Secure Distributed Data Mining

Suresh Gaikwad, Hemant Kumar Gupta
Student, Assistant Professor
Lakshmi Narain College of Technology and Science (RIT), Indore.

Abstract— Handling large dataset is a tedious task. The solution to the present downside is to use parallel or distributed approaches. Through mining, attention-grabbing relations and patterns between variables of enormous information is determined firmly victimization science techniques and therefore the mining algorithms. This paper addresses the matter of secure distributed association rule mining over the horizontally distributed information. Security is that the main downside in association rule mining comes. The performance of data mining algorithm can be accelerated from O(N) to O(N/k) with parallelism, where N = number of data records and k =number of nodes in distributed system [2]. There are several sites in the transaction. This system is predicated on distributed mining algorithmic program, K&C algorithmic program and AES algorithmic program. Distributed mining algorithm proposed here is the distributed version of apriori algorithm. The science technique is employed to produce security so as to reduce the data shared in mining. With projected technique speed up is nonheritable whereas protective the privacy of the info.

Index Terms— AES; K&C; Apriori algorithm; partitioning Distributed mining.

I. Introduction

The problem of securely mining association rule in distributed environment is studied here. In this system there are several sites that hold databases, these databases are distributed horizontally over different sites participating in transaction for experimentation. The goal is to mine these datasets for finding all association rules with support count at least s and confidence count at least c by pruning using minimal support count s and confidence size c, also hold for the unified database. The important objective of the proposed algorithm is to minimize the information disclosed about the private database held by the sites using encryption mechanism. The information that is protected here is individual transactions information in the different database at each site, and also global information like association rules supported locally by each of those database at different sites [1]. Here the design of an alternative protocol has been proposed and implemented for securely computing the union of private subsets. The system offers simplicity and efficiency as well as privacy. In addition the system does not depend on commutative encryption [4], [5].The objectives for implementing this system are multiple, first is to handle big data sets, second is to acquire speed by utilizing resources available in distributed system and last objective is to preserve data privacy using encryption mechanism.

II. Literature survey

Data mining technology has emerged as a means of identifying patterns with 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 applying data mining algorithm on that data. However, privacy concerns can prevent building a centralized warehouse, in case of distributed system data may be distributed among several custodians, none of which are allowed to transfer their data to another site. Here homogeneous databases are assumed. All sites have constant schema, however every web site has distributed data. The goal is to produce association rules that hold globally, while limiting the information shared from each site.

Previous work in privacy preserving and data mining has two approaches. In first approach data owner and the data miner are two different things, and in second approach the data is distributed among several sites in system. Its aim is to jointly perform mining on the unified corpus of data held by those entities. Kantarcioglu and Clifton [8] proposed the protocol for secure computation of the union of private subsets that are held by the different sites. The private subset of a given site includes the item sets that are s-frequent in his partial database. This part of the protocol is costly and its implementation depends upon cryptographic techniques such as commutative encryption, oblivious transfer.

Yao [9] proposed the protocol for securely computing the union of private subsets at each site. The authors proposed a multi-party computation, which is the costly part of the system and in its implementation cryptographic techniques like encryption, decryption, commutative encryption, and hash functions are used. The use of such cryptographic techniques improves communication cost and computation cost. In the existing systems discussed so far these techniques causes some leakage of information. Therefore this system [9] is not perfectly secure. The proposed system overcomes this problem of information leakage.

In the existing systems [1], [5] the protocol for securely computing the union of private subsets at each site in the transaction is suggested. Here a multi-party- computation is considered and in its implementation cryptographic techniques like encryption, decryption, commutative encryption, and hash functions are used. In these systems it is hard to mine association rules through security assumptions in addition it reveals the data during the mining process. It is not possible to mine globally valid results from distributed data without revealing private information. Secure distributed association rule mining is costly in terms of computational cost and communication.

