Règles d'association avec une prémisse composée : Mesure du gain d'information.

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

La communauté de fouille de données a développé un grand nombre d'indices permettant de mesurer la qualité des règles d'association (RA) selon diverses sémantiques (Guillet, 2004). Cependant ces sémantiques, qui permettent d'interpréter les règles simples, s'avèrent d'utilisation trop complexe pour un expert dans le cas de règles à prémisse composée. Notre objectif est donc de sélectionner les règles à prémisse composée de type AB→C qui apportent une information supplémentaire à celle des règles simples A→C et B→C. Pour cela nous définissons un indice de gain d'une règle composée par rapport aux règles simples.

Dans l'application présentée, nous extrayons des RA de résultats de classifications pour en faciliter l'analyse . Le gain a permis de filtrer des règles d'interprétation simple.

1 Calcul du gain

Afin de 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 gauches et droites sans que les supports des sous-motifs ne changent. L'intervalle de variations obtenu a un centre, et nous décidons que le gain d'information correspondant aux motifs de support central est nul. Plus le support du motif s'éloigne de ce centre, plus la valeur absolue du gain augmente. Cela donne la formule suivante pour le gain : **g=2^(L-1)*(s-c)**, où **s** est le support du motif **M**, **L** la longueur de ce motif et **c** le centre de l'intervalle de variation.

Le gain de la règle fait partie des indices de qualité au même titre que le support, la confiance et la plupart de ceux dont on peut trouver la définition dans 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. Nous avons défini précédemment des RA floues sur des propriétés numériques (Cadot et Napoli, 2004). Le calcul du gain se prolonge sans problème à ces RA floues, les valeurs du support et du centre n'étant plus nécessairement entières.

2 Application

Le corpus traité est constitué de 3203 notices bibliographiques extraites de la base PASCAL sur le thème de la géotechnique et indexées manuellement. Nous avons calculé

quatre classifications avec la méthode des K-means axiales (Lelu et François 1992) en paramétrant 20, 30, 40, 50 classes. Si nous calculons toutes les RA à prémisse composée d'une même classification, nous avons 1548 règles. Avec un gain supérieur à 30, il reste 12 règles aisément interprétables. Par exemple la règle :

C50 Pression pores, C50 Champ pétrole → C20 Inélasticité

de support 16,83 de confiance 0,91, et de gain 30,04, 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. 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 apporte ainsi un gain d'information par rapport aux règles simples puisqu'elle lie les notions de pression de pores (donc de roches poreuses plus ou moins saturées) et de distribution des contraintes dans des roches magasins (roches poreuses plus ou moins saturées) avec la notion de champ de contraintes dans le domaine élastoplastique.

3 Conclusion

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 significatif, sont également simples d'interprétation car elles renforcent l'information tirée des premières. Au final, l'ensemble des règles obtenu est de taille réduite. Malgré tout, le filtrage par ce gain laisse quelques règles incohérentes. La construction d'un test permettant d'établir la significativité du gain est en cours afin de les éliminer.

Références

Cadot M., A. Napoli (2004) Règles d'association et codage flou des données. *SFC'04.* Bordeaux, 130-133.

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.

Lelu A., C. François (1992). Information retrieval based on a neural unsupervised extraction of thematic fussy clusters, *Neuro-Nîmes 92,* Nîmes, France.

Summary

In order to filter set of Association Rules with complex premises, we define a criteria which measures the improvement of information supported by the rule AB→C compared to the simple rules A→C or B→C. Application to clustering results.