ijournals International Journal of Software & Hardware Research in Engineering
ISSN No: 2347-4890 Volume 2 Issue 2, February 2014

A Study of Data Perturbation Techniques For Privacy Preserving Data Mining

Aniket Patel[1], HirvaDivecha[2]
Assistant Professor
Department of Computer Engineering
U V Patel College of Engineering
Kherva-Mehsana, India

Samir Patel
Assistant Professor
Department of Information Technology
Sigma Institute Of Engineering,
Baroda,Gujarat

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 [13]. The major challenge of data perturbation is balancing privacy protection and data quality, which normally considered as a pair of contradictive factors. Geometric data perturbation technique is a combination of Rotation , Translation and Noise addition perturbation technique. It is especially useful for data owners to publish data while preserving privacy –sensitive information. Typical examples include publishing micro data for research purpose or outsourcing the data to the third party that provides data mining services. In this paper we try to explore the latest trends in Geometric data perturbation technique.

Keywords

Data mining;Privacy preserving; data perturbation; randomization; cryptography;Geometric Data Perturbation

1. INTRODUCTION

Huge volumes of detailed personal data are regularly collected and analyzed by applications using data mining. Such data include shopping habits, criminal records, medical history, credit records, among others. On the one hand, such data is an important asset to business organizations and governments both to decision making processes and to provide social benifits, such as medical research, crime reduction, national security, etc.[2] The threat to privacy becomes real since data mining techniques are able to derive highly sensitive knowledge from un classified data that is not even known to database holders. Worse is the privacy invasion occasioned by secondary usage of data when individuals are unaware of “behind the scenes” use of data mining techniques [3]

The challenging problem that: how can we protect against the abuse of the knowledge discovered from secondary usage of data and meet the needs of organizations and governments to support decision making or even to promote social benefits? We claim that a solution for such a problem requires two vital techniques: anonymity to remove identifiers (e.g. names, social insurance numbers, addresses, etc.) in the first phase of privacy protection, and data transformation to protect some sensitive attributes (e.g. salary, age, etc.) since the released data, after removing identifiers, may contain other information that can be linked with other datasets to re-identify individuals or entities [4].

We cannot effectively protect data privacy from naive estimation. The rotation perturbation and random projection perturbation are all threatened by prior-knowledge enabled Independent Component Analysis Multidimensional-anonymization is only designed 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].

1.1 Need For Privacy in Data Mining

Information is today probably the most important and demanded resource. We live in an internetworked society that relies on the dissemination and sharing of information in the private as well as in the public and governmental sectors. Governmental, public, and private institutions are increasingly required to make their data electronically available[5][6].To protect the privacy of the respondents (individuals, organizations, associations, business establishments, and so on).Although apparently anonymous, the de-identified data may contain other data, such as race, birth date, sex, and ZIP code, which uniquely or almost uniquely pertain to specific respondents (i.e., entities to which data refer) and make them stand out from others[7].By linking these 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.

2. DATA PERTURBATION

The methods based on the data-perturbation approach fall into two main categories, which we call the probability distribution category and the fixed data perturbation category [8]. The probability distribution category considers the database to be a sample from a given population that has a given probability distribution. In this case, the security control method replaces the original data by another sample from the same distribution or by the distribution itself. In the fixed data perturbation category, the values of the attributes in the database, which are to be used for computing statistics, are perturbed once and for all. The fixed data perturbation methods have been developed exclusively for either numerical data or categorical data [9].

Within the probability distribution category, two methods can be identified. The first is called “data swap- ping” or “multidimensional transformation” In this method, the original database is replaced with a randomly generated database having approximately the same probability distribution as the original database [10]. As long as a new entity is added or a current entity is deleted, the relationship between this entity and the rest of the database has to be taken into consideration when computing a new perturbation. There is a need for a one-to-one mapping between the original database and the perturbed database. The precision resulting from this method may be considered unacceptable since the method may in some cases have an error of up to 50%. The second 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 this function. (2) Generate a sample series of data from the estimated density function of the confidential attribute. The new sample should be the same size as that of the data- base. (3) Substitute the generated data of the confidential attribute for the original data in the same rank order. 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.

3. DIFFERENT METHODS OF DATA PERTURBATION

3.1 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.

3.2 Condensation-based Perturbation:

The condensation approach is a typical multi-dimensional 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.

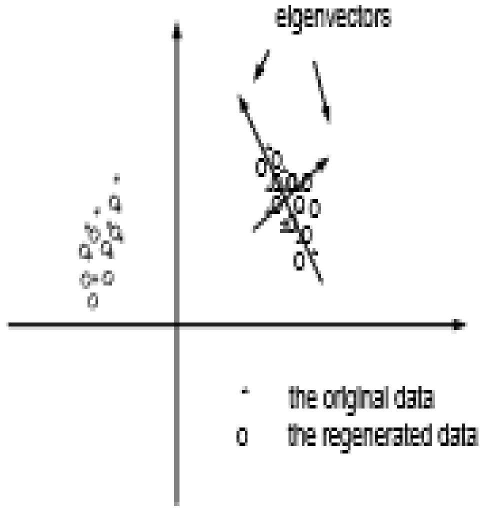

Figure 1 Eigen values of each group, as shown in

3.3 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.

3.4 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 + Ψ + Δ [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 [15].

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 [15].

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 [16].

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 [17].

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

Figure 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 [21].

4. Conclusions

We present a random geometric perturbation approach to privacy preserving data classification. Random geometric perturbation, G(X) = RX + Ψ + Δ, 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. [18]

Geometric perturbation perturbs multiple columns in one transformation, which introduces new challenges in evaluating the privacy guarantee for multi-dimensional perturbation. We propose a multi-column privacy evaluation model and design unified privacy metric to address these problems. [18]

5. 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 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] H.Chhinkaniwala & S.Garg,” Tuple -Value Based Multiplicative Data Perturbtion Approach to preserve

privacy in data stream mining", IJDKP,Vol3,No.3 May 2013.

[15] Mr.Kiran Patel," Privacy Preserving Data Stream Classification: An Approach using MOA framework",GIT Vol-6,2013

[16] A. Bifet, R. Kirkby, P. Kranen and P. Reutemann, Massive Online Analysis Manual. May 2011.

[17] [17] Stanley R. M. Oliveira and Osmar R. Za¨ıane Privacy Preserving Clustering by Data Transformation, Journal of Information and Data Management, Vol. 1, No. 1, February 2010

[18] Aniket Patel, Samir Patel,"A Survey On Geometric Data Perturbation in Multiplicative Data Perturbation, Vol 1 Issue 5,Dec.2013.

[19] Stanley R. M. Oliveira, Osmar R. Zaiane, Privacy Preserving Clustering by Data Transformation, February 2010

[20] Jie Liu, Yifeng XU, Privacy Preserving Clustering by Random Response Method of GeometricTransformation, 2009

[21] Keke Chen, Ling Liu ,Geometric Data Perturbation for PrivacyPreserving Outsourced Data Mining