AN EFFICIENT FDM-KC ALGORITHM FOR SECURE MINING IN HORIZONTALLY DISTRIBUTED DATABASES USING ASSOCIATION RULES

C M Sumana[1], Rajeswari R P[2]

[1]B.E., M.Tech, [2]Assistant Professor, Dept. of Computer Science and Engineering, RYMEC, Ballari (India)

ABSTRACT

In recent years, Data mining is the most fast growing area in which data mining is the process of finding correlations or patterns among various fields in large relational databases in which it is used to extract important knowledge from large datasets, but sometimes these datasets are split among various parties.We propose a protocol for secure mining of association rules in horizontally distributed databases. The current integral protocol is that of Kantarcioglu and Clifton well known as K&C protocol. This proposed protocol is based on an unsecured distributed version of the Apriori algorithm termed as Fast Distributed Mining (FDM) algorithm of Cheung et al.

The main constituents in this protocol are two novel secure multi-party algorithms one that computes the union of private subsets that each of the interacting players hold and another that tests the whether an element held by one player is included in a subset held by another. This protocol offers enhanced privacy with respect to the protocol. This is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost.

Keywords: Privacy Preserving Data Mining, Distributed Computation, Frequent Item sets, Association Rules.

I. INTRODUCTION

Data mining is the extraction of interesting patterns or knowledge from huge amount of data. Today, we have far more information than we must handle from business transactions and scientific data, to satellite pictures, text reports and military intelligence. Thus, Data mining is the computational process of discovering patterns in large data sets. In simple words, it is the process of analyzing data from different perspectives and summarizing it into useful information. A distributed database can reside on network servers on theInternet, on corporate internets or extranets, or on other company networks. Because they store data across multiple computers, distributed databases can improve performance at end-user worksites by allowing transactions to be processed on many machines, instead of being limited to one.

Homogeneous distributed database all sites have identical software and are aware of each other and agree to cooperate in processing user requests. Each site surrenders part of its autonomy in terms of right to change schema or software. A homogeneousDDBMS(Distributed database management system)appears to the user as a single system. The homogeneous system is much easier to design and manage. The following conditions must be satisfied for homogeneous database:

1. The operating system is used, at each location must be same or compatible.
2. The data structures used at each location must be same or compatible.
3. The database application used at each location must be same or compatible.

Data miningis ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

The limitations of frequent or rare itemset mining motivated to develop a secure based mining approach, which allows a user to conveniently express his or her perspectives concerning the usefulness of itemsets as secure values and then find itemsets with high utility values higher than a user-specified threshold.

In the literature we have studied the different methods proposed secure mining from large datasets.That goal defines a problem of secure multi-party computation. In such problems, there are *M* players that hold private inputs, *x*1, . . . , *xM*, and they wish to securely compute $y = f(x1, \ldots, xM)$ for some public function *f*. If there existed a trusted third party, the players could surrender to him their inputs and he would perform the function evaluation and send to them the resulting output. In the absence of such a trusted third party, it is needed to devise a protocol that the players can run on their own in order to arrive at the required output *y*. Such a protocol is considered perfectly secure if no player can learn from his view of the protocol more than what he would have learnt in the idealized setting where the computation is carried out by a trusted third party. Thus to overcome this challenges the efficient algorithm is presented in this paper.

The main aim of this proposed protocol is to achieve the following aspects:

- Reducing the number of scans in the original database.
- Distributed databases remain up-to-date and current replication and duplication.
- Minimize memory utilization (Reducing the search space).
- Reducing the total execution and computation time.
- Reducing the resource utilization.
- Increase the performance in terms of time and space complexity.

The proposed protocol improves upon that in terms of simplicity and efficiency as well as privacy. We propose an alternative protocol for the secure computation of the union of private subsets. In particular, our protocol does not depend on commutative encryption and oblivious transfer (what simplifies it significantly and contributes towards much reduced communication and computational costs). While our solution is still not perfectly secure, it leaks excess information only to a small number (three) of possible coalitions, unlike the protocol of that discloses information also to some single players. In addition, we claim that the excess information that our protocol may leak is less sensitive than the excess information leaked by the protocol.

