Privacy Preserving and Enhancing Security in Association Rule Mining using Robfrugal Algorithm

M.Saravanan

Department of Computer Science and Engineering,
SRM Institute of Science and Technology,
Ramapuram Campus,
Chennai, Tamilnadu, India.

ABSTRACT

*Security of the data or the database is the prime concern in today's technology. Association rule mining recognizes the frequent items based on the market-basket data analysis. Frequent item-set rule mining is been used for identifying the fake or the duplicate rules in the mining process. The cryptographic based technique is given priority for the privacy preserving of the database. Privacy Preserving Data Mining (PPDM) is a solution for privacy threats in data mining. A connection is been made between the data owner and the cloud through centralized database, where the unprocessed data is been processed using the solutions given by the cloud. Ho*lomorphic encryption scheme and secure comparison scheme are used to ensure the data privacy.

Keyword: *Privacy preserving data mining, association rule mining, frequent item-set rule mining, holomorphic encryption scheme and secure comparison scheme.*

I. INTRODUCTION

Privacy preserving data mining (PPDM) is generally used to maximize the analysis and to maximize the disclosure of individuals or an organization's private data. The two main techniques used in data mining are association rule mining and frequent item-set rule mining. The data is been shared between the client and the cloud server in order to process an unprocessed data using the solutions given by the cloud. The unprocessed data contains defects and these defects are been identified and the identified defects are visible to the cloud from where the solutions are reported back to the client. The whole process is done by creating a centralized database or a joint database to which the data owner transforms his/her data to protect cooperate privacy and then ships it to the server. The server identifies the defects or the queries and sends the queries to the cloud by protecting the private information of the data owner. The queries are received by the cloud and the cloud undergoes identification and computation process after which it is going to give the results to the server. The server then receives the results or the solutions which is passed to the data owner through which the data is been processed. Privacy preserving association rule mining is divided into three techniques 1. Reconstruction based technique 2.Heuristic based technique 3.Cryptographic based technique. The main purpose of preserving privacy in data mining is to protect sensitive knowledge and to make minimum changes possible in main database in a manner that sensitive data cannot be changed after running data mining process. For this process to happen reconstruction based technique is used where the main data does not directly go through changes.

Association rule hiding is one of the techniques which hide sensitive rules which are generated by association rule generation algorithm before releasing database. Heuristic based technique has a heuristic algorithm which provides privacy to sensitive rules while ensuring data quality. Cryptographic techniques are used for data security by the encryption and decryption processes. The most commonly used encryption algorithm is Advanced Encryption Scheme (AES) encryption algorithm for securing the data which avoids leakage of the sensitive information when combined with the association rule mining and frequent item-set rule mining.

At first the data to be partitioned in order to secure the data according to the privacy concerns of the data owners. If the data owner has one or more rows in the joint database then that database is called as horizontal database. If each data owner has one or more columns in the joint database then it is called as vertically partitioned database. In horizontally partitioned database, each site possesses different set of tuples for the same set of attributes whereas in vertically partitioned database each site possesses the common set of transactions for distinct set of attributes. The main aim of the paper is to achieve high privacy level including high performance level. The evaluation of the performance is been done with the help of threshold (Ts). Association rule mining mainly performs two operations: Frequent item generation: This is the item set which satisfies minimum threshold by identifying fake and duplicate rules. Rule generation: It generates rules for the database to be secured and is also called as strong rules as it has high confidence. The mining of association rules plays an important role in various data mining fields, such as financial analysis the retail industry and business decision making. This paper has the following sections 1.Introduction 2.Background and related work 3.Proposed system 4.Cryptographic techniques 5.Conclusion.

II. BACKGROUND AND RELATED WORK

The database can be divided into two types 1.Centralized Database and 2.Distributed Database. The centralized database can be called as a Data warehouse where all the datasets are collected at one central site and any mining operations can be performed in it. The various techniques used in centralized database are Data perturbation, Data blocking and Reconstruction based technique. Whereas, the distributed database the data can be partitioned into two categories 1.Vertically partitioned database and 2.Horizontally partitioned database. As the users do not wish to disclose their information to other users but are interested in achieving aggregate results from the dataset this distributed data is been used for dividing the data among the users.

