ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

Privacy Preserving Two Party Distributed Association Rule Mining by FP Growth on Horizontally Partitioned Data

Patil Suraj K, Gadage Shrinivas

Department of Computer Engineering GHRCEM, Wagholi, Pune, India

Department of Computer Engineering GHRCEM, Wagholi, Pune, India

ABSTRACT: Data mining has to be complete in distributed data situation in many applications. Data owners may be worried with the mistreatment of data in such situations. The privacy is not protected or preserve. Due to this reason they do not want their data to be mined particularly when these contain responsive information. Protect the data privacy in the route of data mining is the goal of the PPDM that is Privacy-Preserving Data Mining. Privacy-preserving distributed association rules mining protocols have been developed for horizontally partitioned data situation with more than two participating parties. However they depend on an amalgamation calculation and secure multi-party précis. When number of participating parties is two then it cannot assurance security. To apply the protocols for the privacy-preserving two parties distributed mining of association rule mining. We also analyzed the protocols security and performance. Also we design a protocol to reduce communication cost and times as well as we also design a multiparty data mining protocol with malicious model.

KEYWORDS: Commutative encryption technique, Partitioning technique, Data mining techniques, Association rule mining, Semi honest model, Privacy preserving.

I. INTRODUCTION

For data owners, the data misuse is the main concern in various applications. Therefore, they oppose to the mining of their sensitive information. Therefore they provide their data for any such data mining related activities. Still somehow, Data mining might provide more approaching from data. So it will bring the huge social and cost efficient benefits. PPDM makes transactions between the data privacy and the data mining contributions. Carrying out the mining process effectively and efficiently is the. So, it does not to tamper with sensitive data. Initially, The PPDM was researched by the two different papers [16], [17], [16] Concentrated on the PPDM tasks in centralized data storage scenario. In the individual records values that had been disconcerted, it used a decision tree classifier. To use a new reconstruction procedure to estimate the distribution, and use this distribution to develop a classifier with similar accuracy was the fundamental idea. The data mining technique like:

1) Decision Support System (DSS).
2) Artificial Intelligence (AI).
3) Machine Learning.
4) Business Intelligence (BI)

The data mining involves:

1) Anomaly detection: Detection of changing patterns and detection of unusual data while data is mining.
2) Association rule mining: It finds the relationship between different elements in the database by sing patterns the rules are generated and make decisions to find attributes is done.
3) Clustering: Based on features of data groups are generated.
4) Classification: The data set is trained using classifiers and a test data is given to the rules generated to classify unlabeled data.
5) Regression: In regression it finds function which generates data with minimum amount of error.
6) Summarization: It gives the minimum summary of the data set used in the database.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

Two different research directions were represented by these two studies in PPDM. The first one used the Randomized Perturbation Technique (RPT), while, the Cryptography-based technique was used by the other one. The first method was applied to the centralized storage of data; the distributed data storage scenario the latter one is used. Some of the main phases of Privacy-Preserving two-party distributed association rules mining on horizontally partitioned data are Association Rules Mining, Secure Distributed Association Rules Mining and Distributed Association Rules Mining; and the security infrastructures. These are needed to be used to get the best result from the distributed data with maintaining the privacy. The Association Rules Mining is briefly explained in [15]. The fundamental difficulty with the mining association rule is that, one should always keep in mind all rules, where the confidence is always greater than the minimum confidence threshold.

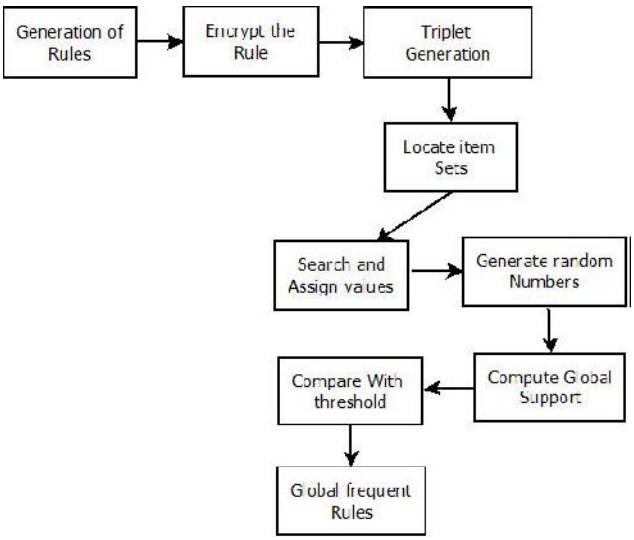

Figure.1 Overview of privacy preserving in data mining

The structure of the paper is as follows:
Section II describes the literature survey, Section III gives description of the proposed system, Section IV gives the mathematical model, Section V the experiments are performed and results are predicted, lastly Section VI gives conclusion.

