Qualité et complexité en évaluation des mesures d'intérêt

Bruno Crémilleux*, Arnaud Giacometti**, Arnaud Soulet**

*Université de Caen Basse-Normandie, GREYC UMR 6072, Campus 2, 14000 Caen
bruno.cremilleux@unicaen.fr
**Université François-Rabelais de Tours, LI EA 6300, Campus de Blois, 41000 Blois
prenom.nom@univ-tours.fr

Résumé. Remplacer des hypothèses sur le modèle de données par des informations mesurées sur les données réelles est l'une des forces de la fouille de données. Cet article étudie cet ajustement entre les données et les méthodes de découverte de motifs pour en évaluer la qualité et la complexité. Nous formalisons ce lien entre données et mesures d'intérêt en identifiant les motifs *liés* qui sont ceux nécessaires pour l'évaluation d'une mesure ou d'une contrainte. Nous formulons alors trois axiomes que devraient satisfaire ces motifs liés pour qu'une méthode d'extraction se comporte bien. En outre, nous définissons la complexité en évaluation qui quantifie finement l'interrelation entre les motifs au sein d'une méthode d'extraction. A la lumière de ces axiomes et de cette complexité en évaluation, nous dressons une typologie de multiples méthodes de découverte de motifs impliquant la fréquence.

1 Introduction

La découverte de motifs locaux introduite par Agrawal et Srikant (1994) consiste à extraire des informations pertinentes décrivant une portion des données. Evaluer et garantir la qualité des motifs extraits demeure une problématique très ouverte malgré le nombre important de propositions (Giacometti et al., 2013). Chacune de ces propositions repose explicitement ou implicitement sur une mesure d'intérêt dont la qualité dépend de la complexité du modèle sous-jacent et de son ajustement aux données. Le modèle repose en général sur des fondements statistiques dont la complexité et la compréhension sont bien connues. A l'inverse, l'ajustement aux données reste une notion difficile à appréhender. Pourtant, c'est probablement cette notion qui distingue la fouille de données des statistiques traditionnelles. L'ajustement aux données est souvent connoté négativement et synonyme de sur-apprentissage par rapport aux données. De notre point de vue, l'ajustement aux données n'est pas un biais d'apprentissage mais un moyen pour lever certaines hypothèses sur le modèle en les remplaçant par des mesures sur les données. Nous proposons d'étudier l'ajustement aux données à travers les interrelations entre motifs lors de l'évaluation d'une mesure d'intérêt ou d'une contrainte d'extraction.

La qualité d'une mesure repose sur sa capacité à isoler un motif singulier qui dévie des autres motifs communs. Pour cette raison, une mesure se doit de mettre en relation le *motif évalué* avec d'autres motifs, dits *motifs liés*. Par exemple, la confiance de la règle d'association $X \rightarrow Y$ met en relation la fréquence de $X \cup Y$ (motif évalué) par rapport à la fréquence