Association rule mining:

Data mining procedure extract the required information from the huge database. Association rule mining plays a major role in the extraction of this information by using association rule generation algorithm. Four efficient namely secure sum, secure set union, secure size of set intersection and scalar product for privacy preserving data mining are introduced.

Association rules are if-then statements that help uncover relationships between seemingly unrelated data in a relational database. An example of an association rule mining would be “If a student is from English medium school and attendance > 80% then his/her result is pass”. Association rules are created by analyzing data for frequent if-then patterns and then using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if-then statements have been found to be true.

For an association rule X=>Y the support and confidence is

$$\text{Support }(X) = (\text{Support\_count}(X) / n) * 100 \tag{1}$$

$$\text{Confidence}(X\text{-> }Y) = (\text{Support}(XY) / \text{Support}(X)) * 100 \tag{2}$$

Frequent item-set rule mining:

The rules generated by the association rule mining are to be checked to make sure that there are no fake or duplicate rules present. For this process frequent item-set rule generation is been used which identifies the duplicate or fake rules from the generated set of rules.

Requirements of a PPDM Algorithm:

Accuracy: The accuracy is been identified by the loss of data, the less the data loss the better is the information quality.

Scalability: Scalability depicts the proficiency patterns when information sizes increment. It is an alternative critical perspective to the execution.

Data Quality: High quality information that has been arranged particularly for information mining assignments will bring valuable information mining models. On the other hand, low quality information has a noteworthy negative effect on the utility of information mining results.

Security: The major role is played by the privacy or the security without loss of data and wrongdoing.

The existing system has the high privacy level as an advantage but the performance of the process is low as the privacy is high and vice versa. The solutions which are been proposed are done by an assumption that the attacker has no knowledge about the encryption process; this is the major drawback of the existing system. The data base is been outsourced to the cloud by the data owners using the server, during this process there is a leakage of data and cooperate privacy level is not up to the mark. Due to which the trust of the clients on the server is a doubt. To overcome this mining techniques like association rule mining, frequent item-set rule mining and Rob frugal algorithm is been used for the extraction and encryption of the database in the proposed system.

III. PROPOSED SYSTEM

Due to fast improvement in the field of information technology a serious issue has been rise about huge information storage. To overcome these problems data mining techniques play a vital role. Data mining is capable of analyzing vast amount of information within a minimum amount of time. In this paper, the problems encountered in the existing system are tried to overcome by using the association rule mining and advance encryption techniques. The rules are been generated for the extraction of data from a huge database. The rules are been checked to avoid fake or duplicate rules with the help of frequent item-set rule mining. These two techniques and algorithms play a vital role in the mining process.

The main aim of the process is to create a centralized database or a joint data base for the data owner and the cloud in which they can share the required information without any leakage of the data and with cooperate privacy. For achieving the cooperate privacy we are going to encrypt the database using the Rob frugal encryption process by which the private data of the data owner is not visible to the cloud server. The queries of the data owner are only visible to the cloud server to which the cloud server gives a response with solutions. The high privacy and high performance together are tried to achieve in this proposed system.

