Privacy Preserving Association Rules in Unsecured Distributed Environment Using Cryptography

Ashish C. Patel[1] , Udai Pratap Rao[2], Dhiren R. Patel[3]
Department of Computer Engineering,
Sardar Vallabhbhai National Institute of Technology,
Surat, Gujarat, India-395007
[1] ashish615@gmail.com
[2] upr@coed.svnit.ac.in
[3] dhiren@coed.svnit.ac.in

Abstract—**In this paper, we propose algorithm to mine association rule using elliptic curve cryptography technique over horizontally partitioned data. Here we consider unsecured distributed environment. Our proposed algorithm provides security against involving parties and intruder and also provides authentication between involving parties. Finally we analyze the privacy and security provided by our proposed algorithm.**

Keywords- Association rule; Datamining; Privacy preseving; Cryptography.

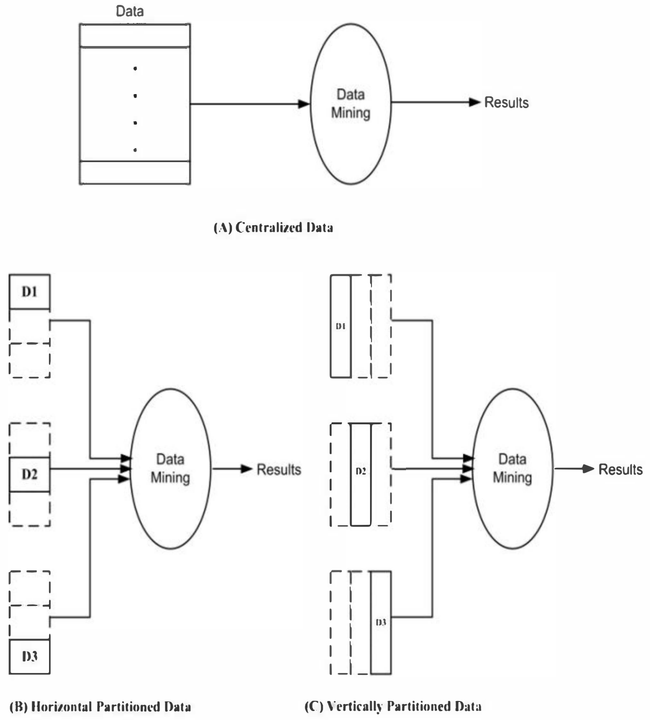
I. INTRODUCTION

An association rule is a rule shows certain relationship among set of attribute in a database. Finding association rule may useful for finding interesting patterns in marketing, statistical analysis, decision making, medical diagnosis etc. Mining association rules requires iterative scanning of database, which is quite costly in processing .These techniques can be demonstrated in centralize [1,2] as well as distributed environment [3,4] where data can be distributed among the different sites. Distributed database scenario can be classified in horizontally partitioned data and vertically partitioned data. In horizontally partitioned data, each site contains same set of attributes with different number of transactions. In vertically partitioned data each site contains different number of attributes with same number of transaction. Figure 1 shows difference between centralize database, horizontally partitioned database and vertically partitioned database mining.

A. Motivation

In commercial applications organizations are very much concerned with privacy issue. Most commercial organization collects their information from individual party as their specific need. They find it essential to share information to each other. In such cases each unit want to be sure that privacy of individual party must not violated.

If government want to survey about some medical disease, individual hospital don't want to reveal personal information of its patient, it is against law. This information is regarded as private and we want to avoid exposing confidential information of patient.



(A) Centralized Data

(B) Horizontal Partitioned Data

(C) Vertically Partitioned Data

Figure 1: Data mining in various environments

In this paper we propose cryptography approach to mine association rules. We have used elliptic curve based digital signature algorithm (ECDSA) which provides authentication between two parties in unsecured environment and elliptic curve Integrated Encryption Scheme (ECIES) which provides privacy against involving parties.

The remaining paper is organized as follows. In Section 2, theoretical background and related work are discussed. In Section 3, technique for finding association rule is proposed. The analysis of proposed algorithm and conclusion are presented respectively in Sections 4 and 5.

