RESEARCH ARTICLE OPEN ACCESS

A Brief Survey of Privacy Preserving Data Mining

Shilpa M.S, Shalini.L
Department of Computer Science and Engineering
Mohandas College of Engineering and Technology
Anad, Trivandrum
Kerala – India

ABSTRACT
Data Mining is used for mining useful information from large databases. The enormous amount of information obtained through mining finds its relevance in many applications, but an increasing concern is about the privacy threats possessed by Data mining. The paper focuses on an emerging research topic in data mining i.e privacy and has been extensively researched in recent years. The main goal of Privacy Preserving Data Mining (PPDM) is to implement data mining algorithms efficiently without disclosing the sensitive information contained in the data. In addition a brief discussion about certain privacy preserving techniques are also presented.
Keywords:- Mining, Information Extraction, Big Data , Privacy Approaches.

I. INTRODUCTION

With the advancement of technology enormous documents become available. Almost 85% of business information lives in the form of text. The major challenge is knowledge discovery from this enormous amount of structured and unstructured data without disclosing the secret of sensitive information contained in the data. The privacy implications of data mining technologies tend to be two-fold. Mining of personal information was the major concern. Personal information includes both identifying and non-identifying information. So the sensitive raw data should not be used for mining and the mining results whose disclosure will result in privacy violation should be excluded. The evolvement of data mining has lead to serious impact on the privacy. The percentage of difficulty in addressing privacy issues with respect to data mining was increased by the following:

• The cost of data mining tools is less while its availability is high.
• Most of the data is digitized and it is impossible for the humans to manually preprocess the data.
• Aggregation of data is increased.
•The readily available nature of data mining tools to extract patterns that go beyond actual data and its ability to predict the repetitive nature of patterns.

Multiple applications make use of data warehouses as central repositories. The main concern with aggregating such personal information and mining it is that profiles of

individuals can be created using information held in separate systems located both in the commercial and government sectors. In recent years a number of ways have been proposed in such a way so as to preserve privacy. A survey on some of the techniques used for privacy preserving data mining will be discussed in this paper. The key directions in the field of data mining are as follows:

Privacy-Preserving Data Publishing: Different transformation methods associated with privacy are studied here. Randomization, k-anonymity, and l-diversity are some methods included in transformation technique. Another related issue is how. The use of perturbed data in conjunction with classical data mining methods such as association rule mining was another issue. The other problems includes studying of privacy preserving methods to find how useful ,the different meanings associated with privacy and its efficiency in different scenarios.

Modifying the data mining results to ensure privacy: The privacy of data can be compromised by the results of data such as association rule or classification rule mining . These lead to the development of privacy preserving data

mining which used certain mining algorithms to protect the confidentiality of the sensitive information. An example is Association rule hiding which uses certain heuristics algorithms to hide the sensitive association rules.

Query Auditing: These methods are similar to the previous case of modifying the results of data mining algorithms. Either we modify or restrict the results of queries. Offline auditing or Online auditing can be used.

Distributed Privacy using Cryptographic Methods: Mostly the data will be distributed across many sites and the authors of these data may require to compute the same function These scenarios make use of cryptographic protocols to communicate with various sites in order to ensure secure function computation without revealing the sensitive information.

High Dimensionality that incorporates theoretical challenges: Real data sets are usually extremely high dimensional, and thus privacy-preservation process becomes extremely difficult both from a computational and effectiveness point of view. The technique of optimal k-anonymization is NP-hard. Furthermore, the technique is not even effective with increasing dimensionality, as it may be combined wih information to reveal the identity of the actual owner of the record.

II. PRIVACY PRESERVING DATA MINING TECHNIQUES

1) Data Perturbation:

Data Perturbation is a technique for modifying data using random process. In this technique sensitive data values are distorted by changing them by adding, subtracting or any other mathematical formula.Different data types can be handled by this technique: character type, Boolean type, classification type and integer. In the case of discrete data the original data set is preprocessed. The preprocessing of data is classified into attribute coding and obtaining sets coded data set. The method of average region to disperse the continuous data is used here. Discrete formula prescribed is: A (max) - A (min)/n = length where A is continuous attribute, n is number of discrete valves, and length is the length of the discrete interval. The technique only reconstructs the distribution, and does not reconstruct the original data value.

