Privacy Preserving in Horizontally Partitioned Data Based on association Rules Mining

Dharmik Makwana[1], Mr.Krunal Panchal[2]

[1] PG Student, Information Technology, LJIET, Ahmedabad, Gujarat, India
[2] Assistant Professor, Information Technology, LJIET, Ahmedabad, Gujarat, India

ABSTRACT

The advances of data mining techniques played an important role in many areas for various applications. In context of privacy and security issues, the problems caused by association rule mining technique are recently investigated. The misuse of this technique may disclose the database owner's sensitive information to others. Hence, the privacy of individuals is not maintained. Many of the researchers have recently made an effort to preserve privacy of sensitive knowledge or information in a real database. In this paper, we have modified EMHS Algorithm to improve its efficiency by using Elliptic Curve Cryptography. Analysis of the experiment on various datasets show that proposed algorithm is efficient compared to EMHS in terms of computation time.

Keyword : Data Mining, Elliptic Curve Cryptography, EMHS, Privacy, Privacy Preserving Association Rule Mining

1. INTRODUCTION

Data mining or knowledge discovery techniques such as association rule mining, classification, clustering, sequence mining, etc. have been most widely used in today's information world [1]. Successful application of these techniques has been demonstrated in many areas like marketing, medical analysis, business, Bioinformatics, product control and some other areas that benefit commercial, social and humanitarian activities.

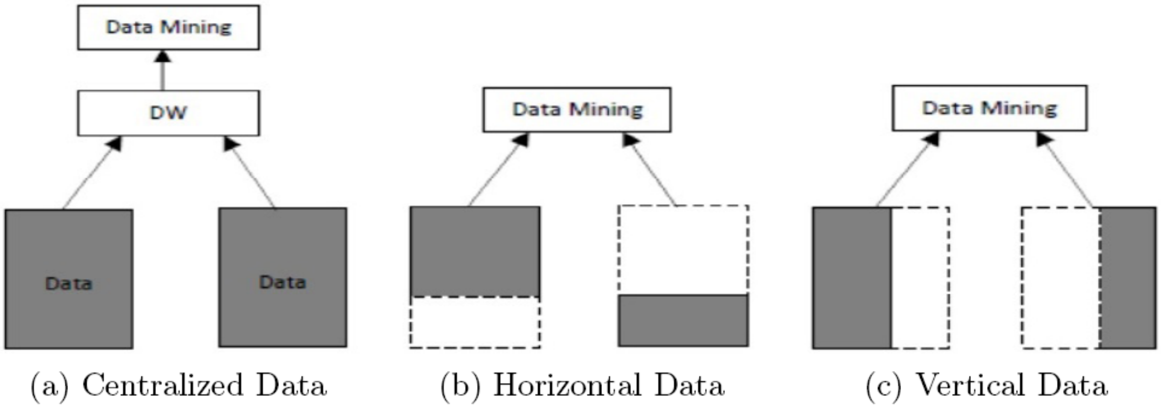

Fig.1 Different Database Environment

These techniques have been demonstrated in centralized as well as distributed environments. In centralized environment, all the datasets are collected at central site (data warehouse) and then mining operation is performed, as shown in Fig.1a, 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. As shown in Fig.1b and Fig.1c Data are distributed among two sites which wish to find the global mining result. The horizontal partitioned data shown in Fig.1b and Fig.1c shows vertical partitioned data.

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 [1]. Recently these techniques are investigated in terms of privacy and security issues and it is concluded that these techniques threat to the privacy of individuals information. That means one (e.g. adversary or malicious user) can easily infer someone's sensitive information (or knowledge) by mining technique. So, sensitive information should be hidden in database before releasing. For distributed mining it should be protected from the involving parties (or sites) who wish to find global mining result [2]. Therefore, to preserve privacy for sensitive knowledge, privacy preserving data mining (PPDM) become a hot directive in data or knowledge engineering field.

