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

Privacy-Preserving Mining of Association Rules in Cloud

Vishal Ravindra Redekar[1], Dr. K.N.Honwadkar[2]

Smt. Kashibai Navale College of Engineering, Pune, Maharashtra, India

Abstract: ***With the appearance of cloud computing and its domains for IT services based on the internet and big data centers, the outsourcing of data and computing services is acquiring a huge relevance. The interest in the area of data mining, as a service, has been the main stay; because of the encouraged development in the various fields, such as Cloud Computing. A third party service provider, the server, comes in the frame, when a company, the data owner, who lacks in expertise or the resources, outsources its mining needs. However the data owner thinks both the items and the association rules of the outsourced Database as a confidential property. The server stores the data and ships transformed by the data owner. Then the data owner sends mining queries to the server, and the server returns the extracted patterns. From these patterns, the owner recovers the true patterns. Within corporate privacy-preserving frameworks, the problem of outsourcing the association rule mining responsibilities in the cloud environment is studied in this paper. An innovative approach ensures that every transformed item from data owner to server is interchangeable with compared to the background knowledge of attackers is proposed in the paper. Our methods are scalable, effectual and protect privacy on an extremely huge and real transaction database representative our complete algorithm. This approach also proposes to provide the privacy-preserving mining over cloud. We can assume that a conventional model where the adversary knows the area of items and their exact occurrence and can employ this information to identify cipher items and cipher item sets.***

Keywords: Association rule mining, Privacy-preserving mining, Database outsourcing, Cloud environment, Extracted patterns.

1. Introduction

As the cloud computing and the model of the same emerged for the IT services based on the internet and the data centers, the data and computing service's outsourcing is obtaining huge relevance, and these are expected to attract the large number of researchers in few years. Services like, business intelligences and the knowledge discovery including, advanced analytics based on the data mining, are going to be amongst the services acquiescent to be outsourced on the cloud, because of the data intensive nature, that of the complexity of the algorithms for data mining. Therefore, the pattern of data mining and data management will presumable to grow as the popularity of the cloud raises [1]. This data mining-as-a-service pattern is intended to enable organizations with partial computational capitals and data mining expertise to outsource the needs of data mining to a third-party service provider [2], [3]. The main security drawback is that the server can always access the valuable data from the owner and can learn sensitive information from it. But, the transactions and these mined patterns both are and will always be the property of data owner. And these must remain in safety at server [4]. Protection of sensitive information in the situation of our research includes two essential goals: knowledge protection and privacy preservation. The earlier is related to privacy preserving association rule mining, whereas the final refers to privacy-preserving clustering. An appealing feature involving knowledge protection and privacy preservation is characteristics that they have common. For example, in knowledge protection, an organization is the owner of the data so it must protect the sensitive knowledge discovered from such data, while in privacy preservation individuals are the owner of their personal information [5].

In this paper, we have proposed a method with the goal in mind that to develop an encryption scheme that will enable formal privacy guarantees and will validate this model with large-scale-real-life transaction databases (TDBs). The client encrypts the data, based on encrypt-decrypt (E/D) module that is essential to be treated as a black box. The transformation of the input data into the encrypted database is done by this module. The data mining and the encryption pattern sending to owner is conducted by the server. The recovery of true identity of the returned patterns and their true supports is done by the E/D module. It is insignificant to explain that when data are encrypted by using 1–1 substitution ciphers without making use of fake transactions, many ciphers and thus, the transactions and prototypes can be broken down by the server by high probability with launching the frequency-based attack. Therefore, the prime focus of this paper is to propose a new encryption schemes so as to the guarantees of the formal privacy can be proved against the attacks carried out by the server. The server might use the background knowledge, at the same time controlling the resource requirements. The research done on Privacy-Preserving Database Mining (PPDM) has attracted lot of attention in few years. The main approach provided here is, the private data is collected from multiple owners by a collector, called server, for the prime purpose of combining the data and conducting the mining on these combined data. The data are subjected to an arbitrary perturbation, as the collectors are not trusted with protection of the privacy, as it is collected. Many techniques have been invented for disturbing the data so that the privacy will be preserved, at the same time ensuring the mined patterns and other systematic properties are adequately close to the mined patterns from original data. This method works as pioneered by [6] and some papers have been following this since [7]. But these approaches are insufficient for corporate privacy.

