Chapter 8.4
A Distributed Algorithm for Mining Fuzzy Association Rules in Traditional Databases

Wai-Ho Au
Microsoft Corporation, USA

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 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.

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; Liu, Hsu, & Ma, 1998), to be tackled more effectively.

The problem of discovering interesting associations in databases is originally defined over binary or Boolean data (Agrawal, Imielinski, & Swami, 1993). It is then extended to cover many real-life databases comprising both discrete- and continuous-valued data (Srikant & Agrawal, 1996). An association that is considered interesting is typically expressed in the form of rule $X \rightarrow Y$, where X and Y are conjunctions of conditions. A condition is either $A_i = a_i$, where a_i is a value in the domain of attribute A_i if A_i is discrete, or $a_i \in [l_i, u_i]$, where l_i and u_i are values in the domain of attribute A_i if A_i is continuous. The association rule $X \rightarrow Y$ holds with support, which is defined as the percentage of tuples satisfying X and Y, and confidence, which is defined as the percentage of tuples satisfying Y given that they also satisfy X.

An example of an association rule is

$$Gender = Female \wedge Age \in [20, 25] \wedge Income \in [15\,000, 20\,000] \rightarrow Occupation = Cashier,$$

which describes that a woman whose age is between 20 and 25 and whose income is between $15,000 and $20,000 is likely a cashier.

To handle continuous attributes, many data mining algorithms (e.g., Liu et al., 1998; Srikant & Agrawal, 1996) require their domains to be discretized into a finite number of intervals. These intervals may not be concise and meaningful enough for human experts to obtain comprehensive knowledge from the discovered association rules. Instead of using intervals, many researchers propose to employ fuzzy sets to represent the underlying relationships hidden in the data (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:

$$Gender = Female \wedge Age = Young \wedge Income = Small \rightarrow Occupation = Cashier,$$

where *Gender*, *Age*, *Income*, and *Occupation* are linguistic variables, and *Female*, *Young*, *Small*, and *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 further enhanced by taking advantage of the scalability of parallel or distributed computer systems. Because of the increasing ability to collect data and the resulting huge data volume, the exploitation of