2. ASSOCIATION RULE MINING

Association Rule Mining is a popular technique in data mining for discovering interesting relations between items in large databases. It is purposeful to identify strong rules discovered in the databases using different available measures. Based on the concept of strong rules, Rakesh Agrawal et al[3]. Described association rules for discovering similarities between products in large scale transaction data in supermarkets. For example, the rule (Bread, Butter) ⇒ (Milk) found in the sales data of a shop would indicate that if a customer buys bread and butter together, he or she is likely to also buy milk. Such information can be used in decision making about marketing policies such as, e.g., product offer , product sales and discount schemes. In addition to the above mentioned example association rules are used today in many application areas including Web usage mining, Intrusion detection, Continuous production, and Bioinformatics [3]. As opposed to sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions.

The problem of association rule mining [3] is defined as: Let I= $\{i_1, i_2, . . . , i_n\}$ be a set of n binary attributes called items. Let D=$\{t_1, t_2, . . . , t_m\}$ be a set of transactions called the database. Each transaction in database D has a unique transaction identity ID and contains a subset of the items in I [3].

A rule is defined as an implication of the form X ⇒ Y where X,Y is subset of I and $X \cap Y = \emptyset$. The sets of items X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule respectively. The Support and Confidence [3] of the rule X ⇒ Y is calculated using following equation:

$$Support(X \Rightarrow Y) = \frac{(X \cup Y).count}{n}$$

$$Confidence(X \Rightarrow Y) = \frac{(X \cup Y).count}{X.count}$$

The most famous application of association rules is its use for Market Basket Analysis [4]. Association Rules are helpful in many fields like Telecommunication and Medical records for retrieving some desired results. Association rules has been used in mining web server log files to discover the patterns that accesses different resources continuously or accessing particular resource at regular interval. Association rules are also useful in mining census data, text document, health insurance and catalog design [4].

3. RELATED WORK

To understand the background of privacy preserving in association rule mining, we present different techniques and algorithm in the following subsections.

3.1 Secure Multiparty Computation with Trusted Third Party

This technique worked as a client server system where one site is a server responsible for the generating global result and all remaining sites are client sites which sends its encrypted data to the server to retrieve global result[5]. An example to SMC with trusted third party was PPDM-ARBSM algorithm[6]. This algorithm has mainly two servers: Data Mining Server and Cryptosystem Management Server[6]. A disadvantage of this algorithm was that the failure of third party fails the communication.

3.2 Secure Multiparty Computation with Semi Honest Model

This technique assumes all the sites as honest. One site acts as an initiator [7] and all others as sites. All the sites send their encrypted data to the next site in queue. Finally the last site sends all data to initiator which finds the global result [7]. An example to SMC with semi honest model was Fast Private Association Rule Mining for Securely Sharing algorithm [8]. The detailed description is mentioned in[8]. The limitation of this algorithm was the increase in computation time with the increase in the number of sites.

3.3 MHS Algorithm for Privacy Preserving on Horizontal Partitioned Database

MHS algorithm worked on minimum 3 sites. One site acts as an Initiator, one site acts as Combiner [9]. This algorithm used RSA cryptosystem. All sites find its frequent itemsets, encrypt it using RSA public key and send it to Combiner. The task of the combiner is to merge all the data with its own data and send it to the initiator. The task of the initiated was to decrypt all the data and generate global results[9]. As this algorithm was based on the concept of frequent itemsets, the limitation was the increase in computation time with the increase in the database size and number of sites.

3.4 EMHS Algorithm for Privacy Preserving Association Rule Mining on Horizontally Partitioned Database

EMHS algorithm was implemented in 3 phases. In the first phase, RSA cryptosystem was used. While in the second and third phase Homomorphic Paillier cryptosystem was used. The results showed better performance in the mining process as compared to other algorithms.

4. PROPOSED METHOD

Proposed system will generate sequential frequent pattern and rule which is generated by novel approach. The relationship between the pattern and the prediction of the occurrence of pattern can be identifying. The proposed algorithm is more efficient in terms of time and memory consumption.

