International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Privacy-Preserving Two-Party Distributed Association Rules Mining on Horizontally Partitioned Data

Patil Suraj Kakaso[1], Gadage Shriniwas[2]

[1]Pune University, G. H. Raisoni College of Engineering and Management, Pune, Maharashtra, India

[2]Professor, Pune University, G. H. Raisoni College of Engineering and Management, Pune, Maharashtra, India

Abstract: ***The need of data mining in various fields has been increased with the3 growth in storage of data. Searching the specific data is a tedious task if a wrong method is chosen. Many current applications, the data is stored in diverse locations, called as 'Distributed scenario'. And in such scenarios, the data mining must be done for various reasons. Simultaneously, the security is very important. Opening the data for multiple users will make it more vulnerable to attacks. The misuse of the data is prime concern of the data owners within the distributed data scenario. Therefore, they won't donate their data for mining purpose and if the data is enclosed with some perceptive information, the data sharing goes low. The Privacy-Preserving data mining (PPDM) endeavors to protect the privacy of the data in the course of data mining. More effective and efficient protocols are needed because of the current generic protocols have no practical implementation in this type of case. If the participating sites are more than two, a protocol is required to be developed for the horizontally partitioned data situations occurred. Because of this reason, the Privacy Preserving distributed association rules mining protocols were developed. Yet, they tend to be dependent on a secured multi-party summary and unification computation. But these will not guarantee the security in the case of two participating parties. The implementation of the protocols for the privacy-preserving two-party distributed mining of association rule mining, a secure division computation protocols and use the commutative encryption as the foundation approach has been proposed in this paper.***

Keywords: Distributed data scenario, data mining, PPDM, PPDM protocols

1. Introduction

In the recent years, the data mining has progressively become a basic a foundation technology for decision making. Thought, it faces many challenges, it still happens to be widely accepted [7]. One of the key challenges is the 'Privacy Preservation' [10]. 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 don't 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. In the meanwhile, [17] deal with the PPDM tasks in the Distributed data situations. By building the efficient cryptographic protocols featuring secure multi-party computation, it focused on the matters of ID3 decision tree learning. 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. The support is always exceeds the support threshold's minimum value. The Distributed Association Rule mining in the distributed scenario is studied in this paper. 'Fast algorithm' is a widely used algorithm, for distributed association rule mining is given in the [5]. We have ensured that the disclosure will be limited by using Secure Distributed Association Rules mining. This means that the contents of the transactions took place at one side, won't be available for the other sides, except the particular side discloses the same information for others. [11] Explained this issue in brief. Though, for the two-party scenario, on the other hand, others will never know the information of support and database size. In addition, the side doesn't reveal whether or not a locally rule is globally supported. A commutative encryption system is necessary for the implementation of the security in multi-party scenario; same is defined in the [14]. Side 1 signify that the given cipher text is compositely commutative encryption is alike, despite the order of encryption. The side 2 will signify that the similar encrypted message will be never produced

by two dissimilar plain messages. The side 3 ensures that the encryption is secure and safe. The next part is the security in the semi-honest models, which is based on the hypothesis saying that the participated parties are semi-honest. They are bound to follow the rules of the protocol using the correct inputs, but they are liberated to utilize whatever they see during the execution of the protocol. [12] Thoroughly explains the composition theorem. This is used for secure multiparty computation.

The important issue that researchers have to concentrate on is the security of data over the distributed data scenarios. Besides this, the remaining paper can be summed as: Section 2 gives association rule mining, secure association rule mining, as well as the distributed association rule mining also. Later in the section, we have reviewed some security related issues. And finally, we have quickly examined commutative encryption system. The section 3 includes briefly reviewed conclusion and some future work that is needed to be done.

2. Literature Review

In recent years, PPDM has been developed as a major research direction in the field of data mining. A significant amount of results are achieved in many areas 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 belongs to the area of privacy preservation data publishing techniques. To prevent indirect identification of data from the public databases is the core purpose of this method. It is so; because, the combination of record attributes might be used to correctly identify every single record. However, k-anonymity is alleged to be defenseless. Some privacy problems will be faced by K-anonymity, only in the case if individual attributes contains little diversity or the attacker have some background knowledge about the victim [1].

