ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 6, Issue 1, January 2017

A Survey on: Techniques of Privacy Preserving Mining Implementation

Priya Gupta, Sini Shibu

***Abstract*— With the increase in the data mining algorithm knowledge extraction from the large data is getting easy. But at the same time this lead to new problem of Privacy of the knowledge from the stored data at various servers. So it is required to provide privacy of the sensitive data from the data miners. This paper focus on various approaches implement by the miners for preserving of information at individual level, class level, etc. A detail description with limitation of different techniques of privacy preserving is explained. This paper explain different evaluation parameters for the analysis of the preserved dataset.**

***Index Terms*— Privacy Preserving Mining, Association Rule Mining, Data Perturbation, Aggregation, Data Swapping.**

I. INTRODUCTION

As the data miners are gathering information from the large dataset base on useful patterns, trends, etc. This is useful for helping crime understanding, any kind of terrorist activity can also be learn by the data mining approach. With this bright side of the Data mining if miners opt to find information about individual then it lead to harm the privacy of the person, place, class. So it is required to provide privacy from such kind of miners activity by applying privacy preserving mining techniques.

This is very useful for the security of data which contain some kind of medical information about the individual, financial information of family or any class. As this make some changes on the dataset, so present information in the dataset get modify and make it general for all class or rearrange so that miner not reach to concern person.

So privacy preserving mining consist of many approaches for preserving the information at various level form the individual to the class of items [3, 4]. But vision is to find the information from the dataset by observing repeated pattern present in the fields or data which can provide information of the individual, then perturb it by different methods such as suppression, association rules, swapping, etc.

Mostly when data is place on the server then miner can get the access of the whole information, so many researchers are working for the access of the data. If data is successfully achieved then it is possible for miner to get all kind of information present in it. Considering this problem people are working for providing security against large number of privacy attacks. Here before placing the data on the puplic server it get perturb so that unfavourable information or negative data is suppress. This lead to put same data with some modification on the server and it will not affect the overall privacy [5]. So it is hard to required that protection of data is done in prior steps by hiding important information like name of person, address, mobile number, date of birth, etc. But this kind of protection is not sufficient for many cases where data mining algorithm is apply as it directly or indirectly fetch information from the raw data. Although utilization of same for the ethical purpose is very helpful in all the data privacy measure. So data mining implies on data where terrorism activity can be involve. In order to understand thought from the literature data mining can be apply and if those thought lead to unfair activity then privacy of those information is done by privacy preserving mining.

Manuscript received Feb, 2013.

***Priya Gupta**, Computer Science & Engineering, NRI, Bhopal, India, 8839498814.*

***P. Sini Shibu** Computer Science & Engineering, NRI, Bhopal, India.*

II. PRIVACY PRESERVING TECHNIQUES

Use of personal data for any activity without any information to the concern is term as Public concern. In order to understand this thing consider a user never want to share there personal information with other without his permission, but it is done by the information holding organization, then this is called as public concern. Mostly it is divide into few category such as bank, health care departments are most trustful for customer information privacy. While in case of credit card company, some of social site they are least trust companies. Privacy preserving techniques can be classified based on the protection methods used by them.

Data Perturbation

Data is directly modified in this technique so it come under data modification category. It is a category of data modification approaches that protect the sensitive data from intruders. Here selected portion of the dataset is consider as the sensitive information which need to be hide by modifying those values or information. So released data is contain inaccurate data where sensitive information is modify. While doing modification it is required to do perturbation in the information having same statistics, as different values get directly act as outliers. So perturbation divide into two main category first is probability distribution approach and the other is value distortion approach. The approach of probability distribution, replaces the data with same data from the distribution of value present in original.

Pros: In this technique one replacement of the data make direct perturbation in the dataset.
Cons: Choosing of same value for replace can be easily be identified by the miner. Replacing different value tends to lead the selection difficult.

Noise Addition

This technique is apply on numeric data only as noise can be produce by some noise producing function such as Gaussian function. Here data quality is maintain by the technique so it look like original, while privacy of the information is maintain. [8].

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, Probabilistic Perturbation Technique, Random Perturbation Technique, All Leaves Probabilistic Perturbation Technique.

Pros: Addition of noise is simple as base on some kind random function such as Gaussian.
Cons: This can only perturb the numeric attributes while textual data remain same.

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 [1]. 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. 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 [2]. It preserves the most classification rules and also obtained different classification algorithms. 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.

Pros: Here originality of the dataset is maximum. This is better approach to increase the frequency of the non frequent patterns.
Cons: Here Generation of paaterns is time taken then t-order is required to be accurate to hide sensitive data.

Aggregation

Generalization of the available data is also term as aggregation. As this provide privacy of individual information before the release by replacing a group of information with a single. In other words aggregation replace k number of session by its representative session. Such as attribute value in the dataset is derived by taking the average of the bunch of same attribute information. Now this raise one new issue where replacement of k number of original records by a aggregated one make information loss. So in order to reduce this loss of information clustering of the records is required where size of cluster get reduce for decreasing the information loss [3]. But by doing this intruder can estimate of the balance information loss in the data. So overall risk of the of the disclosed data is remain same. 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.

