Cost Estimation Model in Horizontal Partitioned Data Using Secure Association Rule Mining

Yesupogu Sandhya Rani.[1] , Lakshmi Ramani Burra[2] , Praveen Tumuluru[3]

[1]Dept. of CSE, Prasad V Potluri Siddhartha Institute of Technology, Vijayawada, A.P, India

[2]Assistant Professor, Prasad V Potluri Siddhartha Institute of Technology, Vijayawada, A.P, India

[3]Assistant Professor, Prasad V Potluri Siddhartha Institute of Technology, Vijayawada, A.P, India

ABSTRACT: Data mining and data warehousing go hand-in-hand: most tools operate by gathering all data into a central site, then running an algorithm against that data. Traditionally we proposed a protocol Fast Distributed Mining (FDM) algorithm for secure mining of association rules in horizontally distributed databases that improves significantly upon the current leading protocol in terms of privacy and efficiency. FDM protocol computes a parameterized family of functions, which we call threshold functions. In this paper we introduce a method that access services from FDM for satisfying computational and communicational cost process. Our method follows the general approach of the FDM algorithm, with special protocols replacing the broadcasts of LLi (k) and the support count of items in LL (k). It is possible to mine globally valid results from distributed data without revealing information that compromises the privacy of the individual sources. Our experimental results show efficient cost estimation process in both communication and computational costs in secure association rule mining with privacy.

Index Terms: Rule Mining, FDM, Multi Party , Distributed Mining.

2. INTRODUCTION

We consider here the issue of secure mining of affiliation governs in evenly divided databases. In that setting, there are a few locales (or players) that hold homogeneous databases, i.e., databases that have the same mapping yet hold data on distinctive elements. The objective is to discover all affiliation tenets with backing at any rate s and certainty in any event c, for some given insignificant help size s and certainty level c, that hold in the brought together database [1], while minimizing the data uncovered about the private databases held by those players.

Private Association Rule Mining Overview

Our strategy takes after the fundamental methodology plot on Page 2 aside from that values are passed between the nearby information mining destinations as opposed to an unified combiner. The two stages are finding hopeful item sets (those that are successive on one or more destinations), and figuring out which of the competitor item sets meet the worldwide help/certainty edges. The principal stage (Figure 1) uses commutative encryption. Each one gathering scrambles its own particular continuous item sets (e.g., Site 1 encodes item set C). The scrambled item sets are then gone to different gatherings, until all gatherings have encoded all itemsets. These are gone to a typical gathering to take out copies, and to start decoding.

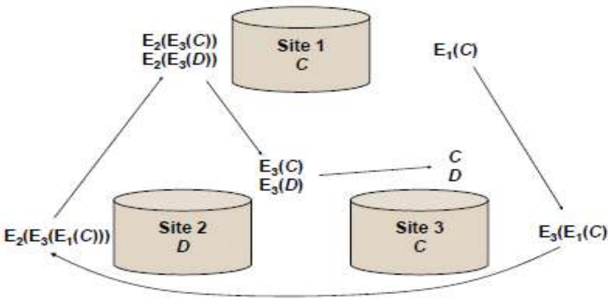

Fig 1: Determining global candidate itemsets

In the second stage (Figure 2), each of the generally underpinned itemsets is tried to check whether it is backed all around. In the figure, the itemset ABC is known to be underpinned at one or more destinations, furthermore each one registers their nearby backing. The principal site picks an irregular worth R, and adds to R the sum by which its backing for ABC surpasses the base help limit. This worth is gone to site 2, which includes the sum by which its backing exceeds the threshold.

