Comparaison des mesures d'intérêt de règles d'association : une approche basée sur des graphes de corrélation

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}@univ-nantes.fr

Résumé. Le choix des mesures d'intérêt (MI) afin d'évaluer les règles d'association est devenu une question importante pour le post-traitement des connaissance en ECD. Dans la littérature, de nombreux auteurs ont discuté et comparé les propriétés des MI afin d'améliorer le choix des meilleures mesures. Cependant, il s'avère que la qualité d'une règle est contextuelle : elle dépend à la fois de la structure de données et des buts du décideur. Ainsi, certaines mesures peuvent être appropriées dans un certain contexte, mais pas dans d'autres. Dans cet article, nous présentons une nouvelle approche contextuelle mise en application par un nouvel outil, ARQAT, permettant à un décideur d'évaluer et de comparer le comportement des MI sur ses jeux de données spécifiques. Cette approche est basée sur l'analyse visuelle d'un graphe de corrélation entre des MI objectives. Nous employons ensuite cette approche afin de comparer et de discuter le comportement de trente-six mesures d'intérêt sur deux ensembles de données a priori très opposés : un premier dont les données sont fortement corrélées et un second aux données faiblement corrélées. Alors que nous attendions des différences importantes entre les graphes de corrélation de ces deux jeux d'essai, nous avons pu observer des stabilités de corrélation entre certaines MI qui sont révélatrices de propriétés indépendantes de la nature des données observées. Ces stabilités sont récapitulées et analysées.

1 Introduction

Dans la dernière décennie, la conception de mesures d'intérêt adaptées à l'évaluation de la qualité des règles d'association est devenue un défi important dans le contexte d'ECD. Bien que le modèle des règles d'association (Agrawal et al., 1993) permette une extraction non supervisée de tendances implicatives dans les données, il produit malheureusement de grandes quantités de règles, ce qui les rend inexploitables sans la mise en oeuvre d'une étape lourde de post-traitement. Le post-traitement doit aider l'utilisateur (un décideur ou un analyste) à choisir les meilleures règles en fonction de ses préférences. Une manière de faciliter la tâche de choix de l'utilisateur consiste à lui offrir des indicateurs numériques sur la qualité des règles d'association : des mesures d'intérêt adaptées à ses buts et aux données étudiées.

Dans les travaux précurseurs sur les règles d'association (Agrawal et al., 1993; Agrawal et Srikant, 1994) , deux premières mesures statistiques sont introduites : le support et la confiance.

Celles-ci sont bien adaptées aux contraintes algorithmiques (cf *apriori*), mais ne sont pas suffisantes pour capturer l'intérêt des règles pour l'utilisateur. Afin de contourner cette limite, de nombreuses mesures d'intérêt complémentaires ont été proposées dans la littérature. (Freitas, 1999) distingue deux types de mesures d'intérêts : les mesures subjectives, et les mesures objectives. Les mesures *subjectives* dépendent des buts, connaissances, croyances de l'utilisateur et sont combinées à des algorithmes supervisés spécifiques afin de comparer les règles extraites avec ce que l'utilisateur connaît ou souhaite (Padmanabhan et Tuzhilin, 1998; Liu et al., 1999). Ainsi, les mesures subjectives proposent de capturer la nouveauté (*novelty*) ou l'inattendu (*unexpectedness*) d'une règle par rapport aux connaissances/croyances de l'utilisateur. Les mesures *objectives*, quant à elles, sont des indices statistiques qui évaluent la contingence d'une règle dans les données. De nombreux travaux de synthèse en récapitulent les définitions et propriétés (Bayardo Jr. et Agrawal, 1999; Hilderman et Hamilton, 2001; Tan et al., 2004). Ces synthèses comparent les mesures d'intérêt selon deux aspects différents : d'une part, la définition d'un ensemble de *principes* qui mènent à la conception d'une bonne mesure d'intérêt et d'autre part, la *comparaison* des mesures d'intérêt à l'aide de techniques d'analyse de données, afin d'en comprendre le comportement, et *in fine* d'aider l'utilisateur à choisir les meilleures.

Dans cet article, nous proposons de comparer des mesures d'intérêt objectives selon une nouvelle approche exploratoire, implémentée dans l'outil ARQAT. Cette approche est basée sur la construction de graphes de corrélation, dont l'intérêt est de permettre à la fois de classer les mesures sur un jeu de données spécifique et de servir de représentation visuelle afin d'aider l'utilisateur à choisir les meilleures mesures et les meilleures règles lors du post-traitement. Plus précisément, nous employons cette approche afin de comparer et de discuter le comportement corrélatif de trente-six mesures d'intérêt sur deux ensembles de données *a priori* très opposés : un premier dont les données sont fortement corrélées et un second aux données sont faiblement corrélées.