4.1 Basic Concepts of New Algorithm

Suppose database D is distributed among n sites (S1,S2,..,Sn) in such a way that database Di (1<=i<=n) containing site Si consists of same set of attributes but different number of transactions. All sites are considered as semi honest. Now the problem is to mine valid global association rules satisfying given minimum support threshold (MST) and minimum confidence threshold (MCT) in unsecured environment, which should fulfill following privacy and security issues.

1. No any involving party should be able to know the contents of the transaction of any other involving parties.
2. Adversaries should not be able to effect the privacy and security of the information of involving parties by reading communication channel between involving parties.

4.2 Elliptic Curve Cryptography

Elliptic curve provides public cryptosystem based on the discrete logarithm problem over integer modulo a prime. Elliptic curve cryptosystem requires much shorter key length to provide a security level same as RSA with larger key length. A detailed overview of elliptic curves and an elliptic curve cryptosystem is given in[11].

4.3 Proposed Communication Protocol

The proposed communication protocol is defined in Fig. 2. Suppose there are 5 sites (namely Site 1, Site 2, Site 3, Site 4, Site 5) which contains datasets D1, D2, D3, D4 and D5. Among them, there are 2 sites, namely Initiator and Combiner. All the parties are semi-honest. Suppose that they want to find the global results without revealing their information to other sites. The proposed communication protocol is same as EMHS algorithm.

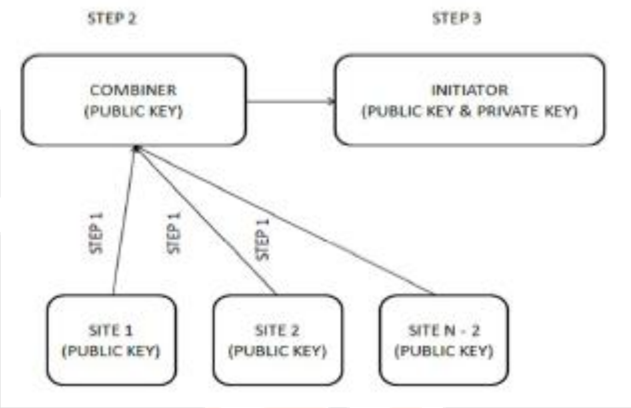

Fig.2 Proposed Communication Protocol

ALGORITHM 1: Improve Privacy

PHASE 1:

Step 1: The INITIATOR shares ECC public key Epu with all the sites also generates ECC private key Epk.

Step 2: Each sites computes its local FI (Frequent Itemsets) using Single Scan Algorithm.

Step 3: All the sites except INITIATOR and COMBINER encrypt its Local FI using ECC Public Key Epu and sends it to the Combiner.

Step 4: The encryption of the local support count of candidate X at site Si is denoted by E($X.supi$).

Step 5: With each X, combiner computes:
E($X.sup_{Combiner}$) = E($X.sup_{Combiner}$) * $\prod_{k=1}^{n-2}$ E($X.supk$)

Step 6: After this, encrypted data is sent to Initiator.

Step 7: Initiator decrypts the data using ECC Private Key(Epu) and generates a global support count of each candidate X as:
X.sup = D(E($X.supCombiner$)) + $X.supInitiator$

Step 8: Each site together computes:
$| DB | = \sum_{k=1}^{n} | DBi |$

Step 9: Initiator generates the global association rules and send/Publish to all other sites.

ALGORITHM 2: SINGLE SCAN

```
Input: Transaction Database and Min_Sup.

Output: Frequent Itemset.

Step 1: Create root for tree ({})
root → null

Step 2: For Each Transaction from DB
            Do
            Select the Item from Transaction and sort Alphanumeric order → Insert into tree and also add one
            count to I-list
            End
            End

Step 3: Discard items from I-list
         Support(I) < min_sup

Step 4: Arrange remaining items of I-list into descending order and Restructure the tree

Step 5: Mine the frequent item set from restructured tree doing Projection of conditional base tree
```

5. ANALYSIS OF THE PROPOSED ALGORITHM

Eclipse is used for compiling and executing purpose. Eclipse is an integrated development environment (IDE) with using SPMF tools. It contains a base workspace and an extensible plug-in system for customizing the environment which mostly written in Java. Experiment was carried out on real-life datasets having varied characteristics. Those datasets are Mushroom , Pumsb and Retail. In experiment the proposed system is compared with the existing algorithm. We evaluate EMHS and Proposed algorithm in terms of privacy and computation time.

