Abstract

The association rule mining (ARM) is one of the core techniques in the field of data mining which identifies and analyses the hidden interesting relationships between the set of items in a large database. The discovered association rules are useful to take business decisions in many application areas such as e-commerce, financial data analysis, medical diagnosis, etc. The quality of an association rule in conventional ARM is typically based on support and confidence measures. However, the usefulness of the results obtained using these measures is limited by the quality and quantity of the generated rules. As a result, many of them are redundant to other rules, and they are practically useless from the business point of view. To address the problems with the quality and quantity of the generated rules, many models have been proposed based on different interestingness measures, logic-based models and constraint-based models, depending on application domains and the nature of the dataset. Although these models improve the quality of the discovered rules and decrease the number of rules but they cannot eliminate the redundancy. To address this problem, a determined effort focused on defining a reduced set of rules from which all redundant rules can be generated without losing any information. The rules in the reduced set are known as non-redundant association rules, which are alternatively called concise or condensed representation of association rules. A concise representation is essential as it only keeps the non-redundant rules that cover all association rules while preserving information as much as possible. Several concise representations for conventional ARM have been proposed in the literature, mainly by defining the notion of redundancy and the corresponding method to discover them. However, the existence of redundancy would logically obtain according to the sense of minimal knowledge based on some order relation. Moreover, the rules obtained using logic-based models and constraint-based models contain redundancy. This thesis deals with both the conceptual and algorithmic aspects of concise representations of association rules with and without using utility constraint. The conceptual aspects aimed at providing the notion of redundancy and the algorithmic aspects concentrate on the design of scalable, efficient algorithms to discover them in three rules discovery framework: support-confidence framework, coherent rules mining framework, and utility based itemset mining. Several pruning strategies to reduce the itemset search space are proposed that accomplish the mining efficiency of the methods. The evaluation of methods is performed by carrying out several experiments conducted on real, dense, and sparse benchmark datasets, mostly used in other existing mining approaches reported in the literature. The obtained results demonstrate that the proposed methods in each framework outperform the other existing related methods in terms of execution time and compactness of the non-redundant rule set.

Keywords: Data mining, Association rule mining, Concise representation, Frequent itemset, Frequent closed itemset, Utility mining, High utility itemset, Non-redundant association rules, Coherent rules, Generators.