Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Engineering

Elixir Comp. Engg. 101 (2016) 43974-43977

ISSN: 2229-712X

Efficient & Secure Mining for Vertical Distributed Database

Jinkal Patel, Deepak Ganta, Harish Dogiparthi and KiranLoudya
Department of Computer Science, Northwestern Polytechnic University, Fremont, CA-94538,USA.

ARTICLE INFO

Article history:
Received: 2 November 2016;
Received in revised form:
2 December 2016;
Accepted: 15 December 2016;

Keywords
Distributed database,
Privacy Preserving Data Mining,
Association rules.

ABSTRACT

This Privacy preserving is most popular study for the research field. Privacy means gives the protection to the private information at preserving time. In Market place, discovery of frequent item sets using association rules mining is one of most important tasks in mining. Association rules is very helpful for finding frequent item sets to predicate about which item sets purchased together in a market and generate qualitative information that is useful for decision making. For distributed environment, database may be distributed as horizontally, vertically or mixed in computer network. The main problem in secure mining with the help of association rules is that transactions are distributed as vertically and the various sites want to find frequent item sets by participating themselves without discovering their individual data. The proposed method will find frequent item sets for vertical distributed database with the help of data miner using encryption based technique. Each sites prepare matrix with the local frequent item sets as per minimum support and encrypted it than send to other sites. The Scalar product with Boolean matrix is used for finding frequent item sets with secure computation between multiple sites without disclosing private input which improved efficiency and privacy of system.

I. Introduction

Data Mining means extract the knowledge from huge number of data. Data mining is most useful for finding the acute information from huge data storage, data warehouses, or other information repositories [1]. Privacy preserving data mining [2], is a novel research direction in data mining, where data mining algorithms are analysed for the side-effects they incur in data privacy. The problem of mining the vertical distributed databases with the algorithm based on association rules with security has been studied here for the several sites which are holding the heterogeneous databases. The goal is to find association rules based on support and confidence. An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in data. A consequence is item that found in combination with the antecedent. Association rules are created by analysing the data for frequent patterns and using the criteria support and confidence to identify the most important relationships. Association rule mining is a process of mining the databases based on rules. This rule is purely based on support and confidence. The concept of privacy preserving data mining involves in preserving personal information from data mining algorithms. PPDM technique [3] is a research area in data mining and statistical databases where mining algorithms are analysed for the side effect they acquire in data privacy. The objective of privacy preserving in data mining is to build algorithms for transforming the original information in secured/unsecured way, so the private data and private knowledge remains confidential even after the mining process [4].

In centralized environment, all the datasets are collected at central site (data warehouse) and then mining operation is performed, where in distributed environment, data may be distributed among different sites which are not allowed to send their data to find global mining result. There are two types of distributed data considered. One is horizontally partitioned data and another is vertically partitioned data. Data are distributed among two sites which want to find the global mining result. In horizontal partitioned data, each site contains same set of attributes, but different number of transactions wherein vertical partitioned data each site contains different number of attributes but same number of transactions [5].

Here some problem in secure mining of association rules in vertically distributed databases. In such a setting, there are several databases where transactional attributes are distributed across the databases. With vertical approach, some of the columns of a relation are apportioned into a base relation at one of the databases, and other columns are assigned into a base relation at another database. The relations at each of the sites must share a common domain so the original table can be reconstructed. In “Market-Basket” example, one database may contain grocery items and other one have clothing purchases. Using a key such as transaction date, transaction id or credit card details, we can join these to identify relationships between purchases of clothing and groceries. Horizontal partitions support an organizational design in which functions are repeated, often on a regional basis, whereas vertical partitions are typically applied across organizational functions with reasonably separate data requirements [6].

II. Privacy Preserving Data Mining

Privacy preserving data mining (PPDM) techniques have been introduced with the aim of preventing the discovery of sensitive information from data mining algorithms. Privacy refers to extraction of sensitive information using data mining process. The concept of PPDM involves in preserving personal data using data mining algorithms. PPDM technique is important research area for data mining.

Tele:
E-mail address: jinkalpatel77@yahoo.com

The main objective of PPDM is to provide algorithms for converting the original data in secure or unsecure way, so the private information will remain confident after the end of mining procedure [7].

The classification of privacy preserving technique [8]:

A. The Randomization Method

The randomization method provides data distortion technique to develop the private records. Two types of perturbation describe for the randomization methods:

- Additive perturbation: Random noise add to records of data.Multiplicative
- Perturbation: Random rotation or random projection is used on data.

Advantages*:* It is easy to implement and other knowledge is not required. There is no need of server. It is useful for hiding individual sensitive data.