In UNIFY-KC algorithm the fake item set is added and then removed from item sets. It adds overhead in computation, whereas this overhead is reduced in AES algorithm [6-8]. In this paper for experimentation the data has been partitioned horizontally so that it can be distributed on different sites. Data partitioning techniques are suitable for dealing with the problems in handling large data sets. Round robin partitioning, range partitioning and hash partitioning are some of available horizontal data partitioning techniques [14]. Round robin is the partitioning strategy that partitions dataset with balanced class distribution.

III. Design Issues

In proposed algorithm we have made minor modification. AES encryption algorithm is used which is found to be more secure, it is used instead of UNIFY-KC as proposed in [1]. The distributed mining algorithm is used for distributed mining of association rules.

Distributed Mining Algorithm:

The DM algorithm is the distributed version of apriori algorithm, this algorithm proceeds as follows:

1) Initialization
2) Site Item Sets Generation - Each site will generate its frequent item set. Check weather frequent item set is locally frequent and globally frequent.
3) Local Pruning-Retains Locally frequent item sets.
4) Identification of the candidate item sets – Each site broadcasts its item set.
5) Computation of local supports - Compute local supports of all item sets.
6) Broadcast Mining Results - Each locally frequent item is subset of globally frequent item set. Algorithm Proceeds until it finds no (k+1) item are longest globally frequent item sets. Here k is number of item sets [5],[13].

AES Algorithm

AES is an symmetric block cipher, which means that:

- AES works by continuation a similar outlined steps multiple times.
- AES is a secret key encryption algorithm.
- AES operates on a hard and fast range of bytes.
- AES further as most secret writing algorithms ar reversible.

The AES algorithmic program operates on bytes, which makes it simpler to implement. The key is divided into individual sub keys, a sub key for each operation round. This process is called KEY EXPANSION. As mentioned before AES is an iterated block cipher that is the same operations are performed many times on a fixed number of bytes. These operations can easily be broken down to the following functions:

1. **Sub Bytes -** This is a non-linear substitution step where each byte is replaced with another byte according to a lookup table data.
2. **Shift Rows -** This is a transposition step where each row of the state is shifted circularly for a certain number of steps.
3. **Mix Columns -** This is a mixing operation which operates on the columns of the states, combining the four bytes in each column.

Add Round Key - *In this approach, each byte of the state is combined with the round key using bitwise xor Rounds.*

(K & C) Kantarcioglu and Clifton Algorithm

The step number 5 of DM algorithm is implemented using K & C algorithm. The K & C algorithm is used for Unifying lists of locally Frequent Item sets, it works as follows:

1) Each site adds his private item set.
2) Sites jointly compute the encryption of their private subsets.
3) Each site adds his own layer of encryption using his private secret key.
4) Every item set in each subset is encrypted by all of the sites.
5) Sites compute the union of those subsets in their encrypted form.
6) Sites decrypt the union set.

IV. METHODOLOGY

In experimentation datasets from KDD community, extended bakery dataset, frequent Itemset Mining Dataset Repository, Bioinformatics Data Set, IBM Almden Quest research group etc. are used. The datasets are namely chess, connect, algebra and test. Figure 1 presents Proposed secure distributed association rule mining approach. In implementation of system the database is distributed horizontally among various sites in the transaction. Round robin technique is employed for Horizontal distribution of knowledge sets to scale back the info skew. The data at each site is encrypted using Advanced Encryption System algorithm and association rule mining is performed. Each site is having its private key. Decryption is performed at centralized server using private key, the K & C algorithm is applied for effectively mining global association rules. In the implementation, server is considered as "Master" of the process or system. In addition security should be maintained while doing this mining process.

