Scientific Journal Impact Factor (SJIF) : 3.518
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745 ISSN (PRINT) : 2393 - 8161

A Survey on Different Issues of Privacy Preserving Mining and Techniques

Umesh Kumar [1],Diamond Jonawal[2]
[1]M.Tech (CTA) Research Scholar, [2]RGTU University, Bhopal, India

Abstract— as data mining is used to extract valuable information from large amount of data. But this is harmful in some cases so some kind of protection is required for sensitive information. So privacy preserving mining has emerged with the goal to provide protection from mining.There are many research branches in this area. This paper focus on analyzing different techniques of privacy persevering and specify there requirement for special type of cases. Basic concepts related of association rule mining, analyzed, summarized the general methods and techniques of privacy preserving.Then paper discusses some of the problem in different approaches.
Keywords:-Privacy Preserving Mining, Association Rule Mining, Data Perturbation, Aggregation, Data Swapping.

I. INTRODUCTION

Data mining concept is the process of extracting valuable information from large databases. Data mining concept is the process of discovering new and valuable patterns from large set of datasets which provides advantages for research in space research programs, marketing analysis, medical diagnosis, atmospheric forecast etc. Most of the time data mining is under attack by privacy advocates. The reason behind is that there is misunderstanding about what it actually is and how it is done. This has raised concerns that our personal data may be used for various intrusive or malicious purposes.
Privacy preserving data mining help us in fulfilling data mining goals without sacrificing the privacy of the users and without learning underlying data values. Association rule mining is one of the techniques used in data mining that finds the consistencies in large volume of data. With the help of Association rule mining which is one of the techniques of data mining we can find and reveal hidden information which is private for a user or organization. Use of association rule for privacy preserving refers to the area of data mining that protects sensitive information from unsolicited or unsanctioned disclosure.

In our research, protecting sensitive information encompasses two important goals- knowledge protection and privacy preservation. Knowledge protection is related to privacy preserving association rule mining, while the privacy preservation refers to privacy-preserving clustering. An interesting fact between knowledge protection and privacy preservation is that they have a common characteristic. For example, in knowledge protection, a company or an organization is the owner of the data due to which it should protect the sensitive knowledge identified from such data, while in privacy preservation individuals users are the owner of their personal information.

Besides this, knowledge protection and privacy preservation have a unique characteristic. Knowledge protection is concerned for the protection of implicit data, i.e., patterns discovered from the data while privacy preservation is related to the protection of explicit data (e.g., salary, address,pan number). One of the constraints with the approach of knowledge protection is that the sensitive knowledge should be known in advance by the data owners. In such case, owners of data have to mine their databases and use interesting measures (e.g., support and confidence) with the desire of finding the

important patterns, i.e, the sensitive knowledge. Accordingly, data owners hide the sensitive knowledge by using the algorithms. The database released after this is then shared for mining. Another drawback of the approach of knowledge protection is that we do not concentrate on protecting against correlations between variables, for example age and salary. Rather, we are concerned to protect specific binary rules (e.g., X->Y), where X and Y represent items purchased in a store or attributes with specific values. But, these rules are private to the company or organization who owns the data and it must be protected because it can provide competitive advantage in the business world.

II. PRIVACY PROTOCOL

There are mainly three protocols for building a privacy-preserving data mining system. These three protocols are shown below [2].

- *Data collection:* It controls privacy during data transmission between the data providers and data ware-house server.
- *Inference control:* It protects privacy between the data warehouse server and data mining servers.
- *Information Sharing:* It controls on information shared among the data mining servers in different systems.

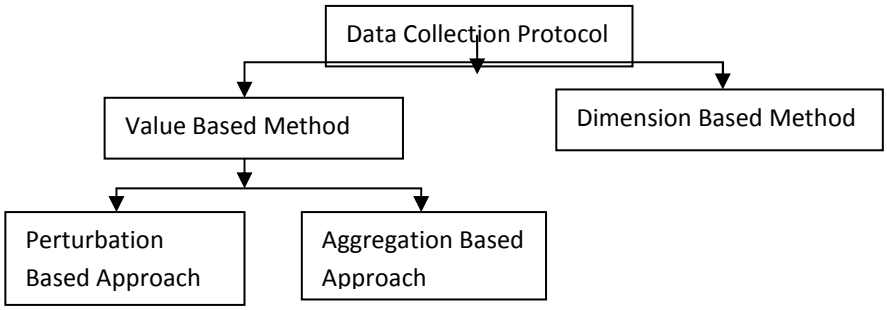

Fig 1: Data Collection Protocol Taxonomy

The main objective of these protocols is to output minimum private information with accuracy for data mining from one entity to another to build accurate data mining models.