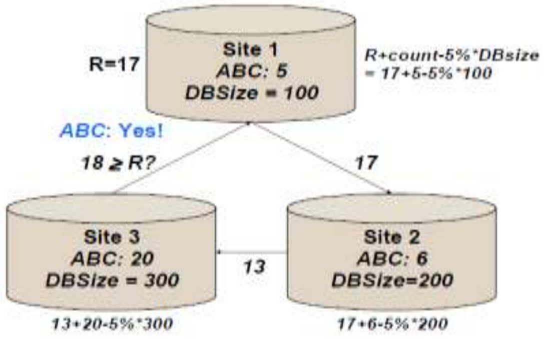

Fig2: **Determining if item set support exceeds 5% threshold**

3 REVIEWS ON LITERATURE

There are a few fields where related work is happening. We first depict other work in protection safeguarding information mining, then go into point of interest on particular foundation deal with which this paper assembles.

Past work in security saving information mining has tended to two issues. In one, the point is safeguarding client protection by mutilating the information values [4]. The thought is that the destroyed information does not uncover private data, and hence is "protected" to use for mining. The key result is that the destroyed information, and data on the dissemination of the arbitrary information used to bend the information, can be utilized to create a rough guess to the first information circulation, without uncovering the first information values. The circulation is utilized to enhance mining comes about over mining the twisted information specifically, basically through determination of part indicates "container" constant information.

All the more as of late, the information twisting methodology has been connected to Boolean affiliation guidelines [6], [7]. Once more, the thought is to change information values such that reproduction of the qualities for any individual exchange is troublesome, yet the tenets adapted on the contorted information are still legitimate. The information twisting methodology addresses an alternate issue from our work.

The other methodology utilizes cryptographic instruments to assemble choice trees. [8] In this work, the objective is to safely construct an ID3 choice tree where the preparation set is conveyed between two gatherings.

Mining of Association Rules

The association rules mining problem can be defined as follows: [1] Let I = {i1, i2, . . . , in} be a set of items. Let DB be a set of transactions, where each transaction T is an item set. In this improved meaning of the affiliation guidelines, missing things, negatives and amounts are not considered. In this admiration, exchange database DB can be seen as 0/1 framework where every segment is a thing and each one column is an exchange. In this paper, we utilize this perspective of affiliation principles.

Distributed Mining of Association Rules: The above problem of mining association rules can be extended to distributed environments.

A fast algorithm for distributed association rule mining is given in Cheung et. al. [2]. Their procedure for fast distributed mining of association rules (FDM) is summarized [2].

Secure Multi-party Computation:

Considerable work has been carried out on secure multi-party reckoning. The key result is that a wide class of processing can be registered safely under sensible suppositions. We give a short diagram of this work, focusing on material that is utilized later as a part of the paper. The definitions given here are from Goldreich [9].

Security in semi-honest model: It takes after the principles of the convention utilizing its right enter, yet is allowed to later utilize what it sees amid execution of the convention to bargain security.

A running example

Let D be a database of $N = 18$ itemsets over a set of $L = 5$ items, $A = \{1, 2, 3, 4, 5\}$. It is partitioned between $M = 3$ players, and the corresponding partial databases are:

$D1 = \{12, 12345, 124, 1245, 14, 145, 235, 24, 24\}$
$D2 = \{1234, 134, 23, 234, 2345\}$
$D3 = \{1234, 124, 134, 23\}$.

For example, $D1$ includes $N1 = 9$ transactions, the third of which (in lexicographic order) consists of 3 items — 1, 2 and 4. Setting $s = 1/3$, an itemset is s-frequent in D if it is supported by at least $6 = sN$ of its transactions. In this case,

$F1s = \{1, 2, 3, 4\}$
$F2s = \{12, 14, 23, 24, 34\}$
$F3s = \{124\}$
$F4s = F5s = \emptyset$,

and $Fs = F1s \cup F2s \cup F3s$. For example, the itemset 34 is indeed globally s-frequent since it is contained in 7 transactions of D. However, it is locally s-frequent only in $D2$ and $D3$.

