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