II. RELATED WORK

Now days, PPDM is research direction developed in the field of data mining. A valuable results are achieved in many field of data mining, like, association rules [8], decision tree[17], clustering[9], and outlier detection[10] etc. Moreover, k anonymity is another PPDM technique [13]. This is a part of the area of privacy preservation data publishing techniques. To avoid indirect identification of data from the public databases is the main work of this method. It is so; due to , the combination of record attributes might be helped to correctly find out every single record. However, k-anonymity is declared to be defenseless. Some privacy problems will be faced by K-anonymity, only in the case if distinct attributes include small diversity or the attacker sound about the victim [1]. In recent few years, differential privacy [2], [3], [6] has diverted to considerable number of researchers for PPDM. Differential privacy model is used for giving privacy to statistical queries and pattern mining. It also gives means to increase the sharpness of queries or data mining, and reduces the chances of identification of records. The differential privacy are implemented by the exponential mechanism [6] and the Laplace mechanism [3]. under the situations of discrete outputs and numeric outputs . All final globally frequent rules are unveiled in [11] to all participant sites. Knowing a rule is not maintain at one as site but is maintained globally in the two party scenarios. This discloses that the rules are being supported by the other site. To skip this leakage is not possible. [1] Follows the general approach of the FDM algorithm [4]. Here some special protocol changes the broadcast of LLi(k) and the LLk item as support count. The rules of the starting are not revealed, it

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

gives a technique for finding the union of locally supported rules. It also gives a method for steady testing the support is higher than threshold. [11] Follows the two-phase approach in description. However, uniting the locally generated rules and support count is made by somehow giving the encrypted values between the sites. The two phases decides the candidate rules. And also decide that these meet the thresholds of global support confidence. The first phase makes use of commutative encryption. Every party encrypts own frequent rules. These encrypted rules are then passed to other parties, unless encryption of the rules is perform by each sharing party. The rules which are locally supported are tested globally in the another phase. The important fact in this method are association rule mining, distributed association rule mining, secure association rule mining, secure infrastructure and commutative encryption system.

A. Association Rule Mining:
Describe the association rule mining in short. The mining association rule told that , one has to maintain all rules in mind, because in this, the confidence is always greater than the minimum confidence threshold, and the support will always exceed the minimum support threshold.

B. Distributed Association Rule Mining:
The proposed paper deals with the Distributed Association Rule mining in the distributed scenario. [5] Gives the Fast Algorithm is a mostly used algorithm for distributed association mining.

C. Secure Association Rule Mining:
With help of Secure Distributed Association Rules mining, we ensure that the exposure will be restricted. The contents of the transactions took place at one part, will not be available for another part, unless the same data is checked by respective sides for others. [11] Explained it in short. But somehow, for the two side scenario, the data of support and database size is not exposure to other sides. As well as, the side does not reveal whether or not a locally rule is globally maintained.

D. Commutative Encryption System:
[14]This has been provide the commutative encryption system is important for the implementation of the security in multiparty scenario. The side 1 verify that the given cipher text is compositely commutative encryption is alike, notwithstanding of the order of encryption. The side 2 verify that the two similar encrypted messages will never be created by the two dissimilar plain messages. The secure and safe encryption is guaranteed by the site 3. 5. Secure Infrastructure: The next part is the protection in the semi-honest models. The method of semihonest is depend on the hypothesis saying that the participated parties are semi-honest. In simple words, they will follow the rules of the protocol by using the appropriate inputs, but are free to use whatever they see during the execution of the protocol. The composition theorem, used for secure multiparty computation has thoroughly described in the [12].