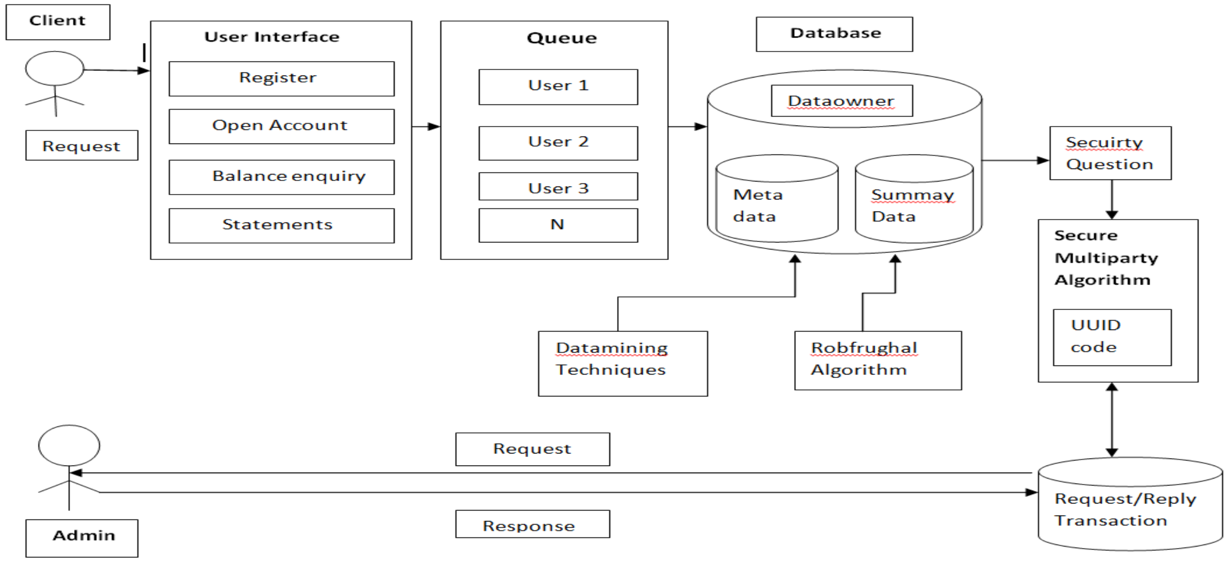

Fig. 1.1 System Architecture

The centralized system or the joint system is where the database given from the server is stored. It can be divided into visible and non-visible data where the visible data could be viewed by the parties I.e., the server and cloud whereas the non-visible data is to be hidden from cloud as this would consist of sensitive information of the data owner. Privacy is to be provided to the non-visible data using a Rob frugal encryption scheme. The original TIDs of some databases may contain sensitive information. To hide such information, the TIDs in the outsourced databases are replaced by the hash values of the original TIDs. Because of the pre- image resistance property of cryptographic hash function, the cloud cannot recover original TIDs from the TIDs used in the outsourced databases. The main aim would be to get the defects or problems being encountered, analyse and send it to the cloud to get the required solution. Once the solution is given it is applied and tested and the process continuous. Considering the hospital scenario, the patients details such as name etc., are kept confidential while sending it to the cloud I.e., they are kept under non-visible data. The rest would be used to find the particular solution for the problem

IV. CRYPTOGRAPHIC PROCESS

The encryption and decryption process plays a major role in this particular privacy preserving data mining. To achieve the high level privacy the Rob frugal encryption technique is been used where the encryption or decryption process is done before transforming it to the cloud server. For this process five main steps are followed to acquire maximum privacy level.

1. A attack model is been defined for the adversary and to make precise the background knowledge the adversary may possess.
2. An encryption scheme is been proposed called Rob frugal algorithm, that the encryption and decryption module can employ to transform client data before it is been shipped.
3. To allow the Encryption or Decryption module to recover true patterns and their correct support.
4. A formal analysis is been conducted based on the attack proposed to prove the probability of each transaction.
5. An experimental analysis is been conducted to show the encryption scheme is effective, scalable and achieve the desired level of privacy.

(i) Encryption Scheme

An encryption scheme is introduced which transforms a TDB D into its encrypted version D. Our scheme is parametric with respect to k > 0 and consists of three main steps:

- using 1-1 substitution ciphers for each plain item
- using a specific item k-grouping method
- using a method for adding new fake transactions for achieving k-privacy.

the constructed fake transactions are added to D (once items are replaced by cipher items) to form D and transmitted to the server. Homomorphic encryption scheme allows one or more plain text operations to be carried out on the cipher texts. If the addition operation is allowed, then the scheme is called as additive homomorphic encryption. If the multiplication operation is allowed, then the scheme is called as multiplicative homomorphic encryption. In an additive homomorphic encryption scheme, the cipher text of the sum of two plaintexts,m1+m2,can be obtained using some computation "." on the cipher texts of m1 and m2, without first decrypting m1and m or requiring the decryption key. Let Epk() be the function of encrypting with the public key, and "." Is modular multiplication in rob frugal.