L'article est structuré en cinq parties. Dans la section 2, nous présentons les travaux relatifs aux mesures d'intérêt objectives pour des règles d'association. Puis, dans la section 3, nous présentons les mesures complémentaires II et sa version entropique (EII). Dans la section 4, nous présentons notre approche de classification basée sur des graphes de corrélation. La section 5 est consacrée à une étude spécifique sur deux ensembles de données opposés afin d'extraire d'éventuels comportements stables.

2 Travaux relatifs sur des mesures d'intérêt objectives

De nombreuses synthèses intéressantes sur les mesures d'intérêt objectives peuvent être trouvées dans la littérature. Ils traitent principalement deux aspects différents, la définition de l'ensemble de principes d'une bonne mesure d'intérêt, et leur comparaison selon une approche analyse de données afin d'aider l'utilisateur à choisir les meilleures.

Dans la perspective d'établir les principes d'une bonne mesure d'intérêt, Piatetsky-Shapiro (1991) présente une nouvelle mesure d'intérêt, appelé Rule-Interest, et propose trois principes fondamentaux pour une mesure sur une règle $a \Rightarrow b$: (P1) valeur 0 quand a et b sont indépendants, (P2) croissant avec $a \wedge b$, (P3) décroissant avec a ou b. Hilderman et Hamilton (2001) ont proposé cinq principes : *minimum value, maximum value, skewness, permutation invariance, transfer*. Tan et al. (2004) ont défini cinq principes d'intérêt : *symmetry under variable*

permutation, row/column scaling invariance, anti-symmetry under row/column permutation, inversion invariance, null invariance. Freitas (1999) propose un principe de "surprise" d'attribut. Bayardo Jr. et Agrawal (1999) conclut que les meilleures règles selon toutes les mesures d'intérêt doivent résider le long d'une frontière de support/confiance. Kononenco (1995) utilise onze mesures pour estimer la qualité des attributs à valeurs multiples, et montre que les valeurs des mesures : information-gain, j-mesure, gini-index, et relevance tendent à augmenter linéairement avec le nombre de valeurs d'un attribut. Zhao et Karypis (2001) utilisent huite critères et proposent des algorithmes pour optimiser un des critères. Ils montrent qu'une partie du critère de fonctions nouvellement proposé mène aux meilleurs résultats globaux. Gavrilov et al. (1999) ont étudiés la similitude des mesures afin de les classer. Gras et al. (2004) proposent un ensemble de dix critères : croissant avec un ou plusieurs des indicateurs prédéterminés, la décroissance doit respecter certaines attentes sémantiques, constraintes à respecter, décroissance avec la trivialité des observations, suffisamment souple et analytiquement générale, résistance discriminative à la croissance du volume de données ou une fonction discriminante, couplée avec sa contraposée $\overline{b} \Rightarrow \overline{a}$, la comptabilité des propriétés analystiques, la formule et les algorithmes, l'indépendance entre les deux variables a et b.

Certaines de ces synthèses abordent également la comparaison des mesures d'intérêt en adoptant un point de vue analyse de données. Hilderman et Hamilton (2001) ont utilisé les cinq principes proposés pour ranger des résumés de bases de données en employant seize mesures de diversité et montrer que : (1) six mesures ont satisfait cinq des principes proposés, (2) neuf des mesures restantes ont satisfait au moins un des principes proposés. En étudiant vingt et une mesures, Tan et al. (2004) montrent qu'aucune mesure n'est adaptée à tous les cas et que la corrélation des mesures augmente avec la diminution du support. Vaillant et al. (2004) évaluent vingt mesures selon huite critères d'intérêt et identifient quatre principaux groupes de mesures. Lallich et Teytaud (2004) utilisent quinze mesures et proposent des critères pour les évaluer. Blanchard et al. (2005) classent dix-huit mesures objectives en quatre groupes selon trois critères : indépendance, équilibre, et caractère descriptif ou statistique. Finalement, Huynh et al. (2005b) présentent le premier résultat d'une nouvelle approche de clustering pour classifier trente-quatre mesures d'intérêt en onze clusters avec la corrélation positive.

Il existe aussi deux outils d'expérimentation sont HERBS (Vaillant et al., 2003) et ARQAT (Huynh et al., 2005a). ARQAT (Association Rule Quality Analysis Tool) est un outil graphique développé à l'école polytechnique de l'université de Nantes, écrit en Java, avec une interface web. Les caractéristiques principales de cet outil sont : (1) l'analyse des ensembles de règles, (2) l'analyse de corrélation et de cluster, (3) l'analyse des meilleure règles, (4) l'analyse de la sensibilité, (5) l'analyse comparative. L'outil supporte différents formats pour l'exportation/importation des règles d'association et les mesures calculées : PMML, CSV et ARFF (employés par WEKA). Tous les résultats illustrés dans cet article issus de cet outil.