2.1 DATA COLLECTION PROTOCOL

The data collection protocol, provide the minimum private information to build accurate data mining models and ensures that they send only that part of the information to the data warehouse server.

Basic requirements for the data collection protocol; First, it must be scalable; because a data warehouse server can deal with thousands of data providers like online survey system. Second, the computational costing to data providers must be less because and a higher cost could discourage them from participating in data mining. Lastly, the protocol must be sturdy; it must deliver relatively accurate data mining results while protecting data provider's privacy, even if data providers have lacking consistency. For example, if some data providers in an online survey system deviate from the protocol or submit meaningless data, then it must be control the influence of such erroneous behavior and provide surety that global data mining results remain sufficiently accurate.Fig1shows data collection protocol taxonomy based on two data collection methods.

2.2 VALUE-BASED METHOD

With the value-based method, a data provider manipulates the value of each data attribute or item independently using one of two approaches. The *perturbation-based* approachadds noise directly to the original data values, such as changing age 25 to 35 or Texas to London. The *aggregation-based* approach generalizes data according to the relevant domain hierarchy, such as changing age 27 to age range 25-30 or Texas to the UK.

The perturbation-based approach is recommended for random data, while the aggregation-based approach depend on knowledge of the domain hierarchy[2], but can be effective in guaranteeing the data's anonymity k-anonymity, means that each perturbed data record is identical from the perturbed values of at least k-1 other data record.

The value based method considers that it would be difficult, but not impossible, for the data warehouse server to rediscover the original private data from the changed values but that the server would still be able to recover the original data distribution from the perturbed data. So easily construct the accurate data mining models.

2.3 DIMENSION-BASED METHOD

With the dimension based method data to be mined usually has many attributes or dimension. It removes the private information from the original data by reducing the numbers of dimensions. This method could result in information loss. So preventing data mining servers from constructing accurate data mining models.

III. PRIVACY PRESERVING TECHNIQUES

The concern of public is mainly caused by the so-called secondary use of personal or private information without the consent of the subject. In other words, users feel strongly that their personal information should not be sold to other organizations without their prior consent. The majority of respondents in society are concerned about the possible misuse of their personal information. Also shows that, when it comes to the confidence that their personal information is properly handled, consumers have most trust in banks and health care providers and the least trust in credit card agencies and internet companies. So we can classify privacy preserving techniques based on the protection methods used by them.

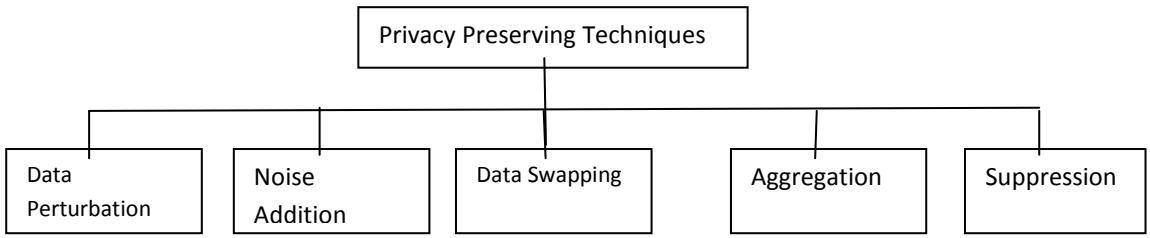

Fig 2: Different Privacy preserving Techniques

3.1 Data Perturbation

It is a kind of data modification technique that protects the sensitive data contained in a dataset by changing carefully selected part of attribute values pairs of its transactions. When the modification is done, the released values are inaccurate, thus protecting the sensitive data. Also achieving preservation of the statistical properties of the dataset. The perturbation method used should be such that data computed on the perturbed dataset do not differ significantly from the statistics that would be obtained on the original dataset. Mainly two categories for data perturbation namely probability distribution approach and the value distortion approach. The approach of probability distribution, replaces the data with another specimen from the same (estimated) distribution or by the distribution itself. The approach of value distortion disturbs the values of attributes or data elements directly by some additive or multiplicative noise before it is released to the data miner.

