A Survey on techniques used in Privacy Preservation for Hiding Sensitive Rules in Data Mining

Nidhi Bhatt[1] Ms. Reema Patel[2]
[1]M.E. Student [2]Assistent Professor
[1,2]Department of Computer Engineering
[1,2]Silver Oak College of Engineering & Technology, Ahmedabad

Abstract— Data Mining is a process of discovering information or knowledge from the data warehouse. In today's world, people are concerned about their privacy and secrecy. Though data mining has emerged as a significant technology, individual privacy concerns has been growing with the use of this technology. This problem is addressed with a new branch of data mining, known as privacy preserving data mining, which incorporates the mechanism for protecting sensitive information. Privacy preserving data mining deals with hiding an individual's sensitive identity without sacrificing the usability of data. Many organizations and businesses wants to secure their data from illegitimate access. In that priority is given to data privacy and security concerns. It must be necessary for them to share their information about data for the sake of getting satisfactory results. The problem is how these individuals or parties can compare their data or share it without disclosing the sensitive data to each other.

Key words*:* Privacy Preserving, Association Rule Mining, Data Mining

I. INTRODUCTION

Data Mining refers to extracting or “mining” knowledge from large amounts of data. It is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. By performing data mining, interesting knowledge, regularities, or high-level information can be extracted from database and viewed or browsed from different angles. The discovered knowledge can be applied to decision making, process control, information management, and query processing.

In today’s world privacy is the major concern to protect the sensitive data. People are very much concerned about their sensitive information which they do not want to share with anyone. Privacy plays an important role by securing and protecting the sensitive data values from being used by unauthorized access and thus it is different from any other field of data security such as data security and access control which prevents information disclosure against illegitimate means.

Privacy Preserving Data Mining(PPDM) was originally meant to extend traditional data mining techniques to work with data modified to hide sensitive information, but the major issue was how to modify data and how to recover data mining results from such modified data.The main goals of a PPDM algorithm include:(1) Preventing discovery of sensible information.(2) Being resistant to various data mining techniques.(3) Being uncompromising in access and use of nonsensitive data.(4) Being usable on large amounts of data.(5) It should have less exponential computational complexity.

Privacy preserving association rule mining is one of the most popular pattern discovery methods in the new and rapidly emerging research area of privacy preserving data mining.

II. PRIVACY PRESERVING ASSOCIATION RULE MINING

Privacy preserving association rule mining needs to prevent disclosure not only of confidential personal information from original or aggregated data, but also to prevent data mining techniques from discovering sensitive knowledge. It is known that each strong rule extract from frequent itemsets. To prevent sensitive rules being mined in the process of association rule mining, many methods are developed, all of which are based on reducing the support and confidence of rules that specify how significant they are.

Let R be a set of rules extracted from the transaction database with *min_sup* and the *min_conf*. Let SR be a set of the itemset A □ B below the *min_ sup* threshold, or decrease the confidence below the *min_conf* threshold while giving as little harm as possible to the remaining non-sensitive rules to keep the data quality as high as possible.

A. Classes Of Association Rule Hiding Algorithms:

Association rule hiding algorithm can be divided into three distinct classes, namely border-based approaches, exact approaches and heuristic approaches:

1) Border-based Approaches: These approaches consider the task of sensitive rule hiding through modification of the original borders in the lattice of the frequent and the infrequent patterns in the dataset.
2) Exact Approaches: Exact approaches are typically capable of providing superior solutions but at a high computational cost. They achieve this by formulating the sanitization process as a constraint satisfaction problem and by solving it using an integer/linear programming solver.
3) Heuristic Approach: These approaches involve efficient, fast algorithms that selectively sanitize a set of transactions from the database to hide the sensitive knowledge. Due to their efficiency and scalability, the heuristic approaches have been the focus of attention for the vast majority of researchers in the knowledge hiding field.

III. RELATED WORKS

Our main focus is hiding of sensitive or crucial data. There is a lot of work which is done in the field of preserving privacy of data mining. In literature, different authors have proposed different techniques of privacy preserving data mining.[6] Randomization method is a popular method in current privacy preserving data mining studies. It masks the values of the records by adding noise to the original data. Encryption method mainly resolves the problems that people jointly conduct mining tasks based on the private inputs they provide. These mining tasks could occur between mutual un-trusted parties, or even between competitors, therefore, protecting privacy becomes a primary concern in distributed data mining setting.

Blocking based technique which replaces known values with unknown by randomization and the problem is to guess the unknown values[‘?’] and easy to crack original value behind the unknown values[1]. Anonymization method aims at making the individual record be indistinguishable among a group records by using techniques of generalization and suppression. The representative anonymization method is k-anonymity[6].

Data Perturbation is a technique for modifying data using random process. This technique apparently distorts sensitive data values by changing them by adding, subtracting or any other mathematical formula. This technique can handle different data types: character type, Boolean type, classification type and integer[3]. In blocking based technique [3], authors state that there is a sensitive classification rule which is used for hiding sensitive data from others. In this technique, there are two steps which are used for preserving privacy. First is to identify transactions of sensitive rule and second is to replace the known values to the unknown values (?).

Another approach used is Condensation approach .It builds constrained clusters in the data set and after that produces pseudo-data. The basic concept of the method is to contract or condense the data into multiple groups of predefined size. For each group, certain statistics are maintained.