In order for privacy preservation, before the records are shared, the information records can be de-identified. By deleting some unique identity fields, like name, passport

number etc, it can be accomplished. But even after deleting this information, some other kinds of information like, behavioral or personal information, is still available. The information may contain Dob, gender, number, zip code, postal address, number of accounts. And if this information is with any other dataset, it would identify the subjects. To avoid this disclosure of important information, a privacy preservation algorithm is essential in association rule mining. There are many different techniques have been proposed to tackle this problem. Where, most of the methods cause information-loss or some other side effects. The side effects may include the falsely generated bogus rules or mistakenly hiding non-sensitive rules. So it is essential to sort out these papers into classes in such an approach to distinguish the benefits and faults of diverse principle concealing procedures.

The important issue that researchers used to concentrate was the use PPDM in cloud. This issue is solved in this paper by the proposed technique. The main issue to be researched in near future is to find a more secure technique for the storage of the data on the cloud. As the techniques will grow, so does the different types of attacks will. More research is needed to be done in that area. The remaining paper can be sorted out as: Section 2 gives Pattern mining task that is the base for this method. Later in the section, we have reviewed privacy model on which, this method is based. Adversary knowledge and attack model is also described there. Finally, we have quickly examined the Encryption/Decryption scheme. This scheme includes the Encryption, Decryption and the Grouping of the items. In section 3, is briefly reviewed conclusion.

2. Literature Review

Evmievski et al. [8] proposed an approach, for conducting the privacy preserving association rule mining. Kargupta et al. [9] proposed a method based on random matrix spectral filtering to recover original data from the perturbed data. Huang et al. [10] proposed further, the two data reconstruction methods, first PCA-DR and second, MLE-DR.

In accumulation, different distribution reconstruction algorithms have been proposed in association to vary randomization operators [11-13]. The base for these algorithms was the estimation of the original data distribution based on the randomization operator as well as the randomized data, by using Bayesian network.

The first person to propose the Randomized Response (RR) was Warner [14].The RR scheme was initially developed in the statistics community. It used to collect the information from individuals such that, the survey interviewers and the data processors do not know which of the two alternative questions are respondent have answered. In data mining, The method of randomization is a simple technique, can be very easily applied at data collection time. It was a useful technique for hiding individual data in privacy reserving data mining. The randomization method is more efficient. Though, it results in high information loss.

The literature on Privacy-Preserving Mining of Association rules can be classified into Pattern mining task, privacy model, and finally Encryption/Decryption scheme.

2.1 Pattern Mining Task

It is assumed that the reader is familiar with the basics of association rule mining. The most known and frequent pattern mining problem is explained in [13] as:

Given a transaction database ‘D’ and a support threshold ‘x’, then with support of ‘D’ and with at least ‘x’, find all the item sets. It is confined with the study of a Privacy Preserving Outsourcing Framework for Frequent Pattern Mining, in this paper.

2.2 Privacy Model

For the protection of the identification of the individual data items, the data owner encrypts the original dataset, and transforms it in an encrypted database. Items in original database can be called as Plain items, whereas items in encrypted database can be called as Cipher items.

The server or an intruder may have background knowledge of the encrypted database to attack it. Therefore, the proposed scheme is based on the two main points: first, replacing each item in database with 1 to 1 substituted ciphers, and second, adding some fake transactions in the encrypted database.

2.3 Encryption/Decryption Scheme

2.3.1 Encryption

In this step, we have added an encryption algorithm called, “RobFrugal”; this is used to transform a transaction database into its encrypted version. It consists of three prime phases:

1. For each plain item, using a 1-1 substitution cipher text.
2. A specific item grouping method need to be used.
3. A method is needed to add fake transactions.

