Extraction des itemsets fréquents à partir de données évidentielles : application à une base de données éducationnelles

Mohamed Anis Bach Tobji *, Boutheina Ben Yaghlane**

Laboratoire LARODEC, Université de Tunis, Institut Supérieur de Gestion
41 avenue de la Liberté, Cité Bouchoucha, Le Bardo 2000, Tunisie
*anis.bach@isg.rnu.tn **boutheina.yaghlane@ihec.rnu.tn

Résumé. Dans cet article, nous étudions le problème de l'extraction des itemsets fréquents (EIF) à partir de données imparfaites, et plus particulièrement ce qu'on appelle désormais les données *évidentielles*. Une base de données évidentielle stocke en effet des données dont l'imperfection est modélisée via la théorie de l'évidence. Nous introduisons une nouvelle approche d'EIF qui se base sur une structure de données de type arbre. Cette structure est adaptée à la nature complexe des données. La technique que nous avons conçue, génère jusqu'à 50% de la totalité des itemsets fréquents lors du premier parcours de l'arbre. Elle a été appliquée sur des bases de données synthétiques ainsi que sur une base de données éducationnelles. Les expérimentations menées sur la nouvelle méthode, montrent qu'elle est plus performante en terme de temps d'exécution en comparaison avec les méthodes existantes d'EIF.

1 Introduction

L'extraction des itemsets fréquents (EIF) à partir des données (Agrawal et al., 1993) a reçu une attention particulière de la part des chercheurs puisqu'elle constitue l'étape coeur de plusieurs méthodes de fouille de données à l'instar de l'extraction des règles d'associations, la construction de certains classifieurs, l'extraction des séries temporelles et d'autres techniques de fouille de motifs (dits aussi patterns). Néanmoins, la majorité des techniques d'EIF (Agrawal et Srikant, 1994; Han et al., 2000; Lucchese et al., 2003; Zhao et Bhowmick, 2003) ne prennent pas en considération l'aspect imparfait des données générées par les applications du monde réel. Ce problème d'imperfection touche en effet plusieurs systèmes d'informations. En sciences expérimentales, les chercheurs se basent sur des expérimentations dont les données générées sont sauvegardées et par la suite traitées. Ces valeurs observées ou mesurées sont la plupart du temps imprécises ou incertaines (Kwan et al., 1996). En médecine, les docteurs se trouvent souvent dans l'obligation d'émettre un diagnostic en présence de symptômes imprécis voire incertains (Konias et al., 2005). Les données générées par certains systèmes à base de capteurs sont aussi imparfaites. Les capteurs d'un système unique peuvent produire des informations à différents niveaux de confiance. En plus, chacun d'eux peut produire une information incertaine, imprécise ou incomplète (Vaughn et al., 2005). L'imperfection des données,