Privacy Preserving Mining Techniques and Precaution for Data Perturbation

J.P. Maurya[1], Sandeep Kumar[2], Sushanki Nikhade[3]
[1] *Asst .Prof. C.S Dept., IES BHOPAL*
[2]*Mtech Scholar, IES BHOPAL*
[1]jpeemaurya@gmail.com , [2]kumar.sandeep@msn.com, [3]sush.nfs@gmail.com

Abstract - **Most distributed methods for privacy preserving data mining is to allow computation of useful aggregate statistics over the entire data set without compromising without compromising the privacy of the individual data sets within the different participants. The enhancement of data mining research will be the development of techniques that incorporate privacy concerns. The goal of privacy preserving, is to mine the potential valuable knowledge without leakage of sensitive records, in other words, use accurate data with privacy. The problem of outsourcing the association rule mining task within a corporate privacy-preserving framework. This paper focus on concepts related to association rule mining analyzed and summarized the general methods and techniques of privacy preserving association rule mining.**

Keywords:-Privacy Preserving Mining, Association Rule Mining, Data Perturbation, Aggregation, Data Swapping.

I. INTRODUCTION

DATA mining is to extract information from large databases. Data mining is the process of discovering new patterns from large data sets which gives advantages for research, marketing analysis, medical diagnosis, atmosphere forecast etc. Data mining is under attack from privacy advocates because of a misunderstanding about what it actually is and a valid concern about how it's generally done. This has caused concerns that personal data may be used for a variety of intrusive or malicious purposes. Privacy preserving data mining help to achieve data mining goals without scarifying the privacy of the individuals and without learning underlying data values. Association rule mining is a technique in data mining that identifies the regularities found in large volume of data. Due to this technique identify and reveal hidden information that is private for an individual or organization. Privacy-preserving data mining using association rule refers to the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure.

Protecting sensitive information in the context of our research encompasses two important goals: knowledge protection and privacy preservation. The former is related to privacy preserving association rule mining, while the latter refers to privacy-preserving clustering. An interesting aspect between knowledge protection and privacy preservation is that they have a common characteristic. For instance, in knowledge protection, an organization is the owner of the data so it must protect the sensitive knowledge discovered from such data, while in privacy preservation individuals are the owner of their personal information [2,3].

On the other hand, knowledge protection and privacy preservation also have a unique characteristic. Privacy preservation is related to the protection of explicit data (e.g., salary), while knowledge protection is concerned with the protection of implicit data, i.e., patterns discovered from the data. One limitation with the approach of knowledge protection is that the sensitive knowledge should be known in advance by the data owners. In this case, data owners have to mine their databases and use interestingness measures (e.g., support and con_dence) with the purpose of _ending the valuable patterns, i.e, the sensitive knowledge. Subsequently, data owners hide the sensitive knowledge by using the algorithms. The released database is then shared for mining. Another limitation of the approach of knowledge protection is that we do not focus on protecting against correlations between variables, such as salary and age. Rather, we protect speci_c binary rules (e.g., X->Y), where X and Y represent items purchased in a store or attributes with speci_c values. Again, these rules are private to the company or organization owning the data and must be protected since they can provide competitive advantage in the business world.

II. PRIVACY PROTOCOL

Mainly three protocols govern privacy for building a privacy-preserving data mining system. The three protocols entities are shown below[2].

- *Data collection,* manages privacy during data transmission between the data providers to the data ware-house server.
- *Inference control,* protects privacy between the data warehouse server and data mining servers.
- *Information sharing, gives the* control on information shared among the data mining servers in different systems.

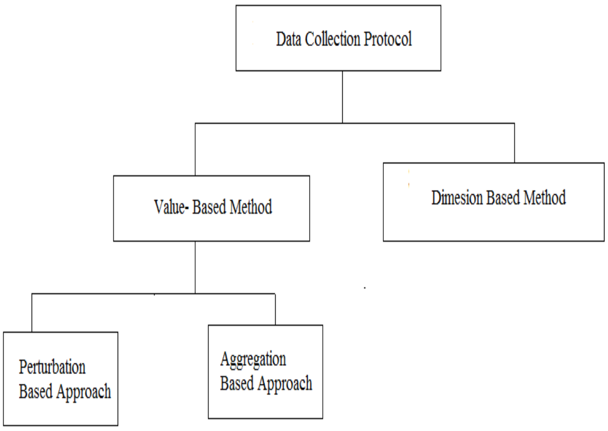

Figure: 1 Data Collection Protocol Taxonomy

An aim of these protocols is to produce minimum private information with accuracy for data mining from one entity to another to build accurate data mining models.