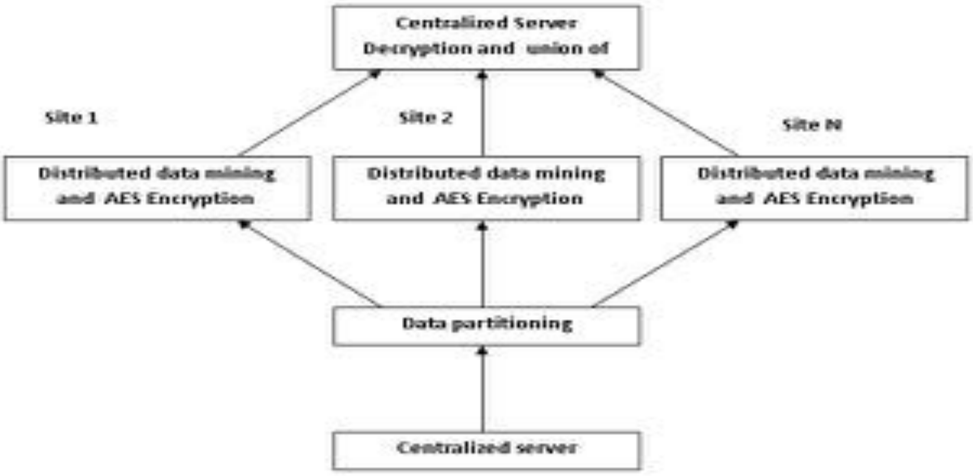

Centralized Server
Decryption and union of
Site 1
Site 2
Site N
Distributed data mining and AES Encryption
Distributed data mining and AES Encryption
Distributed data mining and AES Encryption
Data partitioning
Centralized server

Figure 1 Proposed secure distributed association rule mining approach.

V. Experimental Setup

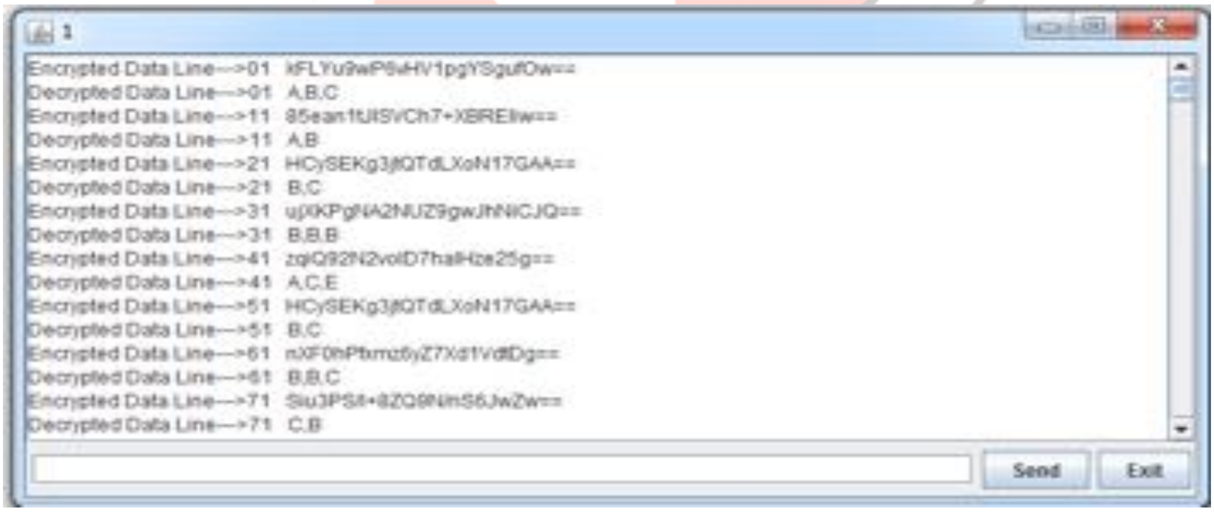
In this system the performance of secure implementations of the DM algorithm is compared. In implementation the DM algorithm is implemented in the secure manner. We have tested the implementations with respect to some measures which are enlisted below:

1. Total computation time of the complete algorithms (DM and AES) over all sites. That live includes the Apriori computation time, and also the time to spot the globally s-frequent item sets.
2. Total computation time of the unification process only over all sites.

Experimental Outcome

- Can handle Big Data sets.
- Speed up is acquired in computation process by utilizing resources available in distributed system.
- Provides security in distributed computing environment.

