Significativité des niveaux d'une hiérarchie orientée en analyse statistique implicative

Régis Gras*, Pascale Kuntz*
Jean-Claude Régnier**

*Laboratoire d'Informatique de Nantes-Atlantique FRE 2729
Site Ecole Polytechnique de l'Université de Nantes
La Chantrerie – BP 60601
44306 Nantes cedex 3
regisgra@club-internet.fr
pascale.kuntz@polytech.univ-nantes.fr
**EA 3727 Savoirs, Diversité et Professionnalisation
86, rue Pasteur
69365 Lyon cedex 07
Jean-claude.regnier@univ-lyon2.fr

Résumé. Dans le cadre de l'analyse statistique implicative développée à l'origine par R. Gras, nous avons proposé le modèle de « hiérarchie orientée » pour structurer des règles partielles de type $a \rightarrow b$ et des règles de règles, appelées *R*-règles, issues d'un corpus de données décrites par des attributs binaires. Dans cet article, nous proposons un nouveau critère de significativité[1] des niveaux de la hiérarchie orientée basé sur des comparaisons de préordres. Une application à un questionnaire d'opinions illustre l'intérêt de la démarche.

1. Introduction

Introduites en Extraction des Connaissances dans les Données au début des années *90* par R. Agrawal, [Agrawal et al., 1993], pour exprimer simplement des tendances implicatives $A_i \rightarrow A_j$ entre des sous-ensembles d'attributs A_i et A_j d'une table relationnelle, les règles d'association ont rapidement connu une utilisation intensive. Contrairement aux approches initiales de l'analyse combinatoire des données et de la classification conceptuelle, la condition d'inclusion $I(A_i) \subset I(A_j)$ entre le sous-ensemble d'individus $I(A_i)$ décrit par A_i et le sous-ensemble $I(A_j)$ décrit par A_j est ici relaxée, et l'on considère que la tendance générale à posséder A_j quand on a A_i n'est pas rejetée par la situation, fréquente pour des données réelles, où l'on observe quelques rares contre-exemples.

De nombreux algorithmes, dont le plus célèbre est certainement *Apriori*, ont été proposés dans la littérature. Cependant, il est bien connu dans la pratique qu'ils engendrent un nombre prohibitif de règles pour une analyse directe *in extenso*. Il est devenu alors nécessaire de

[1] Dans ce texte, le mot « significativité » aura un sens plus général que statistique. Il exprimera : « …qui est révélateur d'un phénomène d'intérêt sémantique majeur », même si la référence à une échelle de probabilité restera généralement présente.