Disadvantage*:* It treats all the records equally and reduces the utility of the data.

B. Anonymization

The anonymization method used to protect individual identity when releasing sensitive information, data holders often remove explicit identifiers. There are two types of disclosures as follow:

- Identity disclosure: It means the individual data is identifying uniquely from the issue data.
- Attribute disclosure: It means the individual data can infer from the issue data.

There are various popular techniques for identity disclosure and attribute disclosure problem that solve by preserving private data method as follow: K-anonymity method, L-diversity method, T-closeness method

C.Distributed Privacy Preservation

In distributed data mining the dataset partitioned as horizontally, vertically or mixed. Each site has all the data about set of record in horizontal distributed database. For vertical distributed database, each site has different attributes. The distributed techniques are based on cryptography which contains various methods like secure multi-party computation. The encryption/decryption method used to transform data for privacy and accuracy purpose. The SMC provide some basic models is as follow:

- Semi-Honest model: The opponent will follow the protocol trustworthy. But opponent always try to infer in private data at execution process.
- Malicious Model: The opponent does every try to infer the private data. For example, send any spoof messages or spurious messages and also abort work at any time.

III. Privacy Preserving data mining in Association rule mining

Association rule mining (ARM) is a technique in data mining that identifies the regularities in large number of data. Privacy preserving association rule mining needs to stop for disclose confident information or private data from original data and also prevent mining methods from finding sensitive data. The common approaches used in association rule hiding algorithms as follow [9]:

- Heuristic methods: It used to modify the selected data of database to hide sensitive information.
- Border-based method: The sensitive left or right rule hiding can done using modification in originals of frequent or infrequent patterns in the data set.
- Exact method: It is non-heuristic techniques which hides the process. The constrain rules satisfaction problem is solved using linear programming or integer programming.

IV. Problem Statement

Privacy preserving is one of the major problems in data mining. The existing method is based on horizontal distributed database that use fast distributed mining technique for finding the frequent item sets and secure multi-party computation algorithm for securing the data held by the users. The main problem of secure mining where transaction of data is distributed as vertical in that each site have some of attributes of each transaction records and the various site want to participate with global valid frequent item sets. In this situation, scalar product algorithm with Boolean matrix is useful for find frequent item sets in vertical environment. The main goal of this work is to find global association rules efficient by using scalar product with Boolean matrix and secure multiparty algorithm for the privacy purpose.

V. Related Work

Santhana Joyce, Nirmalrani et al (2015) proposed Protocol is based on Fast Distributed Mining (FDM) Algorithm, which is the current leading protocol, which overcomes the disadvantages of various other algorithms such as apriori, FP tree etc. This protocol improves simplicity and efficiency as well as privacy. AES (Advanced Encryption standard) algorithm is used to ensure security and to encrypt and decrypt the data while inserting and retrieving.

P.Kalaivani, D.KeranaHanirex et al (2015) paper represents association rule mining technique based on apriori and secure multi-party algorithm for vertical distribute environment. The method finds frequent item sets as per minimum support with confidence and generate strong global association rules [10].

Gayatri K et al (2014) proposed Secure Mining of Association Rules in Horizontally Distributed Databases Using FDM and K&C Algorithm. In their work they used FDM algorithm and Unifi-KC algorithm. Also the efficiency, computational cost and communication cost were compared in that paper. Future work is to devise an efficient protocol for inequality verifications that uses the existence of semi-honest third party and another in the implementation of the techniques to the problem of distributed association rule mining in vertical setting.

TamirTassa (2013) proposed 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 inclusion of an element held by one player in a subset held by another. It is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost.The techniques in thispaper are not implementing for the vertical setting [11].

Feng Zhang,ChunmingRong (2013) proposed Privacy preserving distributed association rules mining protocols horizontally partitioned data with more than two participating parties. It depends on a secure multi-party summary and union computation, which cannot guarantee security while the number of participating parties is two.

VI. Proposed Work

The existing system works on horizontally distributed databases for secure mining protocol in PPDM. The protocol is designed based on SMC and FDM for secure mining process. It provides RSA 1024-bits encryption method for secure computation of private data [11]. The FDM algorithm is unsecure where the sites broadcast the local frequent, and they broadcast the local supports of candidate item sets. Unifying lists of locally Frequent Item sets - Kantarcioglu and Clifton protocol is used to compute the local item sets securely

in FDM algorithm for horizontal setting. It used RSA 1024 Bits encryption method for secure union. The propose work is to find association rules for vertically distributed databases with data miner with efficient and secure manner. The vertical distributed sites interested to find frequent item sets by participating themselves but they don't want to disclose their individual private data. It used scalar product computation and RSA 1024-bits encryption algorithm for distributed environment. Here, the proposed work used the concept of fast distributed mining which define local frequent and global frequent.

