SPoID : Extraction de motifs séquentiels pour les bases de données incomplètes

Céline Fiot, Anne Laurent, Maguelonne Teisseire

Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier
{fiot, laurent, teisseire}@lirmm.fr

Résumé. Les bases de données issues du monde réel contiennent souvent de nombreuses informations non renseignées. Durant le processus d'extraction de connaissances dans les bases de données, une phase de traitement spécifique de ces données est souvent nécessaire, permettant de les supprimer ou de les compléter. Lors de l'extraction de séquences fréquentes, ces données incomplètes sont la plupart du temps occultées. Ceci conduit parfois à l'élimination de plus de la moitié de la base et l'information extraite n'est plus représentative. Nous proposons donc de ne plus éliminer les enregistrements incomplets, mais d'utiliser l'information partielle qu'ils contiennent. La méthode proposée ignore en fait temporairement certaines données incomplètes pour les séquences recherchées. Les expérimentations sur jeux de données synthétiques montrent la validité de notre proposition aussi bien en terme de qualité des motifs extraits que de robustesse aux valeurs manquantes.

1 Introduction

Les données issues du monde réel sont souvent entâchées d'imperfections. En particulier, il est très courant de disposer de nombreuses données incomplètes (pannes, erreur de format, oubli humain, ...). Or la présence de valeurs manquantes induit de très sérieux problèmes, les données contenant des valeurs manquantes étant souvent éliminées lors du processus de fouille de données. C'est notamment le cas pour l'extraction de motifs séquentiels. Cette technique de fouille de données, présentée comme une extension des règles d'association prenant en compte l'information temporelle des bases de données historisées, ne permet en effet que l'analyse des données complètes, sans tenir compte des enregistrements incomplets, ce qui constitue une grande perte d'information. Par ailleurs, les solutions de remplacement des valeurs manquantes sont souvent soit trop simplistes pour produire des résultats intéressants, soit trop coûteuses pour être mises en oeuvre sur de gros volumes de données.

Or, s'il existe à ce jour des techniques robustes aux valeurs manquantes pour l'extraction de règles d'association, il n'existe aucune méthode générique pour l'extraction de motifs séquentiels. En effet, dans le contexte de la recherche de motifs séquentiels, les valeurs manquantes n'ont pas été considérées jusqu'ici, l'application principale, les bases de données de supermarchés, n'en comportant quasiment jamais. Désormais, les motifs séquentiels sont utilisés afin d'extraire des connaissances d'applications industrielles (analyse de processus, web access logs, ...) qui contiennent inévitablement des données incomplètes.