3 Mesure d'intensité d'implication

Dans le cadre de sa théorie statistique implicative, considérant les limites des mesures de support et de confiance, Gras (1996) a proposé une nouvelle mesure statistique appelée l'Intensité d'Implication (II). L'idée principale de cette mesure se fonde sur l'évaluation de la rareté des contre-exemples. L'intérêt d'une règle est mesurée en comparant le nombre effectif de contre-exemples au nombre de contre-exemples prédits par un modèle probabiliste.

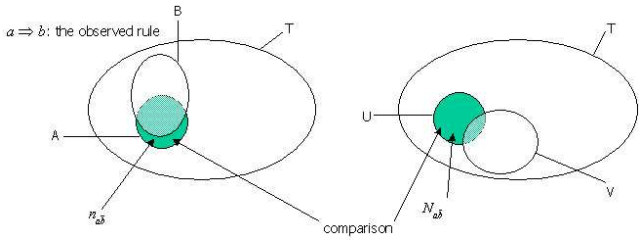


FIG. 1 – *Les cardinalités d'une règle.*

Définition 1 . Soit une règle d'association $a \Rightarrow b$ où a et b sont deux ensembles d'items disjoints (appelés itemsets). L'itemset a (resp. b) est associé à un sous-ensemble de transactions $A = T(a)$ (resp. $B = T(b)$) avec $T(a) = \{t \in T, a \subseteq t\}$ et T est l'ensemble de toutes les transactions. La règle peut être décrite par quatre cardinalités $(n, n_a, n_b, n_{a\bar{b}})$ où $n = |T|$, $n_a = |A|$, $n_b = |B|$, $n_{a\bar{b}} = |A \cap \overline{B}|$. La cardinalité $n_{a\bar{b}}$ correspond au nombre effectif de contre-exemples. Soient U et V deux ensembles aléatoires de transactions respectant les contraintes $|U| = |A|$ et $|V| = |B|$. Nous définissons le *nombre de contre-exemples prédits* comme la variable aléatoire $N_{a\bar{b}} = |U \cap \overline{V}|$. La mesure II est alors définie par la probabilité :

$$II(a \Rightarrow b) = 1 - proba(N_{a\bar{b}} \leq n_{a\bar{b}})$$

La variable aléatoire $N_{a\bar{b}}$ peut être modélisée selon plusieurs lois (hyper-géométrique, binomiale, ou poisson). En choisissant la loi hyper-géométrique, nous obtenons $proba(N_{a\bar{b}} = k) = \frac{C_{n-a}^{n_{\bar{b}}-k} C_{n_a}^{k}}{C_n^{n_b}}$. La formule globale de II peut être efficacement calculée avec une formule récursive.

Une extension de l'II, appelée l'intensité d'implication entropique (EII) a été proposée par (Blanchard et al., 2003) afin d'améliorer l'II dans le contexte de bases de données volumineuses. (Blanchard et al., 2004) ont aussi proposé une mesure complémentaire : TIC. D'autres extensions ont également été proposées pour les variables floues, ordinales, intervalles.

4 Approche par graphe de corrélation

4.1 Principe

Soient $R(D) = \{r_1, r_2, ..., r_p\}$ un ensemble de p règles d'association extraites d'un ensemble de données D. Chaque règle $a \Rightarrow b$ est décrite par ses deux itemsets (a, b) et ses cardinalités $(n, n_a, n_b, n_{a\bar{b}})$. Soit $M = \{m_1, m_2, ..., m_q\}$ un ensemble de q mesures d'intérêt. Chaque mesure est calculée par une fonction numérique sur les cardinalités d'une règle : $m(a \Rightarrow b) = f(n, n_a, n_b, n_{a\bar{b}})$.

Pour chaque mesure $m_i \in M$, nous pouvons construire un vecteur $m_i(R) = \{m_{i1}, m_{i2}, ..., m_{ip}\}, i = 1..q$, où m_{ij} correspond à la valeur de la mesure m_i pour une règle donnée r_j.