VI. Resut analysis



```
Encrypted Data Line--->01   kFLYu9wP9vHV1pgYSgufOw==
Decrypted Data Line--->01   A,B,C
Encrypted Data Line--->11   85ean1fUISVCh7+XBREIiw==
Decrypted Data Line--->11   A,B
Encrypted Data Line--->21   HCySEKg3jfQTdLXoN17GAA==
Decrypted Data Line--->21   B,C
Encrypted Data Line--->31   ujXKPgNA2NUZ9gwJhNICJQ==
Decrypted Data Line--->31   B,B,B
Encrypted Data Line--->41   zqiQ92N2volD7halHze25g==
Decrypted Data Line--->41   A,C,E
Encrypted Data Line--->51   HCySEKg3jfQTdLXoN17GAA==
Decrypted Data Line--->51   B,C
Encrypted Data Line--->61   nXF0hPbmz6yZ7Xd1VdfDg==
Decrypted Data Line--->61   B,B,C
Encrypted Data Line--->71   Siu3PS#+8ZQ9NlnS6JwZw==
Decrypted Data Line--->71   C,B
```

Figure 2 Applying Encryption on site in transaction

The graph shown in figure 3 shows the time required for mining association from datasets by sequential and the proposed approach. Datasets are tested for all the algorithms i.e. Apriori, AES and K &C algorithm . In this system distributed mining algorithm is used for data mining task. It results of increase in speed up and computation time of association rule mining system. Security provided by system results in better performance gain.

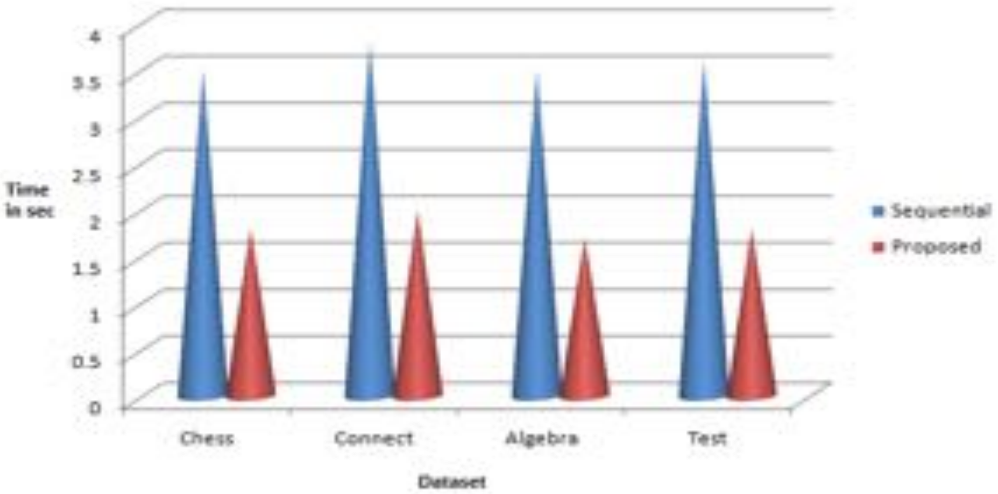


Fig 3. Time required for mining association rules from datasets by sequential and the proposed approach

Let Ts is time required by sequential system for mining and Tp is time required by proposed system for mining. The speed up is calculated by equation 1.

$$\text{Speed up} = \frac{Ts - Tp}{Ts}$$

By evaluation the average speed up for chess dataset is 50%.The average speed up for connect dataset is 55%, the average speed up for algebra dataset is 51%, the average speed up for test dataset is 50% and the average speed up for disease dataset is 55%. Table I Shows the speed up of sequential and proposed system.

