IPEE : Indice Probabiliste d'Ecart à l'Equilibre pour l'évaluation de la qualité des règles

Julien Blanchard, Fabrice Guillet
Henri Briand, Régis Gras

LINA – FRE 2729 CNRS
Polytech'Nantes
La Chantrerie – BP 50609
44306 – Nantes cedex 3 – France
julien.blanchard@polytech.univ-nantes.fr

Résumé. La mesure de la qualité des connaissances est une étape clef d'un processus de découverte de règles d'association. Dans cet article, nous présentons *IPEE*, un indice de qualité de règle qui a la particularité unique d'associer les deux caractéristiques suivantes : d'une part, il est fondé sur un modèle probabiliste, et d'autre part, il mesure un écart à l'équilibre (incertitude maximum de la conclusion sachant la prémisse vraie).

1 Introduction

Parmi les modèles de connaissances utilisés en Extraction de Connaissances dans les Données (ECD), les règles d'association (Agrawal et al., 1993) sont devenues un concept majeur qui a donné lieu à de nombreux travaux de recherche. Ces règles sont des tendances implicatives $a \rightarrow b$ où a et b sont des conjonctions d'items (variables booléennes de la forme $attribut = valeur$). Une telle règle signifie que la plupart des enregistrements qui vérifient a dans les données vérifient aussi b.

Une étape cruciale dans un processus de découverte de règles d'association est la validation des règles après leur extraction. En effet, de par leur nature non supervisée, les algorithmes de data mining peuvent produire des règles en très grande quantité et dont beaucoup sont sans intérêt. Pour aider le décideur (expert des données étudiées) à trouver des connaissances pertinentes parmi ces résultats, l'une des principales solutions consiste à évaluer et ordonner les règles par des mesures de qualité (Tan et al., 2004) (Guillet, 2004) (Lallich and Teytaud, 2004) (Lenca et al., 2004). Nous avons montré dans (Blanchard et al., 2004) qu'il existe deux aspects différents mais complémentaires de la qualité des règles : l'écart à l'indépendance et l'écart à ce que nous appelons l'équilibre (incertitude maximum de la conclusion sachant la prémisse vraie). Ainsi, les mesures de qualité se répartissent en deux groupes :

- les indices d'écart à l'indépendance, qui prennent une valeur fixe quand les variables a et b sont indépendantes ($n.n_{ab} = n_a n_b$) ;
- les indices d'écart à l'équilibre, qui prennent une valeur fixe quand les nombres d'exemples et de contre-exemples sont égaux ($n_{ab} = n_{a\bar{b}} = \frac{1}{2} n_a$).

Les mesures de qualité peuvent également être classées selon leur nature descriptive ou statistique (Lallich and Teytaud, 2004) :