We propose here computes a parameterized family of functions, which we call threshold functions, in which the two extreme cases correspond to the problems of computing the union and intersection of private subsets. Those are in fact general-purpose protocols that can be used in other contexts as well. Another problem of secure multiparty computation that we solve here as part of our discussion is the set inclusion problem namely, the problem where Alice holds a private subset of some ground set, and Bob holds an element in the ground set, and they wish to determine whether Bob’s element is within Alice’s subset, without revealing to either of them information about the other party’s input beyond the above described inclusion.

II. METHODOLOGY

Process Design

Consider D be a transaction database. The database is partitioned horizontally between P1, P2 . . . , Pm players, denoted 1 M. Player Pm holds the partial database Dm that contains Nm = |Dm | of the transactions in D, $1 \leq m \leq M$. The unified database is $D = D_1 U \cdots U DM$. An itemset X is a subset of A. Its global support, supp(X), is the number of transactions in D that contain it. Its local support, sup (X), is the number of transactions in D_m that contain it.

Support

The rule $X \Rightarrow Y$ holds with support s if s% of transactions in D contains $X \cup Y$. Rules that have a s greater than a user-specified support is said to have minimum support or threshold support. The support of rule is defined as,sup(X) = no of transactions that contain X / total no of Transactions.

Confidence:

The rule $X \Rightarrow Y$ holds with confidence c if c% of the transactions in D that contain X also contain Y. Rules that have a c greater than a user-specified confidence is said to have minimum confidence or threshold Confidence. The confidence of a rule is defined as, conf(X =>Y) = sup(X U Y)/ supp(X).

2.1 Apriori Algorithm

Apriori is designed to operate on databases containing transactions. The purpose of the Apriori Algorithm is to find associations between different sets of data. It is sometimes referred to as "Market Basket Analysis". Each set of data has a number of items and is called a transaction. The output of Apriori is sets of rules that tell us how often items are contained in sets of data.

2.2 Algorithm - Fast Distributed Mining (Fdm)

The FDM algorithm proceeds as follows:

(1) Initialization

(2) Candidate Sets Generation

(3) Local Pruning

(4) Unifying the candidate item sets

(5) Computing local supports

(6) Broadcast Mining Results

III. SYSTEM ARCHITECTURE

Data mining consists of various techniques that are applied for secure mining of association rules in horizontally distributed database. In the figure 1 shows two novel secure multiparty algorithmis to provide enhanced privacy, security, and efficiency. In this paper we propose a protocol for secure mining of association rules in horizontally distributed database. This protocol is based on FDM Algorithm which is an unsecured distributed version of the Apriori algorithm. In this protocol two secure multiparty algorithms are involved:

1. Computes the union of private subsets that each interacting players hold.

2. Tests the inclusion of an element held by one player in subset held by another.

Mapper is a database management and processing system. It is a software tool that enables end-users to share computer power in a corporation. Mapper is mapping between database attributes to the java object.Make the

given itemsets are to be frequent itemsets in horizontally distributed database by using association rules, privacy preserving techniques and FDM algorithm to provide secure mining.