T_S	Tp	Speed up
80	40	50
90	42	53
82	37	54
85	42	50

Table 1 speed up

Speed up acquired during mining process

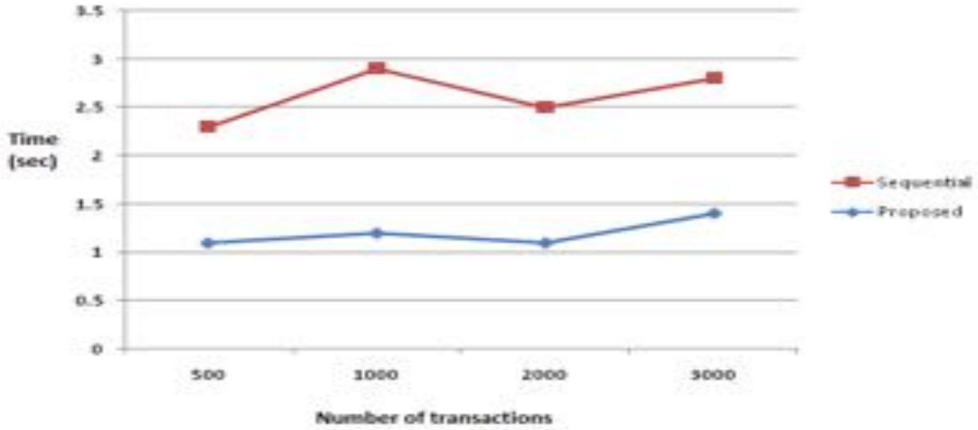


Fig 4. Time required for generating frequent item sets

VII. CONCLUSION

In this paper the secure implementation of mining process is discussed. The implementation consist of distribution of database on different sites, application of distributed mining algorithm for mining of frequent itemset, K & C algorithm for merging of frequent itemsets and AES encryption algorithm. The cryptological algorithmic program like AES permits U.S. for firmly playing association rule mining. The interesting properties between locally frequent and globally frequent itemsets are observed The distributed association rule mining is done efficiently through security assumptions and strong rules are found. Thus mining of globally valid results from distributed data without revealing private information are obtained using security assumptions. Due to use of these techniques of distributed mining, speed up is acquired using distributed association rule mining is done with a reasonable cost.

REFERENCES

[1] T. Tassa, Secure Mining of association rules in horizontally distributed Database"proc, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA EN GINEERING, VOL. 26, NO. 4, APRIL 2014.

[2] G. Alex and A. Freitas, Scalable, high-performance data mining with parallel processing,in Principles and Practice of Knowledge Discovery in Databases,(Nantes, France),1998.

[3] R. Agrawal and R. Srikant, Fast algorithms for mining association rules in largeDatabase". In VLDB, pages 487499, 1994.

[4] A.V. Ev_mievski, R. Srikant, R. Agrawal, and J. Gehrke, Privacy preservingmining of association rules.In KDD, pages 217228, 2002.

[5] D.W.L. Cheung, J. Han, V.T.Y. Ng, A.W.C. Fu, and Y. Fu, a quick distributed rule for mining association rules. In PDIS, pages 3142, 1996.

[6] R.L. Rivest, A. Shamir, and L.M. Adleman, a way for getting Digital Signatures and Public-Key Cryptosystems, Comm.ACM, vol. 21, no. 2, pp. 120-126, 1978.

[7] Ben-David, N. Nisan, and B. Pinkas, Fair play MP - A System for Secure Multi-Party Computation ,Proc. 15th ACM Conf Computer and Communication Security(CCS), pp. 257-266, 2008.

[8] M. Kantarcioglu and C. Clifton, Privacy-preserving distributed mining of association rules on horizontally divided data", IEEE Transactions on Knowledgeand Data Engineering, 16:10261037, 2004.

[9] T. Tassa and E. Gudes. Secure distributed computation of anonymized views ofshared database", Transactions on Database Systems, 37, Article 11,2012.

[10] 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.

[11] J. Brickell and V. Shmatikov, Privacy-preserving graph algorithms within the semi honest model", In ASIACRYPT, pages 236252, 2005.

[12] J.S. Park , M. Chen, and P.S. Yu , An effective Hash Based Algorithm for Mining Association Rules, "Proc. 1995 ACM SIGMOD Int'l Conf. Management of Data , ACM Press, 1995, pp. 175-186.

[13] C. Ray, Distributed database systems. Pearseron Education India, 2009.