III. PROPOSED SYSTEM

This paper proposed the distributed mining of association rules on horizontally partitioned data in a two-party case. The protocol can tell the site whether or not its itemsets are globally frequent under the union transactions of other sites and itself while guaranteeing the related data privacy. We believe that the two-party protocol is significant as such application scenarios exist in reality, where one party wants to know whether or not its itemsets (rules) are globally frequent but does not want to reveal its supports and other private data to the other party. The proposed system consists of follow two steps:

Algorithm 1 Privacy preserving Protocol
Input: Two sites say, Site1 and Site2 contain rules generated using association rule mining algorithm.
Output: Triplets containing
1: For site i
2: Generate rules: FP growth();
3: Encrypt the rules Fei(rules of site i) and send to site (i%2)+1
4: End for
5: For each site i
6: Receive rules from site (i%2)+1

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

7: Encrypt F*ei*(rules of site (i%2)+1) and send to site (i%2)+1
8: End for
9: For each site i
10: Create triplets: {rules, F*ei*F*ei*%2+1 (rules of site i),support}
11: End for

Algorithm 2 Global rules generation
1: For site 1
2: Consider rules, F_2F_{e1} (rules of site1),support
3: Send $F2F_{e1}$ (rules of site 1) to site2
4: End
5: For site2
6: Find F_1F_{e2} (rules of site 1) exact similar to F_2F_{e1} (rules of site 1)
7: If found then
8: Set support2=rule. Support ()
9: Else
10: Support2=0
11: End
12: On site1
13: Generate two non-zero random numbers $rand^1_1$, $rand^2_1$

$$rand1=\frac{rand_1^2}{rand_1^1}$$

14: End
15: On site2
16: Generate two non-zero random number $rand^1_2$; $rand^2_2$

$$rand1=\frac{rand_2^2}{rand_2^1}$$

17: End
18: On site1
19: Compute Z'

$$Z' = \frac{support1+support2}{D1+D2}$$

20: If Z'>=minimum support
21: Global rule. Add(rules)
22: Else
23: Discard that rule
24: End
25: Repeat steps 1 to 5 for site 2
26: Output: From site 1 and site 2 the global rules are generated.

FP Growth:
FP growth uses divide and conquer rule as: It shrinks the data set having frequent items an divide the shrink data base in particular condition with each frequent database and starts mining the database.

FP growth works as follows:
1) Scan the database and compute the count of the items in that certain database. Set a support threshold and choose the items above the threshold assigned. Sort the items selected in descending order.
2) Initialize the FP tree and make a node to which all the frequent items as nodes will be joined. The database is scanned again on new items selected and the item from selected and sorted items is connected to the node.
3) Starting from the least freq item, a freq finder is recursively called. The support of pattern is found and is shown if freq.
The FP growth algorithm is given below:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

Algorithm 3 FP Growth Algorithm
Input: constructed FP-tree
Output: complete set of frequent patterns
FP-growth (FP-tree, a)
1: If tree has one and only path
2: Consider each combination
3: Generate pattern a ∪ b with support as minimum support
in b.
4: Else
5: For every header do
6: Generate pattern b= ai ∪ a considering support as ai.support.
7: Construct b patterns
8: If tree b= null.
9: FP-growth (Tree b,b);

TABLE I. TABLE1 WILL GIVE THE COMPARISON BETWEEN APRIORI ALGORITHM AND FP GROWTH ALGORITHM

Parameters	Apriori Algorithm	FP growth Algorithm
Process	The entire process is dependent upon the pruning & joining property	The process generate FP tree and conditional based tree using the data set satisfying the min support assigned
Memory	Large number of candidate sets are generated and hence utilizes huge amount of memory.	No candidate set generation and hence less Memory usage.
Scan	Scans until all candidate set are generated.	Scans the dataset twice
Time	Takes more time as the candidate set are generated.	Requires less time for computation than Apriori algorithm

A. Mathematical Formulae
Let S, be a system such that,
S = {I, e, X,Y, T, fme, DD, NDD, friend, MEM shared, CPUCoreCnt, Φ}
Where,

- S: Proposed System
- I: Initial state at T<init> i.e. providing the transaction database to the system.
- e: End state of obtaining global frequent item set.
- X: Input of System i.e. transaction database

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

- Y: Output of System i.e. global frequent item set
- T: Set of serialized steps to be performed in pipelined machine cycle. In a given system serialized steps are input, generate rules, encrypt data, generate triplets, global frequent item sets, etc.
- Fme: Main algorithm resulting into outcome Y, mainly focus on success defined for the solution. In given system, FP-Growth algorithm and Security algorithm.
- DD: Deterministic Data, it helps identifying the loadstore function or assignment function. e.g. i= return i. Such function contributes in space complexity. In a given system deterministic data will be Triplet generated used for the global frequent item sets.
- NDD: Non Deterministic Data of the system to be solved. These being computing function or CPU time or ALU time function contribute in time complexity. In a given system we need to find global frequent item set generated by the two parties. Friend: Set of encrypted rules.
- MEM shared: Memory required to process all these operations, memory will allocated to every running process.
- CPUCoreCnt: More the number of counts double the speed and performance.
- Φ: Null value if any.