Pros: It perfectly hide the detail information of the individual. Generating information from the common data is not possible.
Cons: Here Aggregation is done but it tend to make a information loss. This loss of information make that dataset useless.

Suppression

In suppression technique, sensitive data values are deleted or suppressed prior to the release of a data [4]. 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. 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.

Pros: Here sensitive information is completely hide from the dataset. Intruder has no chance for any kind of mining.

Cons: In this method removing of attribute tends to high information loss and this kind of method is rarely use for perturbation.

III. RELATED WORK

In [6] k-anonymity technique is use as it give direct protection for the individual before releasing the data. This can be understand as let a person having salary then that is replace by the range of salary from ten thousand to twenty thousand. In the similar fasion age of a person is replace by range. So by this overall confusion of the data is increase while rest of value remain same. So they simply give range to the age, income. Let age = 24 then its range is 20-30. Then this paper find hidden information from the data with the help of Association Frequent rules. As for finding the pattern of purchasing of item from the transaction frequent pattern need to be generate with the help of association rules. A-prior algorithm is use for the generation of rules, where by specifying the minimum threshold set of items are select. This can be understand by Let D is dataset and I is number of items in the dataset transaction. Now let rule be Ix→ Iy then generate confidence and support of the rule. As support and confidence are the two important measure for the generation of association rules.
Support can be calculated by

Support (Ix→ Iy) =(Ix U Iy) / N
Where Ix U Iy is number of transaction where both item Ix, Iy is present and N is total number of transaction present in the dataset.

While confidence is calculated by

Confidence (Ix→ Iy) = (Ix U Iy) / X

Similarly Ix U Iy is number of transaction where both item Ix, Iy is present and X is total number of transaction where item Ix is present in the dataset.

Finally hiding of sensitive data items. Original dataset is consist of the collection of transaction that produce fruitful knowledge, now one need to hide or remove that knowledge from that, this is the main goal of this work. For removing association rule one has to reduce the support or confidence of the specified rule below minimum values. To decrease the confidence of a rule, there is two approach:

(1) Increase the support of X, the left hand side of the rule, but not support of X → Y.

(2) Reduce the support of the rule set X →Y. In this second case, if we only decrease the support of Y, the right hand side of the rule, it would reduce the confidence faster than simply reducing the support of X → Y.

Issues: Here it only reduce the RHS item Y of the rule correspondingly. So for the rule Bread→ Milk can generate reduce the support of Y only. Now it need to find that for how many transaction this need to be done. So calculation of that number is done by

((Rule_confidence – Minimum_confidence) * X_support* Total_transaction)
Above formula specify the number of transaction where one can modify and overall support of that hiding rule is lower then the minimum confidence.

In [8] multilevel privacy is provide by the author, basic concept develop in this paper is separate perturbed copy of the dataset for different user. Here user are divide into there trust level so base on the trust level dataset is perturbation percentage get increase. Here paper resolve one issue of databse reconstruction by combing the different level perturbed copy then regenerate into single original database. So to overcome this problem perturbation of next level is done in perturbed copy of previous one. In this way if lower trust user get combine and try to regenerate original dataset then only one higer perturbed copy can be regenerate. The distribution of the entries in such a matrix looks like corner-waves originated from the lower right corner.

Issue: In this paper they not focus on the textual data hiding. For perturbation they have just include one noise generation function that generate some random values and those can be add or sub from the original dataset values. But if one can know its trust level and parameter for the noise generation function then it can be regenerate.

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.

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

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.

In [1] perturbation of dataset is done for providing security of the data on server. As some of cooperative store data is store on server for regular updation in price, category, etc. Dataset need protection from unauthorized user. So proper solution for this problem is develop in this paper by perturbing the data before uploading it on server. Then proper algorithm is develop for the de-perturbing the uploaded perturbed copy as if authorized user again read data then it should get original copy. Here by the use of association rule sensitive information or pattern of items is obtained. Now those rule which are above the threshold of minimum support are perturbed by adding fake transaction in the dataset so that

overall support get reduce and dataset get perturb by these fake transaction. Placement of these transactions is done by modulus table. As this modulus remember the fake position in the dataset. In order to increase perturbation Items are replace by chipper text where each text will specify one item in the original dataset. In [14] similar work for outsourcing is done but the algorithm is calculation is unknown to the client and server.

Issue: As the addition of fake transaction in dataset for perturbation size of dataset increases. Because of this number of transaction in dataset increases and individual information is not hide in this work.

In [9, 12] paper cover a new issue for the direct indirect discrimination prevention in the dataset. Here it will collect discriminate item set which help in producing the association rule for identifying the direct or indirect rules. Then hide the rules which are above the threshold value by converting the X→Y to X→Y' where X is a set of discriminating item this tend to hide the information which will generate only those rules that not give any discriminating rule. Here Y is change to Y' means an opposite value is replace at few attributes.