In the first round of the FDM algorithm, the three players compute the sets $C1,m\ s$ of all 1-itemsets that are locally frequent at their partial databases:

$C1,1\ s = \{1, 2, 4, 5\}$, $C1,2\ s = \{1, 2, 3, 4\}$, $C1,3\ s = \{1, 2, 3, 4\}$. Hence, $C1$
$s = \{1, 2, 3, 4, 5\}$. Consequently, all 1-itemsets have to be checked for being globally frequent; that check reveals that the subset of globally s-frequent 1-itemsets is $F1\ s = \{1, 2, 3, 4\}$.

In the second round, the candidate itemsets are: $C2,1\ s = \{12, 14, 24\}$
$C2,2s = \{13, 14, 23, 24, 34\}$
$C2,3s = \{12, 13, 14, 23, 24, 34\}$.

(Note that 15, 25, 45 are locally s-frequent at $D1$ but they are not included in $C2,1\ s$ since 5 was already found to be globally infrequent.) Hence, $C2\ s = \{12, 13, 14, 23, 24, 34\}$.

Then, after veryfing global frequency, we are left with $F2\ s = \{12, 14, 23, 24, 34\}$.

In the third round, the candidate itemsets are: $C3,1\ s = \{124\}$, $C3,2\ s = \{234\}$, $C3,3\ s = \{124\}$.

So, $C3\ s = \{124, 234\}$ and, then, $F3\ s = \{124\}$. There are no more frequent itemsets.

4. EXISTING SYSTEM

Secure data mining of association rules in horizontally distributed data bases is the major task in transactional data base generation. For doing above process efficiently privacy-preserving keyword search is one of the protocols for accessing efficient association rules in distributed data bases. Past work in protection safeguarding information mining has considered two related settings: In the first setting, the thought is that the bothered information can be utilized to surmise general patterns in the information, without uncovering unique record data. In the second setting, the objective is to perform information mining while ensuring the information records of each of the information holders from the other information managers. Protection necessities in server procedure are not proficient in present working status report era for homogeneous conveyed information bases. Process execution is additionally less in security saving decisive word hunt process applications. So the better system was required for significance distributed data bases efficiently.

PROBLEM STATEMENT

Let i>=3 be the number of sites. Each site has a private transaction database DB*i*. We are given support threshold s and confidence c as percentages. The goal is to discover all association rules satisfying the thresholds, as defined in Section II-A.1. We further desire that disclosure be limited: No site should be able to learn contents of a transaction at any other site, what rules are supported by any other site, or the specific value of support/confidence for any rule at any other site, unless that information is revealed by knowledge of

One's own data and the final result.

4 PROPOSED SYSTEM

We proposed a protocol Fast Distributed Mining (FDM) algorithm for secure mining of affiliation administers in evenly circulated databases that enhances fundamentally upon the current heading convention as far as security and effectiveness. FDM which is an unsecured appropriated form of the Apriori-algorithm. Its fundamental thought is that any s- visit thing set must be additionally generally s-visit in no less than one of the locales. The principle fixings in our convention are two novel secure multi-party calculations, one that figures the union of private subsets that each of the communicating players hold, and an alternate that tests the incorporation of a component held by one player in a subset held by an alternate. Proposed convention does not rely on upon commutative encryption and neglectful [6][7]. FDM convention [8] processes a parameterized group of capacities, which we call edge capacities, in which the two compelling cases relate to the issues of processing the union and convergence of private subsets. In that we specified clear separation between distributed data bases in the form secure association rules generation in data item. But In the improvement of the FDM we don't present the communication and computational cost for transaction with distributed data bases. So the better system was required for during above process efficiently.

The FDM algorithm proceeds as follows:

(1) Initialization: It is assumed that the players have already jointly calculated $Fs(pwr(k-1))$. The goal is to proceed and calculate $Fs(pwr(k))$.

