Privacy-Preserving Outsourced Association Rule Mining on Vertically and Horizontally Partitioned Databases Using Twofold Encryption

[#1]Abhishek Kurkut, [#2]Sarang Bora

[1]kurkutabhishek28@gmail.com
[2]sarangbora@gmail.com

[#12]Department of Computer Engineering, SKNCOE
Pune ,Maharashtra, India

ABSTRACT

Data mining can extract important knowledge from large data collections, but sometimes these collections are split among various parties. Privacy concerns may prevent the parties from directly sharing the data and some types of information about the data. This research addresses secure mining of association rules over partitioned data. The methods incorporate cryptographic techniques to minimize the information shared, while adding little overhead to the mining task. This research tries to proposed desirable algorithm for both vertically and horizontally partitioned data. We also propose a technique for solving a main problem of privacy preserving association rule mining in two party databases. We are performing Horizontal Partitioning as well as Vertical Partitioning of Data, Also Double Encryption is used to increase the security of dataset

Keywords— Association Rule Mining, Frequent Itemset Mining, Outsource, FP-Growth, Ecc..

ARTICLE INFO

Article History

Received: 29th May 2017

Received in revised form :

29th May 2017

Accepted: 31st May 2017

Published online :

3rd June 2017

I. INTRODUCTION

The association rule mining task inside of an organization privacy preserving system is outsourcing. An amazing body at work seems to have been done on privacy preserving information mining in a scope of connections. The end indication of a good some previously considered systems is that the patterns mined from the input (which could possibly be distorted, scrambled, anonymized, or you can't change) are proposed as shared with parties except for the criticism proprietor. The critical thing refinement between such groups of work and our issue is that, in the last, both the hidden information and along these lines the mined answers are not only to share and must remain private in the feedback owner we propose Mining of Association Rules from Outsourced Transaction Databases using Two fold encryption. We use FPGrowth algorithm, to generate association rules. The details of our improve schemes are presented, it includes following algorithms: Homomorphic Encryption and Encrypted Curve Cryptography, FP-Growth Algorithm.

The Main objective of this project is to provide computational complexity, communication complexity and less storage cost of our association rule mining and frequent item set mining solutions and with the help of FP-growth algorithm we are trying to achieve it. The Comparison is shown below in table.

With the help of Twofold Encryption we are Processing frequent itemset mining and association rule mining for high privacy requirements. Compared with most solutions, our solutions achieve a higher privacy level.

The Input for Project is real time dataset such as retail dataset from UCI Machine Learning Repository, the dataset includes transactions and while the single transaction contain Item set, in which each item is comma or space separated. The output of project is Association Rule based on the Mining query sent by the user. The mining query is threshold value which varies from 0 to 1.

Frequent item set mining and association rule mining have been employed in applications such as market basket analysis, health care, web usage mining, bioinformatics and Prediction. So the owners of Chains of retail shop or uni retail shop, Chief Website Developer of an E-commerce business etcetera could make use of this very project.

II. LITERATURE SURVEY

1) D. H. Tran, W. K. Ng and W. Zha, ”CRYPPAR: An efficient framework for privacy preserving association rule mining over vertically partitioned data,” TENCON 2009 - 2009 IEEE Region 10 Conference, Singapore, 2009, pp. 1-6. In paper [1] authors have proposed CRYPPAR, a full-fledged framework for privacy preserving association rule mining based on cryptographic approach over vertically partitioned data. We also conducted empirical evaluation on

CRYPPAR. The results indicated that the method of building it is efficient and may become a general way to do PPDM in real life.

2) L. Li, R. Lu, K. K. R. Choo, A. Datta and J. Shao,"Privacy-Preserving-Outsourced Association Rule Mining on Vertically Partitioned Databases," in IEEE Transactions on Information Forensics and Security, vol. 11,no. 8, pp. 1847-1861, Aug. 2016. In paper [2] authors proposed a privacy-preserving outsourced frequent item set mining solution for vertically partitioned databases.This allows the data owners to outsource mining task on their joint data in a privacy preserving manner. Based on this solution, authors built a privacy-preserving outsourced association rule mining solution for vertically partitioned databases.

3) B. Dong, R. Liu and W. H. Wang, "Integrity Verification of Outsourced Frequent Itemset Mining with Deterministic Guarantee," 2013 IEEE 13th International Conference on Data Mining, Dallas, TX, 2013,pp. 10251030. In paper [3] authors we presented an efficient result integrity verification approach that can provide deterministic guarantee for outsourced frequent item set mining. The key idea of the approach is to construct cryptographic proofs of all (in) frequent item sets. They discussed how to optimize the number of proofs to improve the performance.