Epk(m1),Epk(m2) and the public key used in the encryption, one can compute Epk(m1+m2) by performing a modular multiplication of Epk(m1) and Epk(m2). Similarly, given Epk(m1),m2 and the public key, one can compute Epk(m1*m2) by performing a modular exponentiation

*Epk(m1)m2 Epk(m1+m2) = Epk(m1)*Epk(m2) Epk(m1*m2)*
*Epk(m1)*Epk(m2) Epk(m1*m2) = (Epk(m1)Epk(m1)*... Epk(m1)) / (m2 multiplications) = Epk(m1)m2*
In the remainder, denotes homomorphic addition.

(ii) Decryption Scheme

When the client requests the execution of a pattern mining query to the server, showing a minimum support threshold σ, the server send back the computed frequent patterns from D. Clearly, for every item set S and its corresponding cipher item set E, we have that supp D(S) ≤ supp D_(E). For every cipher pattern E returned by the server together with supp D_(E), the E/D module restores the corresponding plain pattern S. It needs to remake the exact support of S in D and decide on this basis if S is a continuous pattern. To obtain this goal, the E/D module adjusts the support of E by removing the effect of the fake transactions.

Supp D(S) = supp D_(E)−supp D_\D(E). Finally, the "S" pattern with adjusted support is kept in the output if suppD(S) ≥ σ. The calculation of supp D_\D(E) is performed by the E/D module using the synopsis of the fake transactions in D \ D. V. ROBFRUGAL ALGORITHM An attack model was generated based on following criteria such as based on assumption that the service provider (who can be an attacker) is semi honest in the sense that although he does not know the details of the encryption algorithm, he can be curious and thus can use his background knowledge to make inferences on the encrypted transactions.

It has been assumed that the attacker always returns (encrypted) item sets together with their exact support. Rob Frugal algorithm provides privacy to the database. By Rob Frugal algorithm, true support of mined patterns can be recovered. Rob Frugal algorithm involves one to one substitution, k grouping methods and Fake transactions. Rob Frugal encryption converts a Plain Transaction Database (TDB) into an encrypted Database D. At the time of pattern mining, the patterns are generated for given query with the high possibilities of spurious or fake patterns that probably degrades the accuracy of generated patterns.

(i). One to One Substitution Data owner will encrypt the original transaction database in one to one substitution method. It helps to encrypt the plain text into cipher text using private key.

(ii). Grouping Method, The Frugal method consists of grouping together cipher items into groups of k adjacent items in the item support table in decreasing order of support, starting from the most frequent item.

(iii). Encryption and Decryption Scheme The following steps can be applied according to the Rob Frugal scheme.

• The new transactions in TDB are inserted into the prefix tree T , obtaining a cumulative representation of TDB. Also, a cumulative item support table IST is constructed by adding the support of each item in IST and IST.

• In particular, for each item ei∈ IST∗ the support of ei is added to the support of ei∈ IST. Clearly, IST could both: a) not contain some item belonging to IST

In case a, the support of these items in the cumulative item support table IST is equal to the support of them in IST; while in case b the support of these items in IST is equal to their support in IST. Note that when the cumulative item support table IST is constructed the method keeps the order of the items in the IST. When an item only belongs to the IST, then this item is appended to the list. Clearly, the balance of support in each group is now generally destroyed by the new item supports, and it is needed to add new fake transactions to restore the balance.

• The old grouping is checked for robustness with respect to the overall prefix-tree T and the existing synopsis, which is equivalent to checking against to D ∗ ∪ F .

• If the check for robustness fails, then a new grouping is tried out with swapping, until a robust grouping is found. Notice that the new grouping is robust with respect to the new fake transactions, as the most frequent item of each group does not occur in any fake transaction.

