www.ijiarec.com

ISSN:2348-2079

Volume-5 Issue-1

International Journal of Intellectual Advancements and Research in Engineering Computations

Privacy and Security Ensured Rule Mining under Partitioned Databases

Ms. K.Dhivyalakshmi, Ms. V.Kaviya and Mr. N.Kaviyarasu, Final Year BTech (IT).,
Mrs. L.Viji ME, Assistant Professor(Sr. Gr.),
Department of Information Technology
Velalar College of Engineering and Technology, Erode, Tamilnadu, India
dhivyalakshmi5796@gmail.com

Abstract - The distributed database environment supports partitioned database management operations. The partitioned data values are stored and maintained in different databases. Horizontal and vertical partitions are managed under the data providers or parties. Two or more parties have their own private data under the distributed environment. The parties can collaborate to calculate any function on the union of their data. Association rule mining techniques are used to fetch frequent patterns. Centralized and distributed rule mining models are applied to discover the frequent patterns under the distributed environment. Trusted nodes are used to perform rule mining in centralized environment.

Privacy preserved data mining techniques are adapted to perform the knowledge discovery process with sensitive attribute protection mechanism. Anonymization methods are applied to secure the sensitive attributes I the public data values. Privacy-preserved mining is performed on vertically partitioned databases under different data owners. The association rule mining process is carried out on collective data sets. Homomorphic encryption scheme and secure comparison scheme are adapted to ensure the data privacy. The rule mining operations are performed under the Cloud aided environment.

Cloud aided framework is constructed to support data sharing and mining on outsourced data from multiple data owners. Horizontal and vertical partition based rule mining operations are supported in the system. All the partitioned data values are collected and integrated by the cloud server environment. The rule mining operations are carried out under the cloud server environment. Resource scheduling is integrated for the data and tasks under the Cloud Server. Data leakage control is improved with utility analysis mechanism. The sensitive attributes are protected with K-anonymity technique. Data communication security is ensured with the RSA algorithm.

1. Introduction

Data represent an important asset. Often concerning individuals and use them for various purposes, ranging from scientific research, as in the case of medical data, to demographic trend analysis and marketing purposes. Organizations may also give access to the data they own or even release such data to third parties. The number of increased data sets that are thus available poses serious threats against the privacy of individuals and organizations. Because privacy is an important concern, several research efforts have been devoted to address issues related to the development of privacy-preserving data management techniques.

The privacy preservation is provided when data are to be released to third parties. In this case, data once are released are no longer under the control of the organizations owning them. Therefore, the organizations that are owners of the data are not able to control the way data are used. The most common approach to address the privacy of released data is to modify the data by removing all information that can directly link data items with individuals; such a process is referred to as data anonymization. It is important to note that simply removing identity information, such as names or social security- numbers, from the released data may not be enough to anonymize the data. There are many examples that show that even when such information is removed from the released data, the remaining data combined with other information sources may link the information to the individuals it refers to. The generalization techniques are applied to overcome the problems, the most well known of which is based on the notion of k-anonymity.

The privacy-preservation techniques are also applied to provide privacy and security in the context of data mining. Data mining techniques are very effective today. Thus, even though a database is sanitized by removing private information, the use of data mining techniques may allow one to recover the removed information. In general, all approaches are based on modifying or perturbing the data in some way; for example, techniques specialized for privacy-preserving mining of association rules modify the data so to reduce the confidence of sensitive association rules. A problem common to most of these techniques is the quality of the resulting database; if data undergo too many modifications, they may not be useful any longer. To address these problems, techniques have been developed to estimate the errors introduced by the modifications; such estimates can be used to drive the data modification process. A different technique in this context is based on data sampling. The idea is to release a subset of the data, chosen in such a way that any inference that is made from the data has a low degree of confidence. Finally, in the area of data mining, techniques have been developed, mainly based on commutative encryption techniques, whose goals is to support distributed data mining processes on encrypted data. In particular, the addressed problem deals with situations when the data to be mined is contained at multiple sites, but the sites are unable to release the data. The solutions involve algorithms that share some information to calculate correct results, where the shared information can be shown not to disclose private data.