B. Deterministic finite automata:

1) Commutative Encryption: Figure3. Shows the DFA to encrypt the rules. DFA consists of five tuples Q, ∑,σ, q0, F

- Q: No. of states. fS1, S2, S3, S4g.
- S1: Association rule algorithm.
- S2: Commutative encryption algorithm on Site1.
- S3: Site 2.
- S4: Commutative encryption algorithm on Site2.
- ∑: Input Data Base.
- σ: Transition function.
- q0: Initial State. Here, S1.
- F: Final State. Here, S1.

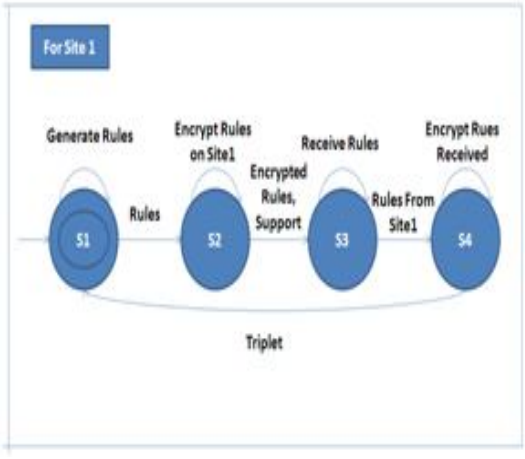

Figure.3 Deterministic finite automata for global rules extraction.

2) Global rules: Figure4. Shows the DFA for generating rules where the input is Triplets.

- Q: No. of states. S1, S2, S3, S4.
- S1: Search and assign.
- S2: Random generator.
- S3: Calculate global support.
- S4: Commutative encryption algorithm on
- Site2.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

- Σ: Input-Triplets.
- σ: Transition function.
- q0: Initial State. Here, S1.
- F: Final State. Here, S4.

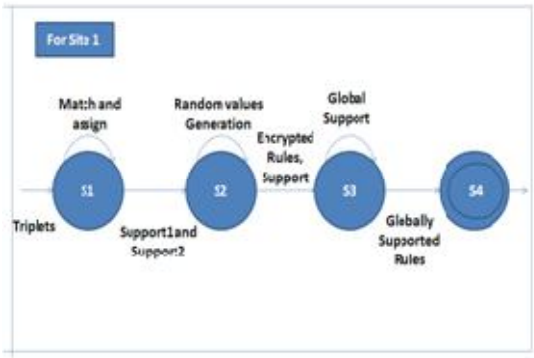

Figure.3 Deterministic finite automata for global rules extraction.

IV. EXPERIMENTS AND RESULTS

The results and the graph have been generated while input given to the system proposed. Table.2 shows the time required to make rules for various instances in the data set.

TABLE II. TIME REQUIRED TO RULES FOR DIFFERENT INSTANCES IN THE DATA SET (SUPPORT=0.0 and CONFIDENCE=0.0).

Number Of Instances	Apriori (time in sec)	FP growth(time in sec)
100	55.32	0.359
200	85.42	0.25
300	168.73	0.359
400	361.60	0.406
500	521.82	0.437
700	734.73	0.437
1000	1081.175	0.842

Table.3 denotes the time required to generate rules for support=0.0 and confidence=0.0. The graphical representation is shown in Figure.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

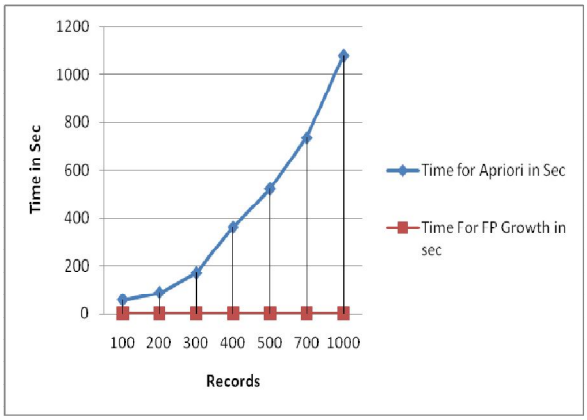

Figure.4 Time required to rules for different instances in the data set for support=0.0 and confidence=0.0.

