# DATA PRIVACY PRESERVATION USING DATA PERTURBATION TECHNIQUES

[1]THANGA REVATHI S, [2]N.RAMARAJ

[1]MNM Jain Engineering College, Chennai
[2]vignan University, Guntur
E-mail: [1]thangarevathi84@gmail.com, [2]ramaraj_gm@yahoo.com

**Abstract-** Data mining is the extraction of valuable knowledge from large data. Nowadays Data mining is done on Dynamic data rather than the traditional static data. The Data Mining techniques used today consider one main issue, the privacy of the critical and sensitive Data. There are various techniques involved in the privacy preservation of critical data. Data Perturbation is an important method to preserve privacy of data. Data perturbation is the data security technique that modifies the database to preserve the privacy and confidentiality. It is used for both data privacy and accuracy. In this paper we are going to discuss about the various perturbation techniques that can be used to preserve data privacy and discuss about the implications of the techniques.

**Keywords-** Data Perturbation, Confidentiality, Data Mining, Privacy Preserving

## I. INTRODUCTION

Today data from all the organizations are collected about the entire organization structure, Human resource, work flow etc., And the organizations suffers to get the full information or knowledge from the data. Sophisticated organizations make use of the Data mining algorithms to extract the unknown patterns or knowledge from the data, which may also access the confidential data stored in the Database. This creates the need for the Database administrator to protect the confidential record about the individual in the organizational database from improper disclosure. Some of the commonly used techniques for privacy preserving in cloud [12] are

1) Reconstruction method
2) Anonymization method
3) Cryptographic method

### The Reconstruction Method

Reconstruction is popular method for the Privacy Preserving Data Mining technique. The sensitive data are transformed or masked by adding additional data to the original data[12].

### The Anonymization Method

In this method the individual record is made indistinguishable using Suppression and Generalization techniques. K- anonymity is a common algorithm for this process. K-anonymity conceals the records which are commonly used as the unique identifier for the data.

### The Cryptographic Method

This method is mainly used when data mining is mutually done by different parties on the same data. This case requires preserving the privacy data mining of the parties. Several algorithms have been developed for this privacy preserving data mining [12].

### Data Perturbation

A data perturbation procedure can be simply described as follows. Before the data owners publish their data, they change the data in certain ways to disguise the sensitive information while preserving the particular data property that is critical for building meaningful data-mining models. Perturbation techniques have to handle the intrinsic trade-off between preserving data privacy and preserving data utility, as perturbing data usually reduces data utility [3].

### Types of Data Perturbation Techniques:

The different type of Perturbation techniques falls into two main categories. i) Probability Distribution category and ii) Fixed data perturbation category. In the Probability Distribution Category a sample of the whole population is considered and while performing the perturbation this sample is replaced by another set of samples in the database[13]. In the Fixed method, the Data is replaced by other set of data and it is done only once. It is not dependent on the samples, it is done individually and it is done once for all. The Probability Distribution method can be either by using data swapping or Probability Distribution. In the data Swapping method the original database is replaced by randomly generated database with the same attributes and same size. In the next method the density function of the attributes is identified and the estimate the values of the functions. Then a series of samples is generated from the estimated values which replaces the actual database of the same rank order and same size[2]. Data perturbation approach is classified into two methods. First is the Probability Distribution were the original database is replaced by the sample in the distribution or the same distribution itself. Next is the Value Distortion approach were the data is perturbed directly by adding noise or multiplicative noise or by some randomized noise[1].

The Data Perturbation approach is classified as

**1)** Rotation Perturbation
**2)** Projection Perturbation
**3)** Geometric Data Perturbation.

**Rotation Perturbation**

In this method the value of the two attributes in the matrix is rotated but the meaning of the value is protected. The pair of attributes is first selected then the value distortion technique is applied for those values [7]. Suppose original dataset have d column and N records then it is represented as X d×n, the rotation perturbation of dataset X will be defined as **G(X)=RX**,
where R d×d is a random rotation orthonormal matrix. [9]

**Projection Perturbation**

The data perturbation is done by moving the data value in high dimensional space to the lower dimensional space randomly. This can be projected either column or row wise [8]. Let Pk×d be a random Projection matrix, where P's rows are orthonormal.

