Data Perturbation Techniques in Privacy Preserving Data Mining

N Sangavi [1], R Jeevitha[2], P Kathirvel[3], Dr. Premalatha[4]

[1,2,3]Pg Student, Bannari Amman Institute of Technology, Tamil Nadu, India
E-Mail : sangavi.cs18@bitsathy.ac.in , jeevitha.cs18@bitsathy.ac.in

[4]Professor, Bannari Amman Institute of Technology, Tamil Nadu , India

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

1. 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, 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].

A. Need for Privacy in Data Mining

Presumably the most important and demanded resource today is information. We live in an online society that relies on the dissemination and sharing of information in both the private and the public and government sectors. Increasingly, state, public and private entities are required to make their data available electronically [5][6]. To protect the privacy of respondents (individuals, groups, associations, companies, etc.). Though ostensibly anonymous, de-identified data may include other data, such as race, date of birth, gender and ZIP code, which may be unique or almost unique. identifying characteristics to publicly available databases associating these characteristics to the respondent's identity, the data recipients can determine to which respondent each piece of released data belongs or restrict their uncertainty to a specific subset of individuals.

B. Data Perturbation

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

Two methods can be defined within the probability distribution group. The first is called "data swap-ping" or "multidimensional transformation" This approach replaces the original database with a randomly generated database with approximately the same distribution of probabilities as the original database [10].As long as a new entity is added or a current entity is deleted, consideration must be given to the relationship between this entity and the rest of the database when calculating a new perturbation. A one-to - one mapping is required between the original database and the perturbed database. The precision resulting from this method may be considered unacceptable, as the method may have an error of up to 50 per cent in some cases. The second method is called method of distribution of probabilities. The method consists of three steps: (1) Identify the attribute values ' underlying density function and estimate the parameters of this function. (2) Generate a sample sequence of confidential attribute data from the approximate density function. The latest sample would have to be the same size as the database. (3) remove such general that is, the smallest value of the new sample should replace the smallest value in the original data, and so on.

Data perturbation is a popular technique for privacy preserving data mining. The major challenge of data perturbation is balancing privacy protection and data quality, which are normally considered as a pair of contradictive factors [11]. In this approach, the distribution of each data dimension reconstructed independently. This means that any distribution based data mining algorithm works under an implicit assumption to treat each dimension independently.

Data perturbation approach is classified into two: the probability distribution approach and the value distortion approach. The probability distribution approach replace the data with another sample from the same distribution or by the distribution itself , and the value distortion approach perturbs data elements or attributes directly by either additive noise, multiplicative noise, or some other randomization procedures. There are three types of data perturbation approaches: Rotation Perturbation, Projection Perturbation and Geometric Data Perturbation.

C. Different Methods Of Data Perturbation

Noise Additive Perturbation

The typical additive perturbation technique[13] is column-based additive randomization. This type of techniques relies on the facts that 1) Data owners may not want to equally protect all values in a record, thus a column-based value distortion can be applied to perturb some sensitive columns. 2) Data classification models to be used do not necessarily require the individual records, but only the column value distributions with the assumption of independent columns. The basic method is to disguise the original values by injecting certain amount of additive random 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,...,x_n) from a column to be randomly drawn from a random variable X, which has some kind of distribution. The randomization process changes the original data by adding random noises R to the original data values, and generates a perturbed data column Y, $Y = X + R$. The resulting record (x_1+r_1, x_2+r_2,...,x_n+r_n) and the distribution of R are published. The key of random noise addition is the distribution reconstruction algorithm that recovers the column distribution of X based on the perturbed data and the distribution of R.

Condensation-based Perturbation:

The condensation approach is a typical multidimensional perturbation technique, which aims at preserving the covariance matrix for multiple columns. Thus, some geometric properties such as the shape of decision boundary are well preserved. Different from the randomization approach, it perturbs multiple columns as a whole to generate the entire "perturbed dataset". As the perturbed dataset preserves the covariance matrix, many existing data mining algorithms can be applied directly to the perturbed dataset without requiring any change or new development of algorithms.

It starts by partitioning the original data into k-record groups. Each group is formed by two steps – randomly selecting a record from the existing records as the center of group, and then finding the (k − 1) nearest neighbors of the center to be the other (k − 1) members. The selected k records are removed from the original dataset before forming the next group. Since each group has small locality, it is possible to regenerate a set of k records to approximately preserve the distribution and covariance. The record regeneration algorithm tries to preserve the eigenvectors and eigen values of each group, as shown in Figure 1.

