ABSTRACT

The mining of fuzzy association rules has been proposed in the literature recently. Many of the ensuing algorithms are developed to make use of only a single processor or machine. They can be further enhanced by taking advantage of the scalability of parallel or distributed computer systems. The increasing ability to collect data and the resulting huge data volume make the exploitation of parallel or distributed systems become more and more important to the success of fuzzy association rule mining algorithms. This chapter proposes a new distributed algorithm, called DFARM, for mining fuzzy association rules from very large databases. Unlike many existing algorithms that adopt the support-confidence framework such that an association is considered interesting if it satisfies some user-specified minimum percentage thresholds, DFARM embraces an objective measure to distinguish interesting associations from uninteresting ones. This measure is defined as a function of the difference in the actual and the expected number of tuples characterized by different linguistic variables (attributes) and linguistic terms (attribute values). Given a database, DFARM first divides it into several horizontal partitions and assigns them to different sites in a distributed system. It then has each site scan its own database partition to obtain the number of tuples characterized by different linguistic variables and linguistic terms (i.e., the local counts), and exchange the local counts with all the other sites to find the global counts. Based on the global counts, the values of the interestingness measure are computed, and the sites can uncover interesting associations. By repeating this process of counting, exchanging counts, and calculating the interestingness measure, it unveils the underlying interesting associations hidden in the data. We implemented DFARM in a distributed system and used a popular benchmark data set to evaluate its performance. The results show
A Distributed Algorithm for Mining Fuzzy Association Rules in Traditional Databases

INTRODUCTION

Of the many different kinds of patterns that can be discovered in a database, the mining of association rules has been studied extensively in the literature (see, e.g., J. Han & Kamber, 2001; Hand, Mannila, & Smyth, 2001). It is because the uncovering of the underlying association relationships (or simply associations) hidden in the data can enable other important problems, such as classification (Au & Chan, 2001, 2003; Carrasco, Vila, Galindo, & Cubero, 2000; Chan & Au, 1997, 2001; Delgado, Marín, Sánchez, & Vila, 2003; Hirota & Pedrycz, 1999; Hong, Kuo, & Chi, 1999; Kuok, Fu, & Wong, 1998; Maimon, Kandel, & Last, 1999; Yager, 1991; Zhang, 1999). The association rules involving fuzzy sets are commonly known as fuzzy association rules.

An example of a fuzzy association rule is given in the following:

\[ \text{Gender} = \text{Female} \land \text{Age} = \text{Young} \land \text{Income} = \text{Small} \rightarrow \text{Occupation} = \text{Cashier}, \]

where \( \text{Gender} \), \( \text{Age} \), \( \text{Income} \), and \( \text{Occupation} \) are linguistic variables, and \( \text{Female} \), \( \text{Young} \), \( \text{Small} \), and \( \text{Cashier} \) are linguistic terms. This rule states that a young woman whose income is small is likely a cashier. In comparison to its counterpart involving discretized intervals, it is easier for human users to understand. The use of fuzzy sets also buries the boundaries of the adjacent intervals. This makes fuzzy association rules resilient to the inherent noise present in the data, for instance, the inaccuracy in physical measurements of real-world entities.

Many of the ensuing algorithms, including those in (Au & Chan, 2001, 2003; Chan & Au, 1997, 2001; Delgado et al., 2003; Hong et al., 1999; Kuok et al., 1998; Zhang, 1999), are developed to make use of only a single processor or machine. They can be that it has very good size-up, speedup, and scale-up performance. We also evaluated the effectiveness of the proposed interestingness measure on two synthetic data sets. The experimental results show that it is very effective in differentiating between interesting and uninteresting associations.
Related Content

ERP Systems Impact on Organizations
[www.igi-global.com/chapter/erp-systems-impact-organizations/18552?camid=4v1a](www.igi-global.com/chapter/erp-systems-impact-organizations/18552?camid=4v1a)

Similarity Learning in GIS: An Overview of Definitions, Prerequisites and Challenges
[www.igi-global.com/chapter/similarity-learning-gis/29669?camid=4v1a](www.igi-global.com/chapter/similarity-learning-gis/29669?camid=4v1a)

SQL Code Poisoning: The Most Prevalent Technique for Attacking Web Powered Databases
[www.igi-global.com/chapter/sql-code-poisoning/8025?camid=4v1a](www.igi-global.com/chapter/sql-code-poisoning/8025?camid=4v1a)

The Quality of Data Representations Developed by NonExpert Designers: An Experimental Study
[www.igi-global.com/article/quality-data-representations-developed-nonexpert/51125?camid=4v1a](www.igi-global.com/article/quality-data-representations-developed-nonexpert/51125?camid=4v1a)