GraphMDL : sélection de motifs de graphes avec le principe MDL

Francesco Bariatti, Peggy Cellier, Sébastien Ferré

Univ Rennes, INSA, CNRS, IRISA
Prénom.Nom@irisa.fr

Résumé. Plusieurs algorithmes de fouille de motifs ont été proposés pour identifier des structures récurrentes dans les graphes. Le principal défaut de ces approches est qu'elles produisent généralement trop de motifs pour qu'une analyse humaine soit possible. Récemment, des méthodes de fouille de motifs ont traité ce problème sur des données transactionnelles, séquentielles et relationnelles en utilisant le principe MDL (*Minimum Description Length*). Dans ce papier, nous proposons une approche MDL pour sélectionner un sous-ensemble représentatif de motifs sur des graphes étiquetés. Une notion clé de notre approche est l'introduction de *ports* pour encoder les connections entre occurrences de motifs, sans perte d'information. Nos expériences montrent que le nombre de motifs est drastiquement réduit et que les motifs sélectionnés peuvent avoir des formes complexes.

1 Introduction et état de l'art

Beaucoup de domaines utilisent des données représentées par des graphes. Par exemple, en chimie et biologie, les molécules sont représentées par des graphes avec des atomes et des liaisons ; en linguistique, les phrases sont représentées par des graphes avec des mots et des liens de dépendance ; en web sémantique, la connaissance est représentée par un graphe avec des entités et des relations. En fonction du domaine, le jeu de données est un graphe unique ou une collection de graphes. Ces graphes sont intrinsèquement complexes à analyser pour en extraire de la connaissance, par exemple pour identifier des sous-structures fréquentes.

Dans le domaine de la fouille de motifs, plusieurs approches de *fouille de graphes* pour extraire des sous-graphes fréquents ont été proposées. Les approches classiques de fouille de graphes telles que MoFa (Borgelt et Berthold, 2002), gSpan (Yan et Han, 2002), Gaston (Nijssen et Kok, 2005) ou FFSM (Huan et al., 2003), génèrent tous les motifs possibles par rapport à une fréquence minimale. Le défaut majeur de ce type d'approches est le grand nombre de motifs générés, qui rend difficile leur analyse. Certaines approches comme CloseGraph (Yan et Han, 2003) réduisent le nombre de motifs en ne générant que les *motifs clos*. Cependant, le nombre de motifs reste généralement trop élevé, avec beaucoup de redondance entre motifs. Les algorithmes *s'appuyant sur les contraintes* comme gPrune (Zhu et al., 2007), essayent quant à eux de réduire la quantité de motifs générés en n'extrayant que des motifs qui suivent une certaine règle d'acceptation. Ces algorithmes réussissent généralement à limiter le nombre