Data distortion or data noise are different names for data perturbation. Securing the sensitive data is very critical and important data perturbation plays an important role in preserving the sensitive data. Distortion is done by applying different methods such as adding noise, data transpose matrix, by adding unknown values etc. Preserving the original data is very difficult in some perturbation approaches. Some of these are distribution based techniques. In order to overcome this problem, new algorithms were developed which were able to reconstruct the distributions. This means that for every individual problem in classification, clustering, or association rule mining, a new distribution based data mining algorithm needs to be developed. A new distribution-based data mining algorithm was developed for the classification problem.

The new method was based on singular value decomposition (SVD) and sparsified singular value distribution(SSVD) technique and having the feature of selection to reduce the feature space. Different matrices have been introduced to compare or measure the difference between original dataset and distorted dataset in this method. SSVD approach is an efficient method in keeping data utitlity, SVD also works better than other standard data distortion methods which add noise to the data so that the data becomes perturbed.

The perturbation approach has a drawback. Here the distribution of each data dimension is reconstructed independently. This implies that any distribution based data mining algorithm works under an assumption to treat each dimension independently. In many cases, relevant information for data mining algorithms are hidden in inter-attribute correlations.

2) Blocking based technique

Blocking based techniques involves a sensitive classification rule which is used for hiding sensitive data from others. The technique involvestwo steps for preserving privacy. First is to identify sensitive rule based transactions and second is to replace the known values to the unknown values (?). Here the original database is scanned and transactions supporting the sensitive rules are identified. Then for each transaction, sensitive data is replaced with unknown values. This technique can be applied to those applications in which one can save unknown values for some attributes. It hides the actual values , they replace ‘1’ by ’0’ or ‘0’ by ‘1’ or with any unknown(?) values in a particular transaction. The is not dependent on any rule. This technique enables to preserve the sensitive data from unauthorized access. According to the requirements there may be different sensitive rules. The scanning of original database is done for every sensitive rule. For a transaction to support any rule the left side of the pair of rule has to be a subset of attribute values pair of the transaction and the right hand side of the rule should be same as the attribute class. For every

transaction which supports the sensitive rule, the algorithm replaces unknown values in the place of attribute.These steps will continue till unknown values hides all the sensitive attributes.

3)Cryptographic Technique

Cryptography is an efficient technique to preseerve the data since it provides security and safety of sensitive attributes.. Here the sensitive data is encrypted. There are different cryptographic algorithms available. Inspite of this, the method has many disadvantages. It fails to protect the output of computation. It prevents privacy leakage of computation. This is not efficient when it talks about more parties. It is difficult to apply this algorithm for huge databases. Privacy of individual's record will also be affected.

4)Condensation Approach

Condensation approach is another efficient approach. Constrained clusters are built in the data set and then produces pseudo-data. The basic idea of this method is to contract or condense the data into multiple groups of predefined size. Each group maintains its own statistics. This approach finds its application in dynamic data update such as stream problems. Each group has a size of atleast 'k' and that is referred to as the level of privacy-preserving approach. The higher the level, the higher is the amount of privacy. In order to generate the corresponding pseudo-data, they use the statistics from each group. The approach is simple but it is not efficient since it leads to loss of the information.

5) Hybrid technique

Hybrid technique is a new technique in which two or more techniques are combined to preserve the data. First they randomize the data and then generalize the modified or randomized data. This technique protects private data ensuring better accuracy; also it reconstructs original data and provide data with no information loss. It can combine many other techniques such as Data perturbation, Blocking based method, Cryptographic technique, Condensation approach etc to make a hybrid technique.

III. PRIVACY PRESERVING ASSOCIATION RULE MINING

Association rule mining is an important data mining task which aims at finding interesting patterns and co-relations among data sets. It faces two problems:
When the input to the data is perturbed, determining the association rules on the perturbed data is challenging.