TABLE III. TIME REQUIRED TO RULES FOR DIFFERENT SUPPORT AND 1000 RECORDS

Support	Apriori (time in sec)	FP growth(time in sec)
0.0	1081.175	0.842
0.3	1.747	0.218
0.5	1.388	0.203

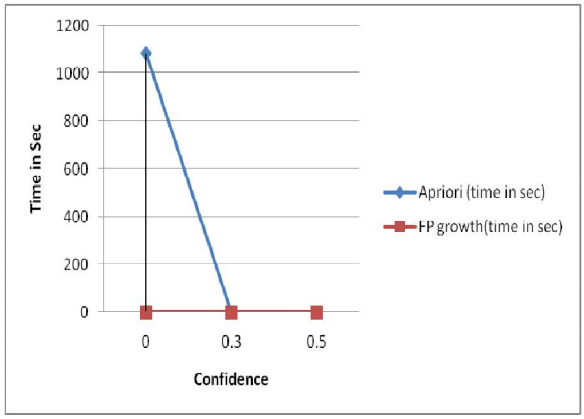

Figure.5 Time required to rules for different Support

VI. CONCLUSION

In this paper, the proposed system has successfully designed a privacy preserving protocol as well as the better association rule mining technique. The experiments are carried out on the crime database and results are predicted on basis of two popular rule mining algorithm viz. Apriori and FP growth I which the FP growth seems to be better than Apriori. The confidential data in the database is protected by the system proposed with more efficient rules. This system can be further made more secure by implementing a new privacy preserving protocol.

REFERENCES

[1] A. Machanavajjhala, D. Kifer, and J. Gehrke, "diversity: Privacy beyond k anonymity", ACMTransactions on Knowledge Discovery from Data (TKDD), 2007.
[2] C. Dwork, "Differential privacy", In: M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, ICALP, Lecture Notes in Computer Science,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

2006.
[3] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis", In: S. Halevi and T. Rabin, editors, TCC, Lecture Notes in Computer Science, 2006.
[4] D. Boneh, "The decision diffie Hellman problem", In: Proceedings of the 3rd Algorithmic Number Theory Symposium, 1998.
[5] D. Cheung, J. Han, V. Ng, "A fast distributed algorithm for mining association rules", In: Proceedings of 1996 Int. Conf. of Parallel and Distributed Information Systems, 1996.
[6] F. McSherry and K. Talwar, "Mechanism design via differential privacy", In: Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS), 2007.
[7] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", 2nd edition, Morgan Kaufmann, 2006.
[8] J. Lin, Y. Cheng, "Privacy preserving itemset mining through noisy items", Expert Systems with Applications. 2009.
[9] J. Sakuma, S. Kobayashi, "Large scale k-means clustering with user centric privacy preservation", Knowledge and Information Systems, 2010.
[10] J. Vaidya, C. Clifton, M. Zhu, "Privacy Preserving Data Mining (Advances in Information Security)", New York: Springer Verlag, 2005.
[11] M. Kantarcioglu, and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned Data", IEEE Transaction on Knowledge and Data Engineering, 2004.
[12] O. Goldreich, "Foundations of Cryptography (Volumn 2)", Cambridge University Press. 2004.
[13] P. Samarati, "Protecting respondents identities in micro data Release", IEEE Transaction on Knowledge Data Engineering, 2001.
[14] R. Agrawal, A. Evfimievski, and R. Srikant, "Information sharing across private databases", In: Proceedings of 2003 ACM SIGMOD Int. Conf. on Management of Data, 2003.
[15] R. Agrawal, R. Srikant, "Fast algorithms for mining association rules", In: Proceeding of the 20th Int. Conf. on Very Large Data Bases, 1994.
[16] R. Agrawal, R. Srikant, "Privacy preserving data mining", In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 2000.
[17] Y. Lindell, and B. Pinkas, "Privacy preserving data mining", In: Proceedings of the 20th Annual Int. Cryptology Conf., LNCS 1880. 2000