2. Related Work

Since the introduction of privacy preserving data mining [3], data perturbation has been used to protect sensitive information when outsourcing data mining of frequent itemsets [8, 9]. This approach aims at protecting the raw data rather than the mining results. Due to the random noise added to the raw data, this approach may have unpredictable impacts on data mining precision. To provide better protection on both raw data and mining results, later works employ simple encryption, which is usually a substitution mapping of raw data items. For example, Wong et al. [50] proposed a solution that encrypts original transactions by a mapping function, and adds random ‘‘fake’’ items to the encrypted transactions. Unfortunately,

Molloy et al. that the random ‘‘fake’’ items can be removed by detecting the low correlations between items, and that top frequent items can be re identified by attackers. Recently, Tai et al. [1] and Giannotti et al. proposed similar methods to achieve k anonymity protection for both raw data and mining results against a knowledgable adversary with certain background knowledge. However, none of these works assume that the adversary at server has the capability of launching chosen plaintext attacks; that is, they do not provide semantic security, which means that the adversary with knowledge on chosen plaintext–ciphertext pairs is able to infer partial information about the raw data and mining results. We clarify that our work is different from the line of work on secure multiparty computation (SMC) [4] and its application to association rule mining [2]. SMC provides a generic approach to solving problems in the distributed computing scenario, in which two or more parties compute collaboratively on an agreed function from their private inputs. When SMC is applied to association rule mining, it is assumed that multiple parties holding part of original data such as horizontally partitioned data work collaboratively in discovering association rules without revealing each other's data. In our outsourcing scenario, however, the involving parties contribute differently in performing the task – while the original data is provided by data owner only, the mining task is mainly performed at cloud servers on encrypted data.

The notions of privacy-preserving data publishing and association rule hiding are also related to our work. Privacy preserving data publishing [10, 11] enables accurate data patterns such as aggregate statistics and association rules to be discovered from published data while any individual's privacy related to the underlying data is protected in data publication. In association rule mining context, this notion protects raw data (in terms of individual's privacy) but no mining results. On the other hand, association rule hiding [6, 7] is concerned with how to transform the raw data such that certain sensitive associate rules cannot be discovered from the released dataset while other rules can. In other words, it aims at protecting part of mining results only. In comparison, our task is to protect both raw data and mining results at an outsourced server in terms of semantic security.

3. Rule Mining On Vertically Partitioned Databases

Frequent itemset mining and association rule mining, two widely used data analysis techniques, are generally used for discovering frequently co-occurring data items and interesting association relationships between data items respectively in large transaction databases. These two techniques have been employed in applications such as market basket analysis, health care, web usage mining, bioinformatics and prediction. A transaction database is a set of transactions, and each transaction is a set of data items with a unique TID (Transaction ID). An itemset Z is regarded frequent if and only if $Supp(Z) \geq Ts$, where Ts is a threshold specified by the data miner. $Supp(Z)$ is Z's support, which is defined as Z's occurrence count in the database. An association rule is expressed using $X \Rightarrow Y$, where X and Y are two disjoint itemsets. $X \Rightarrow Y$ indicates that X's occurrence implies Y 's occurrence in the same transaction with a certain confidence. We will use a supermarket's transaction database as an example, where a transaction is some customer's shopping list. A customer buying "bread" and "butter" will also buy "milk". Then {bread, butter} ⇒ milk is a possible association rule. $X \Rightarrow Y$ is meaningful and useful if the confidence is high and $X \cup Y$ is frequent.

More specifically, $X \Rightarrow Y$ is regarded as an association rule if and only if $Supp(X \cup Y) \geq Ts$ and $Conf(X \Rightarrow Y) \geq Tc$. We define $Conf(X \Rightarrow Y)$ as the confidence of $X \Rightarrow Y$. The latter is the probability of Y 's occurrence given X's occurrence (i.e. $Conf(X \Rightarrow Y) = Supp(X \cup Y)/Supp(X)$). Tc denotes the threshold specified by the data miner. We also remark that the values of Ts and Tc are generally configured based on the type of transactions, the usage of the mining result, the size of database, etc. It is easy to mine association rules after mining frequent itemsets and obtaining their supports. Most association rule mining algorithms are built based on frequent itemset mining algorithms.

Classic frequent itemset mining and association rule mining algorithms, such as Apriori, Eclat and FP-growth, were designed for a centralized database setting where the raw data is stored in the central site for mining. Privacy concerns were not considered in this setting. Vaidya and Clifton and Kantarcioglu and Clifton are the first to identify and address privacy issues in horizontally / vertically partitioned databases. Due to an increased understanding of the importance of data privacy (e.g. in the aftermath of the revelations by Edward Snowden, a former NSA contractor), a number of privacy-preserving mining solutions have been proposed in recent times. In their settings, there are multiple data owners wishing to learn association rules or frequent itemsets from their joint data. However, the data owners are not willing to send their raw data to a central site due to privacy concerns. If each data owner has one or more rows (i.e. transactions) in the joint database, we say that the database is *horizontally partitioned*. If each data owner has one or more columns in the joint database, the database is considered *vertically partitioned*. This paper focuses on vertically partitioned databases, such databases are useful for market basket analysis. For example, different businesses, such as a fashion designer and a luxury watch designer, sell different products to the same community. These businesses collaborate to mine customer buying patterns from the joint database.

