Aide à l'interprétation des règles d'association composées.

Martine Cadot, Pascal Cuxac, Claire François

Aide à l'interprétation des règles d'association composées.

Martine Cadot* , Pascal Cuxac**, Claire François **

* UHP/LORIA, Département Informatique, BP239, 54506 Vandoeuvre-lès-Nancy cedex
martine.cadot@loria.fr
http://www.loria.fr /~cadot/
**INIST-CNRS, 2 allée du Parc de Brabois, 54154 Vandoeuvre-lès-Nancy cedex
pascal.cuxac@inist.fr ; claire.francois@inist.fr

Résumé. L'extraction des règles d'association (RA) est une méthode qui est apparue pour les données type « tickets de caisse ». La création de nombreux indices de qualité a permis sa généralisation à d'autres types de données (Guillet 2004). Nous nous intéressons ici au problème de l'expert qui se trouve confronté à un nombre important de règles pas toujours faciles à interpréter. Les règles formées seulement de deux propriétés, une en partie gauche et une en partie droite s'interprètent aisément une fois l'indice de qualité choisi. Dans le cas de règles composées, c'est-à-dire comportant plus de deux propriétés, ces indices ne suffisent pas à aider l'expert à interpréter le lien entre ces propriétés. Nous proposons un modèle qui permet d'évaluer le gain d'information apporté par les règles de type AB→C et de sélectionner pour l'expert celles qui ajoutent du sens aux règles simples A→C et B→C. L'application de cette méthodologie dans le cadre d'une analyse d'un corpus de textes par classification montre l'aide apportée à l'expert pour l'interprétation de cette classification. Pour faciliter l'exposé, le gain d'information que nous définissons a été appliqué à des règles formées de 3 propriétés, mais il est défini pour un nombre quelconque de propriétés.

1 Introduction

Les RA ont été créées pour extraire de la connaissance à partir de données. A et B étant deux propriétés binaires, la règle d'association est un lien entre A et B noté "A→B" ou "si A alors B". Sa définition varie selon les trois principaux courants initiés par les auteurs suivants : Gras (1979) définit des règles d'implication statistique pour aider les didacticiens à trouver des relations entre les acquisitions de notions élémentaires chez les élèves d'une classe, Guigues et Duquenne (1986) se sont plutôt intéressés à une représentation ordonnée de concepts avec les implications informatives, Agrawal et Srikant (1996) ont privilégié l'extraction optimisée de règles d'association dans de grandes bases de données.

Par la suite, ces formes ont connu des extensions dans plusieurs directions. La binarité des propriétés n'est plus obligatoire, on peut maintenant faire des RA avec des propriétés numériques (Guillaume 2000, Cadot et al. 2004b). Pour éviter l'explosion du temps d'extraction des règles, due à celle de la capacité de stockage des données, des algorithmes plus performants ont été proposés (Pasquier 2000). La sémantique des règles a été affinée grâce à de

nombreux indices de qualité (Guillet 2004), ce qui aide l'utilisateur à choisir les règles les plus adaptées à ses besoins. La navigation ainsi que l'interrogation par un langage adapté ont été mises au point (Botta et al. 2002) pour faciliter l'exploration de cet ensemble de règles.

Toutefois, à notre connaissance, peu de chercheurs en fouille de données se sont intéressés au problème d'interprétation que pose la prémisse composée d'une règle. Les contributions dans ce sens se sont concentrées sur la résolution du problème par élagage des règles qui apportent de l'incohérence au jeu de règles (Cadot et al. 2004a, Zhu H. 1998). Dans cet article, nous nous concentrons essentiellement sur l'ensemble des trois règles $A \rightarrow C$, $B \rightarrow C$ et $AB \rightarrow C$, et nous convenons de garder les deux règles à prémisse simple si elles ont été sélectionnées grâce à un choix d'indices approprié, et de ne garder la règle à prémisse composée que si elle renforce les deux autres, ce renforcement étant mesuré par un indice de gain.

Le but de cet indice de gain est d'aider l'expert à analyser les résultats issus du processus de traitement des données.

2 Le gain d'information

On ne s'intéresse dans cet article qu'aux RA extraites de tableaux numériques de type « SujetsXPropriétés » comportant les valeurs des sujets aux propriétés. Les définitions sont données pour des propriétés binaires et sont ensuite étendues à des propriétés numériques. La qualité d'une règle $A \rightarrow B$ est mesurée par de nombreux indices dont les plus courants sont le ***support***, qui est le nombre d'objets vérifiant les propriétés de A et de B, c'est-à-dire de leur conjonction, appelée ***motif*** AB, et la ***confiance***, qui est le quotient de ce support et du nombre d'objets vérifiant les propriétés de A, c'est-à-dire du support de A.

