Chapter 50
Privacy Preserving Fuzzy Association Rule Mining in Data Clusters Using Particle Swarm Optimization

Sathiyapriya Krishnamoorthy
PSG College of Technology, Department of Computer Science & Engineering, Tamil Nadu, India

M. Rajalakshmi
Coimbatore Institute of Technology, Department of Computer Science & Engineering, Tamil Nadu, India

G. Sudha Sadasivam
PSG College of Technology, Department of Computer Science & Engineering, Tamil Nadu, India

K. Kowsalyaa
PSG College of Technology, Department of Computer Science & Engineering, Tamil Nadu, India

M. Dhivya
SSN College of Engineering, Department of Computer Science & Engineering, Tamil Nadu, India

ABSTRACT

An association rule is classified as sensitive if its thread of revelation is above certain confidence value. If these sensitive rules were revealed to the public, it is possible to deduce sensitive knowledge from the published data and offers benefit for the business competitors. Earlier studies in privacy preserving association rule mining focus on binary data and has more side effects. But in practical applications the transactions contain the purchased quantities of the items. Hence preserving privacy of quantitative data is essential. The main goal of the proposed system is to hide a group of interesting patterns which contains sensitive knowledge such that modifications have minimum side effects like lost rules, ghost rules, and number of modifications. The proposed system applies Particle Swarm Optimization to a few clusters of particles thus reducing the number of modification. Experimental results demonstrate that the proposed approach is efficient in terms of lost rules, number of modifications, hiding failure with complete avoidance of ghost rules.

DOI: 10.4018/978-1-7998-8954-0.ch050

INTRODUCTION

Data or knowledge mining aims at discovering unknown associations among large amount of data items in databases. Association rule mining is a data mining technique to find frequent patterns, associations, correlations among set of items or objects in transactional databases. An association rule is an implication of the form X®Y, where both X and Y are set of attributes (items) from the database. X is called as the body (Left Hand Side (LHS)) of the rule and Y is called as the head (Right Hand Side (RHS)) of the rule. For example, an association rule in a market data may be defined as, In 50% of the transactions, 65% of the people buying pen also buy ink in the same transaction; 50% and 65% represent the support and the confidence, respectively. The importance of an association rule is measured by its support and confidence. Simply, Support is the percentage of number of transactions that contain both X and Y. Confidence is the ratio of the support of XUY to the support of X (Sathiyapriya, Sudhasadasivam & Suganya, 2014).

The famous Apriori algorithm based on the concept of frequent item sets to mine association rules in transaction data was proposed (Srikant & Agrawal, 1995). The apriori algorithm generates large number of candidate item set. The Frequent-Pattern-tree structure (FP-tree) efficiently mines association rules without generation of candidate item sets (Han, & Fu, 1995). Many variations of FP-tree structure and Apriori algorithm have been proposed. The weighted mining was proposed to reflect the importance of different items. Each item was given a numerical value called weight assigned by users. But almost all the algorithms proposed were for binary dataset.

But most databases in real world contains numerical, categorical and integer values. The binary algorithm cannot be applied directly. One way of mining quantitative rules is to treat them like categorical attributes and generate rules for all potential values. This would result in the explosion of number of rules generated and also specific numerical value will not appear frequently. So the domain of each quantitative attribute is divided into intervals and rules are generated from these intervals. This is called discretization (Srikant & Agrawal, 1996). Choosing intervals for numeric attributes is sensitive to support and confidence measures.

As it is possible that the data set may be skewed, intervals cannot be formed randomly. It was shown that if the range of the attribute is divided into equal intervals it leads to two problems of minsupport and minconfidence (Srikant R & Agrawal R, 1996). If we choose to have a large number of small intervals, then the support of some of the interval becomes low. This is called "Minsupport" issue. Building larger intervals results in information loss and rules mined are different from that in original data. This is called "Minconfidence" issue. Another problem with discretization is sharp boundary problem. For example, consider the rule, if experience >=2 and bonus points >= 5000 then credit = approved. If a customer has a job for two years and bonus point of 4,990 then the application for credit approval may be rejected. Such a precise cut-off seems unfair. So, in order to avoid the sharp boundary problem, the data is fuzzified.

In this example, the bonus point can be discretized into categories like low, medium, high and fuzzy logic can be applied to allow fuzzy threshold or boundaries to be defined for each category. Unlike the crisp sets where an element either belongs to a set S or its complement, in fuzzy set theory, each element can belong to more than one fuzzy set. In this example, the bonus point 4,990 belongs to both medium and high fuzzy sets, but to different degrees. In order to avoid sharp boundary problem numerous fuzzy mining approaches have been proposed to mine interesting association rules from quantitative values.