L'extraction de règles de dépendance bien définies entre ensembles de variables multivaluées

Frédéric Pennerath[*,**]

[*]UMI 2958 GeorgiaTech - CNRS
[**]Supélec
2 rue Édouard Belin, 57070 Metz

Résumé. Cet article étudie la faisabilité et l'intérêt de l'extraction de règles de dépendance entre ensembles de variables multivaluées en comparaison du problème bien connu de l'extraction des règles d'association fréquentes. Une règle de dépendance correspond à une dépendance fonctionnelle approximative caractérisée principalement par l'entropie conditionnelle associée. L'article montre comment établir une analogie formelle entre les deux familles de règles et comment adapter à l'aide de cette analogie l'algorithme « Eclat » afin d'extraire d'un jeu de données les règles de dépendance dites bien définies. Une étude expérimentale conclut sur les forces et inconvénients des règles de dépendance bien définies vis-à-vis des règles d'association fréquentes.

1 Introduction

Dans de nombreux problèmes de fouille de motifs, les données à analyser forment une collection d'échantillons décrits par différents attributs multivalués appelés *variables* dans ce qui suit. Un tel jeu de données est illustré par l'exemple simplifié de la figure 1(a). Les motifs

	V_1	V_2	V_3	V_4
D_1	a	1	oui	1
D_2	c	2	oui	1
D_3	b	1	non	1
D_4	c	2	non	1

(a) Avec variables multivaluées

	V_1 = a	V_1 = b	V_1 = c	V_2 = 1	V_2 = 2	V_3 = yes	V_3 = no	V_4 = 1
D_1	×			×		×		×
D_2			×		×	×		×
D_3		×		×			×	×
D_4			×		×		×	×

(b) Avec items ou attributs monovalués

FIG. 1 – *Interprétation mono ou multivaluée d'un même jeu de données.*

les plus souvent recherchés dans ces données sont des *itemsets* : un itemset est un ensemble d'attributs monovalués appelés *items* représentant chacun une des valeurs possibles d'une des variables multivaluées. Un jeu de données défini sur un ensemble d'attributs multivalués peut être représenté de façon équivalente par des items comme illustré sur la figure 1(b). Un itemset *couvre* une donnée si tous les items de l'itemset sont aussi des items de la donnée. La *fréquence* d'un itemset est la proportion des données couvertes par celui-ci. Une question consiste alors