E-ISSN: 2321–9637
Volume 2, Issue 1, January 2014
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

A SURVEY OF PERTURBATION TECHNIQUE FOR PRIVACY–PRESERVING OF DATA

Twinkle Ankleshwaria[1], Prof. J.S. Dhobi[2]
[12]*Computer Science Engineering Department*
G.E.C. Modasa
twi1902@gmail.com jsdhobi@yahoo.com.

ABSTRACT:
In recent years, the data mining techniques have met a serious challenge due to the increased concerning and worries of the privacy, that is, protecting the privacy of the critical and sensitive data Data perturbation is a popular technique for privacy preserving data mining. The approach protects the privacy of the data by perturbing the data through a method. The major challenge of data perturbation is to achieve the desired result between the level of data privacy and the level of data utility. Data privacy and data utility are commonly considered as a pair of conflicting requirements in privacy-preserving of data for applications and mining systems. Multiplicative perturbation algorithms aim at improving data privacy while maintaining the desired level of data utility by selectively preserving the mining task and model specific information during the data perturbation process. The multiplicative perturbation algorithm may find multiple data transformations that preserve the required data utility. Thus the next major challenge is to find a good transformation that provides a satisfactory level of privacy data. A number of perturbation methods have recently been proposed for privacy preserving data mining of multidimensional data records. This paper intends to reiterate several perturbation techniques for privacy preserving of data and then proceeds to evaluate and compare different technologies.

Keywords-- *Privacy Preserving, Perturbation, Data mining.*

1. INTRODUCTION

Knowledge is supremacy and the more knowledgeable we are about information break-in, we are less prone to fall prey to the evil hacker sharks of information technology. Privacy preserving data mining has become increasingly popular because it allows sharing of privacy sensitive data for analysis purposes [1]. So people have become increasingly unwilling to share their data, frequently resulting in individuals either refusing to share their data or providing incorrect data. In turn, such problems in data collection can affect the success of data collection and can affect the success of data mining, which relies on sufficient amounts of accurate data in order to produce meaningful results. In recent years, the wide availability of personal data has made the problem of privacy preserving data mining an important one.

A data perturbation procedure can be simply described as follows. Before the data owner publishes the data, they change the data in certain way to disguise the sensitive information while preserving the particular data property that is critical for building meaningful data mining models.[7].

2. DIFFERENT PERTURBATION TECHNIQUES FOR PRIVACY PRESERVING OF DATA

The current techniques in perturbation approach are classified in two categories based on how they perturb datasets and particular Properties that will be preserved in data: Value-based Perturbation and Multi-Dimensional Perturbation. In the Value-based Perturbation the purpose is to preserve statistical characteristics and columns distribution while Multi-Dimensional Perturbation aims to hold Multi- Dimensional information.

E-ISSN: 2321–9637
Volume 2, Issue 1, January 2014
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

2.1 Value-based Perturbation Techniques

The main idea of this approach is to add random noise to the data values. This approach is actually, based on this fact that some data mining problems do not need the individual records necessarily and they just need their distribution. Since the perturbing distribution is known, they can reach data mining goals by reconstructing their required aggregate distributions. However, due to reconstructing each data dimension's distribution independently, they have the inherent disadvantage of missing the implicit information available in multi-dimensional records and on the other hand it is required to develop new distribution-based data mining algorithms.

2.1.1 Random Noise Addition Technique

This technique is described as follows: Consider n original data X_1, X_2, X_N, where X_i are variables following the same independent and identical distribution (i.i.d). The distribution function of X_i is denoted as F_X, n random variables Y_1, Y_2, Y_N are generated to hide the real values of by perturbation. Similarly, Y_i are *i.i.d* variables. Disturbed data will be generated as follows:

$$w_1, \ldots, w_n \text{ where } w_i = X_i + Y_i \quad i = 1,\ldots, n \qquad (1)$$

It is also assumed that the added noise variance is large enough to let an accurate estimation of main data values take place. Then, according to the perturbed dataset w_1 , w_n , known distributional function F_Y and using a reconstruction procedure based on Bayes rule, the density function f_X' will be estimated by Equation.

