Évaluation et validation de l'intérêt des règles d'association

Stéphane Lallich*, Olivier Teytaud**

*Laboratoire E.R.I.C, Université Lumière Lyon 2
5, avenue Pierre Mendès-France
69676 BRON Cedex – France
stephane.lallich@univ-lyon2.fr

**Artelys
215 avenue Jean-Jacques Rousseau
92136 Issy-les-Moulineaux
olivier.teytaud@artelys.com

Résumé. La recherche de règles d'association intéressantes est un thème privilégié de l'extraction des connaissances à partir des données. Les algorithmes du type Apriori fondés sur le support et la confiance des règles ont apporté une solution élégante au problème de l'extraction de règles, mais ils produisent une trop grande masse de règles, sélectionnant certaines règles sans intérêt et ignorant des règles intéressantes. Il faut disposer d'autre mesures venant compléter le support et la confiance. Dans cet article, nous passons en revue les principales mesures proposées dans la littérature et nous proposons des critères pour les évaluer. Nous suggérons ensuite une méthode de validation qui utilise les outils de la théorie de l'apprentissage statistique, notamment la *VC-dimension*. Face au grand nombre de mesures et à la multitude de règles candidates, l'intérêt de ces outils est de permettre la construction de bornes uniformes non asymptotiques pour toutes les règles et toutes les mesures simultanément.

1 Introduction

L'étude des règles d'association entre attributs booléens est déjà ancienne, liée à l'analyse des tableaux croisés 2×2. Comme le soulignent (Hajek et Rauch 1999), l'une des premières méthodes de recherche des règles d'association est la méthode GUHA initiée par (Hajek, Havel et Chytil 1966), où apparaissent déjà les notions de support et de confiance. L'intérêt pour les règles d'association a été renouvelé par les travaux de (Agrawal, Imielinski et Swami 1993), (Agrawal et Srikant 1994), puis (Srikant et Agrawal 1995) ayant trait à l'extraction de règles d'association à partir des grandes bases de données qui enregistrent le contenu des transactions commerciales.

Dans une telle base, chaque enregistrement est une transaction alors que les différents champs correspondent aux articles susceptibles de composer la transaction. On note n le nombre de transactions et p le nombre d'articles. Dans la mesure où l'on s'intéresse à la présence-absence de chaque article dans les différentes transactions, on