In recent few years, differential privacy [2], [3], [6] has attracted considerable number of researchers for PPDM. Differential privacy model is aimed for providing privacy to statistical queries and pattern mining. It also provides means to increase the accuracy of queries or data mining, and minimizing the chances of identification of records. The differential privacy under the situations of discrete outputs and numeric outputs are implemented by the exponential mechanism [6] and the Laplace mechanism [3].

All final globally frequent rules are unveiled in [11] to all participant sites. Knowing a rule is not supported at one's site but is supported globally in the two party scenarios. This discloses that the rules are being supported by the other site. Avoiding this leakage is not possible. [1] Follows the general approach of the FDM algorithm [4]. Here some special protocol replaces the broadcast of LLi(k) and the LLk item's support count. The rules of the originator are not revealed, it gives a technique for searching the union of locally supported rules. It also provides a method for firmly testing the support is higher than threshold. [11] Follows the two-phase approach in description. However, uniting the locally created rules and support count is done by somehow passing the encrypted values between the sites. The two phases determines the candidate rules. And also determine 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 forwarded to other parties, unless encryption of the rules is done by every participating party. The rules which are locally supported are tested globally in the latter phase.

The Key-points in this method are association rule mining, distributed association rule mining, secure association rule mining, secure infrastructure and commutative encryption system.

1) **Association Rule Mining:**
   [15] Explains the Association Rules Mining in brief. The basic problem with the mining association rule says that, one should always maintain all rules in mind, because in this, the confidence is always higher than the minimum confidence threshold, and the support will always exceed the minimum support threshold.
2) **Distributed Association Rule Mining:**
   The proposed paper studies the Distributed Association Rule mining in the distributed scenario. [5] Gives 'The Fast Algorithm' is a broadly used algorithm for distributed association rule mining.
3) **Secure Association Rule Mining:**
   By using Secure Distributed Association Rules mining, we ensure that the disclosure will be limited. The contents of the transactions took place at one side, will not be available for another side, unless the same information is revealed by respective sides for others. [11] Explained it in brief. But somehow, for the two-party scenario, the information of support and database size is not disclosed to other sides. As well as, the side doesn't reveal whether or not a locally rule is globally supported.
4) **Commutative Encryption System:**
   [14] has given the commutative encryption system is necessary for the implementation of the security in multi-party scenario. The side 1 signify that the given cipher text is compositely commutative encryption is alike, despite of the order of encryption. The side 2 signifies that the two similar encrypted messages will never be generated by the two dissimilar plain messages. The secure and safe encryption is insured by the site 3.
5) **Secure Infrastructure:**
   The next part is the security in the semi-honest models. The method of semi-honest is based on the hypothesis saying that the participated parties are semi-honest. In simple words, they will follow the rules of the protocol using the correct 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 explained in the [12].

3. Conclusion

The significant privacy concerns in the field of data mining will also prevent the unauthenticated use of the data. This paper proposed a protocol with intentionally designed technique that aims to attempt the privacy-preservation

issues of the distributed data mining of the association rules in a two-party case with the horizontally partitioned data. The protocol will inform the sites whether the rules are globally frequent under the transactions of another site or not. And also guarantees the data privacy for respected site. Here the assumption is made that there are few applications, those need the two-party protocol, and a site wants to know whether the item sets (rules) is globally frequent or not. It doesn't intend to share its support and other private data to any other party at the same time.

The proposed technique in this paper is conducted in the semi-honest model, somewhat similar to [14]. The participated parties will follow the given protocol is insured by the semi-honest model. The model will simultaneously try to infer the private information from the messages that they have received during the particular transaction session. But this paper does not claim that, we have solved all the issues related to privacy-preserved two-party association rules mining on horizontally partitioned data, but we certainly have taken some steps towards our destination.

References

[1] A. Machanavajjhala, D. Kifer, and J. Gehrke, "diversity: Privacy beyond k-anonymity", ACM Transactions 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, 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.