2.3.2 Decryption

After the client requests a pattern execution query to be executed to the server, with the specified support threshold, the server will always return the encrypted databases computed frequency patterns. This projected E-D scheme is practical resolution for a Privacy Preserving Pattern Mining rather than the outsourced Transaction database, but the correct and efficient implementation needed to be existed. On the other side, storing the support for every cipher pattern is not practical.

2.4 Grouping items for Privacy

Some of the strategies can be adopted to classify the items into the groups of fixed size, given that the items in supported table. A method called Frugal is used to initiate. We also presume that the item support table is in the descending orderly of support sorted and the cipher items are referred in this order. Given the detail that says, the support of the items will strictly lowers monotonically. And

if the item support table is sorted in the support's descending order, the frugal grouping is optional in this technique.

2.5 Constructing Fake Transaction

If a noise table is given, with the specifying noise that is needed for every cipher item. The fake transactions are generated as follows:

1. Corresponding to the highest common items of each group or to remaining items with equal support to the maximum support of the group, we will drop all the rows with zero noise.
2. The remaining rows will be sorted by the descending order of the noise added.

3. Proposed System

With the help of Rob-Frugal Algorithms as basic idea to develop new system, which is efficient in maintaining Time and Space complexity in feasible criterion as comparing with previous ones.

As we will concern about security of third party server and probably considering data mining-as-a-service, cloud data security will be handled on primary basis. Various issues in previous privacy preserving and security of outsourced data are thus to be considered to handle with new improved and efficient system proposed on behalf of elaborating attacker's behavior pattern .

The proposed system will elaborate privacy preserving of association rules in cloud with emerged standards of:

1) Pattern Matching Task
2) Privacy Model
3) Encryption/Decryption Scheme
4) Elaboration of Attack Models in Future

4. Conclusion

Preserving privacy in data mining activities is a very important issue in many applications. Most randomization based techniques are likely to play an important role in this domain. In this paper, a new approach to solve the problem of privacy preserving data mining in the scenario of outsourced business transaction database has been solved successfully. This approach is efficient and better than many other perturbation and anonymity techniques. Proposed system will reduce the time and space required for execution, as well as false rules problems in effective manner from the previous work.

References

[1] R. Buyya, C. S. Yeo, and S. Venugopal, "Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities", in Proc. IEEE Conf. High Performance Comput. Commun., 2008.
[2] W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis, "Security in outsourcing of association rule mining", in Proc. Int. Conf.Very Large Data Bases, 2007.
[3] L. Qiu, Y. Li, and X. Wu, "Protecting business intelligence and customer privacy while outsourcing data mining tasks", Knowledge Inform. Syst., 2008.
[4] C. Clifton, M. Kantarcioglu, and J. Vaidya, "Defining privacy for data mining", in Proc. Nat. Sci. Found. Workshop Next Generation Data Mining, 2002.
[5] V. Richhariya, P. Chaurey, "A Robust Technique for Privacy Preservation of Outsourced Transaction Database", IJRET, 2014.
[6] R. Agrawal and R. Srikant, "Privacy-preserving data mining", in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2000.
[7] S. J. Rizvi and J. R. Haritsa, "Maintaining data privacy in association rule mining", in Proc. Int. Conf. Very Large Data Bases, 2002.
[8] A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke, "Privacy Preserving Mining of Association Rules", Information System, 2004.
[9] H. Kargupta, S. Datta, Q. Wang, K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques", In Proceedings of the 3rd International Conference on Data Mining, 2003.
[10] Z. Huang, W. Du, B. Chen, "Deriving Private Information from Randomized Data", In Proceedings of the ACM SIGMOD Conference on Management of Data, 2005.
[11] D. Agrawal, C.C. Aggarwal, "On the Design and Quantification of Privacy Preserving Data Mining Algorithms", In Proceedings of the 20th ACM SIGMOD-SIGACTSIGART Symposium on Principles of Database Systems, 2001.
[12] A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrk, "Privacy Preserving Mining of Association Rules", In Proceedings the 8th
[13] ACM SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining, 2002.
[14] S. L. Warner, "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias", J. Am. Stat. Assoc., 1965.