Agrégation de mesures d'intérêt de règles d'association

Jean-Pierre Barthélemy, Angélique Legrain, Philippe Lenca, Benoît Vaillant

Agrégation de mesures d'intérêt de règles d'association

Jean-Pierre Barthélemy*, Angélique Legrain*
Philippe Lenca*, Benoît Vaillant**,*

*GET – ENST Bretagne – Laboratoire TAMCIC, UMR 2872 CNRS
prenom.nom@enst-bretagne.fr
**IUT de Vannes, Département STID
benoit.vaillant@univ-ubs.fr

Résumé. L'un des principaux problèmes posés par l'extraction de règles d'association est l'évaluation de la qualité des règles produites par les algorithmes de type APRIORI. De nombreuses mesures ont été définies afin de pouvoir classer les règles dites *intéressantes*. Très hétérogènes, elles produisent des classements forts variés. C'est pourquoi, plutôt que de privilégier une mesure il paraît intéressant de tenir compte des différentes informations apportées par les mesures. Ainsi, nous avons adopté une nouvelle approche : l'agrégation à l'aide de relations valuées permettant de mesurer le degré d'intensité de préférence d'une règle sur une autre. Elles permettent d'une part de retranscrire la nature numérique des mesures, et d'autre part de réduire les problèmes d'incomparabilité entre les mesures.

Nous avons étudié différents opérateurs d'agrégation. Dans cet article, nous illustrons les résultats obtenus à l'aide d'un exemple jouet en utilisant le plus simple opérateur d'agrégation.

1 Considérations basiques

1.1 Règles d'association

Dans ce papier, nous nous restreignons au cas où les données sont des objets décrits par q attributs binaires : l'objet i satisfait la propriété x (codée en 1), ou non (codée en 0). On note $N = \{1, ..., n\}$ l'ensemble des objets et $Q = \{a, b, ...\}$ l'ensemble des propriétés.

Nous considérons des règles d'association $\mathtt{A} \rightarrow \mathtt{B}$ telles que définies par Agrawal et al. (1993) : si un sous-ensemble de N possède significativement les caractéristiques de l'ensemble $A \subseteq Q$, alors il possédera significativement les caractéristiques de l'ensemble $B \subseteq Q$. Une règle d'association est un 2-uplet (A,B) de sous-ensembles de Q tels que $A \cap B = \emptyset$.

Les algorithmes de type APRIORI (Agrawal et al., 1993) permettant de calculer les règles d'association produisent en général un nombre considérable de règles. Il est ainsi nécessaire de filtrer les règles en fonction de leur intérêt. Nous considérons dans cette étude les mesures d'intérêt dites objectives et basées sur des opérations de comptage dans les données, *i.e.* sur les grandeurs n, n_A, n_B et n_{AB}, où n_A (resp. n_B, n_{AB}) représente le nombre d'objets vérifiant toutes les propriétés de A (resp. B, $\mathtt{A} \cup \mathtt{B}$), selon les notations classiques.

1.2 Mesures de qualité

Le support, $p_{AB} = n_{AB}/n$, et la confiance, $p_{AB}/p_A = n_{AB}/n_A$ sont utilisés comme premier filtre pour extraire un ensemble $\mathcal{R}$ de règles avec les algorithmes de type APRIORI. Il est nécessaire d'utiliser dans un second temps d'autres mesures de qualité. En effet, on considère que le support et la confiance n'ont que peu de bonnes propriétés pour ordonner un ensemble de règles si on les compare à d'autres mesures (Piatetsky-Shapiro, 1991; Tan et al., 2002; Lenca et al., 2003). Nous avons sélectionné et étudié une vingtaine de mesures (Lenca et al., 2003). Dans Lenca et al. (2004), nous proposons une aide à la sélection d'un ensemble de mesures selon les données et les attentes de l'utilisateur. On met en évidence que le choix d'une mesure dépend particulièrement des attentes de ce dernier.

Les mesures constituent un paysage très hétérogène : on peut observer d'importantes variations entre les formules (les mesures ne traduisent pas les mêmes caractéristiques des règles), et de grandes différences dans les co-domaines ([0, 1], [0, +∞[,]-∞, 1], bornes fonction de n_A, n_B, et/ou $n_{AB}\ldots$). Ainsi certaines règles peuvent être très bien classées par une mesure, et mal classées par une autre. La comparaison des préordres totaux induits par les mesures sur une base de règles permet de mettre en évidence cette observation (Vaillant et al., 2004).