(2) Candidate Sets Generation: Each player Pm computes the set of all $(k-1)$-itemsets that are locally frequent in his site and also globally frequent; namely, Pm computes the set $Fs(pwr(k-1))$, ∩ $Fs(pwr(k-))1$. He then applies on that set the Apriori algorithm in order to generate the set $Bk,m\ s$ of candidate k-itemsets.

(3) Local Pruning: For each $X \in Bks(pwr(m,s))$, Pm computes $suppm(X)$. He then retains only those itemsets that are locally s-frequent. We denote this collection of itemsets by $Cs(pwr(k,m))$.

(4) Unifying the candidate itemsets: Each player broadcasts his Ck,ms and then all players compute $CsPwr(k){:}=\cup$m=1$(Pwr(M)\ \ Ck(Pwr(m,s))$.

(5) Computing local supports. All players compute the local supports of all itemsets in $CsPwr(k,m)$.

(6) Broadcast Mining Results: Each player broadcasts the local supports that he computed. From that, everyone can compute the global support of every itemset in $Cspwr(k)$. Finally, $FsPwr(k)$ is the subset

of *CsPwr(k)* that consists of all globally *s*-frequent *k*-itemsets.

Commutative Encryption: Commutative encryption is an important tool that can be used in many privacy-preserving protocols.

Experimental results

Figure 1 demonstrates the estimations of the three measures that were recorded in Section 6.3 as an issue of N. In those analyses, the estimation of M and s stayed unaltered M = 10 and s = 0.1. Figure 2 demonstrates the estimations of the three measures as an issue of M; here, N = 500, 000 and s = 0.1. Figure 3 demonstrates the estimations of the three measures as an issue of s; here, N = 500, 000 and M 10.

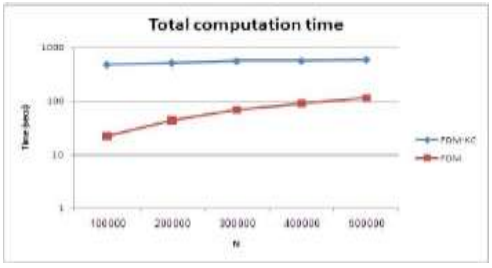

Fig 3: Computation and communication costs versus the number of transactions *N*

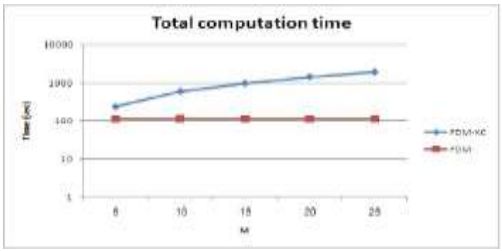

Fig 4: Computation and communication costs versus the number of players *M*.

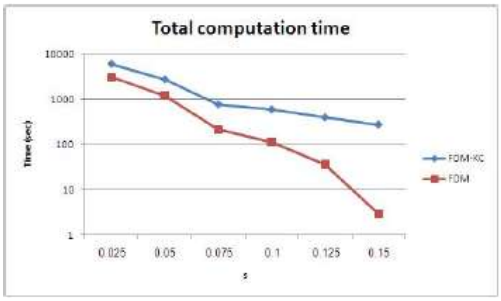

Fig 5: Computation and communication costs versus the support threshold *s*

An alternate answer for the set incorporation issue was as of late proposed in [30], utilizing a convention for negligent polynomial assessment. Data mining technology has emerged as a means of identifying patterns and trends from large quantities of data. Data mining and data warehousing go hand-in-hand: most tools operate by gathering all data into a central site, then running an algorithm against that data. Our method follows the general approach of the FDM algorithm, with special protocols replacing the broadcasts of LLi (k) and the support count of items in LL (k). We now improve the communication and computational costs in distributed association rule mining can be done efficiently under reasonable security assumptions. FDM actually requires an additional factor of N due to the broadcast of local support instead of point-to-point communication. Then FDM efficiently access following things:

- Optimizations in Cost Estimations
- Practical Cost of Encryption

It is possible to mine globally valid results from distributed data without revealing information that compromises the privacy of the individual sources. Our experimental results show efficient cost estimation process in both communication and computational costs in secure association rule mining with privacy.

