

Extraction de motifs graduels clos

Sarra Ayouni^{*,**} Sadok Ben Yahia^{*}, Anne Laurent^{**}, Pascal Poncelet^{**}

^{*}Faculté des Sciences de Tunis, 1060,Campus Universitaire, Tunis, Tunisie
sadok.benyahia@fst.rnu.tn

^{**}LIRMM – CNRS, 161 rue Ada, Montpellier, France
{ayouni,laurent,poncelet}@lirmm.fr

Résumé. La découverte automatique de règles et motifs graduels (“plus l’âge d’une personne est élevé, plus son salaire est élevé”) trouve de très nombreuses applications sur des bases de données réelles (e.g. biologie, flots de données de capteurs). Si des algorithmes de plus en plus efficaces sont proposés dans des articles récents, il n’en reste pas moins que ces méthodes génèrent un nombre de motifs tellement important que les experts peinent à les exploiter. Dans cet article, nous proposons donc une représentation condensée des motifs graduels en introduisant les concepts théoriques associés aux opérateurs de fermeture sur de tels motifs.

1 Introduction

Les outils de la fouille de données permettent d’extraire des motifs et des règles à partir des gros volumes de données. Récemment, l’extraction d’un nouveau type de motif a été étudiée : les motifs graduels (ou itemsets graduels), de la forme *Plus/Moins* $X_1, \dots, \text{Plus/Moins}$ X_n . De même que pour l’extraction d’itemsets classiques, on parle de motif graduel *fréquent* lorsqu’un nombre minimal de données de la base *supporte* le motif considéré. Plusieurs définitions du support ont été proposées dans la littérature. D’une manière générale, il s’agit de trouver à quel point les données de la base peuvent être ordonnées ou non pour que les relations de précedence présentes soient respectées (co-variations sur les attributs X_1, \dots, X_n). Le support sert également à la définition d’algorithmes efficaces fonctionnant par niveau. Cependant ces algorithmes génèrent de très nombreux motifs et restent donc parfois difficilement utilisables par les experts. Face à ce problème, plusieurs méthodes sont envisageables : prise en compte de préférences utilisateurs, ordonnancement des motifs par rapport à des mesures de qualité, ou encore représentations condensées. Dans cet article, nous nous focalisons sur cette dernière méthode et introduisons la notion de motif graduel clos. D’une manière générale, nous appelons motif clos un motif qui n’a pas le même support que tout super-motif. Il concentre ainsi de l’information sur les sous-motifs et permet de définir une représentation condensée et sans perte (Boulicaut et Bykowski (2000); Pei et al. (2000); Pasquier et al. (1999)). L’ensemble des motifs fréquents peut alors être concentré en un sous-ensemble de motifs qui les représentent.

Dans cet article, nous utilisons la théorie de l’analyse formelle de concepts afin d’extraire de tels motifs graduels clos.