1.3 L'agrégation pour faire face à l'hétérogénéité des mesures

Il se pose ainsi naturellement la question suivante : quelle est ou quelles sont les meilleures règles étant donné nos mesures de qualité et les données à traiter ? On peut tenter d'y répondre selon deux voies principales :

(i) position dictatoriale : choisir une mesure préférée et ne pas tenir compte des autres
(ii) position consensuelle : trouver un consensus entre les mesures.

Dans cet article, nous suivons cette seconde piste pour laquelle trois voies apparaissent naturellement :

a) l'agrégation directe des mesures en une seule, en utilisant une sorte de moyenne généralisée. Une difficulté est alors posée par la diversité des échelles de mesures. Comment agréger des mesures dont les co-domaines sont [0,1], [0,+∞[,]-∞,1] ?
b) l'agrégation des rangs induits par les différentes mesures en un rang unique. Les classements ne tiennent alors pas compte des différences dans les évaluations. De plus, une agrégation ordinale implique des problèmes "logiques" périlleux (se référer par exemple au théorème d'Arrow (Arrow, 1951))
c) l'agrégation de relations valuées. C'est la voie que nous explorons dans cette étude. Les classements sont des préordres totaux que nous généralisons sous forme de relations valuées.

Pour cela, nous rappelons qu'une relation valuée (parfois appelée " relation floue ") sur un ensemble S est une transformation R de $S \times S$ dans l'intervalle unité [0,1]. Les relations valuées permettent d'échapper aux effets d'échelles en préservant les différences d'échelles puisqu'elles tiennent compte de telles différences. L'agrégation de relations valuées a été étudiée très précisément par Fodor et Roubens (1994). Elles peuvent être affectées de nombreuses propriétés, parmi lesquelles nous retenons quelques formes particulières de transitivité.

2 Construction d'une relation valuée sur un ensemble de mesures

L'un des avantages de la modélisation de relations valuées est la retranscription des évaluations numériques qu'elle permet. Pour définir des relations valuées il va donc falloir poser certaines conditions. Pour une mesure μ et deux règles r_i et $r_{i'}$, on dira par exemple que $R(r_i, r_{i'})$ est négligeable si $\mu(r_i)$ est faiblement supérieur à $\mu(r_{i'})$. Cette faible différence peut être définie de deux façons : soit par le quotient des évaluations, soit par leur différence.

Nous utilisons trois types de transitivité, la transitivité faible ($\{R(s,t) \geq \frac{1}{2}$ et $R(t,u) \geq \frac{1}{2}\} \Rightarrow R(s,u) \geq \frac{1}{2}$), la min-transitivité ($R(s,u) \geq \min\{R(s,t), R(t,u)\}$) et la max-$\Delta$-transitivité ($R(s,u) \geq \max\{0, R(s,t) + R(t,u) - 1\}$) car elles permettent la préservation de certaines propriétés après agrégation.

Soit $\mathcal{R} = \{r_1, \ldots, r_k\}$ un ensemble de règles, et $\mu_1, \ldots, \mu_m$ les mesures sélectionnées. Chaque mesure μ_j induit une relation valuée R_j sur $\mathcal{R}$. L'idée générale est que $R_j(r_i, r_{i'})$ correspond à une différence normalisée entre les valeurs prises par la mesure μ_j sur les règles r_i, $r_{i'}$ et doit permettre de modéliser un système de préférences sur l'ensemble des règles.

Nous présentons ci-dessous, une des relations valuées que nous avons étudiée et proposée par Brans et Mareschal (1994). C'est une variante de la différence linéaire permettant de lisser les transitions entre "non préférence" et "faible préférence" ainsi qu'entre la "préférence faible" et la "préférence forte" (le paramètre σ_j représente un seuil entre les "préférences faibles" et les "préférences fortes" –point d'inflexion de la courbe) :

$$R_j(r_i, r_{i'}) = \begin{cases} 1 - exp\left(-\frac{(\mu_j(r_i) - \mu_j(r_{i'}))^2}{2\sigma_j^2}\right) & \text{si } \mu_j(r_i) - \mu_j(r_{i'}) > 0 \\ 0 & \text{sinon} \end{cases} \quad (1)$$

3 Agrégation

3.1 Généralités

Un opérateur d'agrégation est une fonction C de $\cup_{m \geq 1}[0,1]^m$ dans [0,1] non décroissante de chaque composant, idempotent, et qui satisfait $C(0, \ldots, 0) = 0$ et $C(1, \ldots, 1) = 1$. Si R^* est un m-uplet $(R_1, \ldots, R_m)$ de relations valuées, un opérateur d'agrégation C va produire une relation de consensus de R^*, notée $C(R^*) : C(R^*)(r_i, r_{i'}) = C(R_1(r_i, r_{i'}), \ldots, R_m(r_i, r_{i'}))$.