II. Theoretical Background and Related Work

A. Data Mining of Association Rule in Distributed Environment

In a distributed database, the transaction are distributed among n different parties $P_1, P_2 \ldots, P_n$ such that each party P_i contains disjoint subset of transactions, $DB = DB_1 \cup DB_2 \ldots DB_n$. The itemset A has a local support count of $A.sup_i$ at party Pi, if $A.sup_i$ of the transaction at P_i contains A. The global support count of A is given as $A.sup = \sum_{i=1}^{n} A.\mathrm{sup}\, i$. An item-set A is globally supported if $A.sup \geq S * |DB| = S * \sum_{i=1}^{n} DBi$. where S is global minimum support. The global confidence of rule $A \rightarrow B$ can be given as $\{A \cup B\}.sup / A.sup$. The aim of mining association rules in distributed databases is to find all rules whose global support and global confidence are higher than the user specified minimum support and confidence. [5]

In term of association rule mining, rule $XY \rightarrow Z$ will be discovered, as long as satisfy these conditions:

1. $\mathrm{Support}(XY \rightarrow Z) = \dfrac{\sum_{i=1}^{sites} Support\_Count\_XYZ(i)}{\sum_{i=1}^{sites} database\_site(i)}$

2. $\mathrm{Support}(XY) = \dfrac{\sum_{i=1}^{sites} Support\_Count\_XY(i)}{\sum_{i=1}^{sites} database\_site(i)}$

3. $\mathrm{Confidence}(XY \rightarrow Z) = \dfrac{Support(XY \rightarrow Z)}{Support(XY)}$

B. Elliptic Curve Cryptography(ECC)

Elliptic curve cryptography (ECC) is an approach to public-key cryptography based on the algebraic structure of the elliptic curve over finite fields [6]. Elliptic curves used in cryptography are defined over two types of finite fields: prime field Fp , where p is a large prime number and binary extension fields $F_2{}^m$.Elliptic curve cryptography require much smaller key compare to the RSA cryptosystem to provide the same level of security. ECDSA public key are smaller than similar strength of DSA keys. These advantages of signature size, bandwidth and computational efficiency make ECC to use in our algorithm. Detail study of ECC given in [7,8].

1) Elliptic Curve Digital Signature Algorithm(ECDSA)

ECDSA algorithm runs in three phases:

- *Key Generation:*
  Party1 select an elliptic curve E over a finite field, say GF(p).The number of points on E should be divisible by large prime n. Then party1 select point $p(x, y) \in GF(p)$ order n. Then selects the integer d in range [1, n-1], where d is a private key of party1. Now to obtain public key of party1 counts Q= dP. Now (E, P, n ,Q) known as public key of party1.
- *Signature Generation:* To sign the message m, first party1 have to select random integer $k \in [1,n-1]$. Then compute $(x_1,y_1)=kP$ then set $r = x_1 \bmod n$, $s = k^{-1}(h(m)+dr) \bmod n$. Signature is included in the message is pair of integers (r,s).
- *Signature Verification:* To verify (r,s) on message m, party2 do the following. First party2 obtain authentic copy of public key of party1 (E, P, n ,Q). Verify the r and $s \in [1,n-1]$. Compute $w = s^{-1} \bmod n$ and h(m). Then $U_1P + U_2Q = (x_0,y_0)$ is calculated where $U_1 = h(m)*w \bmod n$ and $U_2 = rw \bmod n$. finally calculate $v = x_0 \bmod n$ and accept the signature if v=r.

2) Elliptic Curve Integrated Encryption Scheme (ECIES)

ECIES algorithm runs in three phases:

- *Key Generation:* Key generation in ECIES is same as ECDSA
- *Encryption:* To encrypt the message m, first party1 have to select random integer $k \in [1,n-1]$. Then derives shared secret S=Px, where P=(Px,Py)=rd, d is a private key of party1. Party1 use Key Derivation Function (KDF) to derive a symmetric encryption and a MAC keys $k_E||k_M = KDF(S||S_1)$. Now party1 encrypt message $c = E(k_E;m)$. For sending a message party1 prepare a tag of encrypted message and $d = MAC(k_m;c||S_2)$. S_1 and S_2 are optional. Party1 send message R||c||d to party2.
- *Decryption:* To decrypt cipher text R||c||d party2 do the following. Party2 derives shared secret $S = P_x$. where $(P_x ,P_y) = k_BR$. It also derive key the same way as in encryption $k_E||k_M = KDF(S||S_1)$ and output is failed if $d \neq MAC(k_m;c||S_2)$. To decrypt message party2 use symmetric encryption and the message $m = E^{-1}(k_E;c)$.