Another issue is that of output association rule privacy. In this case, we try to ensure that none of the association rules in the output result leads to the leakage of sensitive data. This problem is referred to as association rule hiding. Various approaches have been proposed and they can be catogorized into the following five:

- Heuristic distortion approaches, which resolve how to select the appropriate data setsso that the data can be modified.
- Heuristic blocking approaches, where the degree of support and confidence of the sensitive association rules are reduced by replacing certain attributes of some data items with a specific symbol (e.g. '?').
- Probabilistic distortion approaches, here data is distorted through random numbers generated from a predefined probability distribution function.
- Exact database distortion approaches, it formulate the solution of the hiding problem as a constraint satisfaction problem (CSP), and then apply linear programming approaches to its solution.
- Reconstruction-based approaches, here a database is generated from the scratch and that is compatible with a given set of non-sensitive association rules.

Privacy Attacks.

It is practical to examine the different ways in which one can make adversarial attacks on privacy-transformed data. This enables in designing more effective privacy-transformation methods. Some examples of methods which can be used in order to attack the privacy of the underlying data are SVD-based methods, spectral filtering methods and background knowledge attacks.

Query Auditing and Inference Control.

Querying can be performed on many private databases.This can lead to a situation where security of the results are compromised. For example,In order to narrow down the possibilities for a specific record a combination of range queries can be used. So the results over multiple queries can be combined in order to uniquely identify a record, or at least reduce the uncertainty in identifying it.

IV. PRIVACY PRESERVING DATA MINING ALGORITHMS

1)Hybrid Partial Hiding (HPH) algorithm:

Initially the algorithm reconstructs the support of itemset, and then Apriori algorithm is used to generate frequent itemsets based on which only non-sensitive rules can be

obtained. A heuristic algorithm based on the intersection lattice of frequent itemsets for hiding sensitive rules was proposed by this method. The victim item such that modifying this item causes the least impact on the set of frequent itemsets is first determined by the algorithm. Then, the minimum number of transactions that have to be modified are specified. Then,the victim item is removed from the specified transactions and sanitization of data is done.

2)Decrease Support of Right Hand Side Item of Rule Clusters(DSRRC)

For hiding sensitive association rules.The algorithm aims to hide as many as possible rules at one time so it clusters sensitive rules and hides them based on certain criteria.. A drawback of this algorithm is that it cannot hide association rules with multiple items in antecedent (left hand side) and consequent (right hand side). To overcome thisproblem, an improved algorithm named ADSRRC (Advance DSRRC) was proposed, where the item with highest count in right hand side of sensitive rules are iteratively deleted during the data sanitization process.

3)Inverse Frequent Set Mining (IFM):

It's a reconstruction based approach. A drawback of IFM is that given a collection of frequent itemsets and their supports are known, find a transactional data set such that the data set precisely agrees with the supports of the given frequent itemset collection while the supports of other itemsets would be less than the pre-determined threshold. The approach consists of three steps:

• First, generate all frequent itemsets with their supports and support counts from original data set using frequent itemset mining algorithm.

• Second, identify the itemsets that are related to sensitive association rules and remove the sensitive itemsets.

• Third, use the remaining itemsets to generate a new transactional data set using inverse frequent set mining.

The basic idea of using IFM to reconstruct sanitized data set seems good. However, solving IFM problem is difficult. It has been proved by Mielikäinen that deciding whether there is a data set compatible with the given frequent sets is NP-complete. Reducing the computational cost of searching a compatible dataset is an area which is still being researched. Some representative algorithms include the vertical database generation algorithm the linear program based algorithm, and the FP-tree-based method. Despite all these, the IFM problem does provide us some interesting insights on the privacy preserving issue. Inverse frequent set mining can be viewed as the inverse problem of frequent set mining. The dataminer can use the inverse mining algorithms to customize the data to meet the requirements for data mining results, such as the support of certain association rules, or specific distributions of data categories if the inverse problem can be clearly defined and feasible algorithms for solving the problem can be found.

