Secure Mining of Association Rules in Horizontally Distributed Databases

Shivanand Patil[1] Sukhada Vavhal[2] Pratik Mendre[3] Sagar Pokharkar[4]
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]DR. D Y Patil School of Engineering, Lohegaon

Abstract— We suggest a protocol for secure mining of association rules in horizontally distributed databases. The existing primary protocol is that of Kantarcioglu and Clifton [1]. Our protocol, like theirs, is rely on the Fast Distributed Mining (FDM) algorithm of Cheungetal, which is not a secured distributed version of the Apriori algorithm. The major ingredients in our protocol are two novel safe multi-party algorithms—one that calculates the combination of private subsets that each of the interacting players have, and another that tests the insertion of an element contained by one player in a subset contained by another. Our protocol offers enhanced privacy with respect to the protocol in [1]. In count, it is simpler and is significantly more effective in terms of interaction rounds, communication charge and computational cost. Data mining techniques are used to discover patterns in huge databases of information. But sometimes these patterns can disclose susceptible information about the data holder or persons whose information are the subject of the patterns. The idea of privacy-preserving data mining is to recognize and prohibit such revelations as evident in the kinds of patterns learned using traditional data mining techniques.[5].

Key words: Enhanced Privacy Data Mining, Distributed Computations, Frequent Item Sets (Caching), Association Rules

I. INTRODUCTION

WE study here the issue of secure mining of association rules in horizontally partitioned databases. In that situation, there are several sites that hold uniform databases, i.e., databases that share the identical representation but hold data on various entities. The aim is to find all association rules with maintain at least s and confidence at least c, for some given negligible support size s and confidence level c, that hold in the unified database, while reducing the information revealed about the personal databases held by those players.

The aim of this paper is association rule mining [2], which finds out association patterns from various sets of information and this is one of the most accepted and effective data mining methods. There have been a few privacy preserving solutions for association rule mining. Those solutions can be divided into two approaches: randomization-based and cryptography-based. In the randomization-based approach[2], there is a centralized server who aggregates data from various clients and discovers association rules.

So that to solve the issue of security another protocol is implemented for secure computation of union of private subsets. The planned protocol improves upon that in terms of ease and effectiveness as well as isolation. In particular, our protocol does not rely on commutative encryption and oblivious transfer(what simplifies it significantly and contributes towards much summarized interaction and computational expenses). While our answer is still not entirely secure, it leaks excess data only to a little number (three) of likely coalitions, unlike the protocol of that explores information also to some single players. In addition, we maintain that the extra information that our protocol may disclose is less sensitive than the extra information leaked by the protocol. The objective of new advances in data mining techniques is to effectively find out valuable and non-obvious knowledge from huge databases. The mining of association rules plays an vital role in a variety of data mining fields, such as financial analysis, the retail industry and business management. Modern organisations contain their own databases, situated in different places. Mainly mining techniques assume that the data is centralised or the distributed amounts of data can efficiently shift to a middle site to become a single model. However, organisations may be eager to share only their mining models, not their data. These centralised methods have a high threat of unpredicted information leaks when data is released. Organisations immediately require evaluation to reduce the risk of disclosing information. Privacy-Preserving Data Mining (PPDM) can run a data mining algorithm to get mutually helpful global mining objectives without revealing confidential data. Therefore, PPDM has become an vital subject in many data mining applications. In a number of business environments, the data mining may require to be processed amongst databases. Nevertheless, data may be dispersed among some sites, but not any of the sites is allowed to expose its database to a different site. Kantarcioglu and Clifton proposed a two-phase system for privacy-preserving distributed mining of association rules on horizontally partitioned data. This method transmits and encrypts huge amounts of candidates in the first stage. In the second stage, the Kantarcioglu and Clifton's system has a high risk of collusion among sites. Therefore, this study proposes the improved Kantarcioglu and Clifton's Scheme (EKCS) to speed up the procedure of the first phase and decrease the security risk in the second phase.[6]