C. Related Work

There are many approaches proposed for mining association rules in centralized database as well as in distributed database. Many of them uses cryptography techniques to preserve privacy in such databases.

Authors in [9] develop FDM(Fast Distributed Mining of association rules) efficient algorithm to find association rule in horizontal partitioned data. This algorithm is adoption of apriori algorithm in distributed database. Algorithm generates candidate set at each site. After generation of candidate set local pruning and global pruning take place. To determine whether candidate set is globally supported or not, they exchange support count using polling technique. Although algorithm does not provides privacy.

In [10] authors directly adopt FDM technique. To exchange support, they use cumulative encryption technique. All parties encrypt all itemsets and sends to common party to eliminate duplication. After that it is passed to all parties and each party decrypt given itemset. After decryption itemset is tested if it is globally supported.

In [11] authors introduce secure set union and shows good results then FDM. Here each site encrypt local data and send to other site. Each site re-encrypts that data. They shows duplicate in original itemset is duplicate in encrypted itemset. After encryption is done decryption can be done in any order. An experimental result shows, time spends on communication is reduces then FDM.

Authors in [12] proposed algorithm based on elliptic curve cryptography ECDH and ECDSA and use thirds party to count the global support of the itemset. It counts the global support under the security assumption. What if third party generates the artificial association rules? Therefore we propose an algorithm without use of third party and still it provides privacy and authentication between involving parties.

III. Proposed Technique

A. Problem Definatation

We consider scenario where more than two parties want to corporate for computing association rule on union on their databases. Although databases are private to party and they don't want to reveal their private information to other parties or to the intruder.

We show how involved parties find global association rule efficiently without revealing personal information to other involving parties or intruder. No party learning anything then global association rule.

We use apriori algorithm to generator frequent itemset and for authentication and security we used ECDSA and ECIES under elliptic curve cryptography.

B. Generating Gobal Frequent Itemset

Proposed protocol for finding global frequent itemset is shown in Figure 2. Suppose there are three parties are involved namely party1, party2 and party3 having their homogeneous dataset DB1, DB2, DB3. Involving parties want to count the support of the itemset X without revealing any information to the other involving parties. Here in given protocol we generate global frequent itemset.

1) Algorithm generates elliptic curve based public keys and private keys, and distribute them to the other involving parties.
2) Any involving party can be initiator. Initiator generates random number R, Here we assume party1 as initiator. Party1 initiate the program and sends total count of transactions by adding random number and send to party2 as shown in figure 2. In the end of execution of program party1 gets the sum of total transactions. Party1 remove random number and gets total number of transaction Total_Count.
3) Party1 generates 1-frequent itemset using apriori algorithm.
4) Here we consider party1 as initiator. It calculate the following msg$_1$ = [X$_i$.count + R] as shown in figure. X$_i$.count is the count of item i at party1.In figure 2 A$_i$ and B$_i$ are count of item A at party i and count of item B at party i respectively.
5) Party1 encrypt msg1 using ECIES algorithm, Encrypted message Em= Encrypt(msg1, pubKey$_2$). Now party1 sign Em using ECDSA algorithm, Signed message Sm = Sign (Em, priKey$_1$) and sends to the party2. Here msg1 contains count of all the itemsets.
6) Party2 verify signed message using ECDSA algorithm, Vm=Verify(Sm, pubKey$_1$). After that if message is verified then party2 decrypt the message using ECIES algorithm, msg1= Decrypt(Vm, priKey$_2$).
7) Pary2 add its itemset count to msg$_2$ = [X$_i$.count + msg$_{1i}$] . X$_i$.count is count of item i at party2 and msg$_{1i}$ is count of item i send by Party1.
8) Party2 do as directed in step 5 and send massage to party3.
9) Pary3 verify and decrypt as shown in step 6 and add itemset as shown in step 7. Party3 send the msg$_3$ to the initiator party1.
10) Party1 decrypt and verify message and remove random number. Party1 checks if the Count_of_item_i ≥ S*Total_Count, then add item in 1-frequent item set, which is globally generated itemset. Here S is the minimum support.