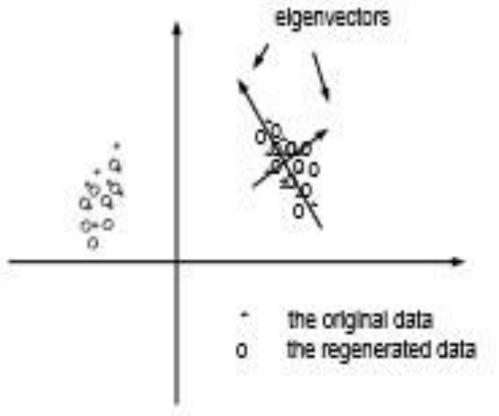

Fig. 1 Eigen values of each group

Random Projection Perturbation:

Random projection perturbation (Liu, Kargupta and Ryan, 2006) refers to the technique of projecting a set of data points from the original multidimensional space to another randomly chosen space. Let $P_{k\times d}$ be a random projection matrix, where P's rows are orthonormal [14].

G(X) = $\sqrt{\frac{d}{k}}PX$ is applied to perturb the dataset X.

Geometric data perturbation:

Def: Geometric data perturbation consists of a sequence of random geometric transformations, including multiplicative transformation (R), translation transformation (Ψ), and distance perturbation Δ.

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

The data is assumed to be a matrix A_{pxq}, where each of the p rows is an observation, O_i, and each observation contains values for each of the q attributes, Ai. The matrix may contain categorical and numerical attributes. However, our Geometric Data Transformation Methods rely on d numerical attributes, such that d <= q. Thus, the p x d matrix, which is subject to transformation, can be thought of as a vector subspace V in the Euclidean space such that each vector vi€ V is the form v_i = (a1; :::; ad),1 <=i<= d, where $\forall i$ a_i is one instance of A_i, a_i€ R, and R is the set of real numbers. The vector subspace V must be transformed before releasing the data for clustering analysis in order to preserve privacy of individual data records. To transform V into a distorted vector subspace V', we need to add or even multiply a constant noise term e to each element v_i of V .

Translation Transformation: A constant is added to all value of an attribute. The constant can be a positive or negative number. Although its degree of privacy protection is 0 in accordance with the formula for calculating the degree of privacy protection, it makes we cannot see the raw data from transformed data directly, so translation transform also can play the role of privacy protection.

Translation is the task to move a point with coordinates (X; Y) to a new location by using displacements(X0; Y0). The translation is easily accomplished by using a matrix representation v' = Tv, where T is a 2 x 3 transformation matrix depicted in Figure 1(a), v is the vector column containing the original coordinates, and v' is a column vector whose coordinates are the transformed coordinates. This matrix form is also applied to Scaling and Rotation.

Rotation Transformation: For a pair of attributes arbitrarily chosen, regard them as points of two dimension space, and rotate them according to a given angle θ with the origin as the center. If θ is positive, we rotate them along anti-clockwise. Otherwise, we rotate them along the clockwise.

Rotation is a more challenging transformation. In its simplest form, this transformation is for the rotation of a point about the coordinate axes. Rotation of a point in a 2D discrete space by an angle is achieved by using the transformation matrix depicted in Figure 1(b). The rotation angle is measured clockwise and this transformation ects the values of X and Y coordinates .

$$\begin{bmatrix} 1 & 0 & X_0 \\ 0 & 1 & Y_0 \end{bmatrix} \quad \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}.$$

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

The above two components, translation and rotation preserve the distance relationship. By preserving distances, a bunch of important classification models will be "perturbation-invariant", which is the core of geometric perturbation. Distance preserving perturbation may be under distance-inference attacks in some situations. The goal of distance perturbation is to preserve distances approximately, while effectively increasing the resilience to distance-inference attacks. We define the third component as a random matrix Δd×n, where each entry is an independent sample drawn from the same distribution with zero mean and small variance. By adding this component, the distance between a pair of points is disturbed slightly.

II. CONCLUSIONS

We present a 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] KekeChen,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,SomayyehSeifi," Classification and Evaluation the PPDM Techniues by using a data Modification -based framework", IJCSE,Vol3.No2 Feb 2011.

[9] VassiliosS.Verylios,E.Bertino,IgorN,"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