An Optimized and Secure Association Rule Mining Algorithm in Parallel and Distributed Environment

Sakharam Kolpe¹, Prof. N.V. Alone²

Department of Computer Engineering, GESCOEMSR Nasik-05, Maharashtra, India

Abstract: Data mining is a computational process of discovering interested patterns in large data sets. Data mining sometimes called data or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information. Handling large data set is a very complex and tedious task. The solution to this problem is to apply distributed or parallel approaches. With help of cryptographic techniques, interesting relations and patterns between variables of large database can be observed securely. This system addresses the problem of secure distributed association rule mining over the horizontally distributed database. This approach can provide security at the time of mining data from various data sources which is useful for business decisions. In data mining Security is the major problem with respect to association rule mining projects. Association rule mining is one of the data mining techniques used in distributed databases these techniques disclose some interesting relationship between locally large and globally large item sets. 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). In distributed association rule mining there are several sites in the transaction. The goal of proposed approach is to find all association rules with support $S$ and confidence $C$ to minimize the information disclosed about the private databases held by those player. The systems will be based on distributed mining algorithm, K&C algorithm and AES algorithm. The Distributed mining algorithm in proposed system 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, Association Rule.

INTRODUCTION

In computer science data Mining is a technique or sometimes also called data or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information it is use to achieve an fascinated prototype from big data set. Advances in computing and networking technologies have resulted in distributed and dynamic sources of data. A distinctive example of such a scenario is found in the warehouses of large national and multinational corporations. Each database is continuously updated with new data as transactions occur. The update rate and additional properties may be unique to a given site. A user may be interested in generating a global model of the database, thus the sites must exchange some information about their local models. However, the information exchange must be made in a way that minimizes the communication overhead. The frequency at which the global model is updated may vary from the frequency at which each local model is updated. Furthermore, in such a distributed scenario, the user may be interested in not only knowing the global model of the database, but
also the differences (or contrasts) between the local models. Analyzing these distributed and dynamic databases requires approaches that make proper use of the distributed resources, minimize communication requirements and reduce work replication. An efficient frequent item set mining approach when data is both distributed and dynamic. (the analysis step of the "knowledge discovery in databases" process, or KDD an interdisciplinary subfield of computer science is the computational process of discovering patterns in large data sets ("big data") involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

The proposed system addresses problem of securely mining association rule in distributed environment. In this proposed system there are several sites or player that hold homogeneous partitioned databases, these databases are distributed horizontally over different sites participating in transaction. The goal of proposed system is to mine these data sets for finding all association rules with satisfy the support count at least s and confidence count at least c. The given minimal support count sand confidence size c, also hold for the unified database. The important objective of the proposed system is to minimize the information disclosed about the private database held by the sites. The information that is protected here is not only individual transactions of every site but also 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 in proposed work the design of an alternative protocol has been proposed and implemented for securely computing the union of private subsets. The systems relay on and offer simplicity and efficiency as well as privacy. In addition this system does not depend on encryption like commutative encryption [4], [5]. The main objectives for implementing this system are as first is to handling an large size of big data sets, second is to acquire and obtain speed by utilizing resources available in distributed system and last objective is to provide more security by using cryptographic technique for preserve data secretly.

