Elliptic Curve Cryptography Based Mining of Privacy Preserving Association Rules in Unsecured Distributed Environment

Chirag N. Modi¹, Udaya Pratap Rao²
Department of Computer Engineering,
Sardar Vallabhbhai National Institute of Technology,
Surat, Gujarat, India-395 007
¹cnmodi.956@gmail.com
²upr@coed.svnit.ac.in

Dhiren R. Patel
Department of Computer Science & Engineering,
Indian Institute of Technology Gandhinagar,
Ahmedabad, Gujarat, India-382 424
dhiren@iitgn.ac.in

Abstract—Distributed data mining techniques are often used for various applications. In terms of privacy and security issues, these techniques are recently investigated with a conclusion that they reveal data or information to each other parties involved to find global valid results. But because of privacy issues, involving parties do not want to reveal such type of data. Recently many cryptography techniques have been found to address privacy problems in distributed mining. In this paper, we propose an elliptic curve cryptography based algorithm to mine privacy-preserving association rules on horizontal partitioned data. Moreover, we have also considered unsecured communication channels in distributed environment. Proposed algorithm provides privacy and security against involving parties and other parties (adversaries) who can reveal information by reading unsecured channel between involving parties. Finally, we analyze the privacy and security provided by proposed algorithm and also discuss the communication and computation cost of proposed algorithm.

Keywords—Data Mining; Association Rules; Distributed Databases; Privacy; Security; Elliptic Curve Cryptography;

I. INTRODUCTION

Data mining techniques are most widely used techniques for various applications like marketing, statistical analysis, decision making etc. These techniques have been demonstrated in centralized or distributed environments. In centralized environment, all the datasets are collected at central 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. They send only required information to each other. There are two types of distributed data considered. One is horizontal partitioned data and another is vertical partitioned data. As shown in Fig. 1 data are distributed among two sites which wish to find global mining result. In horizontal partitioned data, each site contains same set of attributes but different number of transactions where in vertical partitioned data each site contains different number of attributes but same set of transactions. Recently distributed mining techniques are investigated in terms of privacy and security issues and it is concluded that these techniques threat to the privacy of individuals. Recently many cryptography techniques have been found for mining association rules in distributed environment without revealing information to involving parties. The authors in [1] were the first to discuss security and privacy implication of data mining. The authors in [2] presented some suggestions for defining and measuring privacy preservation and showed how these relate to both privacy policy and practice in wider community.

A. Motivation

Suppose the database of one company is distributed among several sites. Each site contains same set of attributes and different number of transactions. Now, for marketing analysis, they want to do joint association rule mining over the databases of all sites. However, it will violate privacy of both customers and sites. The customers do not want to share their private information without confirmation of their privacy protection and each sites also unwilling to share their data because of privacy concern.

Existing approaches mine association rules without revealing information to each involving party. For mining association rules, each site sends some information instead of transaction data to others. They assume that the communication channel between sites is secured. But in real life scenario, channel is unsecured. So, adversary can read communication channel between parties and may affect the
global mining result or privacy of information. So, there is another issue to protect privacy against unauthorized access.

In this paper we propose cryptography approach for mining privacy-preserving association rules. We have used elliptic curve based algorithms named Diffie-Hellman and DSA (Digital Signature Algorithm). Our approach provides privacy and security at certain level against both the involving parties and adversary at reasonable communication and computation cost.

The rest of the paper is organized as follows: In section 2, we discuss some related background. In section 3, we discuss some basic concepts which are used in our proposed technique, while a detailed description of proposed approach is given in section 4. In section 5, we discuss and analyze the proposed algorithm. Finally in section 6, we conclude and define some future direction.

II. RELATED BACKGROUND

A. Association Rule in Distributed Environment

The standard techniques for association rule mining in centralized environment are given in [3]. The standard definition of association rule in distributed environment (on horizontal partitioned data) is given as follows [4]: Let \( I = \{i_1, \ldots, i_m\} \) be a set of items. Database \( D \) is a set of transactions, which is distributed among \( n \) sites named \( S_1, S_2, \ldots, S_n \) in such a way that database \( D_i \) \((1 \leq i \leq n)\) containing at site \( i \) consists of same set of items but different number of transactions.