La valeur de corrélation entre deux mesures m_i et m_j $\{i, j = 1..q\}$ sur l'ensemble de règles R est calculée par le coefficient de corrélation linéaire CC (Saporta, 1990), où $\overline{m_i}, \overline{m_j}$ sont respectivement les valeurs moyennes calculées des vecteurs $m_i(R)$ et $m_j(R)$.

$$CC(m_i, m_j) = \frac{\sum_{k=1}^{p}[(m_{ik} - \overline{m_i})(m_{jk} - \overline{m_j})]}{\sqrt{[\sum_{k=1}^{p}(m_{ik} - \overline{m_i})^2][\sum_{k=1}^{p}(m_{jk} - \overline{m_j})^2]}}$$

Afin d'interpréter la valeur de corrélation, nous introduisons les deux définitions suivantes :

Définition 2. Mesures corrélées (τ-corrélation). Deux mesures m_i et m_j sont corrélées sur l'ensemble de données D si la valeur absolue de leur corrélation est supérieure ou égale à un seuil τ : $|CC(m_i, m_j)| \geq \tau$.

Définition 3. Mesures non-corrélées (θ-noncorrélation). Deux mesures de m_i et m_j sont non-corrélées sur l'ensemble de données D si la valeur absolue de leur corrélation est inférieure ou égale à un seuil θ : $|CC(m_i, m_j)| \leq \theta$.

Pour fixer le seuil de θ-noncorrélation, nous utilisons le seuil de significativité (Ross, 1987) $\theta = 1.960/\sqrt{p}$ où p est le nombre de règles. Les valeurs communes pour α sont : $\alpha = 0.1, 0.05, 0.005$, nous choisirons $\alpha = 0.05$. Nous fixons le seuil de τ-corrélation à $\tau = 0.85$ qui est une valeur commune dans la littérature.

Comme la corrélation est symétrique, les $q(q-1)/2$ valeurs de corrélation peuvent être stockées dans une demi-matrice $q \times q$. Cette matrice de corrélation peut également être vue comme la relation d'un graphe non-orienté et valué, appelé graphe de corrélation, dans lequel un sommet est une mesure d'intérêt et une arête est valuée par la valeur de corrélation entre deux sommets/mesures.

4.2 Graphe corrélé versus graphe non-corrélé

Malheureusement, le graphe de corrélation issu de la matrice de corrélation est complet, et n'est donc pas directement exploitable par l'utilisateur. Nous devons définir deux transformations afin d'extraire des sous-graphes plus limités et plus lisibles. D'abord, en employant la définition 2, nous pouvons extraire *le sous-graphe corrélé partiel (CG+)* : la partie du graphe où nous ne retenons que des arêtes liées à une forte corrélation (τ-corrélation). En second lieu, la définition 3 nous permet de construire le *sous-graphe non-corrélé partiel (CG0)* où nous ne retenons que les arêtes liées aux valeurs de corrélation proches de 0 (θ-noncorrélation).

Ces deux sous-graphes partiels peuvent ensuite être utilisés comme support de visualisation afin d'observer les liaisons corrélatives entre mesures.

On peut également y observer des clusters d'indices correspondant aux parties connexes des graphes.

4.3 Extension à graphe de stabilité

Afin de comparer les corrélations de mesures entre plusieurs jeux de données, nous introduisons une extension des graphes de corrélation aux graphes de stabilité.

Définition 4. Le graphe τ-stable $\overline{CG+}$ (resp. θ-stable $\overline{CG0}$) d'un ensemble de k jeux de règles $R = \{R(D_1), ..., R(D_k)\}$ est défini comme le graphe intersection moyenne des k sous-graphes corrélés (resp. non-corrélés) partiels CG+ (resp.CG0) calculés sur R. Chaque arête retenue est alors valuée par la corrélation moyenne des *k* arrêtes. Ainsi le graphe τ-stable $\overline{CG+}$

(resp. θ-stable $\overline{CG0}$) permettra de visualiser les fortes corrélations (resp. non-corrélations) stables, comme étant communes aux *k* jeux de données étudiés. Leurs complémentaires donnerons les corrélation instables, différentes.

Définition 5. On appellera clusters τ-stable (resp. θ-stable) les parties connexes du graphe τ-stable $\overline{CG+}$ (resp. θ-stable $\overline{CG0}$).

5 Étude du comportement de mesures sur deux jeux de données prototypiques et opposés

5.1 Description de données

Nous avons appliqué notre méthode à deux ensembles de données opposés : D_1 et D_2, afin de comparer les corrélations des mesures d'intérêt et plus précisement de découvrir s'il existe des corrélations stables sur des jeux de données de nature opposée. Le premier jeu de données D_1 est la base de MUSHROOMS issue du dépôt d'Irvine (Newman et al., 1998), et le second D_2 est un ensemble de données synthétiques T5.I2.D10k (T5 : taille moyenne des transactions est 5, I2 : taille moyenne au minimum des grands itemsets potentiellement est 2, D10k : nombre des items est 100). Les ensembles de règles d'association R_1 (resp. R_2) ont été calculés sur D_1 (resp. D_2) en employant l'algorithme *Apriori* (Agrawal et Srikant, 1994).

