Equilibrer l'analyse des motifs fréquents

Arnaud Giacometti, Patrick Marcel, Arnaud Soulet

Université François Rabelais Tours, LI
3 place Jean Jaurès
F-41029 Blois France
prenom.nom@univ-tours.fr

Résumé. Cet article propose une méthode originale d'évaluation de la qualité des motifs en anticipant la manière qui sera utilisée pour les analyser. Nous commençons par introduire le modèle de l'analyse aléatoire d'un ensemble de motifs selon une mesure d'intérêt. Avec ce modèle, nous constatons que l'étude des motifs fréquents avec le support conduit à une analyse déséquilibrée du jeu de données. Afin que chaque transaction reçoive la même attention, nous définissons le support équilibré qui corrige le support classique en pondérant les transactions. Nous proposons alors un algorithme qui calcule ces poids et nous validons expérimentalement son efficacité.

1 Introduction

La découverte de motifs introduite par Agrawal et Srikant (1994) consiste à extraire des informations pertinentes décrivant une partie des données. Depuis une quinzaine d'années, les algorithmes ont gagné en performance et arrivent désormais à extraire rapidement les motifs depuis des données volumineuses. Cependant, évaluer et garantir la qualité des motifs extraits demeure une problématique très ouverte. On distingue dans la littérature deux approches : celles guidées par les données (évaluant l'intérêt des motifs sur les données à analyser) et celles guidées par l'utilisateur (bénéficiant d'informations issues de l'utilisateur). Dans cet article, nous souhaitons adopter une nouvelle approche, dite *guidée par l'analyse*. L'évaluation en amont de l'intérêt des motifs s'appuie alors sur la manière dont les motifs seront analysés.

$\mathcal{D}$

Tid	Items
t_1	A B
t_2	A B
t_3	A B
t_4	C
t_5	C
t_6	C

+

P

Pid	Itemset	Support	Proportion d'analyse
p_1	A	0.5	0.5/2
p_2	B	0.5	0.5/2
p_3	AB	0.5	0.5/2
p_4	C	0.5	0.5/2

$\rightarrow$

Répartition de l'analyse

Tid	Prop. d'analyse
t_1	0.75
t_2	0.75
t_3	0.75
t_4	0.25
t_5	0.25
t_6	0.25

TAB. 1 – *Une analyse* déséquilibrée *du jeu de données $\mathcal{D}$ avec les motifs fréquents P*

Illustrons notre démarche sur un exemple jouet. Le tableau 1 présente un jeu de données $\mathcal{D}$ contenant 6 transactions composées des items A, B et C ainsi que les 4 itemsets présents