Indian Journal of Science and Technology, Vol 9(48), DOI: 10.17485/ijst/2016/v9i48/93737, December 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

# Privacy Preserving Association Rule Mining in Distributed Environments using Fp-Growth Algorithm and Elliptic Curve Cryptography

**T. Nusrat Jabeen[1*] and M. Chidambaram[2]**

[1]Department of Computer Science, Bharathiyar University, Coimbatore, Tamil Nadu, India; nushjabeen@gmail.com
[2]Computer Science Department, Rajah Serfoji Government College, Thanjavur, Tamil Nadu, India; chidsuba@gmail.com

## Abstract

In this paper, we present a privacy preserving association rule mining method in distributed environments. The proposed method uses FP-growth algorithm and combination of elliptic curve cryptography and digital signature. The proposed method is low in cost for both computation and communication. This method uses associative third party, which is the central authority, holding common cryptographic keys of all the sites and responsible for performing distributed ARM. This method maintains privacy regardless of any number of dishonest sites in the distributed environment.

**Keywords:** Association Rule Mining, ARM, Cryptography, Digital Signature, Elliptic Curve, FP-Growth, Privacy in ARM

## 1. Introduction

Association rule mining, abbreviated as ARM, finds frequent patterns, associations and correlations among different items of a data set. The procedure aims to find the occurrence of a specific item based on occurrences of other items in a transaction. Association rule mining plays an important role in applications like market basket analysis, cross marketing and catalogue design[1]. ARM is also widely used in statistical analysis and decision making problems.

Association rule mining can be performed on centralized data environments – all the relevant data are collected in one point and association rule mining process is applied. It can also be done in distributed data environments where data resides in different sites[2]. All the data set may not be supplied for association rule mining process. Only required information is supplied. Here the data set can be of two types: vertically partitioned data horizontally partitioned data[2].

Consider a situation of insurance company performing ARM to find the correlations among its customers insurance amount, age, risk coverage, etc. If association rule mining is performed on multiple databases, located at different sites, a distributed architecture is needed. Different sites contain at most same set of attributes with different volumes of transaction. Owners of different sites are ready to share their data to perform distributed association rule mining[3]. But they want to maintain the privacy of its customers and sites as this information are private and confidential[3].

Generally it is believed that the communication channel in distributed association rule mining process, through which parties share their information, is secure. But this is not always correct and information leakage is encountered in so many situations[4]. Hence it becomes necessary to protect the customer data not only with fellow parties but also with unauthorized third parties. Hence we propose a cryptography based association rule mining algorithm that preserves privacy and maintains security. The algorithm used in our approach is FP-growth algorithm along with combination of Elliptic Curvy Cryptography and Digital Signature.

*Author for correspondence

## 2. Related Work

Privacy preservation in distributed association rule mining method on horizontally partitioned data was proposed[5]. The method works in two phases such as discovering candidate item sets and selecting candidate item sets with global support and threshold value. The first phase uses commutative encryption so that encrypted item sets are passed to other parties.

Privacy preserved association rule mining in vertically partitioned data. It uses component algorithm which securely computes scalar product[6]. Instead of using cryptographic solutions, algebraic solutions are used which hides true values by replacing them with masked random values. Privacy preservation and sharing of data in distributed environments for perturbed data[7]. Here randomization perturbation technique is used which perturbs sensitive data. The altered data is used to perform data mining tasks in distributed environments thereby maintaining privacy. Secure two party association rule mining algorithm[8]. Here homomorphic encryption based solution is provided that produces more accurate results than data perturbation while maintaining accuracy. In this algorithm, there is no need to share support and confidence and they are calculated homomorphically. This improves the privacy and security in association rule mining process. Another forte of this algorithm is that actual data value is reduced to single bit.

An algorithm on privacy preserved association rule mining in unsecured distributed environments[9]. In this algorithm, elliptic curve cryptography is used to maintain privacy over horizontally partitioned data. ECDSA, which is the abbreviation for Elliptic Curve based Digital Signature Algorithm, is used to provide authentication between two parties in unsecured environments. And to provide privacy among parties, Elliptic Curve Integrated Encryption Scheme (ECIES) is used.