Issue: Here it will hide those rules where Y will change only if it is of binary nature. But if Y is of multiple values then replacement of the one will increase the other.

So following are the work still need to be done:

- To develop a algorithm that not only provide privacy to the numeric data but also to the textual data as well.
- Algorithm should provide privacy to the different levels as well in case of multiple distributions.
- Discrimination items should also need to be cover as association leads to find the discriminate items.
- Multiple attribute should provide privacy when handling the association rule suppressing.
- At last individual privacy is on priority before publishing any dataset, as it harms whole organization and the person.

IV. Evaluation Parameters

There are two approaches to evaluate the discriminating algorithm developed which can specify the quality of the work first is Discrimination Removal while second is data quality after the implementation of the algorithm. Normally balancing both is quit difficult as if data quality need to maintain then some of the rules will be unaffected and over all purpose will be not be solve while in case of maintaining discriminating rule less data [11, 13], dataset the quality will definite degrade as it need to either change or remove from the dataset.

Sensitive Item Prevention Degree (SIPD): This measure quantifies the percentage of sensitive rules that are no longer discriminatory in the transformed dataset.

Non Sensitive Item Protection Prevention Degree **(NSIPP)**. This measure quantifies the percentage of the protective rules in the original dataset that remain protective in the transformed dataset.

Since the above measures are used to evaluate the success of the proposed methods in direct and indirect discrimination prevention, ideally their value should be 100%.

Data-Set Originality: As the privacy for the sensitive item is provide by hiding the sensitive item or replacing by other similar value but this lead to make dataset for perturbation. So work which maintain high data quality after prevention is better.

Execution time: Work need time for the effective result but algorithm that generate results in very sort duration of time then much better. So execution time is another evaluation parameter for the same.

Misses Cost (MC): This measure quantifies the percentage of rules among those extractable from the original dataset that cannot be extracted from the transformed dataset (side-effect of the transformation process).

Ghost Cost (GC): This measure quantifies the percentage of the rules among those extractable from the transformed dataset that were not extractable from the original dataset (side-effect of the transformation process). MC and GC should ideally be 0%. However, MC and GC may not be 0% as a side-effect of the transformation process.

V. Conclusion

Mining information from the data is the primary requirement of the data mining out of which privacy preserving mining is opening new emerging field which preserve knowledge from the data. Paper detailed various method like perturbing, swapping, etc. for privacy preserving, where each has its own importance. Researchers works find knowledge in dataset by Aprior and other mining algorithm then apply preserving technique on them. Hiding information at different level is also term as multi-level privacy which provide only numeric data hiding. While in few works both numeric and text data is hide but the time and space required for those algorithm is comparatively high. So a algorithm is still need to develop for the reduced time and space complexity without compromising time and space.

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]. C. Tai, P. S. Yu, and M. Chen, "K-support anonymity based on pseudo taxonomy for outsourcing of frequent itemset mining," in Proc. Int. Knowledge Discovery Data Mining, 2010, pp. 473–482.

[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.

[6]. 1M.Mahendran, 2Dr.R.Sugumar "An Efficient Algorithm for Privacy Preserving Data Mining Using Heuristic Approach" International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 9, November 2012

[7] 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.

[8] Enabling Multilevel Trust in Privacy Preserving Data Mining Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang IEEE transaction on knowledge data engineering, VOL. 24, NO. 9, SEPTEMBER 2012.

[9]. Sara Hajian and Josep Domingo-Ferrer. "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining". IEEE transaction on knowledge data engineering, VOL. 25, NO. 7, JULY 2013.

[10]. Mohamed R. Fouad, Khaled Elbassioni, and Elisa Bertino. A Supermodularity-Based Differential Privacy Preserving Algorithm for Data Anonymization. IEEE transaction on knowledge data engineering VOL. 26, NO. 7, JULY 2014

[11]. F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination Aware Decision Tree Learning," Proc. IEEE Int'l Conf. Data Mining (ICDM '10), pp. 869-874, 2010.

[12]. D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-Aware Data Mining," Proc. 14th ACM Int'l Conf. Knowledge Discovery and Data Mining (KDD '08), pp. 560-568, 2008.

[13] D. Pedreschi, S. Ruggieri, and F. Turini, "Measuring Discrimination in Socially-Sensitive Decision Records," Proc. Ninth SIAMData Mining Conf. (SDM '09), pp. 581-592, 2009.

[14]. Yogachandran Rahulamathavan, Raphael C.-W. Phan, Suresh Veluru, Kanapathippillai Cumanan and Muttukrishnan Rajarajan."Privacy-Preserving Multi-Class Support Vector Machine for Outsourcing the Data Classification in Cloud ".IEEE IEEE transaction on dependable and secure computing, VOL. 11, NO. 5, September 2014.