However, it is presented that privacy breaches as one of the major problems with the random noise addition technique and observed that the spectral properties of the randomized data can be utilized to separate noise from the private data. The filtering algorithms based on random matrix theory are used to approximately reconstruct the private data from the perturbed data. Thus, establishing a balance between Privacy preservation and accuracy of data mining result is hard because more we want privacy preservation, more we should lose information.

In [3], to minimize the information loss of this technique and improve the reconstruction procedure, a new distribution reconstruction algorithm called *Maximizing Algorithm for the Expectation of Mathematics* (EM) is represented. In [2] a new decision-tree algorithm is developed according to this technique. This technique is also used in privacy preserving association rule mining [14, 15].

However, in [16] it is presented that privacy breaches as one of the major problems with the random noise addition technique and observed that the spectral properties of the randomized data can be utilized to separate noise from the private data. The filtering algorithms based on random matrix theory are used to approximately reconstruct the private data from the perturbed data. Thus, establishing a balance between Privacy preservation and accuracy of data mining result is hard because more we want privacy preservation, more we should lose information.

2.1.2 Randomized Responses Technique

The main idea for this technique is to scramble data so that the data collector cannot express, with a probabilities better than of the defined threshold, whether the data sent back by the respondent is correct or not. There are two models in this technique: *Related–Question* and *Unrelated- Question* models. In the former, the interviewer asks every respondent a couple of questions related together, of each the reply is opposite the other one. For example, the questions can be as follows:

1) I have the sensitive attribute A.
2) I do not have the sensitive attribute A.

The respondent will answer randomly and with θ probability to the first question and with $1-\theta$ to the second

E-ISSN: 2321–9637
Volume 2, Issue 1, January 2014
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

question. Although the interviewer finds out the answers (yes or no), he does not know which question has been answered by the respondent; hence, the respondent's privacy preserving will be saved. The collector uses the following equations in order to estimate the percentage of the people who have characteristic A:

$$P^{*}\ A\ \ yes\ \ P\,A\ \ yes\ .\ \ P\,A\ \ no\ .\,1 \qquad (2)$$
$$P^{*}\ A\ \ no\ \ P\ A\ \ no\ .\ \ P\ A\ \ yes \qquad (3)$$

Where $P^{*}\ A\ yes$ (or $P^{*}\ A\ no$) is the ratio of the "*yes*" (or "*no*") replies which are acquired via survey data and $P\ A\ yes$ (or $P\ A\ no$) are estimated ratios of the "*yes*" (or "*no*") answers to sensitive questions. The purpose is to gain P A yes P A no.

Although in this method, information from each individual user is scrambled, if the number of users is significantly large, the aggregate information of these users can be estimated with decent accuracy. Randomized response technique used to provide information with response model, so are used for processing categorical data. Note that the technique can be extended to multi-dimensional, i.e., the techniques are applied to several dimensions altogether.

In [11] the random response technology and the geometric data transformation method are combined. That is called random response method of geometric transformation .It can protect the privacy of numerical data .Theoretical analysis and experimental results showed that at the same time cost , the algorithm can get better privacy protection than the previous algorithms , and would not affect the accuracy of mining results.

2.2 Data Mining Task-based Perturbation Techniques

The purpose of these techniques is to modify the original data so that the properties preserved in perturbed dataset to be task specific information data mining tasks and even a particular model. Thus, it is possible to preserve the privacy without missing any particular information of data mining tasks and make a more suitable balance between privacy and data mining results accuracy. Furthermore, in these techniques, data mining algorithms can be applied directly and without developing new data mining algorithms on the perturbed dataset.

2.2.1 Condensation Technique

The purpose of this technique is to modify the original dataset into anonymized datasets so that this anonymized dataset preserves the covariance matrix for multiple columns. In this technique first the data will be condensed into groups with pre-defined size *K*, and a series of statistical information related to the mean and correlations across the different dimensions will be preserved for each group of records. In the server, this statistical information is used to generate anonymized data with similar statistical characteristics to the original dataset. This technique has been used to create simple classifier for the K Nearest Neighbor (KNN) [4]

However in [5] it is presented that this technique is weak in protecting the private data. The KNN-based data groups result in some serious conflicts between preserving covariance information and preserving privacy.