LITERATURE SURVEY

Survey of various existing approaches for secure distributed data mining are observed here and explained. Data mining technology has huge applications in the field of identifying and observing an interested pattern among tremendous quantities of data. Almost in now days in every application Data mining techniques and data warehousing are used. most of the data mining tools operate by collecting all the data into a centralized site, then applying data mining algorithm on that available data. However, security is main issue in the data which can prevent be building a centralized warehouse. In case of distributed system data may be distributed among several site or systems in that site or system which are allowed to transfer their data to another site. But the problem is what about that privacy of data. in literature survey the homogeneous databases are assumed. in homogenous distributed database All sites have the same schema, but each site has distributed information on same entity. The goal is to produce association rules that hold globally, while giving more concentration on privacy of the information shared from each site. in the mining of association rule the research work is divided into two main settings. In first setting the data owner and the miner are two different entities, and second, which consist of data distribution among several sites or players, these players or sites jointly performs mining on the data held by those sites or players. Kantarcioglu and Clifton [8] have introduced the protocol for securely computing union of each private subsets held by the different sites. The private subset of a given site includes the item sets which are S-frequent in own database. This implementation become more costly and time consuming and in this cryptographic techniques such as commutative encryption, oblivious transfer are used. Yao [9] proposed
the protocol for securely computing the union of private sub-sets at each site. The authors proposed a multi-party computation, in this proposed work more than one site or party are involved in that 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 cause some disclose and loss the security of private information hold by every site. Therefore this system [9] is not perfectly secure. The proposed system overcomes this problem of information leakage and privacy of data. In the existing systems [1], [5] they proposed 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 that implementation of cryptographic techniques is done like encryption, decryption, commutative encryption, and hash functions are used. In that systems it is more difficult to mine association rules through security assumptions in addition it disclose the data during the mining process. It is not possible to mine globally valid results from distributed data without losing security of private information. Secure mining of association rule in distributed environment is costly in terms of computational cost and communication. In UNIFY-KC algorithm the fake item set is added and after the computation is done then it removed from item-sets [1]. It address the problem of overhead in computation, and communication where as this overhead is reduced in proposed AES algorithm [6-8]. In this paper for the data has been partitioned horizontally so that it can be distributed on different sites. Data partitioning techniques are suitable for Related with the problems in handling large data set. the various partitioning technique are Round robin partitioning, range partitioning and hash partitioning are some of the available partitioning technique horizontal data partitioning techniques the Round robin is one of the partitioning strategy that partitions data set with balanced class distribution. Kantarcioglu and Clifton [6] implemented the sub-protocol for the secure computation of the combination or union of private subsets which are held by the different players or sites. (The private subset of a given player or site, includes the item sets that are s-frequent in his partial database.) The use of such cryptographic techniques improves communication cost and computation cost. In the existing system these cryptographic techniques are used it causes some loss the privacy of information; therefore it is not perfectly secure. Thus the union of private subsets is not perfectly calculated, so the proposed system is to overcome with this problem. In the existing system the protocol for securely computing the union of private subsets at each site in the transaction is implemented. Thus interesting pattern on this large volume data becomes slow. In these systems it is hard to mine association rules through security assumptions in addition it reveals the data during the mining process. As explain earlier it is impossible to mine globally frequent valid results from distributed data without disclosing private information mining of association rule securely in distributed manner is costly in terms of computational cost and communication. As per Alex Friedes [7] the speed up of data mining system improves form O(n) to O(n/k) through implementation of distributed approach. Also due to use of distributed or parallel association rule mining techniques the speed can be improve with respect to computation of association rule. Mining in secure way as Rakesh Agrawal [8] proposed technique for mining of association rules using classifier and showing.
PROPOSED SYSTEM ARCHITECTURE

The system architecture is shown in figure 1; it shows how the system works in order to generate globally frequent item sets through mining process. It describes how the system flow is. Initially the centralized server takes transaction data set as input to system. The centralized server performs horizontal data partitioning through the application of distributed mining algorithm, this data is distributed among site1, site2 ..., Site N. While distributing data among sites the data is in the encrypted form so that the privacy is maintained during transfer of data, when this data reaches to particular client, the client again decrypts the data and extracts the original data. The encryption and decryption is performed using AES encryption/Decryption algorithm. The client receives this data in encrypted form, by applying decryption key, client extracts original data, after extraction of original data, client identifies locally frequent item sets by applying distributed mining algorithm, after identifying locally frequent item set it again encrypt these locally frequent item sets and send back to server. At server side these locally frequent item sets are encrypted and for merging of item sets the K&C algorithm is used at server side. This process is repeated until server finds globally frequent item set. Thus the output of system is globally frequent item sets which are obtained at centralized server.