V. CONCLUSIONS

In today's world,preserving the privacy is a major concern. People are very much concerned about their sensitive information and donot wish to share them. The survey in this paper focuses on the existing literature present in the field of Privacy Preserving Data Mining. From the analysis, it has been found that there is no single technique that is consistent in all domains. All methods have different efficiency in performing depending on the type of data as well as the type of application or domain. But still from the analysis, it can concluded that Cryptography and Random Data Perturbation methods perform better than the other existing methods. Cryptography is best technique since it provides encryption of sensitive data. On the other hand Data Perturbation helps to preserve data and hence sensitivity is maintained .In future, we want to propose a hybrid approach of these techniques. In this article, a brief introduction to the field of Privacy preserving data mining was given. The main aim of privacy preserving data mining is developing certain algorithms to hide or provide privacy to certain sensitive information so that they cannot be accessed by unauthorized parties or intruder Privacy and accuracy in case of data mining is a pair of ambiguity and so succeeding one can lead to adverse effect on other. In this scenario an effort was made to review a good number of existing PPDM techniques. Another data mining algorithm could be used called Fast Distributed Mining(FDM) algorithm. FDM algorithm involves generation of candidate set which is similar to that of apriori algorithm. But it uses properties of local and global frequent itemsets to generate reduced set of candidates at each iteration. After the generation of candidate sets reduction strategies are applied to eliminate some candidate set from each side Finally, it can be concluded that with FDM algorithm there is significant reduction in the number of candidate sets and message size .The algorithm needs only O(n) messages to determine whether the candidate set is frequent or not and it is so much less than other mining algorithms.

REFERENCES

[1] J. Han, M. Kamber. "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers.

[2] Charu C. Aggarwal, Philip S. Yu "Privacy-Preserving Data Mining Models and algorithm"advances in database systems 2008 Springer Science, Business Media, LLC.

[3] Vaidya, J. & Clifton, C. W, "Privacy preserving association rule mining in vertically partitioned data," In Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, Edmonton, Canada 2002.

[6] Ahmed HajYasien. Thesis on "PRESERVING PRIVACY IN ASSOCIATION RULE MINING" in the Faculty of Engineering and Information Technology Griffith University June 2007.

[7] R. Agrawal and R. Srikant. "Privacy Preserving Data Mining",ACM SIGMOD Conference on Management of Data, pp: 439-450,2000.

[8] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining", Journal of Cryptology, 15(3), pp.36-54, 2000.

[9] Aris Gkoulalas-Divanis and Vassilios S. Verikios, "An Overview of Privacy Preserving Data Mining", Published by The ACM Student Magazine, 2010.

[10] Stanley, R. M. O. and R. Z Osmar, "Towards Standardization in Privacy Preserving Data Mining", Published in Proceedings of 3rd Workshop on Data Mining Standards, WDMS' 2004, USA, p.7-17.

[6] Tjong Kim Sang, Erik F.; De Meulder, Fien (2003). "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition". CoNLL.

[7] Jenny Rose Finkel; Trond Grenager; Christopher Manning (2005).
"Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling" 43rd Annual Meeting of the Association for Computational Linguistics. pp. 363–370.

[8] McCallum, Andrew; Nigam, Kamal (1998). "A comparison of event models for Naive Bayes text classification". AAAI-98 workshop on learning for text categorization.

[9] Zhang, Harry. "The Optimality of Naive Bayes". FLAIRS2004 conference.

[10] Rish, Irina (2001). "An empirical study of the naive Bayes classifier". IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence.

[11] Rennie, J.; Shih, L.; Teevan, J.; Karger, D. (2003). "Tackling the poor assumptions of Naive Bayes classifiers". ICML

[12] Webb, G. I.; Boughton, J.; Wang, Z. (2005). "Not So Naive Bayes: Aggregating One-Dependence Estimators". *Machine Learning* (Springer) **58** (1): 5–24

[13] Moshe Koppel; Jonathan Schler (2006). "The Importance of Neutral Examples for Learning Sentiment". *Computational Intelligence 22*. pp. 100–109

[14] "Thumbs up? Sentiment Classification using Machine Learning Techniques" *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 79–86.

[15] Bo Pang; Lillian Lee (2004). "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts" *Proceedings of the Association for Computational Linguistics (ACL)*. pp. 271–278.

[16] Ogneva, M. "How Companies Can Use Sentiment Analysis to Improve Their Business". Retrieved 2012-12-13.

[17] Erik Cambria; Catherine Havasi, and Amir Hussain (2012). "SenticNet 2: A semantic and affective resource for opinion mining and sentiment analysis". *Proceedings of AAAI FLAIRS*. pp. 202–207

[18] Rajaraman, A.; Ullman, J. D. (2011). "Data Mining"*Mining of Massive Datasets*. pp. 1–17

[19] Jones KS (1972). "A statistical interpretation of term specificity and its application in retrieval" *Journal of Documentation***28** (1): 11–21