4) M. N. Kumbhar and R. Kharat, "Privacy preserving mining of Association Rules on horizontally and vertically partitioned data: A review paper," Hybrid Intelligent Systems (HIS), 2012 12th International Conference on, Pune, 2012, pp. 231-235. 5. D. Trinca and S.Rajasekaran, "Towards In paper [4] authors tackled issues of privacy preserving association rule mining are addressed here. In particular, privacypreserving algorithms over horizontal and vertical partitioned databases are discussed and results are compared.

5) D. Trinca and S. Rajasekaran, "Towards a Collusion Resistant Algebraic Multi-Party Protocol for Privacy Preserving Association Rule Mining in Vertically Partitioned Data," 2007 IEEE International Performance,Computing, and Communications Conference, New Orleans, LA, 2007, pp. 402-409. In this paper [5], authors concentrate on the situation when the database is distributed vertically, and propose an effective multi-party protocol for evaluating item-sets that preserves the privacy of the individual parties. The proposed protocol is algebraic and recursive in nature, and depends on a recently proposed two-party protocol for the same issue. It is not only appeared to be much faster than similar protocols, additionally more secure. They likewise show a variation of the protocol that is resistant to collusion among parties.

6) Vaidya, Jaideep, and Chris Clifton. "Privacy preserving association rule mining in vertically partitioned data." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002. Privacy considerations [6] often constrain data mining projects. This paper addresses the issue of association rule mining where transactions are distributed across sources. Every site holds some attributes of each transaction, and the sites wish to collaborate to identify globally valid association rules. Notwithstanding, the destinations must not reveal individual transaction information. Authors show a two-party algorithm for efficiently discovering frequent itemises with minimum support levels, without either site revealing individual transaction values.

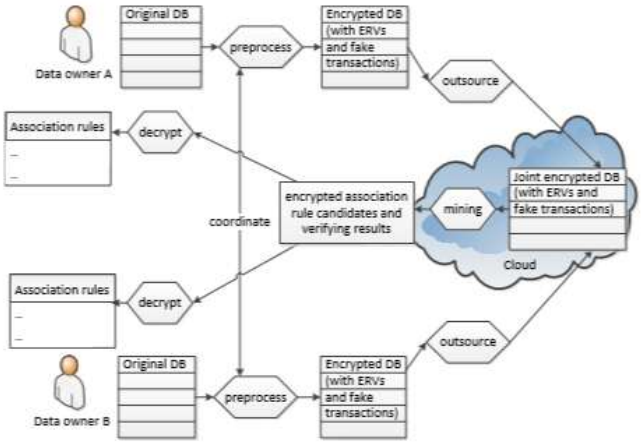
7) Tassa, Tamir. "Secure mining of association rules in horizontally distributed databases." IEEE Transactions on Knowledge and Data Engineering 26.4 (2014):970-983. This paper [7] proposes a protocol for secure mining of association rules in horizontally distributed databases. The present leading protocol is that of Kantarcioglu and Clifton. Their protocol, like theirs, depends on the Fast Distributed Mining (FDM) algorithm of Cheung et al., which is an unsecured distributed variant of the Apriori algorithm. The fundamental ingredients in their protocol are two novel secure multi-party algorithms one that computes the union of private subsets that each of the associating players hold, and another that tests the inclusion of an element held by one player in a subset held by another.

8) G. I. Davida, D. L. Wells, and J. B. Kam. A database encryption system with sub keys. ACM TODS,6(2):312328, 1981.

9) J. He and M. Wang. Cryptography and relational database management systems. In IDEAS, 2001.

10) B. Iyer, S. Mehrotra, E. Mykletun, G. Tsudik, and Y. Wu. A framework for efficient storage security in RDBMS. In EDBT, 2004 There are two approaches that can protect sensitive information. Is to an encryption function that transforms the original data to a completely new format [10] [8]. The second could be to apply data perturbation, which modifies the original raw data randomly [9]. The perturbation approach is less attractive since it could only provide approximate results; nevertheless, the employment of encryption allows the same rules that they are recovered.