A transaction of the database contains the products that a customer had bought from one or more of the participating businesses, and attributes such as the customer credit card number and date of purchase are used as TIDs. Therefore, each of the businesses (i.e. data owners) will own some transaction partitions in the joint database. However, these businesses may not wish to disclose such data, which include trade secrets (e.g. there may be other competing businesses sharing the same joint database) and customer privacy (e.g. due to regulations in existing privacy regime). Therefore, a privacypreserving mining solution must be applied. Other use cases can also be found in areas such as automotive safety and national security.

In this paper, we propose a cloud-aided privacy-preserving frequent itemset mining solution for vertically partitioned databases, which is then used to build a privacy-preserving association rule mining solution. Both solutions are designed for applications where data owners have a high level of privacy requirement. The solutions are also suitable for data owners looking to outsource data storage – i.e. data owners can outsource their encrypted data and mining task to a semi-trusted (i.e. curious but honest) cloud in a privacy preserving manner. To the best of our knowledge, this is the first work on outsourced association rule mining and frequent itemset mining for vertically partitioned databases.

The key underlying techniques in our solutions are an efficient homomorphic encryption scheme and a secure outsourced comparison scheme. The contributions of this paper are three-fold:

- This paper proposes privacy-preserving mining solutions for high privacy requirements. The proposed solutions are uniquely located in the design space. Compared with most solutions, our solutions achieve a higher privacy level, as most existing solutions require the sharing / exposure of raw data or the disclosure of the exact supports to data owners. Such requirements result in the leakage of sensitive information of the raw data. Our solutions are designed to avoid such complications. We note that one of the frequent itemset mining solutions can achieve the same privacy level as our proposed solutions. However, an association rule mining solution cannot be built based on the frequent itemset mining solution. In contrast, we present solutions for both frequent itemset mining and association rule mining. Moreover, our frequent itemset mining solution is 3 to 5 orders faster. Our solution is significantly more efficient due to our customized homomorphic encryption scheme. The introduction of a semi-trusted third party (i.e. the cloud) also allows us to securely compute supports and compare supports with a threshold *Ts* more efficiently comparative summary, respectively.

• This paper proposes an efficient homomorphic encryption scheme and a secure outsourced comparison scheme. To avoid the disclosure of supports/confidences, we design an efficient homomorphic encryption scheme to facilitate secure outsourced computation of supports/ confidences, as well as a secure outsourced comparison scheme for comparing supports/confidences with thresholds. The proposed (symmetric homomorphic) encryption scheme is tailored for the proposed comparison. The scheme only requires modular additions and multiplications, and is more efficient than the homomorphic encryption schemes used in other association rule mining and frequent itemset mining solutions. For example, encryption computing in our scheme is three orders of magnitude faster than and respectively. To the best of our knowledge, the proposed secure comparison scheme is the first scheme based on symmetric homomorphic encryption. The proposed schemes are designed for the data mining solutions outlined in this paper, but they can potentially be adopted in a wide range of secure computation applications.

• This paper proposes a ciphertext tag approach for canceling out fictitious data's effect on mining result. To "hide" the data owner's raw data from the cloud, we adapt the concept outlined in [12] by encrypting items with a substitution cipher, and adding fictitious transactions as a mitigation against frequency analysis attacks on the substitution cipher. To allow secure and accurate computation of supports, we design a ciphertext tag approach to cancel out fictitious transactions in a privacy-preserving manner. Although our approach is designed for the data mining solutions outlined in this paper, it has potential applications in other secure computation contexts, such as secure data aggregation.

4. Problem Statement

The association rule mining methods are applied to detect the hidden knowledge from the databases**.** Apriori algorithm is used to mine association rules in databases. Homogeneous databases share the same schema but hold information on different entities. Horizontal partition refers the collection of homogeneous databases that are maintained in different parties. Fast Distributed Mining (FDM) algorithm is an unsecured distributed version of the Apriori algorithm. Privacy preserved outsourced rule mining on vertical database model is applied to perform the frequent pattern mining on heterogeneous database models. The heterogeneous database model manages the tables with different schemas. The rule mining is carried out on a centralized environment. Homomorphic encryption scheme and secure comparison scheme are adapted in the data security process. All the partitioned data values are transferred to a centralized node to perform the rule mining operations. The following problems are identified from the current partitioned database rule mining systems.