A novel method to maintain privacy during association rule mining over horizontally distributed databases[10]. The drawback of using cryptographic technique is that it needs considerable computation and slower. In the proposed protocol, it combines the advantages of privacy preserving methods. Three methods for randomizing data viz. support values of each L. L. item set, M.L.L. item set and safe computation of support values of each L. L. item set.

### 2.1 Association Rule Mining

Let $I=\{i_1, i_2, i_3, \ldots i_n\}$ be the set of n items and let D be the set of database transaction with each transaction. T represents a set of items such that $T \subseteq I$. Each transaction in D is identified by a unique identifier called $T_{id}$. Let A, B be two set of items; Association rule is of the form $A \rightarrow B$ where $A \subset I$, $B \subset I$ and $A \cap B = \phi$, A is called antecedent of the rule and B is called consequent of the rule.

This rule holds a set of transactions with user specified support and confidence. The support of the rule is defined as Probability $P(A \cup B)$ and confidence of the rule is the conditional probability $P(B \mid A)$. Association rule mining consists of generating all the rules from the transactional database D which are having support and confidence higher than user specified thresholds[11].

## 3. Problem Statement

Let D be the transaction database. If the database is distributed across n sites ($S_1, S_2, \ldots S_n$) such that Di ($1 \leq i \leq n$) and given that all the sites are semi honest, the problem of maintaining privacy from dishonest sites occurs. The requirements of the system is defined as : given minimum support and confidence values, mine all the global association rules having greater than or equal to support and confidence threshold levels, that satisfy the following privacy issues:

- No site is allowed to know the exact transaction of other sites involved in association rule mining process.
- Intermediate adversaries should not be able to view, predict the data by reading the communication channel between involved sites.

### 3.1 Elliptic Curve Cryptography

It is a type of public key cryptography based on elliptic curve theory. It is used to create faster, smaller and efficient cryptographic keys. Elliptic curve cryptography can be combined with other public key cryptosystems like RSA, Diffie-Hellman, etc.[12]. For our proposed method, we use combination of Elliptic Curve Cryptography and Diffie-Hellman (ECDH) algorithm to exchange key among different sites. Another algorithm called ECDSA–Elliptic Curve based Digital Signature Algorithm, for authentication and verification purpose.

Elliptic Curve Diffie Hellman algorithm is used for key agreement between two parties[13]. Suppose two sites $S_1$ and $S_2$, wish to exchange key with each other. First site $S_1$ generates a private key $d_A$. Public key is generated as $Q_A = d_A G$, where 'G' is the generator of the curve. Similarly the site $S_2$ generates a private key $d_B$ and public

key $Q_B = d_B G$. Then if site $S_2$ wants to send its public key to $S_1$, it calculates as

$$d_A Q_B = d_A d_B G$$

And similarly S1 sends its public key to S2 by calculating

$$d_B Q_A = d_A d_B G$$

Here the product $d_A d_B G$ is said to be shared secret and any intruder would only know $Q_A$ and $Q_B$, but would not be able to calculate shared secret. ECDH is faster as it uses addition instead of exponentiation.

Digital signature algorithm based on elliptic curve is said to be ECDSA. It generates digital signature, a pair of numbers for a message. ECDSA works in three phase's viz. key generation, signature generation and signature verification[14]. Let the domain parameters of a site S1, denoted by D such that D = {q, $F_R$, a,b,G,n,h} and E is an elliptic curve defined over $F_q$. The key is generated by selecting a random number for the interval [1, n-1], computing $d_p$ and generating public and private keys. The algorithm for key generation is as follows

- Select a psuedo random integer d between the interval [1, n-1].
- Compute $K_p = x_1, y_1$.
- Computer r = $x_1$ mod n such that value of $x_1$ should be between 0 to q-1.
- Compute $K^{-1}$ mod n.
- Compute S = $K^{-1}$ {h(m)+$d_r$} mod n where h is the hash value from SHA-1.

The signature for the message is the pairs of calculated integers r and s. The same procedure is repeated for another site $S_2$.