De plus, pour une évaluation plus fine du comportement des mesures sur les "meilleures règles", nous avons extrait $R_1^{'}$ (resp. $R_2^{'}$) à partir de R_1 (resp. R_2) comme l'union des 1000 premières meilleures règles ($\approx 1\%$ des 100000 règles disponibles) selon chaque mesure (voir Tab. 1).

Ces deux jeux de données ont été volontairement choisis pour leur caractère caricatural. Les attributs de la base de données D_1 sont fortement corrélés et délivrent de très nombreuses règles sans contre-exemples (confiance à 1). A contrario, la base de données synthétique D_2 est constituée d'attributs faiblement corrélés et délivre peu de règles sans contre-exemple. Ainsi, nous attendons de ce choix qu'il nous amène à ne découvrir que très peu de stabilités corrélatives entre les deux jeux de données.

Ensemble de données	Nombre de variables	Taille moyenne des itemsets	Transaction	Nombre de règles (seuil de support) (seuil de support)	θ	τ	$R(D)$
D_1	118	22	8416	123228 (12%)	0.005	0.85	R_1
				10431 (12%)	0.020	0.85	$R_1^{'}$
D_2	81	5	9650	102808 (0.093%)	0.003	0.85	R_2
				7452 (0.093%)	0.012	0.85	$R_2^{'}$

TAB. 1 – *Description des ensembles de données.*

5.2 Résultats et discussion

Dans notre expérience nous avons utilisé les trente-six mesures d'intérêt (trente-quatre mesures sont définies dans Huynh et al. (2005b) en ajoutant deux measures

$II = 1 - \sum_{k=max(0,n_a-n_b)}^{n_{a\bar{b}}} \frac{C_{n_b}^{n_a-k} C_{n_{\bar{b}}}^{k}}{C_n^{n_a}}$ et $IPEE = 1 - \frac{1}{2^{n_a}} \sum_{k=0}^{n_{a\bar{b}}} C_{n_a}^{k}$). Les mesures EII($\alpha = 1$) et EII($\alpha = 2$) sont deux versions entropiques de la mesure II.

Notre expérience vise à trouver des corrélations stables, *a priori* inattendues, entre les quatre ensembles de règles. A cette fin, nous analysons les résultats produits dans : (1) les quatre graphes $CG0$ et le graphe $\overline{CG0}$ montrant les indices non corrélés stables, (2) les quatre graphes $CG+$ et le graphe $\overline{CG+}$ montrant les indices corrélés stables.

Ensemble de règles	Nombre de corrélations		Nombre de clusters	
	τ-corrélation	θ-noncorrélation	CG+	CG0
R_1	79	2	12	34
$R_1^{'}$	91	15	12	21
R_2	65	0	14	36
$R_2^{'}$	67	17	12	20

TAB. 2 – *Comparaison de corrélation.*

5.2.1 Les graphes $CG+$ et $CG0$

La Fig. 2 présente les quatre $CG+$ obtenus et le tableau Tab. 2 récapitule le nombre des corrélations. Tout d'abord, nous pouvons noter qu'il existe de nombreuses corrélations entre les mesures, et que les mesures de chaque cluster ont tendance à évaluer les règles de la même manière.

Ensuite, nous pouvons observer que les graphes CG+ obtenus sur l'ensemble total des règles ($CG + (R_1)$ et $CG + (R_2)$) et le sous-ensemble des meilleures règles ($CG + (R_1^{'})$ et $CG + (R_2^{'})$) sont très semblables. Ceci nous indique que sur les deux jeux de données les corrélations et les clusters formés demeurent stables lorsque l'on sélectionne les meilleures règles.

D'autre part, comme on l'attendait, on observe un écart important entre les deux jeux de données, ce qui indique une sensibilité des mesures à la nature des données.

En revanche, nous pouvons observer un nombre important de corrélations entre mesures sur le jeu de données R_2 - même s'il est deux fois plus faible que sur R_1 - alors que nous en attendions peu du fait de la faible corrélation des données dans D_2.

La figure Fig. 3 permet de visualiser les mesures non-corrélées, dont le point de vue sur les données diffère.

On y observe un très faible nombre de non-corrélations, ce qui indique que très peu de mesures sont en désaccord fort sur l'évaluation des règles. Toutefois, le nombre de mesures non-corrélées augmente lorsque l'on passe de la totalité des règles aux meilleures. En revanche, contrairement à ce que l'on pouvait attendre, il y a moins de non-corrélations sur le jeu de données synthétique R_2.