5.1 Comparison in Terms of Privacy

EMHS and our proposed algorithm both satisfies semi-honest model. The smaller key size of ECC provides equivalent security as compared to RSA. Thus the privacy remains the same in EMHS and our proposed algorithm.

5.2 Comparison in Terms of Computation Time

Both EMHS and newly Proposed Algorithm are executed with the number of sites, increasing from 3 to 7 on real datasets. Based on the experimental results on 4 different datasets and the number of sites increasing from 3 to 7, the computation time is less as compared to EMHS algorithm. The comparison results of EMHS and Proposed Algorithm is shown in Fig.3a, Fig.3b and Fig.3d respectively.

5.3 Experimental Results

In implementation, each dataset is divided into 3 to 7 parts on the basis of the records. We implemented using Frequent Pattern Mining Framework (FPMF). single scan algorithm is used to find global itmset at each site in the Proposed algorithm.

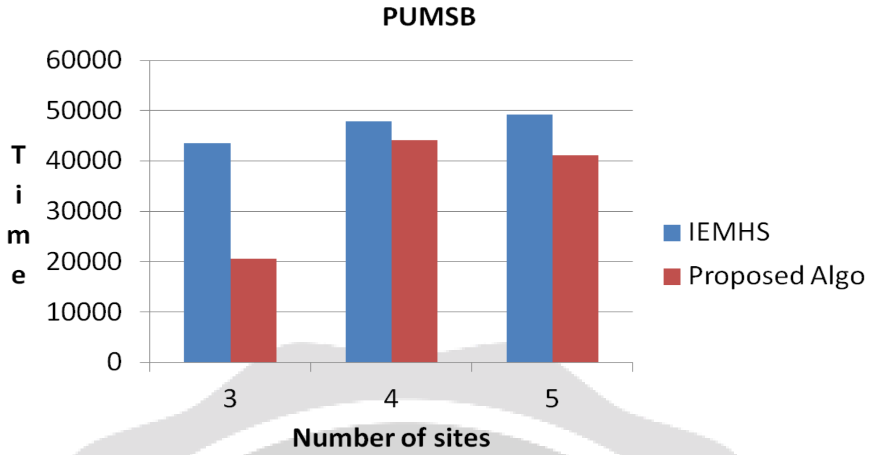

Chart 1: Execution Time comparison of IEMHS and Proposed Algorithm with PUMSB dataset

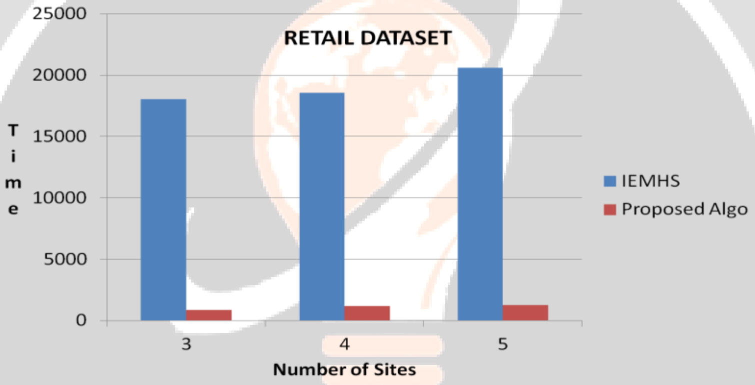

Chart 2: Memory Usage comparison of IEMHS and Proposed Algorithm with PUMSB dataset