Most existing parallel and distributed ARM algorithms are based on a kernel that employs the well-known Apriori algorithm. Directly adapting an Apriori algorithm will not considerably improve performance over frequent item sets generation or overall distributed ARM performance. In distributed mining, synchronization is implicit in message passing, so the goal becomes communication optimization. Data decomposition is very vital for distributed memory. Therefore, the major challenge for obtaining better performance on distributed mining is to discover a good data decomposition between the nodes for better load balancing, and to reduce communication. Securing the privacy of the persons whose private data become visible in those repositories is of supreme value. Al-though identifying attributes such as names and ID numbers are constantly removed before releasing the table for data

mining purposes, sensitive information might still leak due to linking attacks; such attacks may join the public attributes of the published table with a publicly accessible table like the voters registry, and thus reveal private Information of definite individuals. Privacy-preserving data mining has been designed as a paradigm of exercising data mining while securing the privacy of individuals. One of the well-studied models of privacy preserving data mining is k-anonymization. Trusted third party, each site could surrender to that third party his part of the database and trust the third party to calculate an anonymization of the combined database. Without such a trusted third party, the aim is to develop distributed protocols, for the horizontal settings, that allow the data holders to create the operation of a trusted third party and get a k-anonymized and ℓ-diverse view of the union of their databases, without disclosing needless information to any of the other parties, or to any eavesdropping rival.[8]

A. The Fast Distributed Mining Algorithm

The protocol is rely on the Fast Distributed Mining (FDM) algorithm of Cheung et al. [8], which is an unsecured distributed edition of the Apriori algorithm. Its major idea is that any s-frequent item set should be also locally s-frequent in at least one of the sites. Hence, in order to discover all globally s-frequent item sets, every player reveals his locally s-frequent item sets and then the players ensure each of them to see if they are s-frequent also globally.

The FDM algorithm proceeds as follows:

1) Initialization: It is assumed that the players have previously jointly calculated Fk_1s. The goal is to continue and calculate Fk s.
2) Candidate Sets Generation: Each player Pm calculates the set of all ðk _ 1Þ-item sets that are locally frequent in his site and also globally frequent; namely, Pm calculates the set Fk_1;m s \ Fk_1s.
   He then applies on that set the Apriori algorithm in order to produce the set Bk;m s of candidate k-item sets.
3) Local Pruning: For each X 2 Bk;ms, Pm calculates suppmðXÞ. He then returns only those item sets that are locally s-frequent. We denote this set of item sets by Ck;m s.
4) Unifying the candidate item sets: Each player broadcasts his Ck; m s and then all players compute Ck

   s :¼ SM
   m¼1 Ck; ms.
5) Computing local supports. All players calculate the local supports of all item sets in Ck s.
6) Broadcast mining results: Each player broadcasts the local supports that he computed. From that, everyone can compute the global support of every item set in Ck

s. Finally, Fk
s is the subset of Ck
s that consists
of all globally s-frequent k-item sets.
In the first iteration, when k ¼ 1, the set C1;m
s that the mth player computes (Steps 2-3) is just F1;m s, namely, the set of single items that are s-frequent in Dm. The entire FDM algorithm starts by finding all single items that are globally s-frequent. It then proceeds to search all 2-item sets that are globally s-frequent, and so onward, until it finds the highest globally s-frequent item sets. If the length of such item sets is K, then in the ðK þ 1Þth iteration of the FDM it will search no ðK þ 1Þ-item sets that are globally s-frequent, in which case it terminates.

II. RELATED WORK

Previous work in privacy preserving data mining has considered two related settings. One, in which the data owner and the data miner are two different elements, and another, in which the information is distributed among several parties who982 IEEE Transactions On Knowledge And Data Engineering, Computation and communication costs versus the support thresholds. aim to jointly implement data mining on the unified corpus of information that they hold.

In the first setting, the aim is to secure the data records from the data miner. Hence, the data owner aims at anonymizing the data prior to its release. The main aim in this context is to apply data perturbation. The idea is that the perturbed data can be utilized to infer general trends in the data, without revealing original record data.