III.SYSTEM MODEL:



IV. PROPOSED ASSOCIATION RULE ALGORITHM

FP-growth Algorithm : The FP-Growth Algorithm is a way to find frequent itemsets without using candidate generations, thus improving performance. For so much it uses a divide-and-conquer strategy. The core of this method is the usage of a special data structure named frequent-

pattern tree (FP-tree), which retains the itemset association information.

In simple words, this algorithm works as follows: first it compresses the input database creating an FP-tree instance to represent frequent items. After this first step it divides the compressed database into a set of conditional databases, each one associated with one frequent pattern. Finally, each such database is mined separately. Using this strategy, the FP-Growth reduces the search costs looking for short patterns recursively and then concatenating them in the long frequent patterns, offering good selectivity. In large databases, it's not possible to hold the FP-tree in the main memory. A strategy to cope with this problem is to firstly partition the database into a set of smaller databases (called projected databases), and then construct an FP-tree from each of these smaller databases.

Algorithm 1:FP-Tree Construction

Input: A transaction database DB and a minimum support threshold.

Output: FP-tree, the frequent-pattern tree of DB.

Method: The FP-tree is constructed as follows.

- Scan the transaction database DB once. Collect F, the set of frequent items, and the support of each frequent item. Sort F in support-descending order as FList, the list of frequent items.
- Create the root of an FP-tree, T, and label it as "null". For each transaction Trans in DB do the following:
- Select the frequent items in Trans and sort them according to the order of FList. Let the sorted frequentitem list in Trans be [p | P], where p is the first element and P is the remaining list. Call insert tree ([p | P], T).
- The function insert tree ([p | P], T) is performed as follows. If T has a child N such that N.item-name = p.item-name, then increment N 's count by 1; else create a new node N , with its count initialized to 1, its parent link linked to T , and its node-link linked to the nodes with the same item-name via the node-link structure. If P is nonempty, call insert tree (P, N) recursively.

By using this algorithm, the FP-tree is constructed in two scans of the database. The first scan collects and sort the set of frequent items, and the second constructs the FP-Tree.

After constructing the FP-Tree it's possible to mine it to find the complete set of frequent patterns. To accomplish this job, Han in presents a group of lemmas and properties, and thereafter describes the FP-Growth Algorithm as presented below in algorithm.

Algorithm 2: FP-Growth

Input: A database DB, represented by FP-tree constructed according to Algorithm 1, and a minimum support threshold.

Output: The complete set of frequent patterns.

Method: call FP-growth(FP-tree, null).

Procedure FP-growth(Tree, a) {

(01) if Tree contains a single prefix path then { // Mining single prefix-path FP-tree

(02) let P be the single prefix-path part of Tree;

(03) let Q be the multipath part with the top branching node replaced by a null root;

(04) for each combination (denoted as ß) of the nodes in the path P do

(05) generate pattern ß ∪ a with support = minimum support of nodes in ß;

(06) let freq pattern set(P) be the set of patterns so generated; }

(07) else let Q be Tree;

(08) for each item ai in Q do { // Mining multipath FPtree

(09) generate pattern ß = ai ∪ a with support = ai .support;

(10) construct ß' s conditional pattern-base and then ß' s conditional FP-tree Tree ß;

(11) if Tree ß ≠ Ø then

(12) call FP-growth(Tree ß , ß);

(13) let freq pattern set(Q) be the set of patterns so generated; }

(14) return(freq pattern set(P) ∪ freq pattern set(Q) ∪ (freq pattern set(P) × freq pattern set(Q))) }

When the FP-tree contains a single prefix-path, the complete set of frequent patterns can be generated in three parts: the single prefix-path P, the multipath Q, and their combinations (lines 01 to 03 and 14). The resulting patterns for a single prefix path are the enumerations of its subpaths that have the minimum support (lines 04 to 06). Thereafter, the multipath Q is defined (line 03 or 07) and the resulting patterns from it are processed (lines 08 to 13). Finally, in line 14 the combined results are returned as the frequent patterns found.

V.SECURITY MODEL:

Two-fold Encryption:

1.Homomorphic encryption :

We propose a symmetric homomorphic encryption scheme (using only modular additions and multiplications), which is significantly more efficient than asymmetric schemes. The scheme supports many homomorphic additions and limited number of homomorphic multiplications, and comprises the following three algorithms: • Key generation algorithm KeyGen() (s,q,p) ← KeyGen(λ)

The key generation algorithm KeyGen() is a probabilistic algorithm, which takes a security parameter λ as input and outputs a secret key SK = (s,q) and a public parameter p. Both p and q are big primes, and p ≫ q. The bit length of q depends on the security parameter, and s is a random number from Z∗ p. • Encryption algorithm E()

E(SK,m,d) = sd(rq + m) mod p.

