (Fig. 7) qui permet une interprétation rapide des mesures de chaque règle et de leur variation. Alternativement, les règles peuvent aussi y être représentée par leur rang (ordre de classement de la règle selon la valeur de la mesure). Par exemple, la figure 7 présente les dix meilleures règles de la classe C1 (voir Fig. 6) selon une analyse de rang, les rangs des règles sont représentés sur l'axe vertical et les mesures sont affichés sur l'axe horizontal. On y observe des points de concentration sur les rangs faibles (notés 1, 2, 3 sur Fig. 7) correspondant respectivement aux trois mesures Cosine(7), Jaccard (16), et TIC (32) qui permettent trois évaluations proches sur les meilleures règles.

Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

6 Conclusion

Afin d'ouvrir les possibilités d'analyse expérimentales du comportement des mesures d'intérêt sur un jeu de règle spécifique, nous avons conçu et implémenté un outil informatique spécifique, ARQAT, suivant une approche résolument exploratoire, dont nous avons décrit la structure, une partie des 14 vues graphiques, et 3 des 5 tâches offertes.

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 35 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 téléchargeable gratuitement 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.

ARQAT constitue un premier pas vers une plateforme plus complète dediée à l'évaluation de la qualité en fouille de données. Nous allons poursuivre nos travaux selon deux directions. En premier lieu, nous souhaitons améliorer l'analyse de corrélation en quittant le coefficient de corrélation linéaire pour adopter un coefficient plus performant. La deuxième perspective concerne l'amélioration de la classification des mesures en utilisant un opérateur d'agrégation dirigé par les préférences de l'utilisateur afin d'améliorer sa prise de décision pour les sélection des meilleures mesures.

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., Kuntz P., Guillet F. et Gras R. (2003), Implication Intensity: from the basic statistical definition to the entropic version, Statistical Data Mining and Knowledge Discovery, pp 475-493, 2003.

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.

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, 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, pp 229-248, 1991.

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.

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.

Vaillant B., Picouet P. et Lenca P. (2003), An extensible platform for rule quality measure benchmarking, HCP'03, pp 187-191, 2003.

Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

Annexe 1 : Mesures d'intérêts utilisées

N°	Mesures d'intérêt	$f(n, n_a, n_b, n_{a\bar{b}})$
0	Causal Confidence	$1-\frac{1}{2}(\frac{1}{n_a}+\frac{1}{n_{\bar{b}}})n_{a\bar{b}}$
1	Causal Confirm	$\frac{n_a+n_{\bar{b}}-4n_{a\bar{b}}}{n}$
2	Causal Confirmed-Confidence	$1-\frac{1}{2}(\frac{3}{n_a}+\frac{1}{n_{\bar{b}}})n_{a\bar{b}}$
3	Causal Support	$\frac{n_a+n_{\bar{b}}-2n_{a\bar{b}}}{n}$
4	Collective Strength	$\frac{(n_a-n_{a\bar{b}})(n_{\bar{b}}-n_{a\bar{b}})(n_an_{\bar{b}}+n_bn_{\bar{a}})}{(n_an_b+n_{\bar{a}}n_{\bar{b}})(n_b-n_a+2n_{a\bar{b}})}$
5	Confidence	$1-\frac{n_{a\bar{b}}}{n_a}$
6	Conviction	$\frac{n_an_{\bar{b}}}{nn_{a\bar{b}}}$
7	Cosine	$\frac{n_a-n_{a\bar{b}}}{\sqrt{n_an_b}}$
8	Dependency	$\left\lvert\frac{n_{\bar{b}}}{n}-\frac{n_{a\bar{b}}}{n_a}\right\rvert$
9	Descriptive Confirm	$\frac{n_a-2n_{a\bar{b}}}{n}$
10	Descriptive Confirmed-Confidence	$1-2\frac{n_{a\bar{b}}}{n_a}$
11	EII $(\alpha=1)$	$\sqrt{\varphi\times I^{\frac{1}{2\alpha}}}$
12	EII $(\alpha=2)$	$\sqrt{\varphi\times I^{\frac{1}{2\alpha}}}$
13	Example & Contra-Example	$1-\frac{n_{a\bar{b}}}{n_a-n_{a\bar{b}}}$
14	Gini-index	$\frac{(n_a-n_{a\bar{b}})^2+n^2_{a\bar{b}}}{nn_a}+\frac{(n_b-n_a+n_{a\bar{b}})^2+(n_{\bar{b}}-n_{a\bar{b}})^2}{nn_{\bar{a}}}-\frac{n^2_b}{n^2}-\frac{n^2_{\bar{b}}}{n^2}$
15	II	$1-\sum_{k=\max(0,n_a-n_b)}^{n_{a\bar{b}}}\frac{C^{n_a-k}_{n_b}C^k_{n_{\bar{b}}}}{C^{n_a}_n}$
16	Jaccard	$\frac{n_a-n_{a\bar{b}}}{n_b+n_{a\bar{b}}}$
17	J-measure	$\frac{n_a-n_{a\bar{b}}}{n}\log_2\frac{n(n_a-n_{a\bar{b}})}{n_an_b}+\frac{n_{a\bar{b}}}{n}\log_2\frac{nn_{a\bar{b}}}{n_an_{\bar{b}}}$
18	Kappa	$\frac{2(n_an_{\bar{b}}-nn_{a\bar{b}})}{n_an_{\bar{b}}+n_{\bar{a}}n_b}$
19	Klosgen	$\sqrt{\frac{n_a-n_{a\bar{b}}}{n}}(\frac{n_{\bar{b}}}{n}-\frac{n_{a\bar{b}}}{n_a})$
20	Laplace	$\frac{n_a+1-n_{a\bar{b}}}{n_a+2}$
21	Least Contradiction	$\frac{n_a-2n_{a\bar{b}}}{n_b}$
22	Lift	$\frac{n(n_a-n_{a\bar{b}})}{n_an_b}$

Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

23	Loevinger	$1-\frac{nn_{a\bar{b}}}{n_a n_{\bar{b}}}$
24	Odds Ratio	$\frac{(n_a - n_{a\bar{b}})(n_{\bar{b}} - n_{a\bar{b}})}{n_{a\bar{b}}(n_b - n_a + n_{a\bar{b}})}$
25	Pavillon	$\frac{n_{\bar{b}}}{n} - \frac{n_{a\bar{b}}}{n_a}$
26	Phi-Coefficient	$\frac{n_a n_{\bar{b}} - n n_{a\bar{b}}}{\sqrt{n_a n_b n_{\bar{a}} n_{\bar{b}}}}$
27	Putative Causal Dependency	$\frac{3}{2} + \frac{4n_a - 3n_b}{2n} - (\frac{3}{2n_a} + \frac{2}{n_{\bar{b}}})n_{a\bar{b}}$
28	Rule Interest	$\frac{n_a n_{\bar{b}}}{n} - n_{a\bar{b}}$
29	Sebag & Schoenauer	$\frac{n_a}{n_{a\bar{b}}} - 1$
30	Similarity Index	$\frac{n_a - n_{a\bar{b}} - \frac{n_a n_b}{n}}{\sqrt{\frac{n_a n_b}{n}}}$
31	Support	$\frac{n_a - n_{a\bar{b}}}{n}$
32	TIC	$\sqrt{TI(a \to b) \times TI(\bar{b} \to \bar{a})}$
33	Yule's Q	$\frac{n_a n_{\bar{b}} - n n_{a\bar{b}}}{n_a n_{\bar{b}} + (n_b - n_{\bar{b}} - 2n_a)n_{a\bar{b}} + 2n^2_{a\bar{b}}}$
34	Yule's Y	$\frac{\sqrt{(n_a - n_{a\bar{b}})(n_{\bar{b}} - n_{a\bar{b}})} - \sqrt{n_{a\bar{b}}(n_b - n_a + n_{a\bar{b}})}}{\sqrt{(n_a - n_{a\bar{b}})(n_{\bar{b}} - n_{a\bar{b}})} + \sqrt{n_{a\bar{b}}(n_b - n_a + n_{a\bar{b}})}}$

Summary

The problem of choosing interestingness measures to validate association rules has become an important challenge in the context of evaluating knowledge quality in KDD. Many interestingness measures may be found in the literature, and many authors have discussed and compared interestingness properties in order to help the user (analyst, decision-maker) to choose the suitable measures for a given application. As interestingness depends both on the data structure and on the decision-maker's goals, some measures may be relevant in some context, but not in others. Therefore, it is necessary to design new contextual approaches in order to help the decision-maker to select the best interestingness measures. In this paper, we present ARQAT a new tool to study the specific behavior of a set of 35 interestingness measures in the context of a specific dataset and in an exploratory data analysis perspective. The tool implements 14 graphical and complementary views structured on 5 levels of analysis. The tool is described and illustrated on the MUSHROOMS dataset in order to show the interest of both the exploratory approach and the use of complementary views.