In the second setting, the goal is to implement data mining while protecting the data records of every data owners from the other data owners. This is an issue of secure multi-party computation. The usual approach here is cryptographic rather than probabilistic. Lindell and Pinkas showed how to securely build an ID3 decision tree when the training set is distributed horizontally. Lin et al. VOL. 26, NO. 4, APRIL 2014

Discussed safe clustering using the EM algorithm on horizontally distributed data. The issue of distributed association rule mining was studied in the vertical setting, where every party holds a variable set of attributes, and in the horizontal setting. Also the work of considered this issue in the horizontal setting, but they considered large-scale systems in which, on top of the parties that hold the data resources there are also managers which are computers that assist the resources to decrypts messages; another assumption made in that differentiates it from and the current study is that no collisions occur between the different network nodes resources or managers.

The procedure of association rule mining includes two major sub-problems: the first is to find out all frequent itemsets; the second is to utilize these discovered frequent itemsets to create association rules. Since every association rule can easily be derived from the corresponding frequent itemsets, the overall performance of the association rule mining is determined by the first sub-problem. Therefore, researchers usually focus on efficiently discovering frequent itemsets. Agrawal et al. presented the Apriori algorithm to efficiently recognize frequent itemsets.Apriori is a level-by-level algorithm including several passes. In each pass, Apriori creates a candidate set of frequent k-itemsets (frequent itemsets with length k). Each frequent k-itemset is combined from two arbitrary frequent(k-1)-itemsets, in which the first k-2 items are identical.

Then, Apriori scans the complete transaction database to determine the frequent *k*-item sets. The process

is repeated for the next pass until no candidate can be generated. Apriori employs the descending closure property to efficiently create candidates in every pass. The property shows that no subset of a frequent itemset is infrequent; otherwise the itemset is rare. The property can be used to remove useless candidates to speed up the mining procedure. Other measures have been proposed to effectively discover frequent itemsets, such as level-wise algorithms and pattern-growth methods.[6]

III. CONCLUSION

We proposed a protocol for secure mining of association rules in horizontally distributed databases that improves significantly upon the existing leading protocol in terms of security and efficient working. One of the major components in our protocol is a novel security multi-party TASSA: Secure Mining of Association Rules IN Horizontally Distributed Databases protocol for implementing the union (or intersection) of private subsets that every interacting players hold. Another component is a protocol that tests the dependency of an element held by one player in a subset held by another. Those protocols exploit the fact that the underlying issue is of interest only when the number of players is more than two.

IV. FUTURE WORK

For further enhancement, this study suggests the execution of the proposed algorithm or any other advanced algorithm for managing any number of product attributes in a horizontally distributed database environment.

REFERENCES

[1] M.J. Freedman, K. Nissim, and B. Pinkas, "Efficient Private Matching and Set Intersection," Proc. Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT), pp. 1-19, 2004.
[2] T. Tassa, A. Jarrous, and J. Ben-Ya'akov, "Oblivious Evaluation of Multivariate Polynomials," J. Mathematical Cryptology, vol. 7, pp. 1-29, 2013.
[3] H. Grosskreutz, B. Lemmen, and S. R€uping, "Secure Distributed Subgroup Discovery in Horizontally Partitioned Data," Trans. Data Privacy, vol. 4, no. 3, pp. 147-165, 2011.
[4] A. Schuster, R. Wolff, and B. Gilburd, "Privacy-Preserving Association Rule Mining in Large-Scale Distributed Systems," Proc. IEEE Int'l Symp. Cluster Computing and the Grid (CCGRID), pp. 411-418, 2004.
[5] R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining," Proc. ACM SIGMOD Conf., pp. 439-450, 2000.
[6] A.V. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy Preserving Mining of Association Rules," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 217-228, 2002.
[7] M. Kantarcioglu and C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 9, pp. 1026-1037, Sept. 2004.
[8] T. Tassa and E. Gudes, "Secure Distributed Computation of Anonymized Views of Shared Databases," Trans. Database Systems, vol. 37, article 11, 2012.