A. *Data Collection Protocol*

The data collection protocol, provide the minimum private information to build accurate data mining models and ensures that they transmit 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 robust; 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 ensure that global data mining results remain sufficiently accurate. Figure-1 shows data collection protocol taxonomy based on two data collection methods.

B. *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* approach adds 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 a*pproach 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 indistinguishable from the perturbed values of at least *k*-1 other data record.

The value-based method assumes that it would be difficult, but not impossible, for the data warehouse server to rediscover the original private data from the manipulated 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.

C. *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 [9]. This method could result in information loss. So preventing data mining servers from constructing accurate data mining models.

III. PRIVACY PRESERVING TECHNIQUES

Public concern is mainly caused by the so-called secondary use of personal 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 health care providers and banks and the least trust in credit card agencies and internet companies. Privacy preserving techniques can be classified based on the protection methods used by them.

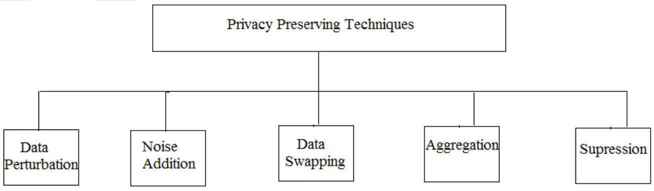

Figure 2: Different Privacy preserving Techniques

A. *Data Perturbation*

It is a category of data modification approaches that protect the sensitive data contained in a dataset by modifying a carefully selected portion of attribute-values pairs of its transactions. When the modification done the released values inaccurate, thus protecting the sensitive data. Also achieving preservation of the statistical properties of the dataset. The perturbation method used should be such that statistics 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 sample from the same (estimated) distribution or by the distribution itself. The approach of value distortion perturbs the values of data elements or attributes directly by some additive or multiplicative noise before it is released to the data miner.

B. Noise Addition

Noise addition techniques 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 distributions of a perturbed data set can be unpredictable if the distributions of the corresponding original data set and/or the distributions 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, Random Perturbation Technique (RPT), Probabilistic Perturbation Technique (PPT) and All Leaves Probabilistic Perturbation Technique (ALPT).

C. 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 introduced 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 [6]. 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 [1]. 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 typically 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.

D. Aggregation

Aggregation is also known as generalization or global recoding. It gives protection of individual privacy in a released data set by perturbing 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 [8]. The cluster size means the number of records in each cluster can produce an appropriate balance of information loss and disclosure risk can be adjust. 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.

E. Suppression

In suppression technique, sensitive data values are deleted or suppressed prior to the release of a data. This technique is used to protect an individual 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 in an attempt to predict a suppressed attribute value. Therefore sufficient number of attribute values should be suppressed in order to protect privacy. However, suppression of attribute values results in information loss [4, 5]. An important issue in suppression is to minimize 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 associated to an identifier, call TID. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subseteq I$, $B \subseteq I$, and $A \cap B = \Phi$. The rule $A \Rightarrow B$ holds in the transaction set D with support, ‘s’ ,where s indicate percentage of transactions in D that contain $A \cup B$. The rule $A \Rightarrow B$ has confidence, ‘c’ in the transaction set D . That is,

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

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

Where |A| is named as the support count of the set of items A in the set of transactions D , as denoted by sup_count(A) . A occurs in a transaction T, if and only if A⊆T . Rules 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 [1, 6]. This can be done by removing or replacing the items from the set then check the minimum support and confidendce of that item. In this way by association rule one can implement privacy preserving.

V. Privacy Preserving Data Mining

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

VI. PROBLEM DEFINATION

The main problem with these method 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.

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

without knowledge of whole method and parameter one can predict approx dataset which is much closer to the original set. One more methos is [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 similar 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.

Data Reconstruction: In order to provide the privacy from the unauthorized users some time it use of privacy preserving techniques but not all the techniques are reversible, so in order to reconstruct the perturbed data one has to choose such method that not only protect the data but also give original dataset. This kind of requirement is develop when firms are storing there data at different servers. In [11] steps of resolving this problem is explained by using association rule method.

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. There are many types of attacks which can be apply to the privacy preserving algorithms for evaluating that the algorithm is how much resist against that attack. One new approach of generating the perturbed data then regenerate it back from the perturbed is emerging which is higly vulnerable for protection concern need to be expand. So a secure method should be develop for perturbation and de-perturbation need to be devepol that only data owner will get the original dataset.

REFERENCES

[1] Fosca Giannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang, "Privacy-Preserving Mining of Association Rules from Outsourced Transaction Databases*" In IEEE Systems 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.Sudha Sadasivam, " 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. Manop Phankokkruad, " Association Rules for Data Mining in Item Classification Algorithm : Web Service Approach ", In IEEE, 2012 pp. 463-468.

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

[7] 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.

[8] 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

[9] 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.

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