The itemset \( X \subseteq I \) has local support count \( li(X) \) at site \( S_i \) if \( X \) is contained in \( li(X) \) of the transactions at site \( S_i \). The global support count of \( X \) is given as \( g(X) = \sum_{i=1}^{n} li(X) \). An itemset \( X \subseteq I \) is globally supported if \( g(X) \geq s \times \sum_{i=1}^{n} |D_i| \), where \( s \) is called minimum support threshold. Now association rule is an implication formula like \( X \Rightarrow Y \), where \( X \subseteq I \), \( Y \subseteq I \). The rule with global support and global confidence is called, if \( g(X \cup Y) \geq s \times \sum_{i=1}^{n} |D_i| \) and \( g(X \cup Y) \geq c \times g(X) \), where \( c \) called minimum confidence threshold. The efficient algorithms for mining distributed association rule are given in [5].

B. Privacy Preserving Association Rule Mining Approaches

There have been several approaches found for privacy preserving association rule mining. For centralized databases, the authors in [6][7][8][9][10] proposed heuristic based algorithms for providing privacy for sensitive association rules. The reconstruction based approaches [11], border based approaches [12], exact approaches [13] are also found for preserving privacy for sensitive association rules in centralized databases. For distributed environment, there are many cryptography techniques have been found. In following we discuss some existing cryptography based approaches for mining of privacy-preserving association rules in distributed environments.

The authors in [14] proposed an approach for sharing association rules when data are vertically partitioned. Proposed approach uses the scalar product over the vertical bit representation of itemsets to compute the frequency of the itemsets. The authors proposed a secure two party algorithm for mining frequent itemset securely. This approach is quite effective in terms of communication cost. But it is very expensive for large datasets.

Kantarcioglu and Clifton [4] addressed the secure mining of association rules over horizontal partitioned data, which uses the secure set union to get the union of candidate association rules without revealing sensitive information to any involving party. Then it uses secure summation and secure comparison which are used to remove candidate itemsets that are not supported globally. Proposed approach is very efficient in terms of communication and computation cost.

Jie Liu et al. [15] proposed an efficient and secure algorithm for mining privacy preserved association rules in distributed environment, which uses hashing of candidate itemsets and filter out the itemset whose support exceeds the minimum thresholds. The authors proposed EP-DMA algorithm which builds a global hash table in every time of iteration for pruning candidate 2 itemset. At the time of building global hash table, cryptography techniques are also applied for security. Proposed algorithm uses secure set union for finding global support count of item securely.

Above presented approaches provide security only against involving parties in distributed environment. If the communication channel is not secured between parties, an adversary can read the channel and may be able to affect the final mining result or privacy. In this paper we propose an elliptic curve cryptography based algorithm for mining privacy-preserving association rule on horizontal partitioned data, which provides privacy for association rules against involving parties and adversary.

III. BASIC CONCEPTS USED IN PROPOSED TECHNIQUE

A. Problem Definition

Suppose database \( D \) is distributed among \( n \) sites \((S_1,S_2,\ldots,S_n)\) in such a way that database \( D_i \) \((1 \leq i \leq n)\) containing site \( i \) 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 affect the privacy and security of the information of involving parties.
by reading communication channel between involving parties.

B. Elliptic Curve Cryptography

Elliptic curve provides public cryptosystem based on discrete logarithm problem over integer modulo a prime. Elliptic curve cryptosystem requires much shorter key length to provide security level same as RSA with larger key length. A detailed overview of elliptic curve and elliptic curve cryptosystem is given in [16][17]. We used elliptic curve based Diffie-Hellman (ECDH) protocol and elliptic curve based Digital Signature algorithm (ECDSA) for key exchange and authentication-verification respectively in our proposed algorithm. In following we give an overview of ECDH protocol and ECDSA.

1) Elliptic curve based Diffie-Hellman (ECDH) protocol for key sharing: Suppose there are two sites named Site 1 and Site 2 which want to share a key securely. They both agreed on elliptic curve E(p,a,b), where p is large prime number and base point g with a high order of n such that ng=O. Site 1 and Site 2 generate random number (their private key) R1 and R2 respectively, where R1,R2 < n. Now Site 1 generates public key (Y) by adding base point (g) R1 times, where Site 2 generates public key (Y') by adding base point (g) R1 times. Then both sites send their public key to each other. Finally, they compute shared key by their private key times adding received public key. ECDH is faster since it uses addition instead of exponentiation and produces common key securely.