3.2 Noise Addition

Noise addition methods were originally used for statistical databases which were supposed to maintain data quality in parallel to the privacy of individuals. Noise addition techniques were also found useful in privacy preserving data mining.

The underlying dispersion of a perturbed data set can be uncertain if the dispersion of the corresponding original data set and/or the dispersion of the added noise is not multivariate normal. In such a case responses to queries involving percentiles, sums, conditional means etc. Some noise addition techniques, Probabilistic Perturbation Technique (PPT) ,Random Perturbation Technique (RPT), and All Leaves Probabilistic Perturbation Technique (ALPT).

3.3 Data Swapping

Data swapping techniques mainly appeal of the method was it keeps all original values in the data set, at the same time the record re-identification is very difficult. Data swapping means replaces the original data set by another one. Here some original values belonging to a sensitive attribute are exchanged between them.

This swapping can be done in a way so that the t-order statistics of the original data set are preserved. A t-order statistic is a statistic that can be generated from exactly t attributes. A new concept called approximate data swap was recommended for practical data swapping. It computes the t-order frequency table from the original data set, and finds a new data set with approximately the same t-order frequency.

The elements of the new data set are generated one at a time from a probability distribution constructed through the frequency table. The frequency of already created elements and a possible new element is used in the construction of the probability distribution. Inspired by existing data swapping techniques used for statistical databases a new data swapping technique has been introduced for privacy preserving data mining, where the requirement of preserving t-order statistics has been relaxed.

The technique emphasizes the pattern preservation instead of obtaining unbiased statistical parameters. It preserves the most classification rules and also obtained different classification algorithms. The noise is added to the class, means the target attribute of a classifier is modified, instead of all other attributes in the data set. As the class is normally a categorical attribute containing just two different values, the noise is added by changing the class in a small number of records. It can be achieved by randomly shuffling the class attributes values belonging to heterogeneous leaves of a decision tree.

3.4 Aggregation

Aggregation is also called as generalization or global re coding. It gives protection of individual privacy in a released data set by changing the original data set prior to its release. Aggregation replaces k number of records of a data set by a representative record.

The attribute value in a representative record is generally derived by taking the average of all attributes values, which belongs to the records that are replaced. Due to the replacement of k number of original records by a representative record aggregation results in some information loss.

The loss can be minimized by clustering the original records into mutually exclusive groups of k records prior to aggregation. This loss results in a higher disclosure risk since an intruder can make a better estimate of an original value from the attribute value of the released record.

The cluster size means the number of records in each cluster can produce an appropriate balance of information loss and disclosure risk can be adjusted. Another method of aggregation or generalization is transformation of attribute values. For example, an exact date of birth can be replaced by the year of birth; an exact salary can be replaced rounded in thousands.

Such a generalization makes an attribute values less informative. Therefore, a use of excessive extent of generalization can make the released data useless. For example, if an exact date of birth is replaced by the century of birth then the released data can become useless to data miners.

3.5 Suppression

In suppression method, sensitive or private data values are deleted or suppressed prior to the release of a data. This method is used to protect an individual or user privacy from intruder's attempts to accurately

predict a suppressed value. A sensitive value is predicted by an intruder through various approaches. For example, a built classifier on a released data set can be used to try to predict a suppressed attribute value. Therefore sufficient number of attribute values should be curbed in order to protect privacy. However, suppression of attribute values results in information loss. An important point in suppression is to reduce the information loss by minimizing the number of values suppressed. For some applications like a medical diagnosis the suppression is preferred over noise addition in order to reduce the chance of having misleading patterns in the perturbed data set.

