

## RANDOM DATA PERTURBATION TECHNIQUES IN PRIVACY PRESERVING DATA MINING

Sangavi N<sup>1</sup>, Jeevitha R<sup>2</sup>, Kathirvel P<sup>3</sup>, Dr. Premalatha K<sup>4</sup>

<sup>1,2,3</sup>PG Scholars, Bannari Amman Institute of Technology, Sathyamangalam

<sup>4</sup>Professor, Bannari Amman Institute of Technology, Sathyamangalam

\*\*\*

**ABSTRACT**-Data mining strategies have been facing a serious challenge in recent years due to heightened privacy concerns and concerns, i.e. protecting the privacy of important and sensitive data. Data perturbation is a common Data Mining privacy technique. Data perturbation's biggest challenge is to balance privacy protection and data quality, which is normally considered to be a pair of contradictory factors. Geometric perturbation technique for data is a combination of perturbation technique for rotation, translation, and noise addition. Publishing data while protecting privacy –sensitive details–is particularly useful for data owners. Typical examples include publishing micro data for research purposes or contracting the data to third parties providing services for data mining. In this paper we are trying to explore the latest trends in the technique of perturbation of geometric results.

**Keywords:** Data mining, Privacy preserving, data perturbation, randomization, cryptography, Geometric Data Perturbation.

### INTRODUCTION

Enormous volumes of extensive personal data are routinely collected and analyzed using data mining tools. These data include, among others, shopping habits, criminal records, medical history, and credit records. Such data, on the one hand, is an important asset for business organizations and governments, both in decision-making processes and in providing social benefits such as medical research, crime reduction, national security, etc. Data mining techniques are capable of deriving highly sensitive information from unclassified data which is not even exposed to database holders. Worse is the privacy invasion triggered by secondary data use when people are unaware of using data mining techniques "behind the scenes"[3].

The daunting problem: how can we defend against the misuse of information that has been uncovered from secondary data use and meet the needs of organizations and governments to facilitate decision-making or even promote social benefits? They claim that a solution to such a problem involves two essential techniques: anonymity in the first step of privacy protection to delete identifiers (e.g. names, social insurance numbers, addresses, etc.) and data transformation to preserve those sensitive attributes (e.g. income, age, etc.) since the release of data, after removal of data. identifiers, may contain other information that can be linked with other datasets to re-identify individuals or entities [4].

We cannot effectively safeguard data privacy against naive estimation. Rotation perturbation and random projection perturbation are all threatened by prior knowledge allowed Independent Component Analysis Multidimensional-anonymization is only intended for general-purpose utility preservation and may result in low-quality data mining models. In this paper we propose a new multidimensional data perturbation technique: geometric data perturbation that can be applied for several categories of popular data mining models with better utility preservation and privacy preservation[5].

### Need for Privacy in Data Mining

Presumably information is the most important and demanded resource today. We live in an online society that relies on dissemination and information sharing in both the private as well as the public and government sectors. State, federal, and private entities are increasingly being required to make their data electronically available[5][6]. Protecting respondent privacy (individuals, groups, associations, businesses, etc.). Though technically anonymous, de-identified data may include other data, such as ethnicity, birth date, gender and ZIP code, which may be unique or nearly unique. Identifying the characteristics of publicly available databases associating these characteristics with the identity of the respondent, data recipients may decide to which respondent each piece of released data belongs, or limit their confusion to a specific sub-set of persons.

### DATA PERTURBATION

Data-perturbation-based approaches fall into two main categories which we call the probability distribution category and the fixed data perturbation category[8]. The probability distribution group considers the aggregation as a sample from a given population with a given probability distribution. In this case, the security check method substitutes for the original data With another sample or by the allotment itself from the same distribution. In the context of fixed data perturbation the values of the attributes in the database to be used to calculate statistics are once and for all disrupted. The perturbation methods of fixed data were developed solely for numerical or categorical data[9].

Within the category of probability distribution two methods can be defined. The first is called "data swap-ping" or "multidimensional transformation" This approach replaces the original database with a randomly generated database of exactly the same probability distribution as the original database[10]. When calculating a new

perturbation, consideration must be given to the relationship between this entity and the rest of the database, as long as a new entity is added or a current entity is eliminated. A one-to - one mapping between the original database and the disrupted database is needed. The Precision resulting from this method may be considered unacceptable, since in some cases the method may have an error of up to 50 per cent. The latter method is called probability distribution method. The method consists of three steps: (1) Identify the underlying density function of the attribute values, and estimate the parameters of that function. (2) Generate a confidential attribute data sample sequence from the approximate density function. The most recent sample would need to be the same size as the database. (3) delete these genera In other words, the lower value of the new sample will replace the lower value in the original data, and so on.