2.2.2 Random Rotation Perturbation Technique

The main idea is as if the original dataset with *d* columns and *N* records represented as $X_{d\,n}$, the rotation perturbation of the dataset X will be defined as $G(X) = RX$,Where $R_{d\,d}$ is a random rotation orthonormal matrix.

A key feature of rotation transformation is preserving the Euclidean distance, inner product and geometric shape hyper in a multi-dimensional space. Also, kernel methods, SVM classifiers with certain kernels and hyper plane-based classifiers, are *invariant* to rotation perturbation, i.e. if trained and tested with rotation perturbed data, will have similar model accuracy to that trained and tested with the original data. [5].

E-ISSN: 2321–9637
Volume 2, Issue 1, January 2014
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

But researches show that having previous knowledge, the random rotation perturbation may become involved in privacy violations against different attacks including Independent Component Analysis (ICA), attack to rotation center and distance-inference attack [6,7].

[10] In this a method of Privacy Preserving Clustering of Data Streams (PPCDS) is proposed stressing the privacy –preserving process in a data stream environment while maintaining a certain degree of excellent mining accuracy. PPCDS is mainly used to combine Rotation –Based Perturbation , optimization of cluster enters and the concept of nearest neighbour, in order to solve the privacy –preserving clustering of mining issues in a data stream environment .In the phase of Rotation –Based Perturbation , rotation transformation matrix is employed to rapidly perturb with data streams in order to preserve data privacy. In the phase of cluster mining, perturbed data is primarily used to establish a micro-cluster through the optimization of a cluster centre, then applying statistic calculation to update the micro-cluster.

2.2.3 Geometric Perturbation Technique

This perturbation technique is a combination of Rotation, Translation and Noise addition perturbation techniques. The additional components T and Δ are used to address the weakness of rotation perturbation while still preserving the data quality for classification modeling. Concretely, the random translation matrix addresses the attack to rotation center and adds additional difficulty to ICA-based attacks and the noise addition addresses the distance-inference attack.[7]

If the matrix $X_{d\,n}$ indicates original dataset with d columns and N records, $R_{d\,d}$ be a orthonormal random matrix, T be a translation random matrix and D be a random noise matrix, (Guassian noise) where each element is

$$G(X) = RX + T+D. \qquad (4)$$

Geometric Data perturbations is the method is totally on distance base for estimating original value from the perturbed data, with addition of Gaussian noise[13]. Main Problem is with accuracy and data loss without use of Gaussian Noise. A Survey of Multi-dimensional Perturbation for Privacy Preserving Data Mining Geometric perturbation exhibits more robustness in countering attacks than simple rotation based perturbation The resulting value becomes the perturbed data for the sensitive attribute which will be considered further in evaluation of classification algorithms. The major cost of perturbation is determined by Eq.4 and a randomized perturbation optimization process that applies to a sample set of data set .The perturbation can be applied to data records in a streaming manner. Based on eq. 1 it will cost $O(d^2)$ to perturb each d-dimensional data record.

This perturbation technique is invariant against geometrical modification and is fixed for Kernel, SVM and linear classifiers. Geometrical perturbation technique also has, rather than Rotation perturbation and condensation, high-great Privacy Preserving guarantees.

In [12] a proposed approach based on geometric data perturbation and data mining –service oriented framework is introduced.GDP had shown to be an effective perturbation method in single –party privacy preserving data publishing. The multiparty framework and the problem of perturbation unification under this framework is presented. Three protocols are proposed which is first effort on applying GDP to multiparty privacy –preserving mining.

In [13] it is shown how several types of well –known data mining models will deliver a comparable level of model quality over geometrically perturbed data set as over the original data set.GDP ,includes the linear combination of three components: rotation perturbation ,translation perturbation , and distance perturbation .GDP perturbs multiple column in one transformation. A multi-column privacy evaluation model is proposed and analysis against three types of inference attacks : naïve-inference ,ICA –based and distance –inference is done.