IV. PRIVACY BY ASSOCIATION RULE

A set of items $I = \{ I_1, I_2 , \ , I_m \}$. Transaction of database be a D where each transaction T is a set of items such that $T \subseteq I$. Each transaction is correlated to an identifier, call TID. A transaction T is said to contain A if and only if $A \subseteq T$. An association law is an implication of the form $A \Rightarrow B$, where $A \subseteq I$, $B \subseteq I$, and $A \cap B = \Phi$. The rule $A \Rightarrow B$ stores in the transaction set D with support, 's' ,where s indicate percentage of transactions in D that contain $A \cup B$. The law $A \Rightarrow B$ has confidence, 'c' in the transaction set D . That is,

$$\text{sup}(A \Rightarrow B) = P(A \cup B) = \frac{|A \cup B|}{|D|} \quad (1)$$

$$\text{conf}(A \Rightarrow B) = P(B | A) = \frac{|A \cup B|}{|A|} \quad (2)$$

Where |A| is called as the support count of the set of items A in the set of transactions D , as denoted by sup_count(A) . A appears in a transaction T, if and only if $A \subseteq T$. Laws that satisfy both a minimum support threshold (min_ sup) and a minimum confidence threshold (min_conf) are called strong. A set of items referred to as an itemset. k -itemset contains k items in that itemset. Itemsets that satisfy min_ sup is named as frequent itemsets. All strong association rules result from frequent itemsets.

By specifying the minimum confidence and support specific items from the dataset can be hide. This can be done by removing or replacing the items from the set then check the minimum support and confidence of that item. In this way by association rule one can implement privacy preserving.

V. PRIVACY PRESERVING DATA MINING

The main goal in most distributed methods for privacy preserving data mining (PPDM) is to allow computation of useful aggregate data over the entire data set without compromising the privacy of the individual data sets within the different participants.

The individual or user may wish to participate in obtaining aggregate results, but may not have full trust on each other in terms of the sharing of their own data sets. The data sets may either be horizontally partitioned or be vertically partitioned for data mining.

The individual or user records are spread out across multiple entities. In horizontally partition, have the same set of attributes in data sets. The individual entities may have different attributes of the same set of

records in vertical partitioning. Both kinds of partitioning have different challenges to the problem of distributed privacy-preserving data mining.

5.1 Algorithms over horizontally partitioned data sets

In horizontally partitioned data sets, a different set of records with the same set of attributes which are used for mining purposes. Horizontally partition your database by splitting a table into complete datasets and placing those datasets into other databases. A field is used to create that separation, such as an ID field, location, age and so on.
We can use approximations of the best splitting attributes. Afterwards, a variety of classifiers have been generalized to the problem of horizontally partitioned privacy preserving mining including the Naïve Bayes Classifier and the SVM Classifier with nonlinear kernels.
A horizontally partitioned case is discussed, in which privacy preserving classification is performed in a fully distributed setting, where every individual have private access to only their own record. A host of other data mining applications have been discovered to the problem of horizontally partitioned data sets. Many applications of data mining can be perform i.e. clustering, filtering and association rule mining.

5.2Algorithms over vertical partitioned data sets

The vertically partitioned have many primitive operations such as computing the scalar product or the secure set size intersection can be useful in calculating the results of data mining algorithms. For example, the methods in discuss how to use to scalar dot product computation for frequent item set counting. The counting process can be achieved by using the secure size of set intersection.
Another method for association rule mining uses the secure scalar product over the vertical bit representation of item set inclusion in transactions, in order to compute the corresponding item sets. This key step is applied repeatedly within the framework of a roll up procedure of item set counting. It has been shown that this approach is quite effective in practice.
Vertically partitioned data is used to perform linear regressions without sharing their data values. The approach of vertically partitioned can be extended to a variety of data mining applications i.e. k means clustering, decision trees, SVM Classification and Naïve Bayes Classifier.