Data perturbation is a popular Data Mining privacy technique. Data perturbation's biggest challenge is to balance privacy protection and data quality, which are normally considered as a pair of contradictory factors[11]. The distribution of in this method Reconstructed independently of every data dimension. This means that any data mining algorithm based on the distribution works under an implicit assumption that each dimension is treated independently.

Approach to data perturbation is divided into two: the approach to probability distribution and the approach to value distortion. The approach to probability distribution replaces the data with another sample from the same distribution or the distribution itself, and the approach to value distortion Disrupts data elements or attributes directly by either additive noise, multiplicative noise, or other procedures of randomization. There are three types of approaches to data perturbation: Rotation perturbation, Projection perturbation and perturbation of geometric data.

**DIFFERENT METHODS OF DATA**

**PERTURBATION**

**3.1 Noise Additive Perturbation**

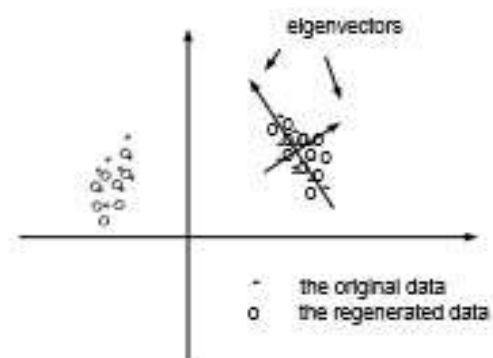
The standard technique of additive perturbation[13 ] is column-based randomisation of additives. This type of techniques is based on the facts that 1) data owners may not want to protect all values in a record equally, so a distortion of the column-based value can be applied to disturb some sensitive columns. 2) The data classification models to be used do not necessarily require the individual records, but only the distribution of the column value assuming separate columns. The basic method is to disguise the original values by injecting some amount of random additive noise, while the specific information, such as the column distribution, can still be effectively reconstructed from the perturbed data.

We treat the original values  $(x_1, x_2, \dots, x_n)$  from a column to be randomly drawn from a random variable X, which has some kind of distribution. By adding random noises R to the original data values, the randomization process changes the original data and generates a disturbed data column Y,  $Y = X + R$ . It publishes the resulting record  $(x_1+r_1, x_2+r_2, \dots, x_n+r_n)$  and the R distribution. The trick to introducing random noise is the algorithm of distribution reconstruction, which restores X's column distribution based on perturbed data and R's distribution.

**3.2 Condensation-based Perturbation:**

The approach to condensation is a standard multidimensional perturbation technique, aimed at maintaining the matrix of covariance for multiple columns. So some geometric properties like decision boundary form are well maintained. Unlike the randomization approach, multiple columns as a whole are disturbed in order to generate the whole "perturbed data set." As for the The perturbed dataset preserves the covariance matrix, and many existing data mining algorithms can be applied directly to the perturbed dataset without requiring algorithm modifications or new development.

It begins by partitioning the original data into groups k-record. The group consists of two steps—randomly choosing a record as the center of the group from the current records, and then identifying the  $(k - 1)$  nearest neighbors of the center as the other  $(k - 1)$  members. Before the next community is created the selected k records are extracted from the original dataset. Since each group has a limited locality, a set of k records may be regenerated to maintain the distribution and covariance roughly. The record replication algorithm aims to maintain each group's ownvectors and values, as shown in the Figure 1.



**Fig. 1** Eigen values of each group

**3.3 Random Projection Perturbation:**

Random projection perturbation (Liu, Kargupta and Ryan, 2006) refers to the technique of projecting a set of data points to another randomly selected space from the original multidimensional space. Let Pk average be a matrix

of random projection, where the rows of  $P$  are orthonormal[14].

$$G(X) = \sqrt{\frac{d}{k}} P X$$

is applied to perturb the dataset  $X$ .