2) Elliptic curve based digital signature algorithm (ECDSA) for authentication and verification: Suppose Site 1 selects an elliptic curve E(p,a,b) defined over finite field Zp. The number points in this field should be divisible large prime number n. Now a large integer d in the interval [1,n-1] is selected as a private key of Site 1 and the point P of order n is selected to compute public key Q=dP of Site 1. So, Site 1 contains public key (E,P,n,Q) and private key d. For authentication, it generates signature for message (suppose M) using following algorithm.

- **Digital signature generation at Site 1:** First a random number k is selected in the interval [1,n-1] and kp=(x,y) calculated. Then taking r=x mod n, s=k^{-1}(h(M)+dr) mod n is calculated, where h(M) is the hash of message M. The (r,s) is taken as a digital signature of message M, which is called authentication of message M, to be sent to Site 2 for verification.

- **Digital signature verification at Site 2:** Site 2 obtains authenticated copy of message M and public key (E,P,n,Q) from Site 1. To verify (r,s) in the interval [1,n-1], first w = s^{-1}mod n and h(M) is computed. Then U_1+U_2Q=(x_1,y_1) is calculated, where U_1=h(M)w mod n and U_2=rw mod n. Finally it calculates v=x_1 mod n and accepts signature, if v=r.

In our proposed algorithm we used ECDH protocol for key sharing and to encrypt support count of itemsets before sending so that any involving will not be able to infer information of others, where ECDSA used for authentication and verification between involving parties to unable adversaries for affecting global result.

IV. PROPOSED PRIVACY PRESERVING RULE MINING ALGORITHM

A. Proposed Communication Protocol

Proposed communication protocol is shown in Fig. 2. Suppose there are three sites (namely Site 1, Site 2 and Site 3) which contain homogeneous dataset D1, D2, and D3 respectively and third party (TP) who is capable for mining global association rules. All the involving parties are called semi-honest parties. Suppose they want to find a global support of an item X without revealing any transaction data to involving sites. The communication channels among involving parties are also unsecured. So, the proposed communication protocol works as following steps to compute global support of an item (X).

![Figure 2: Proposed communication protocol used for unsecured distributed environment.](image)

1) As shown in Fig. 2, firstly sites Si (1 ≤ i ≤ n) share a common key Ki (1 ≤ i ≤ n) with TP. The key Ki is derived by using ECDH algorithm.

2) Site 1 encrypts the count of transaction in D1 by adding key K1 and sends (∑(D1[K1]) to Site 2 after authenticating it by using ECDSA. This step is marked as (1) in Fig. 2.

3) Site 2 verifies signature by using public key of Site 1 and adds in its encrypted count of transactions (∑(D2[K2]) and sends (∑(D1[K1]+D2[K2]) to Site 3 after authenticating it, marked as (2).

4) Like Site 2, Site 3 verifies signature by using public key of Site 2 and adds its encrypted count of transactions (∑(D3[K3]) and sends (∑(D1[K1]+D2[K2]+D3[K3]) to TP after authenticating it,(3).

5) TP verifies signature by using public key of Site 3 and calculates global count of transaction by subtracting summation (K1+K2+K3) of all shared keys. So, TP
gets global g(|D|)=(|D1|+K1+|D2|+K2+|D3|+K3)-(K1+K2+K3).

6) Now, Site 1 sends encrypted local support count of X (l1(x)+K1) to Site 2, after signing it, (4).
7) Site 2 verifies signature by using public key of Site 1 and adds in its encrypted local support count of X (l2(X)+K2) and sends (l1(x)+K1+l2(X)+K2) to Site 3 after authenticating it, (5).
8) Site 3 verifies signature by using public key of Site 2 and adds its encrypted local support count of X (l3(X)+K3) and sends (l1(x)+K1+l2(X)+K2+l3(X)+K3) to TP after signing it, (6).
9) Then TP verifies signature by using public key of Site 3 and calculates global support count of X (g(X)) by subtracting summation (K1+K2+K3) of all shared keys. So, TP gets gs(X)=(l1(x)+K1+l2(X)+K2+l3(X)+K3)-(K1+K2+K3).
10) Finally TP calculates global support of X by dividing gs(X)/g(|D|).