• The E/D module uses both old and new synopses to reconstruct the exact support of a pattern from the server.

Our method extends to the case when simultaneously, a new batch is appended and old batch is dropped; the method also works in the case when new items arrive or old items are dropped. D. Creating Fake Transactions A noise table specifying the noise N(e) needed for each cipher item e, we generate the fake transactions as follows. First, we drop the rows with zero noise, corresponding to the most frequent items of each group or to other items with support equal to the maximum support of a group. Second, we sort the remaining rows in descending order of noise. Let 1 . . . m be the obtained ordering of (remaining) cipher items, with associated noise N (1) . . . N (m).

V. CONCLUSION

The algorithms like association rule mining, frequent item-set rule mining, Rob frugal algorithm are been used for acquiring the privacy level are been used. The maximum drawbacks of the existing system have been overcome in this paper. The performance level and the privacy level are met at the same level with no compromise and leakage of data.

REFERENCES

[1] Lichun Li, Rongxing Lu, Senior Member, IEEE, Kim-Kwang Raymond Choo, Senior Member, IEEE, Anwitaman Datta, and Jun Shao, “Privacy-Preserving Outsourced Association Rule Mining on Vertically Partitioned Databases”, IEEE Transactions on Information Forensics and Security, 2016.

[2] Agrawal R, Imielinski T, Swami A, “Mining association rules between sets of items in large databasesWashington, DC, 1993, pp.207–216.

[3] Chen M S, Yu P S. Data Mining:An Overview from a Database Perspective [J]. IEEE Trans on Knowledge and Data Engineering, 2004,8(6) :866-883.

[4] M K Reiter. Crowds:Anonymity for Web Transactions[J]. The ACM Transactions on Information and System Security,2005.

[5] R. Agrawal, R. Srikant, “Fast algorithms for mining association rules,”In: Proc. 20th Int’l Conf. Very Large Data Bases, 1994.

[6] Shaofei Wu, Hui Wang, “Research On The Privacy Preserving Algorithm Of Association Rule Mining.

[7] T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets, “Using association rules for product assortment decisions: A case study,” in SIGKDD 1999.

[8] S. E. Brossette, A. P. Sprague, J. M. Hardin, K. B. Waites, W. T. Jones, and S. A. Moser, “Association rules and data mining in hospital infection control and public health surveillance,” Journal of the American medical informatics association, vol. 5, no. 4, pp. 373–381, 1998.

[9] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, “Effective personalization based on association rule discovery from web usage data,” in WIDM 2001.

[10] C. Creighton and S. Hanash, “Mining gene expression databases for association rules,” Bioinformatics, vol. 19, no. 1, pp. 79–86, 2003.

[11] X. Yin and J. Han, “Cpar: classification based on predictive association rules.” in SIAM SDM 2003.

[12] J. Zhan, S. Matwin, and L. Chang, “Privacy-preserving collaborative association rule mining,” in DBSEC 2005.

[13] S. Zhong, “Privacy-preserving algorithms for distributed mining of frequent itemsets,” Information Sciences, vol. 177, no. 2, pp. 490–503, 2007.

[14] P. Paillier, “Public-key cryptosystems based on composite degree residuosity classes,” in EUROCRYPT 1999.

[15] R. Cramer, R. Gennaro, and B. Schoenmakers, “A secure and optimally efficient multi-authority election scheme,” European transactions on Telecommunications, vol. 8, no. 5, pp. 481–490, 1997.

[16] National Highway Trac Safety Administration. Firestone tire recall. http://www.nhtsa.dot.gov/hot/Firestone/Index.html, May 2001.

[17] A. Prodromidis, P. Chan, and S. Stolfo. Meta-learning in distributed data mining systems: Issues and approaches, chapter 3. AAAI/MIT Press, 2000.

[18] S. J. Rizvi and J. R. Haritsa. Privacy-preserving association rule mining. In Proceedings of 28th International Conference on Very Large Data Bases.