Procedure:Different n data set located in different sites as site 1, site 2 up to site n. Data sets partition like DB1 to DBn. Every site wants to communicate like site 1 to site 2, site 2 to site 3 up to last Site n with Data miner. The Data miner is communicating with all sites, Site1 to Site n.

Step 1:Data miner sends public key and minimum support to all sites.

Step 2:Each site prepare matrix with local frequent based on minimum support andsend encrypted data to predecessor site.

Step 3:Site 1 and site 2 compute scalar product and encrypt data then prepare newmatrix by appending all matrix as shown in figure 3.1. Encrypt new matrix and sendto other site.

Step 4:End of the procedure, Data miner receives the final results from last site andthen applies decryption algorithm to find frequent item sets so the data miner decryptdata using private key.

Step 5:Data miner finds the global frequent item sets by calculating last receivedmatrix then prepares a list of global frequent item sets with their support.

Step 6:Data miner will broadcast final result to all the sites.

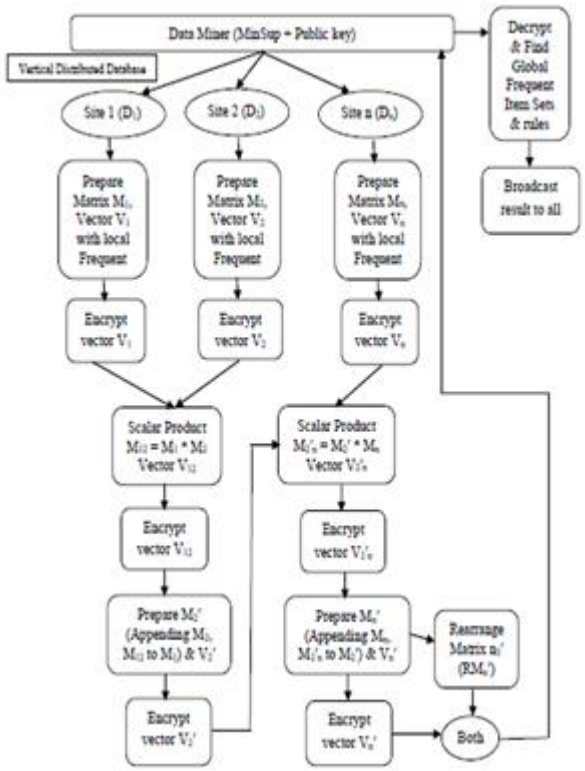

Figure 1. Flow Chart of Proposed System.

VII. Expromental Results and Analysis

The German credit card data set is downloaded from UCI repository. This dataset has 1000 number of records and 22 attributes. The existing is for horizontal database with unify-kc algorithm and proposed for vertical database with scalar product. The proposed and existing system is tested based on three sites.

Partition of data is maintained by three different sites. The existing system distributed horizontally with 200, 300, 500 records and proposed system distributed vertically with 8, 9 and 5 Attributes. The experimental results tested with different intervals.

Execution Time

Execution time and variant supports: The experimental results are based on processing time of protocol with different support values. There are 1000 thousand records and support is variant as 0.2, 0.4 up to 0.6. The processing time is in millisecond

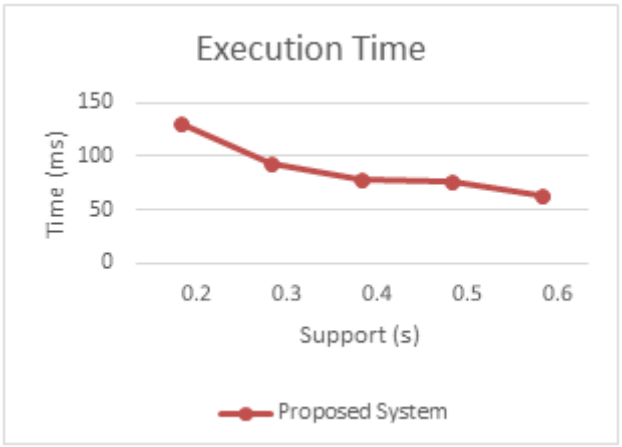

Figure 2. Execution Time Vs. Supports

- Number of records and require time: There are 1000 record of data divided in different intervals. Support is 0.2 for both systems. Execution time is in millisecond for both systems.