Party1 do the step from 1 to 10 until it finds out the globally generated k-itemset.

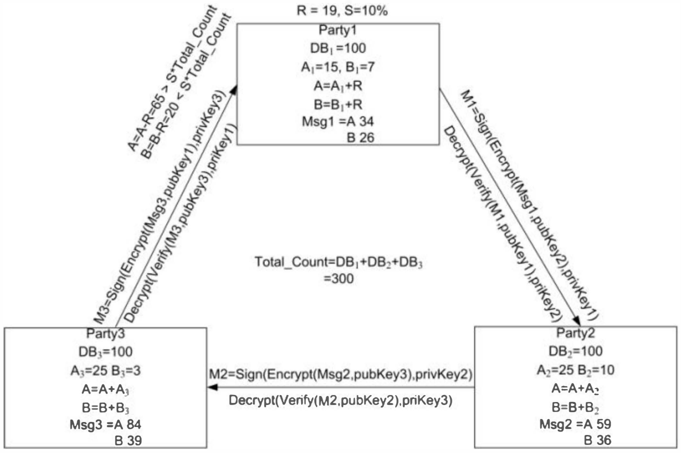


Figure 2 : Determining Global itemsets.

C. Generating Global Association Rule

After generating k-frequent itemset, global association rule can be generated using genRules(F) function of apriori algorithm [13] for centralize database, satisfying minimum support threshold and minimum confidence threshold . Initiator finds all global association rules and broadcast it globally. No other party can learn more than global association rules.

D. *Algorithm for Generating Global Association Rule*

In Table 1 proposed algorithm is given. Our proposed algorithm works on two phases. First, it generates global frequent itemset and second is to generate global association from global frequent itemset.

Table 1: Proposed Algorithm

Input: Databases of involving parties $DB_1, DB_2,\ldots, DB_n$.
    DB= $\sum_{i=1}^{n} DBi$
    Minimum global support (MGS)
    Minimum global confidence (MGC)
Output: Global association rules

(a) Initiator party generates global k-frequent itemset using the step 1-10 as shown in above protocol.
(b) Initiator calculate global association rule like a → b where global confidence gc $\geq$MGC and a, b should be globally frequent itemsets.

IV. Analysis of Proposed Algorithm

A. *Privacy Aginst Involving Parties*

In proposed algorithm initiator party adds random number and encrypts the total count and send to another party. Another party cannot guess the value easily and cannot predict original value and this way it provides privacy against involving parties. Parties cannot read communication channel because massage is passing in encrypted form. Thus using this protocol we provide privacy against involving parties.

B. *Privacy and Security Against Intruder*

If intruder wants to read communication channel or want to alter message, he cannot do that because communication channel is secure. Each party sends massage after digitally signed so intruder cannot alter this massage.

C. *Communication Cost*

If n sites are participating and k numbers of 1- frequent itemsets are found then 2^k-1 itemset are generated. The communication cost for sending message to n site is $n*(2^k - 1)$ which is reduces to 2^k.

D. *Computation Cost*

Computational cost of encryption and decryption cost is constant and independent of parameters itemset k and party count n. Suppose a value of encryption and decryption algorithm is C1 and value of signature algorithm for n site is C2. Cost of generating non empty subset to each site is m^2. Total cost of signature generating and encryption decryption is L*(C1+C2), where L is the level of frequent itemset. So total cost of generating frequent itemset is $O(n*(L*(C1+C2) + m^2))$, which is equal to $O(n* m^2)$. Cost of generating association rule from frequent item is $O(2^{2m}+m^2)$. So, total cost of algorithm is 2^m. This is reasonable for small databases.

E. *Exprimental Evalution*

The proposed algorithm is evaluated on dataset [14]. In our experiment we simulate 3 different parties.

The 100k records of dataset named T10I4D100K [14] are used in our experiment. Database distributed equally in each party. The execution time for different support is shown in figure 3. Result is good for small number of database.