In the proposed system the problem of secure computation of union of private subsets available on sites is addressed and the database is horizontally Partitioning and distributed among various sites or player which are involved in transaction. Round robin technique is used for Horizontal distribution of Data sets to reduce the data skew. The problem of securely mining association rule in distributed environment is implemented here. In this system there are several sites that hold identical databases these databases are distributed horizontally over different sites participating in transaction. The goal is to mine these data sets 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 main target is to design an algorithm to enable handling of large data sets using available computing resources and to design the system to accelerate a mining process on large data sets. Achieving a speed up in computation process by utilizing resources available in distributed Environment provide more security in distributed computing environment. System is concerned with maintaining security while mining of association rules in distributed database, where input is synthetic database and output will be set of association rules. To design secure multi-party algorithms that computes the union of private subsets that each of the interacting sites hold, and to design algorithm to test the inclusion of set of an element held by one site in a subset held by another. The important objective of the proposed algorithm is to minimize the information disclosed about the private database held by the sites.
METHODOLOGY

The proposed approach distributes each transaction \( D \) into partitions and in each local partition frequent item sets are find out at every site. After finding local frequent item sets all local frequent item sets are combined to find candidate item set. In last stage global frequent item sets are found. Data mining on a very large data set is a complex task. The rule discovery on this large volume data becomes slow, since it is done serially on available big data sets. The large quantity of data records may overload a computer’s memory and processor due to this the learning process becomes very slow. Mining on large data set may become impossible because of limitations on processor and memory.

Distributed mining algorithm which is a combination of preserving privacy and fast distributed mining algorithm which is an unsecured distributed version of the Apriori algorithm. Its main idea is that any S-frequent item set must be also locally S-frequent in at least one of the sites. Hence, in order to find all globally S-frequent item sets, each player reveals his locally S-frequent item sets and then the players check each of them to see if they are S-frequent also globally.

The FDM algorithm proceeds as follows:

1. **CandidateSetsGeneration:**
   Each player \( p_m \) computes the set of all \((k-1)\) item sets, \( L_{K-1} \) that are locally frequent and also globally frequent. The intuition behind the candidate set generation is that if an item set \( X \) has

![Figure 1: System architecture](image-url)
minimum support, so do all subsets of X. Hence the player then applies set the Apriori algorithm on LK-1 in order to generate the set of candidate k-item sets.

2. **Local Pruning:**
   The pruning step eliminates the extension of (K-1) item sets which are not found to be frequent. Here, player pm computes supply(X). And retains only those item sets that are locally S-frequent. We denote this collection of item set by Cs k, m.

3. **Computing localsupports:**
   All players compute the local supports of all item sets in Csk, m.

4. **Broadcast mining results:**
   Each player broadcasts the local supports that he computed. From that, everyone can compute the global support of every item set in Csk, m. Finally Fsk, m is the subset of Csk, m that consist of all globally S-frequent k-item set.

**CONCLUSION**

The problem of securely mining association rule in distributed environment is addressed here. In this proposed 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 data sets 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 system 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 The implementation consist of distribution of database on different sites, application of distributed mining algorithm for mining of frequent itemsets, K&C algorithm for merging of frequent itemsets and AES encryption algorithm are used. The Cryptographic algorithm like AES enables us for securely performing association rule mining. Then interesting properties between locally frequent and globally frequent item sets are observed. Also the encryption rate in MB/s and encryption time in ms and speed up acquired during encryption for mentioned cryptographic algorithms are compared, comparison resulted in higher encryption rate for AES.

**ACKNOWLEDGMENT**

We are thankful to PG Co-coordinator Prof Mrs. Archana S. Vaidya and HOD of computer department Prof. N. V. Alone for their valuable guidance and encouragement. We would also like to thank the (Gokhale Education Society’s R.H.Sapat College of Engineering, Nashik- 05) for providing the required facilities, Internet access and important books. At last we must express our sincere heartfelt gratitude to all the Teaching and Non-teaching Staff members of Computer Engineering Department who helped us for their valuable time, support, comments, suggestions and persuasion.
REFERENCES