**G(X) = (Sqrt(d)/k) PX**

is applied to perturb the original data set X. [9]

**Geometric Perturbation:**

This is a hybrid technique with the combination of rotation, translation and adding random noise value to the given data value in the matrix to provide quality of data preservation mainly for clustering [9].

**G(X) = RX+T+D;**

Matrix (X) d*n indicates the original data set with d-number of columns and n records, (R) d*d be a random rotation matrix and D be a random noise matrix and T is the Translation matrix.

**Transformation:**

Translation Transformation is done by adding a constant value to all the attribute values. The matrix can be moved from (X,Y) to (X',Y'). The original data cannot be viewed directly. It is represented by **v' = Tv**, where T is transformation matrix.

Rotation Transformation is performed by rotating a pair of attributes to a certain angle with respect to origin.

## II. SURVEY ON DATA PERTURBATION

**Geometric Perturbation**

Random geometric perturbation[3[ is the linear combination of the three components: rotation perturbation, translation perturbation, and distance perturbation.

**G(X)=RX+T+Δ.**

The Geometric properties are preserved by the modified data. In random rotation perturbation technique, multiple column data values are moved into single column transformation. But it leads to lack of privacy in multi dimensional rotation perturbation. In this paper a unified privacy metric was proposed to address the problem. And this metric was analyzed against the naive- interference attacks, ICA based attacks and distance interference attacks. And also proved for greater accuracy of rotation-invariant classifiers against other techniques. In this paper[2] geometric transformation technique is used including translation, scaling, rotation, which transforms the data by preserving the similarity between data objects. Geometric data translation is a type of data perturbation where the original data is translated to other form by adding some noise. Addition of noise is possible to categorical data and also numerical data. Addition of random noise may cause some loss of data which leads to less accuracy of results. To minimize the loss of data this paper uses Gaussian noise to perturb data which provide less loss in data. The geometric data perturbation was done only for numeric data which can be extended to non-numeric data using K-anonymization technique and for streams using Stream Analysis tool.

**Rotation Perturbation**

In this paper [5] the data matrix is divided vertically into sub-set matrix. For each sub-set matrix the random rotation matrix is used for the perturbation process. The Geometric class boundaries and the accuracy of the classifiers are well maintained in this method. Usually the geometric properties of the data cannot be preserved by the random rotation perturbation. This paper has handled this issue using the sub-set matrix.

Data perturbation approach was used in the single level trust of Data Miners. The perturbation approach uses random perturbation for the individual values to preserve the privacy of the data before it is published. But this is not suitable for the single level trust on data miners. This paper [3] extends the trust level to the multi level trust on data miners, were the higher trust level data miners will have less copies of perturbed data. But this MLT poses challenges based on the perturbation. Because of the different copies of perturbed data available, the data miner can utilize the diversity of the copies and can generate more accurate data as the original data. This is called the Diversity attack. In this paper this challenge in MLT – PPDM. The different copies of the perturbed data is generated by the Random Rotation process which varies in mean and co-variance. Based on these, the data owners will release only the mean and co-variance of the perturbed data for each trust levels to handle the Diversity attack. The rotation perturbation helps to reveal the statistical data instead of the original data. Many Privacy preserving data mining techniques has been proposed to preserve the sensitive information in the original dataset. But these techniques hold only to the static data and not for the data streams. This data stream has many dimensions which differ from the static data. They differ with time, with speed of data, amount of data, need of immediate response and many more. And mainly the accuracy of the data decreases if the transformation is performed on the data streams. This paper [7] has performed the random rotation of the data followed by the clustering process to preserve the sensitive data. This paper has also proved for the accuracy of the data after perturbation process.

**Projection perturbation**

Random projection refers to the technique of projecting a set of data points from a high dimensional space to a randomly chosen lower dimensional space. Projecting the data from higher dimensional space to a lower dimensional random space, will modify the data from original form to perturbed form by preserving much of its distance-related characteristics. This research paper[4] presents experimental results on the accuracy and privacy of the random projection-based data perturbation technique.