De nombreux opérateurs d'agrégation aux propriétés différentes ont été étudiés dans la littérature (cf. Fodor et Roubens (1994)) : moyennes généralisées, opérateurs OWA, intégrales de Choquet et de Sugeno, maximum et minimums pondérés, etc. Nous avons choisi de nous concentrer sur les plus simples d'entre-eux, les moyennes généralisées :

$$M(u_1, \ldots, u_m) = f^{-1}\Big(\sum_{1 \leq j \leq m} w_j f(u_j)\Big)$$

où f est une fonction monotone continue, f^{-1} sa réciproque, et les w_j des poids non négatifs. Ainsi on distingue certains cas particuliers : la moyenne arithmétique pondérée (WMean, $f(u) = u$) la moyenne géométrique pondérée (WGeom, $f(u) = \log(u)$) la moyenne d'ordre α (RPM, $f(u) = u^\alpha, \alpha \in \mathbb{R}^*$), et la moyenne harmonique pondérée (WHarm, $f(u) = 1/u$). Dans cet article, nous présentons des résultats obtenus avec RPM et WMean.

3.2 Comportement vis à vis de la transitivité et de certaines propriétés

Saminger et al. (2002); Peneva et Popchev (2003) ont étudié les propriétés préservées par ces opérateurs. Seules celles qui jouent un rôle dans la modélisation des préférences nous intéressent dans le cadre de ce travail. Nous avons par conséquent éliminé différentes formes de dissimilarités. Nous nous concentrons principalement sur la transitivité, qui est l'une des propriétés qui se rapproche le plus des attentes de l'utilisateur. Elle garantit que si un ensemble de règles est mieux évalué qu'un autre, alors cet ordre sera préservé par la procédure d'agrégation.

4 Résultats expérimentaux

Le tableau 1 contient le rangement de 21 règles pour les 20 mesures que nous avons étudiées (les notations utilisées sont des abréviations du nom des mesures, SUP pour le support, IIE pour l'intensité d'implication entropique, etc., voir Lenca et al. (2004)). Les données "jouet" et le mode de calcul de ces règles sont présentés dans (Barthélemy et al., 2006), ainsi que les évaluations numériques des règles pour chaque mesure.

La relation valuée retenue pour exprimer la préférence individuelle d'une règle sur une autre pour une mesure donnée nécessite de fixer le paramètre σ_j (cf. formule 1). Pour ce faire, nous avons choisi de tenir compte des différences d'évaluations, et utilisons la valeur pour un quantile donné. Par exemple, la valeur prise à un quantile de 0% correspond à la plus petite valeur absolue de la différence (qui est évidemment 0, puisque la différence des évaluations d'une règle sur elle-même est nulle). La valeur prise par un quantile de 100% est la plus grande valeur absolue de la différence d'évaluation entre toute paire de règles et un quantile de 50% amène à la valeur médiane de l'absolue des différences. Dans notre exemple, le quantile est fixé à 60% de toutes les différences absolues, et c'est donc une valeur élevée. Nous expliquons plus loin pourquoi nous avons choisi une telle valeur. Les poids choisis sont tous égaux à $1/20$. D'autres poids ont été proposés dans (Legrain, 2004) grâce à une modélisation de préférences d'utilisateurs (Lenca et al., 2004).

Une fois ces paramètres fixés, la procédure d'agrégation produit une matrice carrée de valeurs entre 0 et 1, chaque valeur représentant la préférence agrégée d'une règle sur une autre. Pour obtenir des résultats lisibles (i.e. binaires), on utilise un seuil de coupe λ et on compare l'indice d'agrégation à λ. Les valeurs plus faibles que λ sont considérées comme égales à 0 et les valeurs supérieures égales à 1. On représente alors les préférences par des graphes (cf. figures 1 et 2, pour RPM et WMean) où les arcs entre deux sommets représentent la préférence d'une règle sur une autre, la flèche pointant sur la règle qui est préférée. On considère qu'aucune règle n'est préférée à elle-même. Il apparaît que les règles qui sont bien classées par toutes les mesures (comme r_{18}) restent bien classées, et celles qui étaient mal classées (comme r_{19}) restent mal classées. Le cas des règles plus controversées, comme r_2 nécessite d'approfondir la méthode. Notons que WHarm et WGeom produisent un effet de véto (Barthélemy et al., 2006). En effet, quand une mesure a la même valeur pour deux règles, l'utilisation du logarithme et de la fonction inverse produit des valeurs "infinies", ce qui fait que toute autre différence n'est pas prise en compte. Ceci explique qu'un grand nombre de règles resteront incomparables en terme d'agrégation de préférences avec de tels agrégateurs. Le fait que SUP divise les règles en seulement deux groupes joue aussi un rôle sur la valeur que nous avons choisie pour fixer σ_j : si nous avions choisi une valeur plus faible que 60%, alors la valeur