Enfin, aucun comportement stable n'apparaît entre les mesures sur les quatre graphes $CG0$, et donc le graphe $\overline{CG0}$ est vide.

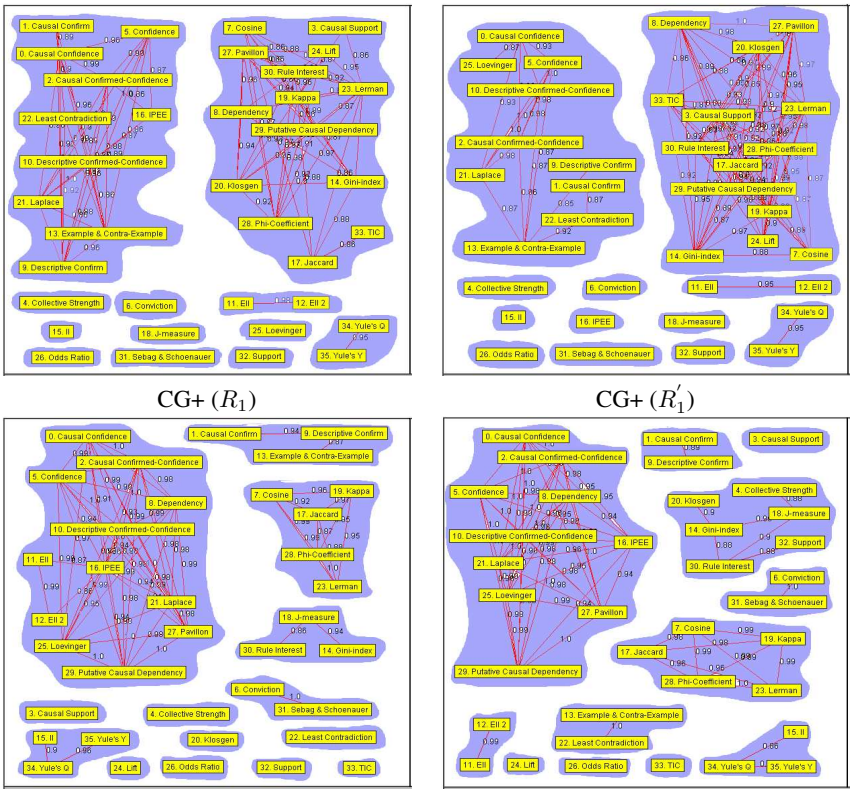


FIG. 2 – *Les 4 graphes CG+ (les clusters sont grisés).*

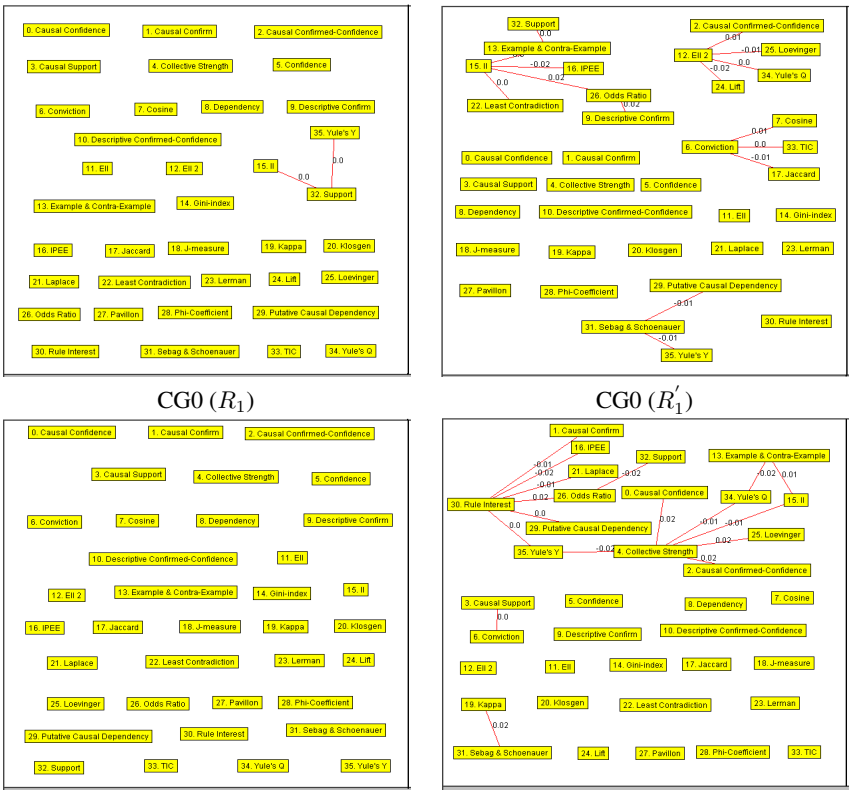
5.2.2 Le graphe $\overline{CG+}$: étude de la stabilité des corrélations

Le résultat le plus surprenant apparait dans le graphe $\overline{CG+}$. En effet, contrairement à notre attente, nous découvrons cinq clusters de mesures τ-stables, c'est-à-dire dont les corrélations demeurent inchangées entre les jeux de données. Ceci dénoterait d'une invariance avec la nature des données !