L'indice de qualité que nous proposons est construit seulement sur les règles à prémisse composée. Il mesure l'apport d'une telle règle par rapport aux règles simples qui la composent. Les règles simples ne permettent pas de calculer le support de la règle composée, mais seulement l'intervalle de ses variations. Le gain d'information mesure l'importance de l'écart entre le support observé et le milieu de l'intervalle de variation qui représente une situation d'équilibre. Notre méthode est à rapprocher de l'IPEE de Blanchard et al (2005) qui mesure l'écart à l'équilibre entre la partie gauche et la partie droite de la règle. Mais notre équilibre concerne toutes les propriétés constituant la règle. Freitas (1999) a également défini un gain d'information qu'il utilise dans son indice « Attribute surprinsingness » portant sur des règles de discrimination (où le membre droit de la règle est une propriété déterminée à l'avance). Son objectif est de mesurer l'apport individuel de chaque propriété à la discrimination, alors que nous mesurons le gain d'information apporté par une propriété supplémentaire.

2.1 Définition du gain

Le principe. Pour mesurer le gain d'information d'une règle, nous nous appuyons sur les variations possibles du support du motif M obtenu en réunissant les propriétés des parties gauche et droite de cette règle. On impose à ces variations de se faire en laissant les supports des sous-motifs de M inchangés. Ainsi ce gain mesure ce que l'association de toutes les propriétés le composant apporte de plus que l'ensemble des diverses associations d'une partie de ces propriétés.

Recherche de l'intervalle de variation. Pour obtenir que le support de M augmente d'une unité, on choisit un sujet qui vérifie toutes les propriétés sauf une, et on lui rajoute cette pro-

priété. Cela a pour conséquence que le support de chaque sous-motif de M contenant cette propriété est également augmenté de une unité. On compense cette augmentation en faisant de nouveaux changements élémentaires, qui vont également devoir être compensés. Si ce processus peut se réaliser, il s'arrête nécessairement au bout de 2^(L-1) changements, L étant la longueur du motif M, c'est-à-dire son nombre de propriétés. Et le support de M peut augmenter d'autant d'unités que de réitérations possibles de ce processus. Pour le faire diminuer, on procède pareillement en faisant les changements inverses de sujets. On crée ainsi l'intervalle de variation du support de M.

Choix de la valeur du gain. L'intervalle obtenu a un centre à partir duquel le support des motifs peut augmenter ou diminuer. Nous décidons que le gain d'information correspondant aux motifs de support central est nul, et cela reste valable si l'intervalle est réduit à une valeur. Puis plus le support du motif s'éloigne de ce centre en se rapprochant des bornes de l'intervalle, plus la valeur absolue du gain augmente. Selon que le support est à droite du centre ou à gauche, le gain est positif ou négatif. Nous mesurons la valeur de ce gain en nombre de sujets dont on doit changer la valeur d'une propriété pour obtenir ce motif en partant d'un motif de support central.

La formule. Pour mesurer le gain d'information g d'une règle, nous calculons le support **s** du motif M sur lequel elle s'appuie, la longueur **L** de ce motif (le nombre de propriétés le constituant), et le centre **c** de l'intervalle décrit par le support de M. Pour valeur de gain, nous choisissons la fonction g=2^(L-1)*(s-c).

2.2 Propriétés du gain d'un motif

La condition préalable au calcul du gain de motifs est qu'ils soient construits avec des propriétés dont on connaît les valeurs pour chacun des N sujets d'un ensemble donné. Bien qu'on utilise par la suite le gain pour des règles à prémisse composée, donc construites sur des motifs de longueur au moins égale à 3, le principe de calcul du gain s'étend sans problème à des motifs de longueur 2 et la formule à des motifs de longueur 1 (notons toutefois que dans ce dernier cas g=s-N/2, ce qui fait qu'on peut obtenir un gain avec des moitiés de sujets).

- prop 1 : Le gain d'un motif M ne peut pas dépasser N/2 en valeur absolue.
- prop 2 : le gain d'un motif de longueur L est un nombre entier de fois 2^(L-2)
- prop 3 : Si a est l'amplitude de l'intervalle de variation du gain g d'un motif M de longueur L, l'intervalle de variation du gain de ses sur-motifs de longueur L+1 a une amplitude inférieure ou égale à a-2|g|. La valeur de a pour les motifs de longueur 1 est N, l'intervalle étant [-N/2 ;N/2].