The encryption algorithm E() is a probabilistic algorithm, which takes a secret key SK, a plaintext m ∈ Fq and a parameter d as inputs. The algorithm outputs a ciphertext c ← E(SK,m,d). The parameter d is a small positive integer called ciphertext degree, and we say the ciphertext is a d-degree ciphertext. Let r denotea big random positive integer, and the bit length of r, |r|, satisfies |r| + |q| < |p|. We say r isthe random ingredient of c. The encryption of a plaintext m is denoted by E(m) for short. • Decryption algorithm D() D(SK,c,d) = (c × s−d mod p) mod q

The decryption algorithm D() is a deterministic algorithm, which takes a secret key SK, a ciphertext c ∈ Fp and the ciphertext's degree d as inputs. The algorithm outputsaplaintext m ← D(SK,c,d). Let s−d denote the multiplicative inverse of sd in the field Fp. The correctness proof of the decryption algorithm is given below.

D(SK,c,d) = (c × s−d mod p) mod q = ((sd(rq + m) mod p) × s−d mod p) mod q = (rq + m) mod q = m.

2.Elliptic Curve Cryptography:

Elliptic curve cryptography (ECC) is an approach to public-key cryptography based on the algebraic structure of elliptic curves over finite fields. ECC requires smaller keys compared to non-ECC cryptography to provide equivalent security. Elliptic curves are applicable for encryption , digital signatures , pseudo-random generators and other tasks. They are also used in several integer factorization algorithms that have applications in cryptography, such as Lenstra elliptic curve factorization

For current cryptographic purposes, an elliptic curve is a plane curve over a finite field (rather than the real numbers) which consists of the points satisfying the equation along with a distinguished point at infinity, denoted ∞.

(The coordinates here are to be chosen from a fixed finite field /of characteristic not equal to 2 or 3, or the curve equation will be somewhat more complicated.) This set together with the group operation of elliptic curves is an Abelian group, with the point at infinity as identity element. The structure of the group is inherited from the divisor group of the underlying algebraic variety.As is the case for other popular public key cryptosystems, no mathematical proof of security has been published for ECC as of 2009.

VI. RESULT

IV. CONCLUSION

In this paper, we proposed a privacy-preserving outsourced frequent itemset mining solution for vertically And Horizontally partitioned databases using twofold Encryption with best association rule mining algorithm. This allows the data owners to outsource mining task on their joint data in a privacy-preserving manner. In future the Privacy-preserving tools for individuals and incorporating privacy protection in engineering process can be generated.

- Privacy-preserving tools for individuals: The privacy preserving techniques in research is proposed only for information holders; however individual record owners should additionally have the rights and obligations to ensure their own particular private information.
- Incorporating security assurance in engineering process: The privacy issue should be considered as an essential necessity in the engineering process of creating new technology. This includes formal detail of protection prerequisites and formal confirmation devices to demonstrate the rightness of a privacy-preserving framework.

ACKNOWLEDGEMENT

We would like to express deep sense of gratitude to project guide **Prof. P.B.Mali** for his inspiring valuable suggestions. We are deeply indebted for giving me a chance to study this subject providing constant guidance throughout this work

REFRENCES

[1] J. He and M. Wang. Cryptography and relational database management systems. In IDEAS, 2001

[2] B. Iyer, S. Mehrotra, E. Mykletun, G. Tsudik, and Y. Wu. A framework for efficient storage security in RDBMS. In EDBT, 2004

[3] Tassa, Tamir. ”Secure mining of association rules in horizontally distributed databases.” IEEE Transactions on Knowledge and Data Engineering 26.4 (2014): 970-983..

[4] D. Trinca and S. Rajasekaran, ”Towards a Collusion-Resistant Algebraic Multi-Party Protocol for Privacy-Preserving Association Rule Mining in Vertically Partitioned Data,” 2007 IEEE International

[5] Performance,Computing, and Communications Conference, New Orleans, LA, 2007, pp. 402-409.

[6] M. N. Kumbhar and R. Kharat, ”Privacy preserving mining of Association Rules on horizontally and vertically partitioned data: A review paper,” Hybrid Intelligent Systems (HIS), 2012 12th International Conference on, Pune, 2012, pp. 231-235.