En analysant plus précisement ces cinq clusters τ-stable, nous notons quelques éléments intéressants.

(C1), le plus grand cluster, (Confidence, Causal Confidence, Causal Confirmed-Confidence, Descriptive Confirmed-Confidence, Laplace) rassemble des mesures dérivées de la mesure de confiance (Confidence). De plus, ce lien est fort, puisque le graphe est complet et les valeurs de corrélation supérieures à 0.97. Ceci indique un très fort accord entre ces cinq mesures.

(C2), ce cluster moins fortement corrélé que le premier, est constitué des mesures Phi-



CG0 (R_1) CG0 (R'_1)

CG0 (R_2) CG0 (R'_2)

FIG. 3 – *Graphes CG0.*

Coefficient, Lerman, Kappa, Cosine et Jaccard. Ce cluster rassemble des mesures partageant les quatre propriétés de : *symmetric under variable permuatation*, *antisymmetric under row/column permutation*, et *null invariance* (Tan et al., 2004).

Les deux mesures Jaccard et Cosine ne partagent que la cinquième propriété (*null invariance*) proposée par Tan et al. (2004).

(C3), rassemble trois mesures concernées par la première propriété (*symmetry/asymmetry under variable permutation*) proposée par Tan et al. (2004). L'existence de ce cluster est nécessaire pour distinguer la règle $a \Rightarrow b$ de $b \Rightarrow a$.

(C4), est un cluster constitué par deux versions de l'intensité de l'implication EII et EII 2, ce qui n'est pas surprenant.

(C5), la stabilité de la corrélation Yule'Q et Yule'Y, est elle aussi sans surprise puisque les deux mesures présentent une dépendance fonctionnelle. Ce cluster est lié à la deuxième propriété (*row/column scaling invariance*) proposée par Tan et al. (2004).

Les résultats du graphe τ-stable, donnent une piste intéressante pour construire une base réduite de mesures dont le point de vue est le plus différent sur les données. Il suffit pour cela de proposer à l'utilisateur de choisir cinq mesures, une parmi chacun des clusters. Nous pourrions aussi calculer automatiquement le meilleur représentant de chaque cluster en fonction des valeurs de corrélation.

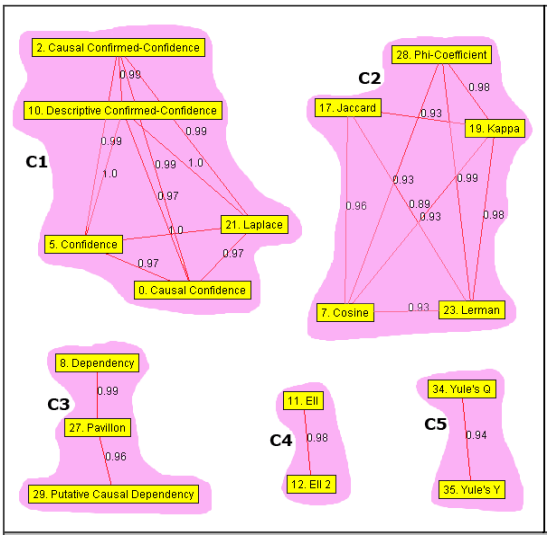


FIG. 4 – *Graphe* $\overline{CG+}$.

6 Conclusion

Nous avons étudié et comparé les diverses mesures d'intérêt décrites dans la littérature afin de mieux comprendre le comportement des mesures d'intérêt, et *in fine* d'aider l'utilisateur lors du post-traitement des règles d'association.

Une nouvelle approche exploratoire basée sur des graphes de corrélation, implémentés et visualisés dans outil d'ARQAT, a été présentée. Cette approche permet à un décideur d'observer le comportement corrélatif des mesures sur son propre jeu de données, et ainsi de l'aider à choisir les meilleurs indices et les meilleures règles.

Contre toute attente, une étude de la stabilité des corrélations entre mesures d'intérêt, a fait apparaître cinq groupes de mesures stables entre deux jeux de données choisis pour leur nature opposée.

