Privacy Protection in Association Rule Mining

Neha Jha
Indian Institute of Technology, Kharagpur, India

Shamik Sural
Indian Institute of Technology, Kharagpur, India

INTRODUCTION

Data mining technology has emerged as a means for identifying patterns and trends from large sets of data. Mining encompasses various algorithms, such as discovery of association rules, clustering, classification and prediction. While classification and prediction techniques are used to extract models describing important data classes or to predict future data trends, clustering is the process of grouping a set of physical or abstract objects into classes of similar objects.

Alternatively, association rule mining searches for interesting relationships among items in a given data set. The discovery of association rules from a huge volume of data is useful in selective marketing, decision analysis and business management. A popular area of application is the "market basket analysis," which studies the buying habits of customers by searching for sets of items that are frequently purchased together (Han & Kamber, 2003).

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of items and D be a set of transactions, where each transaction T belonging to D is an itemset such that T is a subset of I. A transaction T contains an itemset A if A is a subset of T. An itemset A with k items is called a k-itemset. An association rule is an implication of the form A => B where A and B are subsets of I and $A \cap B$ is null. The rule A => B is said to have a support s in the database D if s% of the transactions in D contain AUB. The rule is said to have a confidence c if c% of the transactions in D that contain A also contain B. The problem of mining association rules is to find all rules whose support and confidence are higher than a specified minimum support and confidence. An example association rule could be: Driving at night and speed > 90 mph results in a road accident, where at least 70% of drivers meet the three criteria and at least 80% of those meeting the driving time and speed criteria actually have an accident.

While data mining has its roots in the traditional fields of machine learning and statistics, the sheer volume of data today poses a serious problem. Traditional methods typically make the assumption that the data is memory resident. This assumption is no longer tenable and implementation of data mining ideas in high-performance parallel and distributed computing environments has become crucial for successful business operations. The problem is not simply that the data is distributed, but that it must remain distributed. For instance, transmitting large quantities of data to a central site may not be feasible or it may be easier to combine results than to combine the sources. This has led to the development of a number of distributed data mining algorithms (Ashrafi et al., 2002). However, most of the distributed algorithms were initially developed from the point of view of efficiency, not security. Privacy concerns arise when organizations are willing to share data mining results but not the data. The data could be distributed among several custodians, none of who are allowed to transfer their data to another site. The transaction information may also be specific to individual users who do not appreciate the disclosure of individual record values. Consider another situation. A multinational company wants to mine its own data spread over many countries for globally valid results while national laws prevent trans-border data sharing.

Thus, a need for developing privacy preserving distributed data mining algorithms has been felt in recent years. This keeps the information about each site secure, and at the same time determines globally valid results for taking effective business decisions. There are many variants of these algorithms depending on how the data is distributed, what type of data mining we wish to do, and what restrictions are placed on sharing of information. This paper gives an overview of the different approaches to privacy protection and information security in distributed association rule mining. The following two sections describe the traditional privacy preserving methods along with the pioneering work done in recent years. Finally, we discuss the open research issues and future directions of work.

BACKGROUND

A direct approach to data mining over multiple sources would be to run existing data mining tools at each site independently and then combine the final results. While it is simple to implement, this approach often fails to give

globally valid results since a rule that is valid in one or more of the individual locations need not be valid over the entire data set.

Efforts have been made to develop methods that perform local operations at each site to produce intermediate results, which can then be used to obtain the final result in a secure manner. For example, it can be easily shown that *if a rule has support > m% globally, it must have support > m% on at least one of the individual sites*. This result can be applied to the distributed case with horizontally partitioned data (all sites have the same schema but each site has information on different entities). A distributed algorithm for this would work by requesting each site to send all rules with support at least *m*. For each rule returned, the sites are then asked to send the count of their items that support the rule, and the total count of all items at the site. Using these values, the global support of each rule can be computed with the assurance that all rules with support at least *m* have been found.

This method provides a certain level of information security since the basic data is not shared. However, the problem becomes more difficult if we want to protect not only the individual items at each site, but also how much each site supports a given rule. The above method reveals this information, which may be considered to be a breach of security depending on the sensitivity of any given application. Theoretical studies in the field of secure computation started in the late 1980s. In recent years, the focus has shifted more to the application field (Maurer, 2003). The challenge is to apply the available theoretical results in solving intricate real-world problems. Du & Atallah (2001) review and suggest a number of open secure computation problems including applications in the field of computational geometry, statistical analysis and data mining. They also suggest a method for solving secure multi-party computational geometry problems and secure computation of dot products in separate pieces of work (Atallah & Du, 2001; Ioannidis et al., 2002).

In all the above-mentioned work, the secure computation problem has been treated with an approach to providing absolute zero knowledge whereas the corporations may not always be willing to bear the cost of zero information leakage as long as they can keep the information shared within known bounds. In the next section, we discuss some of the important approaches to privacy preserving data mining with an emphasis on the algorithms developed for association rule mining.

MAIN THRUST

In this section, we describe the various levels of privacy protection possible while mining data and the corresponding algorithms to achieve the same. Identification of privacy concerns in data mining has led to a wide range of proposals in the past few years. The solutions can be broadly categorized as those belonging to the classes of data obfuscation, data summarization and data separation. The goal of data obfuscation is to hide the data to be protected. This is achieved by perturbing the data before delivering it to the data miner by either randomly modifying the data, swapping the values between the records or performing controlled modification of data to hide the secrets. Cryptographic techniques are often employed to encrypt the source data, perform intermediate operations on the encrypted data and then decrypt the values to get back the final result with each site not knowing anything but the global rule. Summarization, on the other hand, attempts to make available innocuous summaries of the data and therefore only the needed facts are exposed. Data separation ensures that only the trusted parties can see the data by making all operations and analysis to be performed either by the owner/creator of the data or by trusted third parties.

One application of data perturbation technique is decision tree based classification to protect individual privacy by adding random values from a normal/Gaussian distribution of mean 0 to the actual data values (Agrawal & Srikant, 2000). Bayes' rule for density functions is then used to reconstruct the distribution. The approach is quite elegant since it provides a method for approximating the original data distribution and not the original data values by using the distorted data and information on the random data distribution. Similar data perturbation techniques can be applied to the mining of Boolean association rules also (Rizvi & Haritsa, 2002). It is assumed that the tuples in the database are fixed length sequences of 0's and 1's. A typical example is the market basket application where the columns represent the items sold by a supermarket, and each row describes, through a sequence of 1's and 0's, the purchases made by a particular customer (1 indicates a purchase and 0 indicates no purchase). One interesting feature of this work is a flexible definition of privacy; for example, the ability to correctly guess a value of "1" from the perturbed data can be considered a greater threat to privacy than correctly learning a "0."

For many applications such as market basket, it is reasonable to expect that the customers would want more privacy for the items they buy compared to the items they do not. There are primarily two ways of handling this requirement. In one method, the data is changed or perturbed to a certain extent to hide the exact information that can be extracted from the original data. In another approach, data is encrypted before running the data mining algorithms on it. While data perturbation techniques usually result in a transformation that leads to loss of information and the exact result cannot be determined,