A Review on Reconstruction Based Techniques for Privacy Preservation of Critical Data

Keyur Dodiya[#1], **Shruti Yagnik***[2]

[#]Department of Computer Engineering, Gujarat Technological University
Gujarat, India

*Department of Information Technology, Gujarat Technological University
Gujarat, India

Abstract— **The internet world today stands on the pillars of the security principles and Cryptography. It is very important to be able to preserve the privacy and confidentiality of critical data. In this paper we address the privacy preservation problem against unauthorized secondary use of information. To do so, we have tried to study the various data perturbation and Reconstruction Based Techniques which ensures that the mining process will not violate privacy up to a certain degree of security. This is done by Perturb ting the data and adding randomization and transforming the data through translating the data, rotating it and adding some noise.**

Keywords — **Randomization, Reconstruction, Rotation, Translation, Cryptography, Unauthorized, Perturbation, Privacy and Confidentiality.**

I. Introduction

Large volumes of detailed personal data are regularly collected and analyzed by applications using data mining. Such data include shopping, criminal records, medical history, credit records, among others [1]. On the one hand, such data is an important asset to business organizations and market-basket analysis both to decision making processes and to provide social benefits, such as medical research, crime reduction, national security, etc.[2]. On the other hand, analyzing such data opens new threats to privacy and autonomy of the individual if not done properly.

We present a new data perturbation technique for privacy preserving outsourced data mining. A data perturbation procedure can be simply described as follows. Before the data owners publish their data, they change the data in certain way 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 tradeoff between preserving data privacy and preserving data utility, as perturbing data usually reduces data utility. Several perturbation techniques have been proposed for mining purpose recently, but these two factors are not satisfactorily balanced. For example, random noise addition approach [3] is weak to data reconstruction attacks and only good for very few specific data mining models. The condensation approach [4].

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.

II. 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.

III. Reconstruction Based Approach

Reconstruction based approaches generate privacy aware database by extracting sensitive characteristics from the original database. These approaches generate lesser side effects in database than heuristic approach [8].Reconstruction based techniques perturb the original data to achieve privacy preserving. The perturbed data would meet the two conditions. First, an attacker cannot discover the real original data from the issuance of the distortion data. Second, the distorted data is still to maintain some statistical properties of the original data, namely some of the information derived from the distorted data are equivalent to data obtained from the original information [9]. Y. Guo proposed a FP tree based algorithm which reconstruct the original database by using non characteristic of database and efficiently generates number of secure databases [10].

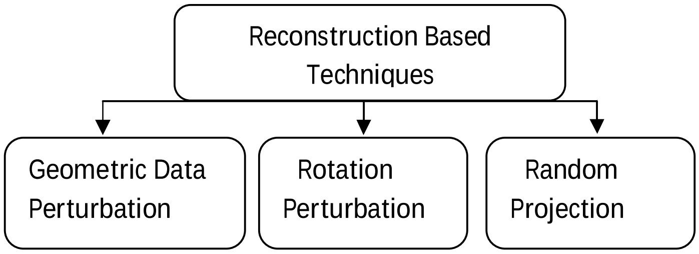

Fig. 1 Reconstruction Based Technique

A. Data perturbation

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 [12].

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 [13] or by the distribution itself [14], and the value distortion approach perturbs data elements or attributes directly by either additive noise, multiplicative noise, or some other randomization procedures [15]. There are three types of data perturbation approaches: Rotation Perturbation, Projection Perturbation and Geometric Data Perturbation.

1) Rotation Perturbation: Rotation perturbation was initially proposed for privacy preserving data clustering . As one of the major components in geometric perturbation, we first applied rotation perturbation to privacy-preserving data classification in our paper and addressed the general problem of privacy evaluation for multiplicative data perturbations. Rotation perturbation is simply defined as G(X) = RX where R d×d is a randomly generated rotation matrix and X d×n is the original data. The unique benefit and also the major weakness is distance preservation, which ensures many modeling methods are perturbation invariant while bringing distance-inference attacks. Distance-inference attacks have been addressed by recent study. we discussed some possible ways to improve its attack resilience, which results in our proposed geometric data perturbation. To be self-contained, we will include some attack analysis in this paper under the privacy evaluation framework. The scaling transformation, in addition to the rotation perturbation, is also used in privacy preserving clustering. Scaling changes the distances; however, the geometric decision boundary is still preserved.