## 4. Proposed Method

The objective of the proposed method is to maintain privacy during association rule mining using elliptic curve cryptography. The proposed method is demonstrated using sample three sites $S_1$, $S_2$ and $S_3$ whose homogeneous datasets are $D_1$, $D_2$ and $D_3$ respectively. The authority responsible for performing global association rule mining is said to be Associative Third Party (ATP), to whom all the sites send their data. The two assumptions that need to be used to demonstrate our proposed method is that 1. All the sites involving in the global association rule mining process are semi honest and 2. The communication channel through which data are transmitted is unsecure. Now the objective is to find the global support of an item without disclosing transaction details to other sites involved. The procedure as follows:

- All the sites generate its common key and send it to ATP.
- Each site, S, encrypt the transaction count of its database (|D|) using its key (K) and sends this count to its neighbour site. The authentication of neighbour is done using ECDSA algorithm.
- The next site receives the transaction count from its predecessor and adds its encrypted count of transactions. Now the transaction count contains ($|D_1|$ + $K_1$ + $|D_2|$ + $K_2$) where $|D_1|$ and $|D_2|$ are the transaction counts of previous and current site, $K_1$ and $K_2$ are the common keys of previous and current site respectively.
- The above procedure is repeated by all the sites in the network and count along with the respective secret key is sent to associative third party
- ATP substantiates the geniuses of the received transaction count by decrypting the last site's key. Hence ATP gets global transaction count.
- The above procedure is repeated for getting local support count of the item.
- Each site encrypts local support count, X, using its Key (K) and sends to its neighbour site.
- Next site receives the local support count of its predecessor and add its encrypted local support count with that.
- Now the local support count contains $L_1(X) + K_1 + L_2(X) + K_2$, where $L_1$, $L_2$ are the local support counts of site 1 and site 2, $K_1$ and $K_2$ are the private keys of site1 and site2 respectively.
- All the remaining sites in the network adopt the above steps to send their local support to the ATP.
- The authenticity of the total local support is verified by ATP by subtracting the summation with all the shared keys.
- Global support of item X is computed by GS (X) | G(|D|) by ATP.

The proposed method ensures security and no site including central authority, ATP, can be able to view the contents of transactions of other sites. Moreover, an intruder cannot alter the message as it is protected by digital signature with encryption. Hence it is secure. The overall working of the proposed system is presented in Figure 1.

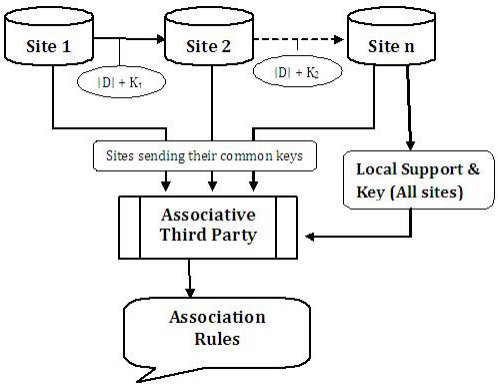


**Figure 1.** Proposed method.

During association rule mining is performed in two steps. Global frequent item sets are mined in the first step and possible association rules from the frequent item sets are generated in the second step. From the list of steps discussed above, ATP knows global count of transactions. It also knows the local support count of each candidate item set. Now ATP calculates the global frequent item sets from items whose support count is above the minimum support specified by the user. Frequent item sets whose value is below the minimum support are eliminated. ATP generates all the possible association rules that satisfied the minimum support and publishes the rules.

## 5. Results and Discussion

Since each site sends the local support only its local support to its next site, no site is able to know the actual transactions. Additionally, each site encrypts and sends the data only to its neighbour after encryption, the data is secure. ATP can also view the total local support, i.e. local support of all the sites but not the individual local support of each site. Thus the proposed method is secure regardless of any number of dishonest sites in the network.

If the intruder intrudes in communication either between two sites or between a site and ATP, he/she cannot view, read or alter the message as it is transmitted after digitally signing it using ECDSA algorithm. Moreover, data is encrypted using private key of sites, intruder cannot decrypt the message.

The communication cost of the proposed method is also less. The cost incurred to send local support and count by the site to its neighbours is alone considered. If each site contains $2^{m-1}$ non empty candidate item sets, then the total cost for each site to send candidate item set to ATP is $n(2^{m-1})$, where n represents number of sites involved. Hence total communication cost for all the sites is $n+n(2^{m-1})$ which is approximately equivalent to $2^m$.