### 3.4 Geometric data perturbation:

Def: Geometric data perturbation consists of a sequence of random geometric transformations, including multiplicative transformation ( $R$ ), translation transformation ( $\Psi$ ), and distance perturbation  $\Delta$ .

$$G(X) = RX + \Psi + \Delta [15]$$

The data is assumed to be an  $A_{p \times q}$  matrix where each of the  $p$  rows is an observation,  $O_i$ , and for each of the  $q$  attributes,  $A_i$ , each observation contains values. The matrix may include both numerical and categorical attributes. However, our methods of geometric data transformation rely on numerical attributes  $d$ , such that  $d \leq q$ . Thus, in the Euclidean space, the matrix  $p \times d$ , which is subject to transformation, can be regarded as a vector subspace  $V$ , so that each vector  $v_i \in V$  is the form  $v_i = (a_1, \dots, a_d)$ ,  $1 \leq i \leq d$ , where  $a_i$  is one instance of  $A_i$ ,  $a_i \in \mathbb{R}$ , and  $\mathbb{R}$  is the set of real numbers. Before releasing the data for clustering analysis, the vector subspace  $V$  must be transformed to preserve the privacy of the individual data records. We need to add or even multiply a constant noise term  $e$  to each element  $v_i$  of  $V$  in order to transform  $V$  into a distorted vector subspace  $V'$ .

**Translation Transformation:** A constant is added for all attribute values. The constant may be a negative or a positive number. Although its degree of privacy protection is 0 according to the formula for calculating the degree of privacy protection, it makes us unable to see the raw data directly from transformed data, so translation transform can also play the role of privacy protection as well.

Translation is the task of moving a point with coordinates  $(X;Y)$  through displacements  $(X_0;Y_0)$  to a new location. Using a matrix representation  $v' = T v$ , where  $T$  is a  $2 \times 3$  transformation matrix depicted in Figure 1(a),  $v$  is the vector column containing the original co-ordinates, and  $v'$  is a column vector whose co-ordinates are the transformed co-ordinates, is easily achieved. This form of matrix is also extended to Scaling and Rotation.

**Rotation Transformation:** Consider them as two-dimensional space points for a pair of arbitrarily selected

attributes and rotate them according to a given angle with the origin as the middle. If it is positive, we must rotate it in anti-clockwise direction. Otherwise we'll rotate them in the clockwise direction.

A more challenging transformation is rotation. This transformation, in its simplest form, is for the rotation of a point around the coordinate axes. Rotation of a point by angle in a discrete 2D space is achieved using the transformation matrix shown in Figure 1(b). The rotation angle is measured clockwise and this transformation affects the values of  $X$  and  $Y$  coordinates.

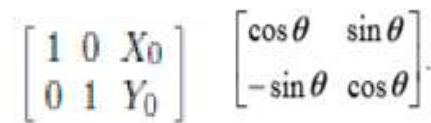


Fig. 2 (a) Translation Matrix (b) Rotation Matrix

The two elements above, translation and rotation maintain the relationship between the distances. A bunch of essential classification models will be "perturbation-invariant" by retaining distances, which is the center of geometric perturbation. In some situations, distance conserving perturbation may be subject to distance-inference attacks. The objective of distance disturbance is to preserve the Distances are approximate, while resilience to distance-inference attacks is effectively increased. We define the third component as a random matrix, where each entry is a separate sample with zero mean and small variance from the same distribution. By adding this component it slightly disturbs the distance between a pair of points.

### CONCLUSIONS

It focuses mainly on random geometric perturbation approach to privacy preserving data classification. Random geometric perturbation,  $G(X) = RX + \Psi + \Delta$ , includes the linear combination of the three components: rotation perturbation, translation perturbation, and distance perturbation. Geometric perturbation can preserve the important geometric properties, thus most data mining models that search for geometric class boundaries are well preserved with the perturbed data.

Geometric perturbation perturbs multiple columns in one transformation, which introduces new challenges in evaluating the privacy guarantee for multi-dimensional perturbation.

**REFERENCES**

1. Chhinkaniwala H. and Garg S., "Privacy Preserving Data Mining Techniques: Challenges and Issues", CSIT, 2011.
2. L.Golab and M.T.Ozsu, Data Stream Management issues-"A Survey Technical Report", 2003.
3. Majid, M.Asger, Rashid Ali, "Privacy preserving Data Mining Techniques:Current Scenario and Future Prospects", IEEE 2012.
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 Geometric Transformation", IEEE 2010
12. Keke Chen, Ling lui, Privacy Preserving Multiparty Collabrative Mining with Geometric Data Perturbation , IEEE, January 2009