

A Survey on Data Perturbation Techniques for Privacy Preserving in Data Mining

Devang J. Patel¹ Prof. Swati Patel²

¹Student of M.E ²Assistant Professor

¹Department of Information Technology ²Department of Computer Engineering

^{1,2}L.D. College of Engineering, Ahmedabad, Gujarat, India

Abstract— In recent years, advance in hardware technology have lead to increase in the capability to store and record personal data about service consumers and individuals. This has lead to concerns that the personal data may be misused for a variety of purposes. It is also very important to find out some useful information from large data sets. So in this paper we address privacy problem during mining large data sets. We introduce a certain transformation methods that deal with the privacy while mining. The goal of our methods are preserving the accuracy of specific data and preserving privacy of original data. Proposed methods provide privacy to only sensitive numerical attributes. This paper focuses on Geometric data perturbation technique for large data sets analyze.

Key words: Data mining; privacy; perturbation; Geometric data perturbation

I. INTRODUCTION

Information explosion now a day is very much. So it is becomes important to find some useful information from the large amount of data sets for the future purpose or for the future prediction. These data are may be belonging to some specific application such as telecommunication, financial, library, sensor networks, retail industry, online transaction and some of individual. These data sets need to be analyzed for identifying patterns which can be used to predict future behaviour. However data owners are not interested to share their original data sets due to privacy reason. So we need to do some procedure on their original data for privacy purpose before it is going to be released for mining [1]. A number of methods have recently been proposed for privacy preserving data mining of multidimensional records.

We are going to present a new data perturbation technique that provides the privacy to outsourced data sets while we mining the data. A Data perturbation approach is simply work in the following manner. Before any data owner is going to publish their original data, they change the original data in such a way that it is becomes very difficult to get the original one [2]. The goal of perturbation technique is twofold. Preserving the accuracy of specific data mining models which are data utility and preserving the privacy of original data [3][4]. The discussion about transformation-invariant data mining models has shown that multiplicative perturbations can theoretically guarantee zero loss of accuracy for a number of data mining models.

II. NEED FOR PRIVACY IN DATA MINING

It is general thinking that when thing is about privacy, it is “Keep the information about me from being unavailable to others”. Most of the time it is concern that it should not be an unauthorized used. Once the information is leaked then it is difficult to prevent from misuses. Privacy is also talked about the organization privacy. In organizational privacy

leakage the bunch of data is leaked rather than individual data item. It is not to be known the birth date, Gender, Zip Code, Security Number or else identity information which uniquely pertain to specific entities and make them stand outs from others [6]. By linking these identical characteristics one can determine all the information about an individual. Essential of the privacy concern is about the releasing information of individual and organization without worry of the misuse of the information while the mining of data is performed [8].

III. RECONSTRUCTION BASED APPROACH

Reconstruction based approaches generate privacy aware database by extracting sensitive characteristics from the original database [12]. These approaches generate lesser side effects in database than heuristic approach. Reconstruction based technique perturb (modify) the original data to achieve the privacy preserving. The modified data should meet the two conditions. First one is the opponent (attacker) cannot get the original data from analyzing the modified dataset. Second condition is that the modified data is still to maintain some statistical properties of the original data [13]. It means that the privacy of the data and accuracy of the data both would be maintained. The work that performed by original data that should be performed by modified data also.

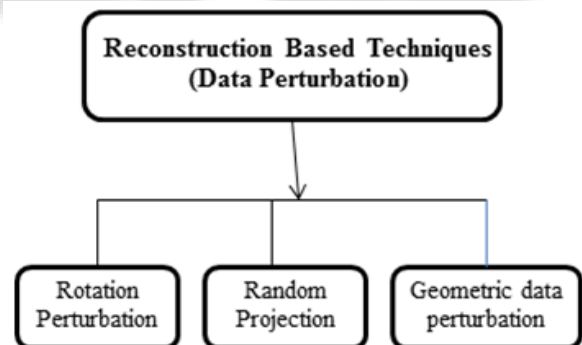


Fig. 1: Reconstruction Based Techniques

A. Data Perturbation:

An ultimate goal for all data perturbation algorithms is to optimize the data transformation process by maximizing both data privacy and data utility achieved [14]. Perturbation techniques are often evaluated with two basic metrics: Level of privacy guarantee and level of model-specific data utility preserved, which is often measured by the loss of accuracy for data classification and data clustering. However, the two metrics are typically representing two conflict goals in many existing techniques.