E-ISSN: 2321–9637
Volume 2, Issue 1, January 2014
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

2.3 Dimension Reduction-based Perturbation Techniques

The main purpose of these techniques is to obtain a compact representation with reduced- rank to the original dataset while preserving dominant data patterns. These techniques also guarantee that both the dimensionality and the exact value of each element of the original data are kept confidential.

2.3.1 Random Projection Perturbation Technique

Random projection [6] refers to the technique of projecting a set of data points from a high-dimensional space to a randomly chosen lower-dimensional subspace.

$$F(X) = P*X \qquad (5)$$

- X is m*n matrix: m columns and n rows
- P is a k*m random matrix, k <= m
- RPP is more resilience to distance-based attacks
- RPP approximately preserves distances
- Model accuracy is not guaranteed

The key idea of random projection arises from the *Johnson-Lindenstrauss Lemma* [17]. According to this lemma, it is possible to maintain distance-related statistical properties simultaneously with dimension reduction for a dataset. Therefore, this perturbation technique can be used for different data mining tasks like including inner product/Euclidean distance estimation, correlation matrix computation, clustering, outlier detection, linear classification, etc.

However, this technique can hardly preserve the distance and inner product during the modification in comparison with geometric and random rotation techniques. It has been also clarified that having previous knowledge about this perturbation technique may be caught into privacy breach against the attacks [7].

3. COMPARISON AND EVALUATION FRAMEWORK

The evaluation framework recommended for assessing and evaluating data perturbation techniques, is in accordance with the following eight criteria:

- ***Privacy Loss*:** is defined as difficulty level in estimating the original values from the perturbed data values.
- ***Information Loss*:** is defined based on the amount of important data information, which needs to be saved after perturbation for data mining purposes.
- ***Data Mining Task*:** is defined based on a data mining task, which contains the possibility to mine it, after applying the privacy preserving techniques.
- ***Modifying the Data Mining Algorithms***: based on the needs, notifies the change for the existed data mining algorithms, in order to mine over the modified dataset.
- ***Preserved Property*:** that is, data information, which was already saved after applying the privacy preservation techniques.
- ***Data Type:*** it points out the types of data, which could be numerical, binary, or categorical.
- ***Data Dimension****:* it is defined based on the purpose of PPDM technique for preserve Dimensional information, which could be single-Dimensional or Multi-Dimensional.

Table below shows comparison of different perturbation techniques for privacy preserving of data.

E-ISSN: 2321–9637
Volume 2, Issue 1, January 2014
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

Table 1.Comparision of different perturbation techniques

		Data Perturbation Techniques of PPDM					
		Value-based Perturbation		Multi Dimensional Perturbation			
				Data Mining Task-based Perturbation			Reduction
Comparison Criteria		*Noise Addition*	*Randomized Response*	*Condensation*	*Random Rotation*	*Geometric*	*Random Projection*
Privacy Loss		Average	Average	Low	Low	Very Low	Very Low
Information Loss		Low	Low	Very Low	Very Low	Very Low	Very Low
Modifying DM Algorithms		Yes	Yes	No	No	No	No
Data Mining Task	*Asso*	√					
	Class	√	√	√	√	√	√
	Clus				√	√	√
Data Dimension		single-Dimensional	Single Dimensional	Multi Dimensional	Multi Dimensional	Multi Dimensional	Multi Dimensional
Preserved property		Values distribution	Values distribution	Covariance structure	Geometrical characteristic	Geometrical characteristic	Corelation between dimension
Data type		-	Categorical	numerical	numerical	numerical	numerical

4. CONCLUSION

The increasing ability to track and collect large amounts of data with the use of current hardware technology has lead to an interest in the development of data mining algorithms which preserve user privacy. Data perturbation techniques are one of the most popular models for privacy preserving data mining. It is especially useful for applications where data owners want to participate in cooperative mining but at the same time want to prevent the leakage of privacy-sensitive information in their published datasets. Typical examples include publishing micro data for research purpose or outsourcing the data to the third party data mining service providers. In this paper we have reviewed different Perturbation Techniques for Privacy Preserving and compared them using different evaluation criteria. According to the data set any one can choose from perturbation techniques available and preserve privacy of data sets.