TAB. 1 – *Ordonnancement induit par les mesures sur une base de règles*

measure	rank 1	rank 2	rank 3	rank 4	rank 5	rank 6	rank 7	rank 8	rank 9	rank 10	rank 11
LIFT	r_{18}	$r_1\ r_2\ r_{17}$	r_{21}	$r_6\ r_7$	$r_3\ r_{16}\ r_{20}$	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$		
CONFCEN	$r_1\ r_{17}$	r_{18}	r_2	r_{21}	$r_3\ r_6\ r_{16}$	r_7	r_{20}	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$
CONF	$r_1\ r_3\ r_{16}\ r_{17}$	$r_6\ r_8\ r_{11}$	$r_{18}\ r_{19}\ r_{20}\ r_{21}$	$r_5\ r_7\ r_{12}\ r_{13}\ r_{14}\ r_{15}$	$r_2\ r_4\ r_9\ r_{10}$						
SUP	$r_5\ r_6\ r_7\ r_8\ r_{11}\ r_{12}\ r_{13}\ r_{14}\ r_{15}$	$r_1\ r_2\ r_3\ r_4\ r_9\ r_{10}\ r_{16}\ r_{17}\ r_{18}\ r_{19}\ r_{20}\ r_{21}$									
IIE	$r_1\ r_{17}$	$r_3\ r_{16}$	r_{18}	$r_6\ r_{21}$	r_{20}	r_2	r_7	$r_4\ r_5\ r_8\ r_9\ r_{10}\ r_{11}\ r_{12}\ r_{13}\ r_{14}\ r_{15}\ r_{19}$			
INTIMP	$r_1\ r_{17}$	r_{18}	r_2	$r_3\ r_{16}$	$r_6\ r_{21}$	r_7	r_{20}	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$
IQC	r_{18}	$r_1\ r_2\ r_{17}$	$r_6\ r_7\ r_{21}$	r_{20}	$r_3\ r_{16}$	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$		
CONV	$r_1\ r_3\ r_{16}\ r_{17}$	r_{18}	$r_6\ r_{21}$	r_2	r_{20}	r_7	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$	
GI	r_{18}	$r_1\ r_2\ r_{17}$	r_{21}	$r_6\ r_7$	$r_3\ r_{16}\ r_{20}$	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$		
-INDIMP	$r_1\ r_{17}$	r_{18}	r_2	$r_3\ r_{16}$	$r_6\ r_{21}$	r_7	r_{20}	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$
IPD	$r_1\ r_{17}$	r_{18}	r_2	$r_3\ r_{16}$	$r_6\ r_{21}$	r_7	r_{20}	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$
LAP	$r_1\ r_3\ r_{16}\ r_{17}$	$r_6\ r_8\ r_{11}$	$r_{18}\ r_{19}\ r_{20}\ r_{21}$	$r_5\ r_7\ r_{12}\ r_{13}\ r_{14}\ r_{15}$	$r_2\ r_4\ r_9\ r_{10}$						
LOE	$r_1\ r_3\ r_{16}\ r_{17}$	r_{18}	$r_6\ r_{21}$	r_2	r_{20}	r_7	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$	
MC	$r_1\ r_3\ r_{16}\ r_{17}$	r_{18}	r_2	r_{21}	r_6	$r_7\ r_{20}$	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$	
PS	r_{18}	$r_1\ r_2\ r_{17}$	$r_6\ r_7\ r_{21}$	$r_3\ r_{16}\ r_{20}$	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$			
R	r_{18}	$r_1\ r_2\ r_{17}$	$r_6\ r_7\ r_{21}$	$r_3\ r_{16}$	r_{20}	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$		
SEB	$r_1\ r_3\ r_{16}\ r_{17}$	$r_6\ r_8\ r_{11}$	$r_{18}\ r_{19}\ r_{20}\ r_{21}$	$r_5\ r_7\ r_{12}\ r_{13}\ r_{14}\ r_{15}$	$r_2\ r_4\ r_9\ r_{10}$						
MOCO	r_{18}	$r_1\ r_{17}$	r_6	$r_7\ r_{21}$	$r_3\ r_8\ r_{11}\ r_{16}$	$r_2\ r_{12}\ r_{13}\ r_{20}$	$r_5\ r_{14}\ r_{15}\ r_{19}$	$r_4\ r_9\ r_{10}$			
TEC	$r_1\ r_3\ r_{16}\ r_{17}$	$r_6\ r_8\ r_{11}$	$r_{18}\ r_{19}\ r_{20}\ r_{21}$	$r_5\ r_7\ r_{12}\ r_{13}\ r_{14}\ r_{15}$	$r_2\ r_4\ r_9\ r_{10}$						
ZHANG	$r_1\ r_3\ r_{16}\ r_{17}$	r_{18}	r_2	r_{21}	r_6	$r_7\ r_{20}$	$r_{12}\ r_{13}$	$r_4\ r_8\ r_9\ r_{10}\ r_{11}$	r_{19}	$r_5\ r_{14}\ r_{15}$	