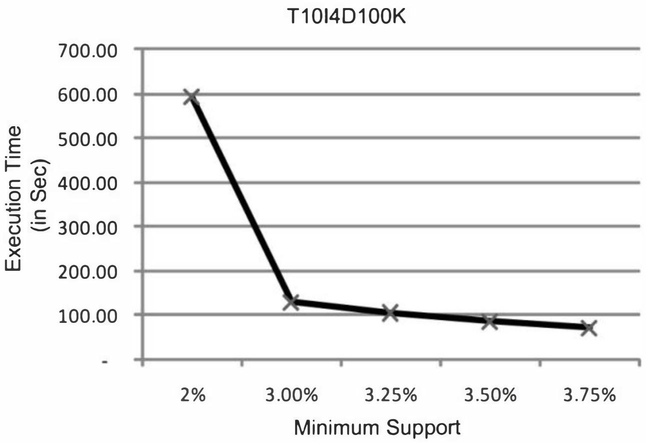


Figure 3 : Execution time v/s Minimum Support graph of T10I4D100K database

V. Conclusion

In this paper we proposed a technique for mining association rules on horizontally partitioned data in unsecured environment using ECDSA and ECIES algorithms. We also show that the global association rule can be generated without loss of privacy to the involving parties or intruder. The communication cost and computational cost also be reasonable for small amount of databases. In future proposed algorithm can be extended to vertically partitioned database and also computational cost can be reduced.

References

[1] Elena Dasseni, Vassilios S. Verykios, Ahmed K. Elmagarmid, and Elisa Bertino, “Hiding Association Rules by using Confidence and Support,” In Proceedings of the 4th Information Hiding Workshop,pp 369–383, 2001.

[2] Vassilios S. Verykios, Ahmed K. Elmagarmid, Bertino Elisa, Yucel Saygin, and Dasseni Elena, “Association Rule Hiding,” IEEE Transactions on Knowledge and Data Engineering ,2003.

[3] Vaidya, J. & Clifton, C.W, “Privacy preserving association rule mining in vertically partitioned data,” In Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, Edmonton, Canada 2002.

[4] Murat Kantarcioglou and Chris Clifton, “Privacy-preserving distributed mining of association rules on horizontally partitioned data,” In Proceedings of the ACM SIGMOD Workshop on Research Isuues in Data Mining and Knowledge Discovery, pp 24–31, 2002.

[5] Liu, J., Piao, X., & Huang, S, “A privacy-preserving mining algorithm of association rules in distributed databases,” Proceedings of the 1st

International Multi-Symposium on Computer and Computational Sciences (IMSCCS),pp 740-750, 2006.

[6] N. Koblitz, A. Menezes, and S. A. Vanstone, “The state of elliptic curve cryptography, Designs, Codes and Cryptography,” pp. 173–193,2000.

[7] N. Gura, A. Patel, A. Wander, “Comparing elliptic curve cryptography and RSA on 8-bit CPUs,” In Proc. Cryptographic Hardware and Embedded Systems: 6th International Workshop, Cambridge, MA,2004.

[8] N.P. Smart, “The exact security of ECIES in the generic group model,” In B. Honary, editor, Coding and Cryptography, Springer-Verlag LNCS 2260,pp 73–84,2001.

[9] D.W.-L. Cheung, J. Han, V. Ng, A.W.-C. Fu, and Y. Fu, “A Fast Distributed Algorithm for Mining Association Rules,” Proc. 1996 Int’l Conf. Parallel and Distributed Information Systems (PDIS ’96),pp. 31-42, 1996.

[10] Hua-jin Wang, Chun-an Hu, Jian-sheng Liu, “Distributed Mining of Association rules Based on Privacy-preserved method,” Information Science and Engineering (ISISE), International Symposium, pp 494 – 497, 2010.

[11] Weiwei Jing, Liusheng Huang, Yonglong Luo, Weijiang Xu, and Yifei Yao, “An algorithm for privacy-preserving quantitative association rules mining,” In DASC Proceedings of the 2nd IEEE International Symposium on Dependable, Autonomic and Secure Computing, pp 315–324, 2006.

[12] Modi Chirag, Rao Udai Pratap, Patel Dhiren R, “Elliptic Curve Cryptography Based Mining of Privacy Preserving Association Rules in Unsecured Distributed Environment,” In proceedings of the International Conference on Advances in Communication, Network, and Computing,pp 94-98,2010.

[13] J. Han and M. kamber, “Data Mining:Concepts and Techniques,”San Fransisco ,2005.

[14] IBM Almaden Quest research group. Frequent Itemset Mining Dataset Repository [Online]. Available : http://fimi.ua.ac.be/data/T10I4D100K.dat