Bien sûr ces résultats préliminaires restent à confirmer sur un ensemble de données plus important. Nous envisageons de prolonger notre travail dans deux directions. En premier lieu, nous souhaitons un indice de similarité entre mesures meilleur que l'indice de corrélation linéaire dont les limites sont soulignées dans la littérature. En second lieu, nous souhaitons amé-

liorer la prise en compte des préférences de l'utilisateur grâce à une agrégation des mesures d'intérêt.

Références

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining association rules between sets of items in large databases. *Proceedings of 1993 ACM-SIGMOD International Conference on Management of Data*, 207–216.

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules. *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases*, 487–499.

Bayardo Jr., R. J. et R. Agrawal (1999). Mining the most interestingness rules. *KDD'99, Proceedings of the 5th ACM SIGKDD International Confeference On Knowledge Discovery and Data Mining*, 145–154.

Blanchard, J., F. Guillet, R. Gras, et H. Briand (2004). Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel TIC. *EGC'04, 4ème Conférence francophone Extraction et Gestion des Connaissances RNTI-E-2*, 278–298.

Blanchard, J., F. Guillet, R. Gras, et H. Briand (2005). Assessing rule interestingness with a probabilistic measure of deviation from equilibrium. *ASMDA'05, Proceedings of the 11th International Symposium on Applied Stochastic Models and Data Analysis*, 191–200.

Blanchard, J., P. Kuntz, F. Guillet, et R. Gras (2003). Implication intensity: from the basic statistical definition to the entropic version (chap. 8). *Statistical Data Mining and Knowledge Discovery*, 475–493.

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

Gavrilov, M., D. Anguelov, P. Indyk, et R. Motwani (1999). Mining the stock market: which measure is best? *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining*, 487–496.

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

Gras, R., R. Couturier, J. Blanchard, H. Briand, P. Kuntz, et P. Peter (2004). Quelques critères pour une mesure de qualité de règles d'association. *Mesures de Qualité pour la Fouille de Données, Cépaduès Editions RNTI-E-1*, 3–31.

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

Huynh, X.-H., F. Guillet, et H. Briand (2005a). ARQAT: an exploratory analysis tool for interestingness measures. *ASMDA'05, Proceedings of the 11th International Symposium on Applied Stochastic Models and Data Analysis*, 334–344.

Huynh, X.-H., F. Guillet, et H. Briand (2005b). Clustering interestingness measures with positive correlation. *ICEIS'05, Proceedings of the 7th International Conference on Enterprise Information Systems 2*, 248–253.

Kononenco, I. (1995). On biases in estimating multi-valued attributes. *IJCAI'95, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1034–1040.

Lallich, S. et O. Teytaud (2004). Evaluation et validation de l'intérêt des règles d'association. *Mesures de Qualité pour la Fouille de Données RNTI-E-1*, 193–217.

Liu, B., W. Hsu, L. Mun, et H. Lee (1999). Finding interestingness patterns using user expectations. *IEEE Transactions on Knowledge and Data Engineering 11(6) 11*, 817–832.

Newman, D., S. Hettich, C. Blake, et C. Merz (1998). UCI repository of machine learning databases [http://www.ics.uci.edu/∼mlearn/MLRepository.html]. *University of California, Irvine, Department of Information and Computer Sciences*.

Padmanabhan, B. et A. Tuzhilin (1998). A belief-driven method for discovering unexpected patterns. *KDD'1998, Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, 94–100.

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

Ross, S. M. (1987). *Introduction to probability and statistics for engineers and scientists*. Wiley.

Saporta, G. (1990). *Probabilité, analyse des données et statistiques*. Edition Technip.

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

Vaillant, B., P. Lenca, et S. Lallich (2004). A clustering of interestingness measures. *DS'04, the 7th International Conference on Discovery Science, LNAI 3245*, 290–297.

Vaillant, B., P. Picouet, et P. Lenca (2003). An extensible platform for rule quality measure benchmarking. *HCP'03, Human Centered Processes*, 187–191.

Zhao, Y. et G. Karypis (2001). Criterion functions for document clustering: experiments and analysis. Technical report, Deparment of Computer Science, University of Minnesota. TR01-40.

Summary

Evaluating association rules with interestingness measures has become an important knowledge quality issue in KDD. Many interestingness measures may be found in the literature, and many authors have discussed and compared interestingness properties in order to improve the choice of the best 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 in selecting the best interestingness measures. In this paper, we present a new approach implemented by a new tool ARQAT for making comparisons. This approach is based on the analysis of a correlation graph presenting the clustering of objective interestingness measures, reflecting the post-processing of association rules. We use this graph-based approach to compare and discuss the behavior of thirty-six interestingness measures on two prototypical and opposite datasets: a strongly correlated one and a lowly correlated one. We focus on the discovery of the stable clusters obtained from the data analyzed between these thirty-six measures.