References

[1] CHHINKANIWALA H. AND GARG S., "PRIVACY PRESERVING DATA MINING TECHNIQUES: CHALLENGES AND ISSUES", CSIT, 2011.
[2] R. Agrawal and R. Srikant. "Privacy-preserving data mining," In Proc. SIGMOD00, 2000, pp. 439-450.
[3] D. Agrawal and C. Aggarwal. "On the design and quantification of privacy pre-serving data mining algorithms", In Proc. of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Santa Barbara, California,USA, May 2001.

E-ISSN: 2321–9637
Volume 2, Issue 1, January 2014
International Journal of Research in Advent Technology
Available Online at: http://www.ijrat.org

[4] AGGRAWAL,C.C, AND YU.PS. ," A CONDENSATION APPROACH TO PRIVACY PRESERVING DATA MINING". PROC . OF INT .CONF. ON EXTENDING DATABASE TECHNOLOGY(EDBT)(2004).
[5] CHEN K, AND LIU, " PRIVACY PRESERVING DATA CLASSIFICATION WITH ROTATION PERTURBATION", PROC.ICDM,2005,PP.589-592.
[6] K.LIU, H KARGUPTA, AND J.RYAN," RANDOM PROJECTION –BASED MULTIPLICATIVE DATA PERTURBATION FOR PRIVACY PRESERVING DISTRIBUTED DATA MINING ." IEEE TRANSACTION ON KNOWLEDGE AND DATA ENGG,JAN 2006,PP 92-106.
[7] KEKE CHEN,GORDON SUN , AND LING LIU. TOWARDS ATTACK-RESILIENT GEOMETRIC DATA PERTURBATION." IN PROCEEDINGS OF THE 2007 SIAM INTERNATIONAL CONFERENCE ON DATA MINING,APRIL 2007.
[8] M. REZA,SOMAYYEH SEIFI," CLASSIFICATION AND EVALUATION THE PPDM TECHNIUES BY USING A DATA MODIFICATION -BASED FRAMEWORK", IJCSE,VOL3.NO2 FEB 2011.
[9] VASSILIOS S.VERYLIOS,E.BERTINO,IGOR N,"STATE –OF-THE ART IN PRIVACY PRESERVING DATA MINING",PUBLISHED IN SIGMOD 2004 PP.121-154.
[10] CHING-MING, PO-ZUNG & CHU-HAO," PRIVACY PRESERVING CLUSTERING OF DATA STREAMS", TAMKANG JOURNAL OF SC. & ENGG,VOL.13 NO. 3 PP.349-358
[11] JIE LIU, YIFENG XU, "PRIVACY PRESERVING CLUSTERING BY RANDOM RESPONSE METHOD OF GEOMETRICTRANSFORMATION", IEEE 2010
[12] KEKE CHEN, LING LUI, PRIVACY PRESERVING MULTIPARTY COLLABRATIVE MINING WITH GEOMETRIC DATA PERTURBATION,IEEE,JANUARY 2009
[13] KEKE CHEN, LING LIU," GEOMETRIC DATA PERTURBATION FOR PRIVACY PRESERVING OUTSOURCED DATA MINING", SPRINGER,2010.
[14] Rizvi, S. J. & Haritsa, J. R. , "Maintaining Data Privacy in Association Rule Mining" . In Proc. of the 28th International Conference on Very Large Data Bases (VLDB'02), Hong Kong, China, 2002,pp. 682–693.
[15] Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J, "Privacy Preserving Mining of Association Rules", In Proc. KDD02, 2002,pp. 217-228.
[16] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques," Proc. IEEE Int'l Conf. Data Mining, Nov. 2003.
[17] W.B. Johnson and J. Lindenstrauss, "Extensions of Lipshitz Mapping into Hilbert Space," Contemporary Math., vol. 26,pp. 189-206, 1984.