Data privacy is commonly measured by the difficulty level in estimating the original data from the perturbed data. Given a data perturbation technique, the higher level of difficulty in which the original values can be

estimated from the perturbed data, the higher level of data privacy this technique support. There are three types of data perturbation approaches: Rotation perturbation, Projection Perturbation and Geometric data Perturbation.

1) *Rotation Perturbation:*

In rotation perturbation approach, the original data set is multiplied with the random rotation matrix before going to publish in order to preserve the data privacy [7]. The advantage of this method is that it can maintain the geometric properties of the data set. When technique is applied to Dataset it must preserve the distances between data points. Suppose original dataset have d column and N records then it is represented as $X d \times n$, the rotation perturbation of dataset X will be defined as $G(X)=RX$, where R $d \times d$ is a random rotation orthonormal matrix. So the key feature of this method is preserving the geometric shape in space, Euclidean distance and inner products.

Principal component analysis (PCA) is used to modify the multi-dimensional data in to the lower dimensional data. PCA states that variability in process can be used in the analysis so it becomes difficult to separate the very important and less important variables [15] [16]. A dataset X_i ($i = 1,2,3,\dots,n$) is summarized as a linear combination of orthonormal vectors:

$$F(x, V) = u + (xV) VT \quad (1)$$

Where $f(x,V)$ is a vector valued function, u is the mean of the data $\{x_i\}$ and V is an $d \times m$ matrix with orthonormal columns. The mapping $Z_i = x_iV$ provides a lower dimensional projection of the vector x_i if $m < d$. Consequently, Principle component analysis (PCA) replaces the original variables of a dataset with smaller number of uncorrelated variables called the principle component.

2) *Projection Perturbation:*

This projection based perturbation technique refers to projecting a set of data points from an original multidimensional space to a randomly chosen lower dimensional subspace [17]. Let $P_k \times d$ be a random Projection matrix, where P's rows are orthonormal.

$$G(X) = (\text{Sqrt}(d)/k) PX \quad (2)$$

is applied to perturb the original data set X. It shows that if we are taking distance pair wise for any two points it gives small number of errors. Quality is preserved. The random projection based technique may be even more powerful when used with some other geometric transformation techniques like scaling, translation and rotation.

3) *Geometric Based Perturbation:*

Geometric data perturbation is developed to overcome the limitation of the projection perturbation technique. Geometric data perturbation, consisting of random rotation perturbation, random translation perturbation and noise addition, aims at preserving the important geometric properties of multidimensional data sets, while providing better privacy guarantee for data classification modelling. The preliminary study has shown that random geometric perturbation can well preserve model accuracy for several popular classification models [8].

$$G(X) = RX+T+D; \quad (3)$$

Matrix (X) $d \times n$ indicates the original data set with d-number of columns and n records, (R) $d \times d$ be a random rotation matrix and D be a random noise matrix, where each element is independently and identically distributed variable [18].

The data is assumed to be matrix Apq , where P rows is observation, O_i and each observation contains value for each of q attributes, A_i . The matrix may contain categorical and numerical attributes. But this method is only deal with numerical attributes, $d \leq q$.

4) *Translation Transformation:*

The positive or negative constant number is going to be added to all value of attributes. We cannot see original data from transformed data directly. So it is important factor in privacy preserving. With translation we are move point with coordinates (X;Y) to a new location by (X1,Y1). It can be easily represented by $v' = Tv$, where T is transformation matrix. Translation achieved by following matrix:

$$\begin{bmatrix} 1 & 0 & X_0 \\ 0 & 1 & Y_0 \end{bmatrix}$$

Fig. 2: Translation Matrix

5) *Rotation Transformation:*

According to given the rotation angle (θ), we are rotating pair of attributes with respect to origin. If the given angle (θ) is positive we rotate them anti-clockwise. Otherwise, we rotate them along the clockwise. Rotation is a more challenging transformation. In its simplest form, this transformation is for rotation of point about coordinate axes. Rotation in 2D space by some angle is achieved by using following given matrix [19] [20].

$$R = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}$$

Fig. 3: Rotation Matrix

The additional components Translation (T) and noise (D) are used to address the weakness of rotation perturbation while still preserving data quality [21]. Random translation matrix addresses the attack to rotation centre and adds additional difficulty to ICA (Independent Component Analysis) based attacks and noise addition addresses the distance-inference attack.