VI. PROBLEM DEFINATION

The main problem with these methods is that they can be regenerated by distortion where Y is perturbed and x is original set. This is done in the case of the numeric set of values.

$$D(x,y) = 1/N(\sum E(y-x)\ ^2)$$

without knowledge of whole method and parameter one can predict approx dataset which is much closer to the original set. One more method [10] Linear Least Squares Error Estimation method

$$X(y) = Kx / Ky((y - \mu) + \mu)$$

Where Kx and Ky are covariance of x & y while μ is mean of x.
In the same fashion if more then one perturbed copy is use for generating the original copy then by finding the pattern between then it is possible to generate.
One more precaution that one has to take that if the attacker has the prior knowledge of the data then chance of regeneration increases accordingly. This is also known as linkages [11] for having the prior knowledge. In order to restrict this k-anonymity has been proposed but still it lacks in many cases. So

above methods are use to find how accurate that algorithm is in terms of the privacy concern, this can be understand as the prediction of original dataset or hiding of original set is the primary concern of the set which one can be evaluate on above methods.

In [12] instead of perturbing anonymization of the numeric data is done. By giving some kind of range to the field's researcher reduce value regeneration. But by this only numeric data gets privacy. So perturbation or suppressing the rule generation from perturbed copy is required. As textual data provide information in form of rules which should need to be suppress.

In [13] privacy of image and numeric type of data is done where same approach for different for privacy is required. Here encryption of the work is done by Paillers method. In this work efforts are required for the same at client and server end. Here also textual data privacy is not done by researcher, so this work is still remaining.

VII. CONCLUSION

As the data mining is a vast field for many researchers out of this privacy preserving mining is the important field of interest for them. As there are many method making the privacy of the dataset but perturbing both the text and numeric data with the single algorithm is not so protected, although the individuals for perturbing either text or numeric is highly protected. So a secure method should be developing for perturbation is required for suppressing all kind of fruitful information.

REFERENCES

[1] FoscaGiannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang, "Privacy-Preserving Mining of Association Rules from Outsourced Transaction Databases" *In IEEESystems Journal, VOL. 7, NO. 3, SEPTEMBER 2013*, pp. 385-395.

[2] N. Zhang and W. Zhao, "Privacy Preserving Data Mining Systems " In IEEE Computer society, 2007 pp. 52-58.

[3] 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, pp. 111–122.

[4] K.Sathiyapriya and Dr. G.SudhaSadasivam, " A Survey on Privacy Preserving Association Rule Mining", In IJKDP Vol.3 No 2– March-2013, pp 119-131.

[5] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proc.ACM SIGMOD Int. Conf. Manage. Data*, 2000, pp. 439–450.

[6] ManopPhankokkruad, " Association Rules for Data Mining in Item Classification Algorithm : Web Service Approach ", In IEEE, 2012 pp. 463-468.

[7] D.Narmadha, G.NaveenSundar and S.Geetha,"*A Novel Approach to Prune Mined Association Rules in Large Databases*‖", IEEE, 2011 pp.

[8] T zung -Pei, Hong Kuo-Tung Yang, Chun-Wei Lin and Shyue-Liang Wang, "Evolutionary privacy preserving in data mining ",In IEEE World Automation Congress conference , 2010 pp.

[9] Z. Yang and R. N. Wright. "Privacy-preserving computation of bayesian networks on vertically partitioned data." In IEEE Trans. on Knowledge and Data Engineering , 2006, pp.1253–1264.

[10] Enabling Multilevel Trust in Privacy Preserving Data Mining Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL.24, NO. 9, SEPTEMBER 2012.

[11] Survey on Privacy Preserving Data Mining Haitao Liu and Jing Ge University of Illinois at Urbana-Champaign, Urbana IL, 61801 USA.

[12]Privacy-Preserving Data Publishing for Multiple Numerical Sensitive AttributesQinghai Liu, Hong Shen_, and Yingpeng Sang. TSINGHUA SCIENCE AND TECHNOLOGY Volume 20, Number 3, June 2015.

[13] Privacy-Preserving Multi-Class Support Vector Machine for Outsourcing the Data Classification in Cloud YogachandranRahulamathavan, Member, IEEE, Raphael C.-W. Phan, Suresh Veluru, KanapathippillaiCumanan, Member, IEEE, and MuttukrishnanRajarajan, Senior Member, IEEE. IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 11, NO. 5, SEPTEMBER/OCTOBER 2014.