- Horizontal partition based rule mining is not supported
- Utility analysis is not performed
- Resource scheduling is not supported
- Communication and computational cost is high

5. Frequent Pattern Mining on Partitioned Databases with Privacy and Security

Data preprocess module is designed to manage partitioned databases and data cleaning operations. The cloud server is build to provide resources for the rule mining process. Data security module is designed to protect the sensitive data attributes. Horizontal partition based rule mining is carried out under the rule mining on HP module. Rule mining on VP module is build to perform association rule mining on vertical partitions.

5.1. Data Preprocess

The German credit card data set is used in the system. Horizontal and vertical partitions are used in the system. The data values are imported from the data files into Oracle database. Data cleaning process is initiated to assign missing values.

5.2. Cloud Server

The cloud server is constructed to provide resources for the partitioned databases. Storage and computational resources are provided for the mining process. Key management process is used to generate and distribute the key values. Partitioned databases are collected and integrated in the cloud server.

5.3. Data Security Process

The partitioned data values are uploaded to the cloud server from the data owners. The data security is provided with privacy and cryptography techniques. Sensitive attributes are protected with the anonymization techniques. The K-anonymity technique is applied to anonymize the sensitive attributes with the user specified threshold value. The data values are anonymized and encrypted before the uploading process. The homomorphic encryption scheme is applied for the data security process. RSA algorithm is used in the system. Secure comparison scheme is applied to analyze the sensitive data values. All the data upload operations are carried out under the data security process module. The data values are received and updated by the cloud server. Local rule mining operations are carried out under the nodes with their current data values.

5.4. Rule Mining on HP

The databases are partitioned two ways. They are horizontal partition and vertical partitions. Homogeneous database schema is used in the horizontal partitions bit in the case of vertical partition heterogeneous database schema is adapted. Candidate set and item set preparation operations are called on the integrated database. Support and confidence values are estimated for the item sets. The frequent rules are filtered with the minimum support and minimum confidence values.

5.5. Rule Mining on VP

The vertically partitioned databases are build with heterogeneous database schemas. Different attribute sets are used in each database. The databases are collected and integrated under the cloud server. The rule mining is applied on the integrated data environment. The rules are filtered with minimum support and minimum confidence values. The rules are listed with its source information. Communication and computational complexity are analyzed in the system.

6. Conclusion

The cloud aided frequent pattern mining model is constructed to fetch the frequent rules on horizontal and vertical partitioned database environments. The centralized mining model is adapted in the system. All the rule mining operations are carried out with privacy and security features. The cloud server is the central authority used for the mining process. Frequent rules are mined with user support levels. Data security is improved in the system. Resource scheduling is adapted in the cloud server environment. Computational complexity is reduced in the system.

References

[1] C.H. Tai, P.S. Yu, M.S. Chen, k-support anonymity based on pseudo taxonomy for outsourcing of frequent itemset mining, in: KDD, pp. 473–482, 2010.

[2] M. Kantarcioglu, C. Clifton, Privacy-preserving distributed mining of association rules on horizontally partitioned data, in: DMKD, 2012.

[3] Junzuo Lai a,b,, Yingjiu Li a, Robert H. Deng a, Jian Wenga,b, Chaowen Guan a, Qiang Yan, "Towards semantically secure outsourcing of association rule mining on categorical data", Information Sciences 267 (2014) 267–286

[4] A.C.C. Yao, How to generate and exchange secrets, in: FOCS, pp. 162–167, 2012.

[5] W.K. Wong, D.W. Cheung, E. Hung, B. Kao, N. Mamoulis, Security in outsourcing of association rule mining, in: VLDB, pp. 111–122, 2015.

[6] Z. Zhu, W. Du, K-anonymous association rule hiding, in: ASIACCS, pp. 305–309. 2014.

[7] V.S. Verykios, A. Gkoulalas-Divanis, A survey of association rule hiding methods for privacy, in: Privacy-Preserving Data Mining, 2008, pp. 267–289.

[8] A. Mohaisen, N.S. Jho, D. Hong, D. Nyang, Privacy preserving association rule mining revisited: privacy enhancement and resources efficiency, IEICE Trans. 93-D (2010) 315–325.

[9] A.V. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke, Privacy preserving mining of association rules, in: KDD, pp. 217–228, 2014.

[10] B.C.M. Fung, K. Wang, R. Chen, P.S. Yu, Privacy-preserving data publishing: a survey of recent developments, ACM Comput. Surv. 42 (2010).

[11] J. Gehrke, D. Kifer, A. Machanavajjhala, Privacy in Data Publishing, in: ICDE, p. 1213, 2012.

[12] F. Giannotti, L. V. S. Lakshmanan, A. Monreale, D. Pedreschi, and H. Wang, "Privacy-preserving mining of association rules from outsourced transaction databases," IEEE Syst. J., vol. 7, no. 3, pp. 385–395, Sep. 2013.