In above described protocol no any site or TP can be able to know the contents of transactions of others and reveal the actual support count of X of any site to other sites. Also if an adversary reads communication channel, he/she cannot alter message or read actual support count of X since digital signature is used over encrypted support count of X. So, this protocol provides privacy of information against involving sites and also provides security and privacy against adversary under some other security assumption.

B. Proposed Algorithm for Privacy-Preserving Mining of Association Rules

In Table I, proposed privacy-preserving mining algorithm is shown. Like apriori algorithm [3], our proposed algorithm works on two phases but using different steps. First phase mines global frequent itemsets, where the second phase generates all possible association rules from the mined frequent itemset.

1) Global frequent itemset mining: First TP calculates global count of transaction by receiving summation of encrypted local count of transaction from last site, according to steps 1 to 5 of proposed communication protocol discussed in previous sub section. Then TP receives messages of local support count of each candidate itemset from last site and finds global support of each itemset by repeatedly following steps 6 to 10 of communication protocol. This step continues until global support of all the candidate itemsets are not computed. Now, TP finds global frequent itemsets whose global support satisfies given minimum support threshold (MST) and removes frequent itemsets whose all non-empty subsets are not frequent, for providing downward closure property. This phase ends by generating all proper global frequent itemsets in given homogeneous environment.

<table>
<thead>
<tr>
<th>Table I: Proposed Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> n homogeneous datasets denoted by ( D_i (1 \leq i \leq n) ), belong to each site ( S_i (i \geq 3) ), minimum support threshold (MST) and minimum confidence threshold (MCT).</td>
</tr>
<tr>
<td><strong>Output:</strong> Global Association rules over ( D ), where ( D = \sum_{i=1}^{n} D_i ).</td>
</tr>
</tbody>
</table>

1) TP calculates global count of transactions (g(|D|)) according to steps 1 to 5 of proposed communication protocol.
2) TP calculates global support count of all candidate itemsets by repeatedly followed steps 6 to 10 of communication protocol.
3) Then TP finds global frequent itemsets whose global support \( g_s \geq MST \).
4) TP removes frequent itemset whose all non-empty subsets are not frequent.
5) Finally, TP finds global association rules like \( X \Rightarrow Y \) whose global support \( g_s \geq MST \) and global confidence \( g_c \geq MCT \), where \( X, Y \) are globally frequent itemsets.

2) Generating possible global association rules from the mined frequent itemsets: According to genRules(F) function of apriori algorithm [3] for centralized database, TP finds all possible global association rules satisfying given minimum support threshold (MST) and minimum confidence threshold (MCT), by using mined global frequent itemsets. Finally, TP publish produced association rules to all involving sites.

The proposed algorithm provides valid global association rules securely as an output. In next section, we discuss privacy and security of information (or transaction data) provided by our proposed work.

V. ANALYSIS OF PROPOSED WORK

A. Privacy Against Involving Sites and TP

In proposed communication protocol each site sends only local support count of items to next site instead of original transaction data. So, next site or any other site cannot be able to know the actual contents of transactions. Moreover each site sends information to next site after encrypting it by using the shared key with TP. So any other site cannot predict the original value. TP also cannot predict local support of an itemset contained at any site because it gets total encrypted support count of itemset from the last site.

Thus, using proposed communication protocol no any involving site or TP can be able to know the actual content of transactions or information of other sites.

B. Privacy and security against adversary

If an adversary reads communication channel between involved any two sites or last site and TP, he/she cannot either alter the message or send fake message, because each site sends message to next site via communication channel after digitally signing it using ECDSA algorithm. Also an adversary can not affect the privacy of sent information because it is encrypted using shared key produced by ECDH.
VI. Conclusion and Future Scope

In this paper we proposed elliptic curve cryptography based algorithm to mine privacy-preserving association rules on homogeneous datasets in unsecured environment. We also analyzed security and privacy of proposed algorithm against involving sites or adversary. Proposed algorithm provides certain level of privacy and security under some other security assumption. The communication cost and computation cost are also reasonable for small databases which contain less number of items.

In future, the communication cost and computation cost of proposed algorithm can be decreased for large database. Proposed algorithm can be modified to mine privacy preserving association rules over vertically partitioned data.

REFERENCES


