ISSN No: - 2456 - 2165

Privacy Preserving Association Rule Mining using ROBFRUGHAL Algorithm

[1]K.Aishwarya chakravarthy, [1]M.Meena , [1]M.Spoorthy , [2]Saravanan
[1]UG Students, SRM University, Chennai,
[2]Assistant professor, SRM University ,Chennai,

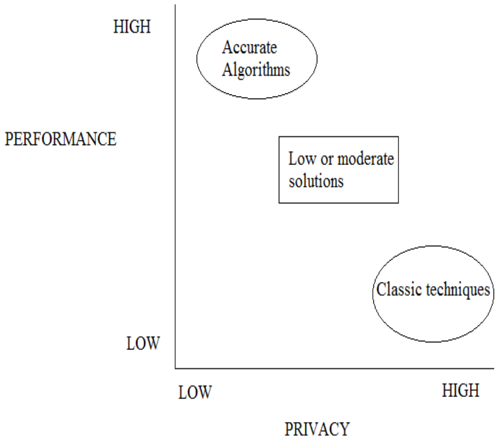
Abstract :- **Concept of data preservation has become a challenging task. Security of the data is one of the prime concerns in today's technology. Privacy Preserving Data Mining (PPDM) deals with the protection of private data. Association rule mining and frequent item-set rule mining are highly used for application purposes. These lead to approximations which have high accuracy, low data leakage and improvement in efficiency. Transparency of the data processing should be limited to the users involved in order to minimize the leakage of sensitive business information. Accuracy and efficiency is being obtained from the rules. Homomorphic encryption scheme is been developed to ensure data privacy. The data is transformed by the owner in order to provide privacy which is then sent to the server. The server receives the mining queries which are then sent to the cloud to get the desired solution. SIN order to protect the sensitive information in the corporate field, the owner would perform ROBFRUGHAL encryption scheme which is transformed to the server.**

Index Terms: *Privacy preserving data mining, Association rule mining, frequent item-set rule mining, Rob frugal encryption scheme algorithm.*

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.

There are many researches that have taken place in securing the privacy of the database but the privacy level and the performance level were equal. In other words, if the privacy of a data base is high then the performance level is low and it's vice versa. This disadvantage is been tried to overcome were both the privacy and performance level are at the same level with no leakage of information and without any compromise.



A centralized database forms a connection between the server and the cloud. Frequent item-set mining and association rule mining algorithms were have a centralized database setting where the raw data is stored in the central site for mining. The database can be vertically or horizontally partitioned, when there are one or more rows it is vertically partitioned and when there are one or more columns it is horizontally partitioned. However, nobody would be willing to send their raw data to a central site due to privacy concerns.

Consider the market basket example where a site may contain grocery purchases, while another has clothing purchases, using a key such as credit card number and date, we can join these to identify relationships between purchases of clothing and groceries. However, this discloses the individual purchases at each site, possibly violating consumer privacy agreements. The aim is to provide security to private information I.e., confidential information or sensitive information of the data owner.

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 every data owner have 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 focus 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). 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.

A. Data Mining

Data mining is the method of determining patterns in large data sets with machine learning, statistics and database systems. Data mining process is used to extract information from a huge volume of data set to have logical structural representation of the data item in the transactional database. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in database. Data mining is the part of the knowledge discovery process.

Association rule mining is a technique in data mining that is used to find regularities in a large volume of data. This technique may identify data which is private from human or an organisation. Frequency item set mining is a technique where the fake rules would be identified. A homomorphic encryption scheme is developed to avoid disclosure of data. The Knowledge Discovery in Databases process comprises of a few steps leading unprocessed data to a processed data:

- Data cleaning: This is a step where noisy and irrelevant data are removed from the collection.
- Data integration: At this stage, multiple data sources, often heterogeneous, may be combined in a common source.
- Data selection: At this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- Data transformation: It is also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- Data mining: It is the crucial step in which clever techniques are applied to extract patterns potentially useful.

B. Privacy Preserving Data Mining

The privacy preserving data mining is to provide a solution for protecting sensitive information by developing a data mining techniques which could be applied on databases without affecting the accuracy of data mining result and without violating the privacy of individuals is the motivation for this research. Privacy preserving data mining with association rule denotes the area of data mining that looks to preserve sensitive information from unnecessary or unlawful disclosure.

Privacy information comprises personal or confidential information in business like social security numbers, home address, credit card numbers, credit ratings etc. The privacy preservation data mining requires guarantee for hiding of sensitive information in efficient manner. The association rule hiding technique protects the sensitive data .

It affects the privacy of rules and the utility of the data mining results. The association rule mining takes right decisions to improve the performance of the business or service origination. But the main threat to the association rule mining is privacy.

It performs data mining on union of two parties. Data stays private that is no party learns anything but output. Assumption is made that it contains large databases-Generic solutions which is not possible and semi-honest parties is present. PPDM deals with protecting the privacy of individual data or sensitive knowledge without sacrificing the utility of the data. 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.

A. 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

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

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

B. 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.

C. 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 whereas, low quality information hasa negative effect on the utility of information mining results.

- **Security**

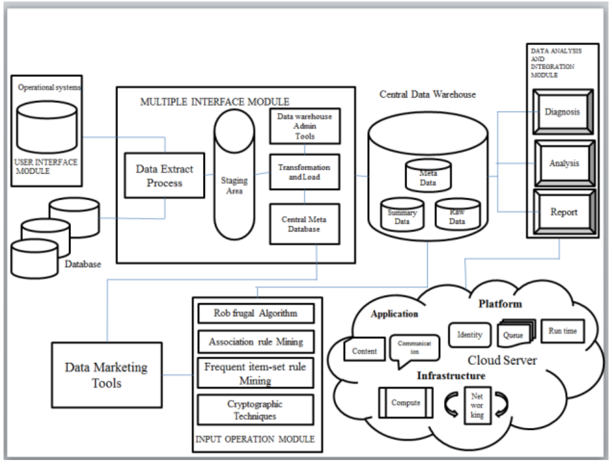
  The major role is played by the privacy or the security without loss of data and wrongdoing.