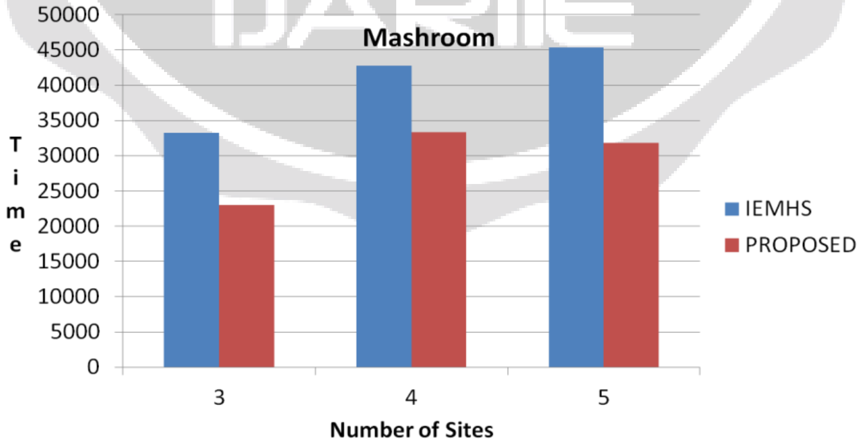

Chart 3: Memory Usage comparison of IEMHS and Proposed Algorithm with PUMSB dataset

6. CONCLUSIONS

In this paper, we proposed an algorithm to improve privacy and performance of EMHS when increasing the number of sites. We maintain the model of EMHS and apply ECC Cryptography. From the experimental results we conclude that the proposed algorithm has better performance than EMHS in dense datasets when increasing the number of sites. In future, we will try to solve the problem of large dataset between Initiator and Combiner.

7. REFERENCES

1. Atallah, M., Elmagarmid, A., Ibrahim, M., Bertino, E., Verykios, V.: Disclosure limitation of sensitive rules. In: Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange, KDEX 1999, pp. 45–52. IEEE Computer Society, Washington, DC (1999)

2. Clifton, C., Kantarcioglu, M., Vaidya, J.: Defining privacy for data mining. In: National Science Foundation Workshop on Next Generation Data Mining, pp. 126– 133 (2002)

3. Agrawal, R., Srikant, R.: Privacy-preserving data mining. In: Proceedings of the ACM SIGMOD Conference on Management of Data, pp. 439–450 (2000)

4. Verykios, V.S., Bertino, E., Fovino, I.N., Provenza, L.P., Saygin, Y., Theodoridis, Y.: State-of-the-art in Privacy Preserving Data Mining (March 2004)

5. Muthu Lakshmi, N.V., Sandhya Rani, K.: Privacy Preserving Association Rule Mining Without Trusted Party For Horizontally Partitioned Databases. Interna-tional Journal of Data Mining and Knowledge Management Process (IJDKP) 2(2) (March 2012)

6. Gui, Q., Cheng, X.-H.: A Privacy-Preserving Distributed Method for Mining As-sociation Rules. In: 2009 International Conference on Artificial Intelligence and Computational Intelligence, pp. 294–297 (2009)

7. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, pp. 227–245. Morgan Kaufmann Publishers Inc., San Francisco (2001)

8. Estivill-Castro, V., Hajyasien, A.: Fast Private Association Rule Mining by a Pro-tocol Securely Sharing Distributed Data. In: Proceedings of the 2007 IEEE Intelli-gence and Security Informatics (ISI 2007), New Brunswick, New Jersey, USA, May 23-24, pp. 324–330 (2007)

9. Hussein, M., El-Sisi, A., Ismail, N.: Fast Cryptographic Privacy Preserving Associ-ation Rules Mining on Distributed Homogenous Data Base. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS (LNAI), vol. 5178, pp. 607–616. Springer, Heidelberg (2008)

10. Xuan, C.N., Hoai, B.L., Tung, A.C.: An enhanced scheme for privacy preserv-ing association rules mining on horizontally distributed databases. In: 2012 IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), pp. 1–4 (2012)

Web Links

11. Stallings, W.: Cryptography and Network Security, 5th edn. (2011)

12. http://fimi.ua.ac.be/data/ 10.1007/BFb0014140.