Principal component analysis (PCA) is used to perturb the multidimensional data into lower dimensions. PCA assumes that all the variability in a process should be used in the analysis therefore it becomes difficult to distinguish the important variable from the less important. A dataset Xi ,(i=1,…,n) is summarized as a linear combination of orthonormal vectors(called principal components):

$$f(x,V)=u+(xV)VT \qquad (1)$$

where f(x,V) is a vector valued function, u is the mean of the data {xi}, and V is an d × m matrix with orthonormal columns. The mapping Zi = xiV provides a low-dimensional projection of the vectors xi if m<d. Consequently, Principle component analysis (PCA) replaces the original variables of a dataset with a smaller number of uncorrelated variables called the principle component. [16]

2)Projection Perturbation: Random perturbation[17] refers to the technique of projecting a set of data points from the original Multidimensional space to another randomly chosen space. Let Pk×d be a random projection matrix, where P's rows are orthonormal.

$$G(X) = \frac{\sqrt{d}}{k} PX \qquad (2)$$

is applied to perturb the dataset X. The rationale of projection perturbation is based onits approximate distance preservation, which is shows that any dataset in Euclidean space could be embedded into another space, such that the pair-wise distance of any two points are maintained with small error. As a result, model quality can be approximately preserved. We will compare random projection perturbation to the proposed geometric data perturbation.

3) Geometric Based Perturbation: The basics of imaging geometry in a 2D discrete space[18]. However, the foundations are scalable to other dimensions. A digital image a[m; n] described in a 2D discrete space is derived from an analog image a(x; y) in a 2D continuous space through a sampling process that is frequently referred to as digitization. The 2D continuous image a(x; y) is divided into N rows and M columns. The intersection of a row and a column is termed a pixel. The value assigned to the integer coordinates [m; n] with m = 0; 1; 2; :::;M - 1 and n = 0; 1; 2; :::;N - 1 is a[m; n].

Definition:

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

$$G(X) = RX + T + \Delta.$$

The data is assumed to be a matrix Apq, where each of the p rows is an observation, Oi, 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 v_i€ V is the form v_i = (a1; :::; ad),1 <=i<= d, where$\forall i$ a_i is one instance of Ai, 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[19].

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 [20].

Translation is the task to move a point with coordinates (X; Y) to a new location by using displacements(X_0; Y_0). 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 [19].

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 [20].

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 effects the values of X and Y coordinates [20].

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

Figure 1 (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].

IV. Conclusion

The increasing ability to track and collect large amounts of data with the use of current hardware technology has led to an interest in the development of data mining algorithms which preserve user privacy. We have carried out a survey of the various approaches of Geometric data Perturbation in data mining and briefly explain each and every approaches and its classification. The work presented in this paper, indicates the increasing interest of researchers in the area of recurring sensitive data and acknowledge from malicious users.

References

[1] R. Agrawal and R. Srikant, "Privacy preserving data mining," In Proceedings of SIGMOD Conference on Management of Data, pp. 439-450, 2000.

[2] D. Agrawal and C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithm," In Proceedings of ACM SIGMOD, pp. 247-255, 2001.

[3] A. Evfimieski, R. Srikant, R. Agrawal and J. Gehrke, "Privacy preserving mining of association rules," In Proceedings of the 8th ACM SIGKDD, pp. 217-228, 2002.

[4] W. Du and Z. Zhan, "Using randomized response techniques for PPDM," In Proceedings of the 9th ACM SIGKDD, pp. 505-510, 2003

[5] K. Liu, H. Kargupta and J. Ryan, "Random projection-based multiplicative perturbation for privacy preserving distributed data mining," IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 1, pp. 92-106, 2006.

[6] J. Ma and K. Sivakumar, "Privacy preserving Bayesian network parameter learning," 4th WSEAS International Conference on Computational Intelligence, Man-machine Systems and Cybernetics, Miami, Florida, November, 2005.

[7] J. Ma and K. Sivakumar, "A PRAM framework for privacy-preserving Bayesian network parameter learning," WSEAS Transactions on Information Science and Applications, vol. 3, no. 1, 2006.

[8] 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, November, pp 99-106, 2003

[9] Z. Huang, W. Du and B. Chen, "Deriving private information from randomized data," In Proceedings of SIGMOD, pp 37-48, USA, 2005.

[10] K. Chen, G. Sun and L. Liu, "Towards attack-resilient geometric data perturbation," In Proceedings of the 7th SIAM International Conference on Data Mining, pp 26-28, USA, 2007.

[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.VidyaBanu 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.VidyaBanu 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 GeometricTransformation, 2009

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