Abstract— Data mining on a large dataset is a complex and time consuming task. The mining process on this large volume of data becomes slow, as it has to be done serially. The solution to the problem is to accelerate the mining process with the help of parallel or distributed approaches. Through mining, interesting relations and patterns between variables of large database can be observed securely using cryptographic techniques and the mining algorithms. This paper addresses the problem of secure distributed association rule mining over the horizontally distributed database. Security is the main problem in association rule mining projects. The solution to the problem is to accelerate the learning process with the help of parallel or distributed approaches. As mentioned earlier the performance of data mining algorithm can be enhanced from $O(N)$ to $O(N/k)$ with parallelism, where $N$ = number of data instances and $k$ = number of nodes [2]. There are several sites in the transaction. This system is based on distributed mining algorithm, K&C algorithm and AES algorithm. Distributed mining algorithm proposed here is the distributed version of apriori algorithm. The cryptographic technique is used to provide security in order to minimize the information shared in mining. With proposed method speed up is acquired while preserving the privacy of the data.

Keywords— AES, K&C, Apriori Algorithm, 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 homogeneous databases, these databases are distributed horizontally over different sites participating in transaction. The goal is to mine these datasets for finding all association rules with support count at least $s$ and confidence count at least $c$. The given 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. 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.

II. HISTORY AND BACKGROUND

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 the same schema, but each site has distributed information. 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 itemsets 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 itemset is added and then removed from itemsets. It adds overhead in computation, where as 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 datasets. Round robin partitioning, range partitioning and hash partitioning are some of the 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 modification to algorithm proposed by T. Tassa[1]. AES encryption algorithm is used which is found to be more secure. 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 ItemSets Generation - Each site will generate its frequent itemset. Check whether frequent itemset 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 itemset.
5) Computation of local supports - Compute local supports of all itemsets.
6) Broadcast Mining Results - Each locally frequent item is subset of globally frequent itemset. Algorithm proceeds until it finds no \((k+1)\) item are longest globally frequent itemsets. Here \(k\) is number of itemsets [5],[13].

AES Algorithm

AES is an symmetric block cipher, which means that:

- AES operates on a fixed number of bytes.
- AES as well as most encryption algorithms are reversible.

The AES algorithm operates on bytes, which makes it simpler to implement. The key is divided into individual subkeys, 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.
4. 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 Itemsets, it works as follows:

1) Each site adds to his private itemset.
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 itemset 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. PROPOSED ALGORITHM

Let \(s\) (be a main set of) \(\equiv \{\text{CSDB, LDB, C, A, S, MR, AR}\}\) where,

- CSDB is the copy of the centralized server database. This database is responsible for storing and processing user information.
- LDB is a set of local database of a user. It consists of data tables having data items related to the products and their sales transactions.
- C is a set of all clients using the server database and mining services from the server. \((C_1, C_2 , C_3, \ldots\ldots, C_n) \in C\).
- A is a set of algorithms applied on the input data for achieving mining results.
- S is the server of the system. The server is responsible for authenticating and providing associations to the end user.
MR is a set of mining rules that are applied on the input dataset provided by the client from his local database LDB
(\(MR_1, MR_2, MR_3, \ldots, MR_n\) \(\in\) MR).

AR is a set of associations rules that are extracted from the input and a form the output of the system. Functionalities:

- \(LDB_n = \) Local data at site \(n\)
- \(ED_n = (Encoded\ data)_{AR_n}\)

**Proposed Algorithm:**

1. Apply Mining\((LDB_n, AR_n)\);
2. Encode\((AR_n, ED_n)\);
3. UPLOAD\((ED_n)\);
4. Decode\((ED_n, AR_n)\);
5. Results = K&C\((AR_n)\)

**V. METHODOLOGY**

In experimentation datasets from UCI repository are used. The datasets are namely mushroom, monks problem, vote, soybean, disease and itemsets. 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 used for Horizontal distribution of data sets to reduce the data skew.

![Proposed Secure Distributed Association Rule Mining Approach](image)

**VI. 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 measure includes the Apriori computation time, and the time to identify 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. RESULT AND ANALYSIS**

The result in terms of graphs shows the performance of the proposed system. The proposed system provides security while doing mining task in the distributed environment. To verify the performance of the system, the encryption/decryption overhead and support overhead is evaluated. In experimentation a datasets from UCI repository are used namely mushroom, monks problem, vote, soybean, disease and itemsets. The graph in figure 2 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 DM algorithm. In this system distributed mining algorithm is used for data mining task. Figure shows speed up and computation time of association rule mining system. Security provided by system results in better performance gain.

Let \(T_s\) is time required by sequential system for mining and \(T_p\) is time required by proposed system for mining. The speed up is calculated by equation 1.

![Graph showing time required for mining](image)

\[
\text{Speed up} = \frac{T_s - T_p}{T_s}
\]  

(1)

By evaluation the average speed up for mushroom dataset is 75%. The average speed up for monks dataset is 80%, the average speed up for vote dataset is 78%, the average speed up for itemsets is 77% and the average speed up for disease dataset is 79%.

This system is purely independent of oblivious transfer and commutative encryption which makes it simple and it also contributes to the relatively less cost of computation and communication and increase in computational time. The graph in figure 3 gives the time required for generating frequent itemsets. The graph shown in figure 4 gives the computational time required for sequential and proposed
system, both figures present the speed up acquired by the proposed system.

**Fig. 2 Computational Time for the Proposed System in Distributed Approach and in Sequential Approach where all Transaction Databases are combined**

**Fig. 3 Time for Generation of Frequent Itemsets with the Proposed and in Sequential Approach**

**Fig. 4 Computation Time as Number of Transactions Increase**

**IV. CONCLUSION**

Thus the interesting properties between locally frequent and globally frequent itemsets are observed. The distributed version of apriori algorithm is applied for distributed mining of association rules. The Cryptographic algorithm like AES enable us for securely performing association rule mining. 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 is implemented using security assumptions. Due to use of these techniques of distributed mining, performance is enhanced from $O(N)$ to lower bound $O(N/k)$, where $N$ = number of data instances and $k$ = number of nodes (that is 3). Secure distributed association rule mining is done with a reasonable cost.

**REFERENCES**


[10] D.W.L. Cheung, J. Han, W.T.Y. Ng, A.W.C. Fu, and Y. Fu, A Fast Distributed Algorithm for Mining Association Rules.