Ces trois propriétés permettent de limiter le coût machine de la recherche du gain d'un motif. Avec les propriétés 1 et 2, dès que sa longueur est telle que 2^(L-2) dépasse N /2 (soit L>1+log(N)/log(2)), le gain est nul. Avec la propriété 3, chaque fois que le gain d'un motif de longueur L est différent de 0, cela réduit l'intervalle de variation des motifs qui le contiennent. Ainsi, au fur et à mesure que la longueur du motif augmente par ajout de propriétés, ses possibilités de variation diminuent ou restent constantes, ce qui limite sa valeur possible de gain. Cet effet est accentué par la propriété 2. Cela est en adéquation avec le fait que dans le cas le plus courant, une fois que l'information essentielle est apportée par quelques propriétés, au fur et à mesure qu'on ajoute de nouvelles propriétés, l'information supplémentaire qui en résulte est de plus en plus petite.

Aide à l'interprétation des règles d'association composées.

2.3 Relation du gain avec les indices de qualité des règles

Le gain de la règle est celui du motif sur lequel elle est construite. Il fait ainsi partie des indices de qualité d'une règle au même titre que le support, la confiance et tous ceux qu'on définit habituellement (Guillet 2004). Toutefois, il ne mesure pas comme les autres indices la qualité intrinsèque d'une règle, mais la valeur additionnelle d'une règle avec prémisse composée par rapport à celles avec prémisses plus simples. Dans l'application que nous en faisons, nous extrayons d'abord les motifs de longueur quelconque ayant un support suffisant, puis les règles ayant une confiance suffisante construites sur des motifs de longueur 2. L'utilisation d'un seuil de support pour l'extraction des motifs se justifie par le besoin de généraliser les règles obtenues. Le choix de la confiance pour les règles sur des motifs de longueur 2 a été justifié a posteriori par l'interprétation satisfaisante des règles obtenues. Le gain permet de sélectionner les règles de longueur 3 qui renforcent les précédentes.

2.4 Le gain des règles d'association floues

Nous avons défini précédemment des supports flous et des RA floues sur des propriétés numériques (Cadot et al. 2004b). Les supports des motifs flous restent positifs, mais leurs valeurs peuvent ne pas être entières. L'intervalle de variation du support d'un motif ne peut plus se construire en déplaçant des sujets entre les 2^L parties délimitées par les valeurs aux propriétés, car leur appartenance à une partie est floue. Pas plus qu'il ne peut se construire en remplaçant la valeur à une propriété par son complément à 1, celui-ci n'ayant de sens qu'en calcul binaire. Toutefois, comme les effectifs de chacune des 2^L parties étaient calculés précédemment à partir des supports des motifs, le calcul peut toujours se faire pareillement, la différence est que les valeurs obtenues ne sont plus entières, ce sont des effectifs flous. Cela ne gêne aucunement le calcul du gain dont il importe peu qu'il soit ou non entier. Nous avions trois propriétés permettant de limiter le coût machine du calcul du gain. La deuxième propriété due au caractère entier des effectifs disparaît, mais les deux autres restent, et il en résulte une perte d'efficacité de l'algorithme du calcul du gain sur un ensemble de motifs flous. Cette perte est compensée par le choix d'un codage flou plutôt que binaire qui permet de ne pas multiplier les propriétés habituellement binarisées avec plusieurs seuils.

Nous avons ainsi défini un gain pour les règles d'association floue qui prolonge celui que nous venons de définir pour les RA classiques.

3 Application

Notre objectif est d'appliquer des RA floues sur des résultats de classifications, afin d'aider l'expert à analyser les résultats. Les règles à prémisses composées permettent de visualiser les classes qui peuvent fusionner entre deux classifications différentes ou au sein de la même classification.

3.1 Méthode de classification

Les classifications sur lesquelles nous travaillons ont été obtenues en utilisant la plateforme Stanalyst® (Polanco et al. 2001) qui permet de traiter des corpus bibliographiques et inclut la méthode des K-means Axiales comme méthode de classification (Lelu 1993).