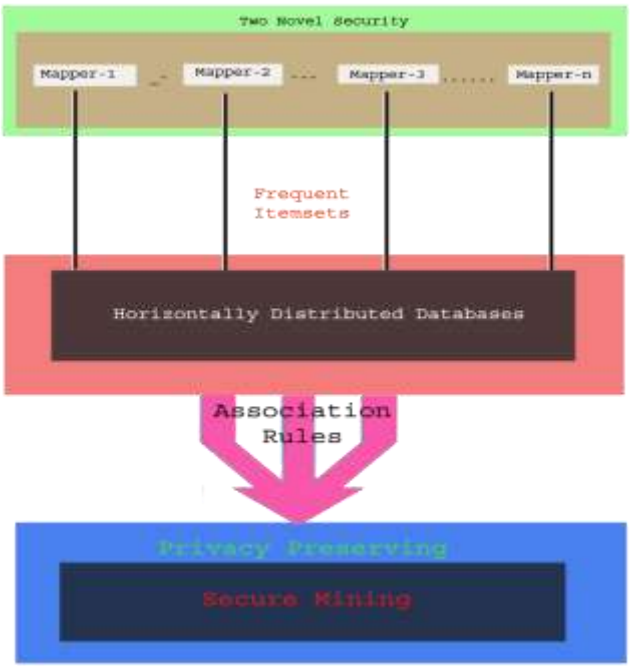

Figure 1: System Architecture for Secure Mining of Association Rules in Horizontally Distributed Data Bases

IV. CONCLUSION

The main problems with the existing methods are the generation a huge set of candidate items and scanning of the original database several times. In this proposed a protocol for secure mining of association rules in horizontally distributed databases that improves significantly upon the current leading protocol in terms ofprivacy and efficiency.In this proposedprotocolhasthepotentialtoovercomeseveral restrictionsinthecurrent data mining model thatcanprevent system'sfunctionalityandlimitations,andmakefurther refinements.

One of the main ingredients in our proposed protocol is a novel secure multi-party protocol for computing the union (or intersection) of private subsets that each of the interacting players hold. Another ingredient is a protocol that tests the inclusion of an element held by one player in a subset held by another. Those protocols exploit the fact that the underlying problem is of interest only when the number of players is greater than two.

Since the algorithm generates few candidate items it takes less time to find the frequent itemsets. And also the memory used is less compared to the existing algorithms. Thus, pruning the itemsets very well at early stages saves the time as well as space.

V. ACKNOWLEDGEMENT

I consider it is a privilege to express my gratitude and respect to all those who guiding me in the progress of my paper.

I wish my grateful thanks to Smt Rajeswari R PM.Tech, project guide, for her invaluable support and guidance.

C M Sumana

REFERENCES

[1]. R. Agrawal and R. Srikant.Fast algorithms for mining association rules in large databases. In VLDB, pages 487–499, 1994.

[2]. R. Agrawal and R. Srikant.Privacy-preserving data mining. In SIGMOD Conference, pages 439–450, 2000.

[3]. D. Beaver, S. Micali, and P. Rogaway. The round complexity of secure protocols. In STOC, pages 503–513, 1990.

[4]. M. Bellare, R. Canetti, and H. Krawczyk. Keying hash functions for message authentication. In Crypto, pages 1–15, 1996.

[5]. A. Ben-David, N. Nisan, and B. Pinkas. FairplayMP - A system for secure multi-party computation. In CCS, pages 257–266, 2008.

[6]. J.C. Benaloh. Secret sharing homomorphisms: Keeping shares of a secret secret. In Crypto, pages 251–260, 1986.

[7]. J. Brickell and V. Shmatikov.Privacy-preserving graph algorithms in the semi-honest model. In ASIACRYPT, pages 236–252, 2005.

[8]. D.W.L. Cheung, J. Han, V.T.Y. Ng, A.W.C. Fu, and Y. Fu. A fast distributed algorithm for mining association rules. In PDIS, pages 31–42, 1996.

[9]. D.W.L Cheung, V.T.Y. Ng, A.W.C. Fu, and Y. Fu. Efficient mining of association rules in distributed databases. IEEE Trans. Knowl. DataEng., 8(6):911–922, 1996.

[10]. T. ElGamal. A public key cryptosystem and a signature scheme based on discrete logarithms.IEEE Transactions on Information Theory, 31:469– 472, 1985.

BIOGRAPHY

C M Sumana,is a student pursuing her Master degree in Computer Science and Engineering department at RYMEC,Ballari, Karnataka, India. Her research interests are Computer Science related aspects such as Data minig technology, Java programming language and web technology.

Rajeswari R P, is an assistant professor in the department of Computer Science and Engineering at RYMEC, Ballari, Karnataka, India. She received her Master degree in computer science and engineering. Her research interests are related to Biomedical image processing and big data analytics in health care.