A. Hiding Association Rules by Using Confidence and Support:

Authors of the paper suggested some rules for hiding sensitivity by changing the support and the confidence of the association rule or frequent item set as data mining mainly deals with generation of association rules. In order to hide an association rule a new concept of 'not altering the support' of the sensitive item(s) has been proposed in this work.

1) Advantages:

- First advantage of proposed algorithm is that support for the sensitive item is unchanged. Instead, only the position of the sensitive itemset is changed.
- The second advantage id the proposed approach uses a different approach for modifying the database transactions so that confidence of the sensitive rules can be reduced but without changing the support of the sensitive item.

2) Disadvantages:

One of the main disadvantage of the existing approaches is that the approach is tries to hide every single rule from a given set of rules without checking if some of the rules could be pruned after modification of some transactions from the set of all transactions.

B. Privacy Preserving Clustering By Data Transformation:

Preserving the privacy of individuals when data are shared for clustering was a complex problem. The challenge was how to protect the underlying data values subjected to clustering without jeopardizing the similarity between objects under analysis. A family of geometric data transformation methods can distorts numerical attributes by scaling, rotations, translations or by the combination of all above transformations. This method was designed to specify privacy-preserving clustering, in context where data owners must meet privacy requirements as well as guarantee valid clustering results.

1) Advantages:

- The geometric data transformation methods that distorts confidential numerical attributes in order to meet privacy protection in clustering analysis.
- End usres are able to use their own tools so that the constraint for privacy has to be applied before the mining process on the data by data transformation.

2) Disadvantages:

- One major disadvantage is that the privacy preservation of individuals when data is shared for clustering is very complex.
- The protection of the underlying data values subjected to clustering without jeopardizing the similarity between objects under analysis is hard to achieve.

C. Perturbation Based Privacy Preserving Data Mining For Real World Data:

The perturbation method has been extensively studied for privacy preserving data mining. In this method, random noise from a known distribution is added to the privacy sensitive data before the data is sent to the miner for data mining. Consequently, the data miner rebuilds an approximation to the original data distribution from the perturbed data and uses the reconstructed distribution for data mining purposes. Unfortunately, recent privacy preserving data mining techniques based on perturbation do not allow the individuals to choose their desired privacy levels.

1) Advantages:

- Simple and efficient technique for building data mining models from perturbed data.
- As the distribution of the added noise is known, the data miner could rebuild the original distribution using various statistical methods and mine rebuild data.

2) Disadvantages:

- Recent privacy preserving data mining techniques based on perturbation do not allow the individuals to choose their desired privacy levels.
- As the noise is added, information loss versus preservation of privacy is always a trade off in the perturbation based approaches.

IV. Conclusion

Here in this paper a wide survey has been done on the various approaches for privacy preserving data mining and analysed different algorithms for data mining with their drawbacks. In order to find out the perfect and efficient solution for privacy preservation the following issues should be studied:

1) Privacy should be achieved with accuracy. So application of various optimizations should be deeply researched.
2) Data sanitization process should be with minimum negative impact.
3) In distributed data mining, more efficient algorithms should be developed to balance all costs such as computation cost, communication cost. Etc
4) Deployment of privacy preserving technique into practical is need to be studied.

REFERENCES

[1] Shweta Taneja, Shashank Khanna, Sugandha Tilwalia, Ankita "A Hybrid C- Tree Algorithm for Privacy Preserving Data Mining " in proceedings of International Journal of Soft Computing and Engineering (IJSCE) , ISSN: 2231-2307, Volume-4, Issue-ICCIN-2K14, March 2014

[2] Sridhar Mandapati, Dr Raveendra Babu Bhogapathi and Dr M.V.P.Chandra Sekhara Rao "Swarm Optimization Algorithm for Privacy Preserving in Data Mining" in proceedings of IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 3, March 2013

[3] Shweta Taneja, Shashank Khanna, Sugandha Tilwalia, Ankita "A Review on Privacy Preserving Data Mining :Techniques and Research Challenges " in proceedings of International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014

[4] Imran Khan, Virendra Kumar, Savita Shiwani "An Efficient Technique Privacy Preserving Association Rule Data Mining using Modified Hybrid Algorithm " in proceedings of International Journal of Science, Engineering and Technology, Volume 02, Issue 06, July 2014

[5] Praveena Priyadarsini, M.L.Valarmathi, S.Sivakumari "Hybrid Perturbation Technique using Feature Selection Method for Privacy Preservation in Data Mining " in proceedings of International Journal of Computer Applications, Volume 58– No.2, November 2012

[6] Pingshui WANG, "Survey on Privacy Preserving Data Mining" in proceedings of International Journal of Digital Content Technology and its Applications Volume 4, Number 9, December 2010

[7] Kshitij Pathak, Narendra S Chaudhari, Aruna Tiwari " Privacy Preserving Association Rule Mining by Introducing Concept of Impact Factor" in proceedings of IEEE, 2011

[8] Praveena Priyadarsini, M.L.Valarmathi, S.Sivakumari "Hybrid Perturbation Technique using Feature Selection Method for Privacy Preservation in Data Mining " in proceedings of International Journal of Computer Applications, Volume 58– No.2, November 2012

[9] Dharmendra Thakur ,Prof. Hitesh Gupta "An Exemplary Study of Privacy Preserving Association Rule Mining Techniques " in proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, November 2013

[10] Arvind Batham, Mr.Srikant Lade ,Mr. Deepak Patel "A Robust Data Preserving Technique by K-Anonymity and hiding Association Rules " in proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, January 2014