du quantile aurait été nulle. Un moyen de palier à une telle situation pourrait être d'ajouter un léger degré de bruit dans les valeurs prises par les mesures. Cette stratégie, connue comme le "jittering", est souvent utilisée dans des outils de visualisation, afin de représenter un nombre important de points, autrement confondus. La représentation graphique a pour but d'illustrer les résultats sur un exemple jouet. Les résultats obtenus sur une base de données réelles sont présentés dans (Legrain, 2004), mais les représentations graphiques nécessitent alors des méthodes plus élaborées. Cette possibilité a notamment été explorée en fouille visuelle de règles d'associations par Lehn et al. (1999) et Blanchard et al. (2003). Notons cependant que l'exploration visuelle n'est pas requise pour sélectionner les règles, c'est un moyen confortable et centré sur l'expert.

5 Conclusion et perspectives

Dans cette étude, nous nous sommes intéressés à la construction d'opérateurs d'agrégation de relations valuées définies à partir des évaluations des mesures d'intérêt de règles d'association. Les attentes de l'utilisateur contraignent les choix possibles. De telles contraintes peuvent

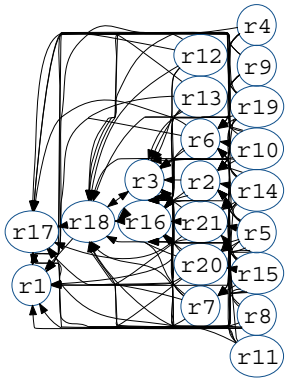

FIG. 1 – *RPM ($\alpha = 2$)*

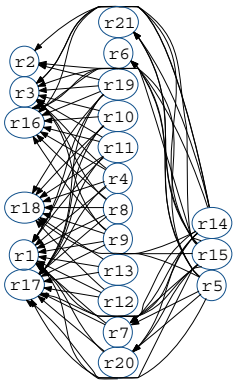

FIG. 2 – *WMean*

être explicitées mathématiquement. Nous nous sommes particulièrement intéressés à la propriété de transitivité. Parmi les opérateurs d'agrégation classiques, peu d'entre eux respectent de telles contraintes, excepté la moyenne généralisée. Un exemple jouet illustre notre approche. Des expériences prometteuses effectuées sur de plus grandes bases de données laissent entrevoir des résultats visuels intéressants. Bien qu'il persiste certains conflits entre le traitement de grandes bases de règles et leur représentation intelligible sous forme de graphe, nous pensons qu'une ouverture possible serait d'autoriser des zooms sur certaines régions significatives.

Une perspective d'étude intéressante est d'élargir le choix des opérateurs d'agrégation à des opérateurs de type compromis, comme l'intégrale de Choquet ou de Sugeno. Le classement final d'une règle pourrait aussi être consolidé si une coalition importante de mesures lui était favorable. Le choix du paramètre σ_j est déterminant. Nous l'avons fixé à l'aide d'une approche expérimentale basée sur une méthode de quantile. D'autres possibilités peuvent être envisagées, par exemple nous proposons de le fixer à l'aide d'un expert des données. Dans tous les cas, une étude de la robustesse de la solution proposée doit être menée.

Références

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

Arrow, K. J. (1951). *Social Choice and Individual Values*. Cowles Foundations and Wiley.

Barthélemy, J. P., A. Legrain, P. Lenca, et B. Vaillant (2006). Aggregation of valued relations applied to association rule interestingness measures. In *MDAI'06 (To be published)*, LNAI. Springer-Verlag.

Blanchard, J., F. Guillet, et H. Briand (2003). A user-driven and quality-oriented visualization for mining association rules. In *Third IEEE ICDM*, pp. 493–496.

Brans, J. et B. Mareschal (1994). The PROMETHEE-GAIA decision support system for multicriteria investigations. *Investigation Operativa 4*(2), 102–117.