Cette méthode est basée sur le principe de classification par centres mobiles, plus connue sous le nom de K-means (Forgy 1965), mais elle réalise une analyse factorielle sphérique sur chaque classe ; les classes sont donc matérialisées par des demi-axes représentatifs des éléments. L'utilisation de ces axes permet de quantifier l'appartenance d'un élément à une classe. De plus, au lieu d'affecter l'élément à la seule classe où sa valeur est la plus grande, on l'affecte également aux classes pour lesquelles cette valeur dépasse un certain seuil. Cet algorithme, paramétré par le nombre maximal de classes désiré et le seuil des coordonnées des éléments et descripteurs sur les axes, permet donc de construire des classes recouvrantes où les individus et descripteurs (documents et mots-clés) sont ordonnés selon un degré de ressemblance au type idéal de la classe.

Le corpus traité est constitué de 3203 notices bibliographiques extraites de la base PASCAL sur le thème de la géotechnique, publiées en 2001 et 2002 et indexées manuellement. Nous avons calculé quatre classifications avec la méthode des K-means axiales en paramétrant 20, 30, 40, 50 classes. Dans la suite de l'article elles sont nommées respectivement C20, C30, C40, C50.

3.2 Règles obtenues

Si nous calculons toutes les RA à prémisse composée nous avons 1548 règles. Le gain calculé permet de filtrer ces résultats : la variation du gain de 5 à 30 permet de passer de 90% (1395) de règles à 7% (105 règles). Pour faciliter l'analyse des résultats nous ne considèrerons dans ce qui suit que les règles où les deux membres de la prémisse appartiennent à la même classification. Avec un gain supérieur à 30 on a les 12 règles résumées dans le tableau suivant (S :support, C :confiance, G : Gain) :

N°	Règle	S	C	G
R1	C20 Barrage, C20 Eau souterraine → C40 Pollution	216,3	0,76	30,56
R2	C20 Inélasticité, C20 Relation σ ε (σ : contrainte, ε : déformation) → C50 Mécanique rupture	25,19	0,52	42,16
R3	C20 Inélasticité, C20 Relation σ ε → C30 Résistance compression	20,7	0,43	31,66
R4	C30 Barrage, C30 Eau souterraine → C40 Pollution	16,13	0,73	30,2
R5	C30 Relation σ ε, C30 Essai sol → C20 Résistance cisaillement	21,37	0,88	37,08
R6	C30 Résistance compression, C30 Mécanique rupture → C20 Inélasticité	18,42	0,91	33,12
R7	C40 Essai sol, C40 Relation σ ε → C20 Résistance cisaillement	19,83	0,86	33,06
R8	C50 Mécanique rupture, C50 Relation σ ε → C20 Inélasticité	19,63	0,91	35,32
R9	C50 Conductivité hydraulique, C50 Pollution → C20 Eau souterraine	23,73	0,97	46,22
R10	C50 Conductivité hydraulique, C50 Pollution → C30 Eau souterraine	23,66	0,97	45,94
R11	C50 Pression pores, C50 Champ pétrole → C20 Inélasticité	16,83	0,91	30,04
R12	C50 Relation σ ε, C50 Essai sol → C20 Résistance cisaillement	18,17	0,91	32,54

TAB. 1 – *Règles avec prémisse composée de deux classes de la même classification et avec g>30 (S=support, C=confiance, G=gain).*

Analysons par exemple la règle R11, constituée des règles simples suivantes :

C50 Pression Pores → C20 Inélasticité
C50 Champ pétrole → C20 Inélasticité

A première vue l'intitulé "Champ pétrole" peut paraître surprenant. L'analyse des données qui sont regroupées dans ces classes (titre des articles, résumés, indexation) permet de comprendre cette règle et de la valider. En effet la classe "Champ pétrole" est essentiellement consacrée aux roches magasins et aux distributions des contraintes dans ces roches. La classe "Inélasticité" est dominée par des aspects liés à l'élastoplasticité et à l'analyse des champs de contraintes. Cette règle est alors plus lisible puisqu'elle lie des articles parlant de la pression de pores (donc de roches poreuses plus ou moins saturées) et des articles sur la distribution des contraintes dans des roches magasin (roches poreuses plus ou moins saturées) avec des articles sur les champs de contraintes dans le domaine élastoplastique.

Cependant, certaines de ces règles, comme par exemple R1, sont difficilement interprétables. Cela est peut-être dû au fait que le gain est gonflé artificiellement par des effets d'effectifs. Ce phénomène est bien connu en statistique et nécessite qu'on sélectionne non seulement les effets les plus importants mais aussi les plus significatifs. En effet un score élevé peut être dû au hasard. La construction d'un test permettant d'établir la significativité du gain est en cours afin de les éliminer.