IV. CONCLUSION

An ultimate goal for all data perturbation algorithm is to optimize the data transformation process by maximizing both the data privacy and the data utility achieving. Proposed approaches focused on data perturbation by various transformation methods to preserve privacy of sensitive attributes. Geometric data transformation will provide better privacy than the rotation and projection perturbation because at a time we are doing random rotation, translation and adding some noise. We conclude that we have reached from reviewing this area by focusing on privacy issues within limits of privacy preserving data mining approaches.

REFERENCES

[1] Prashant Lahane, R K Bedi, Prasad Halgaonkar, "Data Mining", International Journal of Advances in computing and Information Researches, ISSN:2277-4068, Volume 1-No. 1, January 2012.
[2] Ms. Ompriya Kale, Ms. Prachi Patel, "A Survey on Privacy Preserving Data Mining Techniques", Global journal of Advanced Engineering Technologies, ISSN: 2277-6370, vol2, Issue3-2013.

- [3] Hitesh Chhikaniwala, Dr. Sanjay Garg, "Privacy Preserving Data Mining Techniques: Challenges and Issues".
- [4] Keke Chen, Member IEEE and Ling Liu, Senior Member IEEE, "Privacy Preserving Multiparty Collaborative Mining with Geometric Data Perturbation" IEEE Vol. XX, No. XX, January 2009.
- [5] Neha Gupta, Indrajeet Rajput, "Preserving privacy using data perturbation in Data Stream" International Journal of advanced Research in computer engineering & technology (IJARCET) Volume 2, No. 5, May 2013.
- [6] Kun Liu, Hillol Kargupta, Senior Member, IEEE and Jessica Ryan, "Random Projection-based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining", IEEE transaction on knowledge, Vol. 18, No. 1, January 2006.
- [7] Stanley R. M. Oliveria, Osmar R. Zaiane, "Data Perturbation by rotation for privacy preserving Clustering", Technical Report TR 04-17, August 2004.
- [8] Twinkle Ankleshwaria, Prof. J. S. Dhobi, "Geometric Data Perturbation Approach For Privacy preserving in data Stream Mining" Engineering Universe for Scientific research and Management, Impact factor 3.7, Volume 6, Issue 4, April 2014.
- [9] Yabo Xu, Ke Wang, Ada Wai-Chee Fu, Rong She, and Jian Pei, Privacy-Preserving Data Stream Classification, Springer, pp.489-510(2008).
- [10] Ching-Ming Chao, Po-Zung Chen and Chu-Hao Sun, Privacy-Preserving Classification of Data Sets, Tamkang Journal of Science and Engineering, Vol. 12, No. 3, pp. 321-330(2009).
- [11] S. Fienberg and J. McIntyre, "Data Swapping: Variations on a Theme by Dalenius and Reiss," Technical Report, National Institute of Statistical Sciences, 2003.
- [12] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," Technical Report SRI-CSL-98-04, 1998.
- [13] L. Sweeney, k-anonymity, "A model for protecting privacy, International Journal on Uncertain Fuzziness Knowledge Based System," vol. 10, no. 5, pp. 557-570, 2002.
- [14] P. Samarati, "Protecting respondents' identities in microdata release," In IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 13, issue 6, pp 1010-1027, 2001.
- [15] R. Vidya Banu and N. Nagaveni, "Preservation of Data Privacy using PCA based Transformation", in 2009 International Conference on Advances in Recent Technologies in Communication and Computing, in 2009 IEEE computer society, p.43
- [16] R. Vidya Banu and N. Nagaveni, "Preservation of Data Privacy using PCA based Transformation", in 2009 International Conference on Advances in Recent Technologies in Communication and Computing, in 2009 IEEE computer society, p.439.
- [17] Kun Liu, Hillol Kargupta, Senior Member, IEEE, and Jessica Ryan, "Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining", IEEE Transactions on Knowledge and Data Engineering, VOL. 18, NO. 1, JANUARY 2006, p.92.
- [18] C. Keke and L. Ling, Privacy-preserving Multiparty Collaborative Mining with Geometric Data Perturbation, IEEE Transactions On Parallel and Distributed Computing, Vol XX, 2009.
- [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 Geometric Transformation, 2009
- [21] Keke Chen, Ling Liu, Geometric Data Perturbation for Privacy Preserving Outsourced Data Mining.