In the proposed method, each site has to generate the key and send it to ATP using digital signature for secure communication. Since computation is involved, it is necessary to check the communication cost. Let $m^2$ represents the cost for generating non-empty candidate item sets for each site. Let $C_1$ and $C_2$ represent the cost of key generation and digital signature. Then the total computation cost for both key generation and digital signature is given by $O(n(2^m+m^2))$, where m represents number of items and n represents total number of sites. The computational cost for generating global frequent item set and association rules are $O(n^{2m})$ and $O(2^{2m} + m^2)$ respectively, which is reasonably less.

## 6. Conclusion

In this paper, a method for maintaining security and privacy during distributed association rule mining process is presented. The proposed method uses FP-Growth algorithm along with elliptic curve cryptography and elliptic curve based digital signature algorithm for authentication of sites and verification of data. This method is secure and privacy is maintained. The proposed method uses central authority called association third party which collects all the information relating to the association rule mining under secure manner. Since ATP alone knows the public key of all the sites. The intruder cannot view, read and alter the data. Additionally the communication cost and computational cost to implement the proposed method is reasonably low. Therefore, it is concluded that the proposed method is highly secure, maintains privacy regardless of any number of dishonest sites in the network and economically cheap.

## 7. References

1. Yinz Y, Kaku I, Tang J, Zhu J. Association rules mining in inventory database. Data Mining. London: Springer; 2011. p. 9-23.
2. Prakash S, Shanmugam V, Murugesan APP. Privacy preserving combinatorial function for multi-partitioned data sets. International Journal of Computer Applications. 2012; 44(8):8-10.
3. Slavkovic S, Aleksandra B, Nardi Y, Matthew M, Tibbits T. Secure logistic regression of horizontally and vertically partitioned distributed databases. IEEE 7th IEEE

International Conference on Data Mining Workshops (ICDM); Pittsburgh. 2007. p. 1-6.
4. Krishnamurthy K, Balachander B, Craig E, Wills W. On the leakage of personally identifiable information via online social networks. Proceedings of the 2nd ACM Workshop on Online Social Networks. USA. 2009. p. 1-6.
5. Kantarcioglu K, Murat M, Clifton C. Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE Transactions on Knowledge and Data Engineering. 2004; 16(9):1026-37.
6. Vaidya V, Jaideep J, Clifton C. Privacy preserving association rule mining in vertically partitioned data. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; New York. 2002. p. 639-44.
7. Kamakshi P, Babu AV. Preserving privacy and sharing the data in distributed environment using cryptographic technique on perturbed data. Journal of Computing. 2010; 2(4):115-9.
8. Kaosar K, Golam MD, Paulet R, Yi X. Secure two-party association rule mining. Proceedings of the 9th Australasian Information Security Conference Australian Computer Society, Inc. Australia. 2011; 116:15-22.
9. Patel P, Ashish C, Rao UP, Dhiren R, Patee P. Privacy preserving association rules in unsecured distributed environment using cryptography. The Proceedings of the 3rd International Conference on Computing Communication and Networking Technologies (ICCCNT); Coimbatore. 2012. p. 1-5.
10. Alborzi A, Ziaeddin S, Raji F, Mohammad H, Saraee S. Privacy preserving mining of association rules on horizontally distributed databases. International Proceedings of Computer Science and Information Technology; West Lafayette. 2012. p. 1-21.
11. Karthikeyan T, Vembandasamy K. A novel algorithm to diagnosis type II diabetes mellitus based on association rule mining using MPSO-LSSVM with outlier detection method. Indian Journal of Science and Technology. 2015 Apr; 8(S8):1-11.
12. Tseng T, Huei-Ru H, Jan RH, Yang W. A chaotic maps-based key agreement protocol that preserves user anonymity. IEEE International Conference on Communications; Dresden. 2009. p. 1-6.
13. Kayal P, Kannan S. A partial weighted utility measure for fuzzy association rule mining. Indian Journal of Science and Technology. 2016 Mar; 9(10):1-6.
14. Sinha S, Anshuman A. A survey of system security in contactless electronic passports. International Journal of Critical Infrastructure Protection. 2011; 4(3):154-64.