L'application de cette méthodologie nous a donc permis de "filtrer" les règles à prémisse composée pour ne garder que celles porteuses de valeur ajoutée.

4 Conclusion et perspectives

Le gain que nous proposons combine les avantages des indices de qualité des RA, et de l'élagage du jeu de RA. Il garde les règles simples, construites sur deux propriétés qui ont été extraites à l'aide d'un indice de qualité choisi pour sa valeur sémantique, et sont donc aisément interprétables. Les autres règles, qui ne sont gardées que si leur gain est suffisant, voire significatif renforcent l'information tirée des premières. Au final, l'ensemble des règles obtenu est de taille réduite et sans incohérence. Le gain proposé s'étend sans problème aux RA quantitatives codées de façon floue.

Nous avons vu que l'efficacité du gain doit être renforcée par un test qui en assure la significativité. D'autre part, le gain que nous avons utilisé mesure la valeur ajoutée de la règle AB→C par rapport aux règles A→C et B→C en fixant tous les sous-motifs de ABC. Il faudrait peut-être autoriser la variation du motif AB.

Références

Agrawal, R. Srikant, H, (1994) *Fast algorithms for mining association rules in large databases*, Research Report RJ 9839, IBM Almaden Research Center, San Jose, California, June 1994.

Blanchard, J., Guillet, F., Briand, H., Gras, R.(2005) IPEE : Indice Probabiliste d'Ecart à l'Equilibre. pour l'évaluation de la qualité des règles. DKQ 2005 Paris, Atelier EGC 2005 pp. 26-34

Botta, M., Boulicaut J.-F., Masson C., Meo R.(2002). A Comparison between Query Languages for the Extraction of Association Rules. *DaWaK* 2002, p. 1-10

Cadot, M., di Martino, J., Napoli, A. (2004a). Réduction d'un jeu de RA par des méta-règles issues de la logique de "sens commun". *EGC'2004*. (Clermont-Ferrand, France). RNTI, 2004. p.353.

Cadot, M., Napoli, A. (2004b) RA et codage flou des données. *SFC'04.* (Bordeaux). p.130-133.

Forgy E. W. (1965). Cluster analysis of multivariate data : efficiency versus interpretability of classifications, *Biometrics*, vol. 21, n° 3, p. 768.

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

Guigues J.L., Duquenne V. (1986) Familles minimales d'implications informatives résultant d'un tableau de données binaires, *Math. Sci. Hum*. n°95, pp. 5-18

Guillaume S. (2000) *Traitement des données volumineuses, mesures et algorithmes d'extraction de RA et règles ordinales*, Thèse Nantes, 2000.

Guillet F. (2004) Mesure de qualité des connaissances en ECD, *Cours donné lors des journées de la conférence EGC 2004,* Clermont-ferrand, 20 janvier 2004.

Gras R., (1979) *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques* , Thèse, Rennes I, 1979.

Lelu A. (1993). *Modèles neuronaux pour l'analyse de données documentaires et textuelles.* Paris, Thèse de l'Université de Paris VI, 238 pages.

Morineau, A., Nakache, J.-P., Krzyzanowski, C. (1996) *Le modèle log-linéaire et ses applications*, Cisia-Ceresta, Paris 1996.

Pasquier N. (2000), *Data Mining : Algorithmes d'Extraction et de Réduction des RA dans les Bases de Données*, Thèse, Clermont-Ferrand II, 2000

Polanco X., François C., Royauté J., Besagni D (2001). STANALYST: An Integrated Environment for Clustering and Mapping Analysis on Science and Technology, *CSCI 2001*, Sydney, Australia, Proceedings Vol 2, pp. 871 – 873.

Zhu H., (1998) *On-Line Analytical Mining of Association Rules*, Thèse, Simon Fraser University, 1998

Summary

When a simple rule of type A→C or B→C becomes a rule of type AB→C, the two premises A and B are a new premise AB. The fusion of these premises in only one is not a problem in binary logic, which is the foundation of the association rules, but the interpretation of the rule obtained is a semantic problem. We propose a model to evaluate the profit of information brought by rule AB→C and to select for the expert those which reinforce sufficiently the semantics of simple rules A→C and B→C. The application of this methodology to the analysis of a corpus by classification enables us to help the expert to interpret this classification.