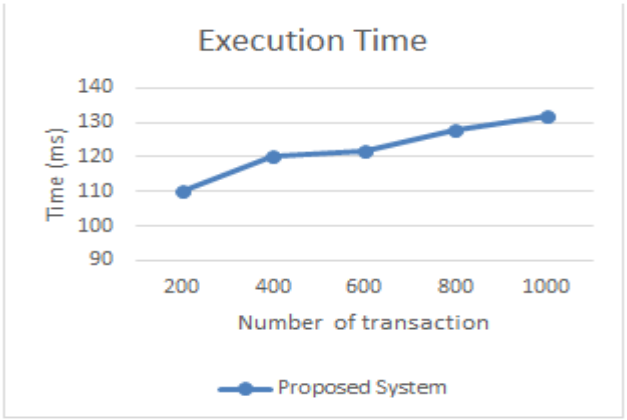

Figure 3. Execution Time Vs. Number of Records

Communication Cost

Sites are communicate with other to find global frequent for that sites transfer local frequent to other sites. Existing perform union and proposed perform scalar product for computation.

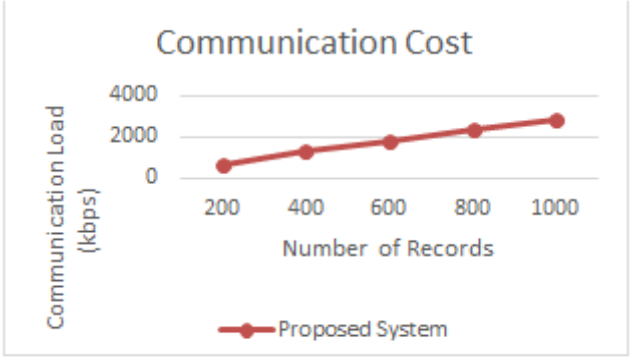

Figure 4. Communication Load Vs. Number of Records.

Conclusion

Privacy is the major concern to protect the sensitive information. This work is for the vertical distributed environment. The Scalar product with Boolean matrix is used for finding frequent item sets with secure computation between multiple sites with data miner which improved efficiency and privacy of system. The analysis result graphs show that the execution time of proposed work is less than the existing system with different number of records and supports hence, improve the efficiency.

References

[1] Chris Clifton and Donald Marks, Security and privacy implications of data mining, In Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (1996), 15–19.

[2] Daniel E. O'Leary, Knowledge Discovery as a Threat to Database Security, In Proceedings of the 1st International Conference on Knowledge Discovery and Databases (1991), 107–516.

[3] Evfimievski, A., R. Srikant, R. Agrawal and J. Gehrke. "Privacy preserving mining of association rules". Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-25, ACM Press, Edmonton, AB., Canada, pp. 1-12,2002.

[4] Pingshui WANG "Survey on Privacy Preserving Data Mining" International Journal of Digital Content Technology and its Applications Volume 4, Number 9, December 2010

[5] R. Agrawal, R. Srikant, "Fast algorithms for mining association rules,"In: Proc. 20th Int'l Conf. Very Large Data Bases, 1994, pp. 487–499.

[6] Vaidya, J. & Clifton, C.W., "Privacy preserving association rule miningin vertically partitioned data," In Proceedings of the eighth ACMSIGKDD international conference on knowledge discovery and datamining, Edmonton, Canada, July 2002.

[7] Pingshui WANG "Survey on Privacy Preserving Data Mining" International Journal of Digital Content Technology and its Applications Volume 4, Number 9, December 2010.

[8] Saranya, K., K. Premalatha, and S. S. Rajasekar. "A survey on privacy preserving data mining." Electronics and Communication Systems (ICECS), 2015 2nd International Conference on. IEEE, 2015.

[9] Zhu Yu-Quan, Tang Yang, Chen Geng, "A Privacy Preserving Algorithm for Mining Distributed Association Rules". IEEE, 2011.

[10] Association Rules Mining in Vertically Distributed Databases, P.Kalaivani,D.Kerana Hanirex, Dr.K.P.Kaliyamurthie, IJIRCCE - March 2015.

[11] Tassa, Tamir. "Secure mining of association rules in horizontally distributed databases." Knowledge and Data Engineering, IEEE Transactions on 26.4 (2014): 970-983.

[12] Chris Clifton,Murat Kantarcioglou,XiadongLin,and Michaed Y.Zhu,"Tools for privacy preserving distributed data mining," SIGKDDExplorations 4 (2002),

Author Profiles

Jinkal Patel Dept. of Computer Science, Northwestern Polytechnic University, Fremont, CA-94538,USA.

Harish Dogiparthi Dept. of Computer Science, Northwestern Polytechnic University, Fremont, CA-94538, USA.

Deepak Ganta Dept. of Computer Science, Northwestern Polytechnic University, Fremont, CA-94538, USA.

KiranLoudya Dept. of Computer Science Northwestern Polytechnic University, Fremont, CA-94538, USA.