Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous Database

Ashraf El-Sisi
Faculty of Computers and Information, Menofyia University, Egypt

Abstract: *Privacy is one of the most important properties of an information system must satisfy, in which systems the need to share information among different, not trusted entities, the protection of sensible information has a relevant role. A relatively new trend shows that classical access control techniques are not sufficient to guarantee privacy when data mining techniques are used in a malicious way. Privacy preserving data mining algorithms have been recently introduced with the aim of preventing the discovery of sensible information. In this paper we propose a modification to privacy preserving association rule mining on distributed homogenous database algorithm. Our algorithm is faster than old one which modified with preserving privacy and accurate results. Modified algorithm is based on a semi-honest model with negligible collision probability. The flexibility to extend to any number of sites without any change in implementation can be achieved. And also any increase doesn't add more time to algorithm because all client sites perform the mining in the same time so the overhead in communication time only. The total bit-communication cost for our algorithm is function in (N) sites.*

Keywords: *Association rule mining, Apriori, cryptography, distributed data mining, privacy, security.*

Received July 9, 2008; accepted November 25, 2008

1. Introduction

Privacy preserving data mining is an important property that any mining system must satisfy. So far, if we assumed that the information in each database found in mining can be freely shared. However, there is now an increasing need for computing association rules across databases belonging to sites in such a way that no more information than necessary is revealed from each database to the other databases that only every site knows its input and final mining results. This need is driven by several trends like security, government agencies need to share information for devising effective security measures, within the same government and across governments. However, an agency cannot indiscriminately open up its database to all other agencies. Also privacy, privacy legislation and stated privacy policies place limits on information sharing. However, it is still desirable to mine across databases while respecting privacy limits.

There are many methods for privacy preserving distributed association rule mining across private databases. These methods try to compute the answer to the mining without revealing any additional information about user privacy. An application that needs privacy preserving distributed association rule mining across private databases, like medical research; imagine a future where many people have their DNA sequenced. A medical researcher wants to validate a hypothesis connecting a DNA sequence D with a reaction to drug G. People who have taken the drug are partitioned into four groups, based on whether or not they had an adverse reaction and whether or not their DNA contained the specific sequence; the researcher needs the information about drug effectiveness on people in each group. DNA sequences and medical histories are stored in databases in autonomous enterprises. Due to privacy concerns, the enterprises do not wish to provide any information about an individual's DNA sequence or medical history, but still wish to help with the research to mine the databases for some pattern help in drug industry. We want the property that the researcher should get to know the association rules about data and nothing else, and any enterprise should not learn any new information about any individual from other enterprise. There are some existing techniques that one might use for building this application, but they are inadequate related to some disadvantages. One from these techniques is trusted third party. The main parties give the data to a "trusted" third party and have the third party do the computation [3]. However, the third party has to be completely trusted, both with respect to intent and competence against security breaches. The level of trust required is too high for this solution to be acceptable.

Data perturbation technique has different idea, the idea is that the distorted data does not reveal private information, and thus is "safe" to use for mining. The key result is that the distorted data, and information on

the distribution of the random data used to distort the data, can be used to generate an approximation to the original data distribution, without revealing the original data values. The distribution is used to improve mining results over mining the distorted data directly, primarily through selection of split points to "bin" continuous data. Later refinement of this approach tightened the bounds on what private information is disclosed, by showing that the ability to reconstruct the distribution can be used to tighten estimates of original values based on the distorted data [2]. Another approach is secure multi-party computation. In this approach given two parties with inputs x and y respectively, the goal of secure multi-party computation is to compute a function f(x, y) such that the two parties learn only f(x, y), and nothing else. In [8] there are various approaches to this problem. In [9] an efficient protocol for Yao's millionaires' problem showed that any multi-party computation can be solved by building a combinatorial circuit, and simulating that circuit. A variant of Yao's protocol is presented in [11] where the oblivious transfers is used to make secure decision tree learning using ID3 with efficient cryptographic protocol there are also two solution of our problem under the secure multi party computation for association rule mining in [7, 10]. In this paper we addresses the problem of computing association rules when databases belonging to sites and each site needing preserving the privacy of users data in databases. We assume homogeneous databases: All sites have the same schema, but each site has information on different entities. The goal is to produce a modification to algorithm in [7] that computes association rules that hold globally while limiting the information shared about each site in order to increase the efficiency of the algorithm.

The organization of this paper is as follows. Section 2 gives an overview about the problem and the related work in the area of privacy preserving association rule mining on distributed homogenous databases. In section 3 the details of the modification for the algorithm of computing the distributed association rule mining to preserve the privacy of users. Section 4 describes implementation and results of our new algorithm verse the old algorithm. Finally, some conclusions are put forward in section 5.

2. Distributed Association Rule Mining Problem and Related Work

Association Rule mining is one of the most important data mining tools used in many real life applications. It is used to reveal unexpected relationships in the data. In this section, we will discuss the problem of computing association rules within a horizontally partitioned database. We assume homogeneous databases. All sites have the same schema, but each site has information on different entities. The goal is to produce association rules that hold globally, while limiting the information shared about each site to preserve the privacy of data in each site.

2.1. Association Rule Mining

Association rule mining finds interesting associations and/or correlation relationships among large sets of data items. Association rules show attributes value conditions that occur frequently together in a given dataset. A typical and widely-used example of association rule mining is market basket analysis. For example, data are collected using barcode scanners in supermarkets. Such market basket databases consist of a large number of transaction records. Each record lists all items bought by a customer on a single purchase transaction. Managers would be interested to know if certain groups of items are consistently purchased together. They could use this data for adjusting store layouts (placing items optimally with respect to each other), for cross-selling, for promotions, for catalog design and to identify customer segments based on buying patterns. Association rules provide information of this type in the form of "if-then" statements. These rules are computed from the data and, unlike the if-then rules of logic, association rules are probabilistic in nature. In association analysis the antecedent and consequent are sets of items (called item-sets) that are disjoint (do not have any items in common). In addition to the antecedent ("if" part) and the consequent ("then" part), an association rule has two numbers that express the degree of uncertainty about the rule. The first number is called the support for the rule. The support is simply the number of transactions that include all items in the antecedent and consequent parts of the rule (the support is sometimes expressed as a percentage of the total number of records in the database). The other number is known as the confidence of the rule. Confidence is the ratio of the number of transactions that include all items in the consequent as well as the antecedent (namely, the support) to the number of transactions that include all items in the antecedent.

In [2] the association rules mining problem can formally be defined as follows: let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of items. Let DB be a set of transactions, where each transaction T is an itemset such that T