The following steps reduce the dimensionality of the data by random projections:

Consider the data set X={x1, …. xn} where each data point is a p dimensional vector such that xi є Rp and the following steps reduces the data onto a q dimensional space such that $1 \leq q < p$.

**1)** Arrange the data into a p × n matrix where p is the dimensionality of the data and n is the number of data points.

**2)** Generate a q × p random projection matrix R* [MATLAB randn (q, p)].

**3)** Multiply the random projection matrix with the original data.

$$\mathbf{X * q \times n = R* q \times p * Xp \times n}$$

Thus the data is projected to the lower random projection space.

This technique allows many algorithms to be applied directly on the perturbed data and perform analysis. The Euclidean distance is preserved in the perturbed data when this random projection technique is used.

## CONCLUSION

Data publishing is gaining more value nowadays; in turn the importance of the Data Privacy has also improved to a greater extent which is the major part in data publishing. In this paper, the different Privacy Preserving Data Mining techniques have been discussed and analysed. In this paper, we have discussed in detail about the various Data Perturbation techniques and have discussed about their merits. We have detailed about the random perturbation technique which has improved the efficiency of the perturbation techniques[4] [5] [1] . From the analysis of all the perturbation techniques, it is considered that the geometric perturbation provides a higher level of privacy and utility of the data to the users.

## REFERENCES

[1] Agrawal, C. and Yu, P.S., “General Survey of Privacy Preserving Data Mining Models and Algorithms”, Privacy-Preserving Data Mining Advances in Database Systems Volume 34, pp 11-52,2008.

[2] Fung, B.C.M., Wang, K., Chen, R., and Yu, P.S., “Privacy-Preserving Data Publishing: A Survey of Recent Developments,” ACM Computing Surveys”, Vol. 42, No. 4, pp. 1-53, 2010.

[3] Keke Chen, Ling Liu, “Privacy Preserving Data Classification with Rotation Perturbation”, Fifth IEEE International Conference on Data Mining, 2005.

[4] Keke Chen, Ling Liu, “Geometric Data Perturbation For Privacy Preserving Outsourced Data Mining”, Springer Knowl Inf Sys, 2010.

[5] Jie Liu, Yifeng XU, “Privacy Preserving Clustering by Random Response Method of Geometric Transformation”, Fourth International Conference on Internet Computing for Science and Engineering (ICICSE), pp. 181-188, 2009.

[6] R. Agrawal and R. Srikant, “Privacy Preserving Data Mining,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’00), 2000.

[7] Swapnil Kadam, Prof. Navnath Pokale, “Privacy Preserving through Data Perturbation using Random Rotation Based Technique in Data Mining “,International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5 Issue 1, January 2016

[8] Stanley, R. M., Oliveira and Osmar R. Za¨iane, “Privacy Preserving Clustering by Data Transformation”, Journal of Information and Data Management”, Vol. 1, No. 1, 2010.

[9] Chhinkaniwala H. and Garg S., “Privacy Preserving Data Mining Techniques: Challenges and Issues”, CSIT, 2011.

[10] Nimpal Patel and Shreya Patel,”A Study on Data Perturbation Techniques in Privacy Preserving Data Mining”, International Research Journal of Engineering and Technology , Volume 02, Issue 09, Dec 2015.

[11] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, “On the privacy preserving properties of random data perturbation techniques,” in Proceedings of the IEEE International Conference on Data Mining, Melbourne, FL, November 2003.

[12] Z. Huang, W. Du, and B. Chen, “Deriving private information from randomized data,” in Proceedings of the 2005 ACM SIGMOD Conference, Baltimroe, MD, June 2005, pp. 37–48.

[13] S. Guo and X. Wu, “On the use of spectral filtering for privacy preserving data mining,” in Proceedings of the 21st ACM Symposium on Applied Computing, Dijon, France, April 2006, pp. 622–626.

[14] Majid,M.Asger,Rashid Ali, “Privacy preserving Data Mining Techniques:Current Scenario and Future Prospects”,IEEE 2012.

★★★