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 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.



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 TECHNIQUES

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,which transforms client data before it is being 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.

A. Encryption Scheme

An encryption scheme is introduced which transforms a TDB D into its encrypted version D. Our scheme is parametric w.r.t. 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 $E_{pk}()$ be the function of encrypting with the public key, and "." Is modular multiplication in rob frugal. $E_{pk}(m1)$,$E_{pk}(m2)$ and the public key used in the encryption, one can compute $E_{pk}(m1+m2)$ by performing a modular multiplication of $E_{pk}(m1)$ and $E_{pk}(m2)$. Similarly, given $E_{pk}(m1)$,m2 and the public key, one can compute $E_{pk}(m1*m2)$ by performing a modular exponentiation $E_{pk}(m1)^{m2}$.

$$E_{pk}(m1+m2)=E_{pk}(m1)*E_{pk}(m2)$$

$$E_{pk}(m1*m2)=(E_{pk}(m1)E_{pk}(m1)*\ldots E_{pk}(m1)) / (\text{m2 multiplications})$$

$$= E_{pk}(m1)^{m2}$$

In the remainder, . denotes homomorphic addition.

B. 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 supp

D(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.

A. 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.

B. K-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.

C. 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∗, and
  - *b)* Contain some new items. 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).

VI. CONCLUSION

The wide range of experiments on existing techniques calculates the relative performance of several privacy preserving techniques and its limitations. For this reason the privacy preservation technique using association rule mining and frequent itemset mining are implemented which the main approach of privacy preservation during association rule mining, a centralized database has been created through which the transactions between client and server are taken place. It avoids data leakage which is caused by data sharing. High privacy has been maintained without any compromise with performance level.

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].N.V.Muthu Lakshmi & K.Sandhya Rani , Research scholar, Professor, Dept of Computer Science, IJ, “Privacy-Preserving Outsourced Association Rule Mining on Vertically Partitioned Databases”, International Journal of Computer Applications , 2012.

[3]. Arpita B. Modh, Dept of Computer Science and Engineering, IJ, “Privacy-Preserving Outsourced Association Rule Mining on Horizontally Partitioned Databases”, International Journal for Research in Emerging Science and Technology, 2015.

[4]. Asavari G. Smart, Student, P.M. Mane, Assistant Professor, IJ, “A Survey on Privacy-Preserving Mining of Association Rule Databases”, International Journal of Science and Research, 2012.

[5]. Agrawal and Srikant, “Privacy Preserving Data mining”, Proceedings of the ACM SIGMOD International Conference on Management of data, 2000.

[6].Majid Bahir Mailk, M. Asger Gahzi, Rashid Ali, “Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects”, IEEE Third International Conference on Computer and Communication Technology, 2012.

[7].Murat Kantarcioglu and Chris Clifton, “Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data”, IEEE Transaction on Knowledge and Data Engineering, volume. 16, issue 9, 2004.

[8].M. B. Malik, M. A. Ghazi and R. Ali, “Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects”, in proceedings of Third International Conference on Computer and Communication Technology, IEEE 2012.

[9]. J. Vaidya and C. Clifton, “Privacy preserving association rule mining in vertically partitioned data”, in The Eighth ACM SIGKDD International conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, CA, July 2002, IEEE 2002.

[10].M. N. Kumbhar and R. Kharat, “Privacy Preserving Mining of Association Rules on horizontally and Vertically Partitioned Data: A Review Paper”, in proceedings of 978-1-4673-51164/12/$31.00_c, IEEE 2012.

[11]. J. Liu, J. Luo and J. Z. Huang, “Rating: Privacy Preservation for Multiple Attributes with Different Sensitivity requirements”, in proceedings of 11th IEEE International Conference on Data Mining Workshops, IEEE 2011.

[12]. Fosca Giannotti, Laks V.S.Lakshmanan, Anna Monreale, Dino Pedreschi and Hui(Wendy) Wang, “Privacy-Preserving Mining of association rules from transaction Databases” IEEE Systems Journal, 2012.

[13]. K. Sathiyapriya and Dr. G. Sudha Sadasivam, “A survey on privacy preserving association rule mining” International Journal of Data Mining & Knowledge Management Process (IJDKP), 2013.

[14]. Pankaj P. Joshi and Prof. R. M. Goudar, “Privacy preserving association rule mining in partitioned databases”, International Journal of Advanced Research in Computer and Communication Engineering, 2016.

[15]. T. Brijis, G. Swinnen, K. Vanhoof and G.Wets, “Using assiciation rules for producct assortment decisions: A case study,” in Proc. SIGKDD, 1999.

[16] P. Samarati. Protecting respondents’ identities in microdata release. In TKDE, volume 13, pages 1010–1027, 2001.

[17] C. E. Shannon. Communication theory of secrecy systems. 28:656–715, 1948.

[18] C. Tai, P. S. Yu, and M. Chen. k-support anonymity based on pseudo taxonomy for outsourcing of frequent itemset mining. In KDD, pages 473–482, 2010.

[19] W. K. Wong, David W. Cheung, Edward Hung, Ben Kao, and Nikos Mamoulis. Security in outsourcing of association rule mining. In VLDB, pages 111–122, 2007.

[20] Dakshi Agrawal and Charu C. Aggarwal, 2001.