CONCLUSION:

We proposed a protocol Fast Distributed Mining (FDM) algorithm for secure mining of association rules in evenly appropriated databases that enhances altogether upon the current heading convention regarding protection and effectiveness. FDM convention registers a parameterized group of capacities, which we call limit capacities, in which the two amazing cases compare to the issues of processing the union and crossing point of private subsets. In FDM we don't present the communication and computational cost for transaction with distributed data bases. In this paper we introduce a method that access services from FDM for satisfying computational and communicational cost process. Our method follows the general approach of the FDM algorithm, with special protocols replacing the broadcasts of LLi (k) and the support count of items in LL (k). Our experimental results show efficient cost estimation process in both communication and computational costs in secure association rule mining with privacy.

REFFERENCES:

[1] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proceedings of the 20th International Conference on Very Large Data Bases*. Santiago, Chile: VLDB, Sept. 12-15 1994, pp. 487–499. [Online]. Available: http://www.vldb.org/dblp/db/conf/vldb/ vldb94-487.html

[2] D. W.-L. Cheung, J. Han, V. Ng, A. W.-C. Fu, and Y. Fu, "A fast distributed algorithm for mining association rules," in *Proceedings of the 1996 International Conference on Parallel and Distributed Information Systems (PDIS'96)*. Miami Beach, Florida, USA: IEEE, Dec. 1996, pp. 31–42.

[3] D. W.-L. Cheung, V. Ng, A. W.-C. Fu, and Y. Fu, "Efficient mining of association rules in distributed databases," *IEEE Trans. Knowledge Data Eng.*, vol. 8, no. 6, pp. 911–922, Dec. 1996.

[4] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proceedings of the 2000 ACM SIGMOD Conference on Management of Data*. Dallas, TX: ACM, May 14-19 2000, pp. 439–450. [Online]. Available: http://doi.acm.org/10.1145/342009.335438

[5] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. Santa Barbara, California, USA: ACM, May 21-23 2001, pp. 247–255. [Online]. Available: http: //doi.acm.org/10.1145/375551.375602

[6] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," in *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, July 23-26 2002, pp. 217–228. [Online]. Available: http://doi.acm.org/10.1145/775047.775080

[7] S. J. Rizvi and J. R. Haritsa, "Maintaining data privacy in association rule mining," in *Proceedings of 28th International Conference on Very Large Data Bases*. Hong Kong: VLDB, Aug. 20-23 2002, pp. 682–693. [Online]. Available: http://www.vldb.org/conf/2002/S19P03.pdf

[8] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in *Advances in Cryptology – CRYPTO 2000*. Springer-Verlag, Aug. 20-24 2000, pp. 36–54. [Online]. Available: http://link.springer.de/link/ service/series/0558/bibs/1880/18800036.htm

[9] O. Goldreich, "Secure multi-party computation," Sept. 1998, (working draft). [Online]. Available: http://www.wisdom.weizmann.ac.il/ oded/ pp.html

[10] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, July 23-26 2002, pp. 639–644. [Online]. Available: http://doi.acm.org/10.1145/775047.775142

[11] A. C. Yao, "How to generate and exchange secrets," in *Proceedings of the 27th IEEE Symposium*

on Foundations of Computer Science. IEEE, 1986, pp. 162–167.

[12] I. Ioannidis and A. Grama, “An efficient protocol for yao’s millionaires’ problem,” in *Hawaii International Conference on System Sciences (HICSS-36)*, Waikoloa Village, Hawaii, Jan. 6-9 2003.

[13] O. Goldreich, “Encryption schemes,” Mar. 2003, (working draft). [Online]. Available: http://www.wisdom.weizmann.ac.il/ oded/ PSBookFrag/enc.ps. i