A Survey On Privacy Preserving Data Mining Using Geometric Data Perturbation Technique

Prof. Gargi Shah
Assistant Professor , Dept of Computer Engineering
Vadodara Institute of Engineering & Research, Vadodara, Gujarat, India

Abstract- *Data perturbation is one of the popular data mining techniques for privacy preserving. A major issue in data perturbation is that how to balance the two conflicting factors – protection of privacy and data utility. This paper proposes a Geometric Data Perturbation (GDP) method using data partitioning and three dimensional rotations. In this method, attributes are divided into groups of three and each group of attributes is rotated about different pair of axes. The rotation angle is selected such that the variance based privacy metric is high which makes the original data reconstruction difficult. As many data mining algorithms like classification and clustering are invariant to geometric perturbation, the data utility is preserved in the proposed method. The experimental evaluation shows that the proposed method provides good privacy preservation results and data utility compared to the state of the art techniques.*

Keywords- data mining, data perturbation, variance, three dimensional rotation, privacy preserving.

I. INTRODUCTION

Data mining efficiently discover valuable, non-obvious information from large datasets, is particularly vulnerable to abuse. A fruitful future research leadership in data mining is the development of technology that incorporates the concern for privacy. A recent survey of web users 17% of respondents as privacy fundamentalists, the unclassified data on a site, even if privacy measures are in place [1].Nowadays organisms around the world are dependent on mining gigantic datasets. These datasets typically contain delicate individual information inevitably all is exposed to the various parties. Consequently privacy issues are constantly in the limelight and the public dissatisfaction which may well threaten the exercise of data mining.

There is much research on privacy-preserving data mining (PPDM) [6] malfunctioning, randomization and secure multi-party system based calculations. More recently, there has been much research on anonymity including k-anonymity and 1-diversity. As a result now have numerous preserving algorithms.

Many government agencies, businesses and non-profit organizations to support their short-and long-term schedule activities, to collect for a way to store, analyze and report data on persons, households or businesses looking. Information systems therefore contain confidential information such as social security numbers, income, credit ratings, type of illness, customer purchases, etc., that 'need to be adequately protected. With the Web revolution and the emergence of data mining, have privacy concerns provided technical challenges fundamentally different from those that occurred before the information age [3].

The understanding of privacy in data mining requires understanding how privacy can be violated [5], the can means clustering and clustering on the prevention of invasion of privacy. Usually carries a significant factor in breach of Private in data mining: the misuse of data. Users' privacy can be violated in different ways and with different intentions. Although data mining can be very useful in many applications, it is also on the lack of adequate safeguards may violate informational privacy. Privacy can be violated are when personal data for other purposes after the original transaction brokers are an individual and an organization if the information was collected, used. Malthus when personal data are exposed to mining, the attribute values associated with private persons and must be protected from disclosure.

Users' privacy can be violated in different ways and with different intentions. Although data mining can be extremely valuable in many applications it can also, in the absence of adequate safeguards, violate informational privacy. Privacy can be violated if personal data are used for other purposes subsequent to the original transaction between an individual and an organization when the information was collected. When personal data are subjected to mining, the attribute values associated with individuals are private and must be protected from disclosure [4].

One of the sources of privacy violation is called data magnets. Data magnets are techniques and tools used to collect personal data. Examples of data magnets include explicitly collecting information through on-line registration, identifying users through IP addresses, software downloads

that require registration, and indirectly collecting information for secondary usage. In many cases, users may or may not be aware that information is being collected or do not know how that information is collected. In particular, collected personal data can be used for secondary usage largely beyond the users' control and privacy laws. This scenario has led to an uncontrollable privacy violation not because of data mining itself, but fundamentally because of the misuse of data.

Securing against unauthorized accesses has been a long-term goal of the database security research community and the government research statistical agencies. Solutions to such a problem require combining several techniques and mechanisms. In an environment where data have different sensitivity levels, this data may be classified at different levels, and made available only to those subjects with an appropriate clearance.

Clustering is a well-known problem in statistics and engineering, namely, how to arrange a set of vectors (measurements) into a number of groups (clusters). Clustering is an important area of application for a variety of fields including data mining, statistical data analysis and vector quantization. The problem has been formulated in various ways in the machine learning, pattern recognition optimization and statistics literature. The fundamental clustering problem is that of grouping together (clustering) data items that are similar to each other. Given a set of data items, clustering algorithms group similar items together. Clustering has many applications, such as customer behavior analysis, targeted marketing, forensics, and bioinformatics [7].

Small companies have recognized the value in data, especially with the introduction of the knowledge discovery process. However, small companies do not have enough expertise for doingdata analysis, although they have good domain knowledge and understand their data.

II. GEOMETRIC DATA PERTURBATION

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

$$G(X) = RX + \Psi + \Delta$$

The data is assumed to be a matrix A[p×q], 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 vi€ V is the form vi = (a1; :::; ad),1 <=i<= d, where∀iai is one instance of Ai, ai€ 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 vi 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 acts the values of X and Y coordinates.

$$\begin{bmatrix} 1 & 0 & x0 \\ 0 & 1 & y0 \end{bmatrix}$$

Fig -1(a): Translation Matrix

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

Fig -1(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 $\Delta d \times n$, 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.

III. CONCLUSIONS

This paper presents a novel privacy preserving data transformation technique that can be used with different types of data mining models. Moreover in geometric data perturbation technique data is not much affecting the data mining capability of the data mining model because of preservation of Euclidean distance. Also the data mining accuracy of the original and perturbed data are nearly same and has high variance, which shows its high privacy preserving capability.

REFERENCES

[1] Majid Bashir Malik and M. Asger Ghazi and Rashid Ali ;"Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects"; Third International Conference on Computer and Communication Technology; 978-0-7695-4872-2/12

[2] Hitesh Chhinkaniwala and Dr. Sanjay Garg "Privacy Preserving Data Mining Techniques: Challenges & Issues" in Proceedings of International Conference on Computer Science & Information Technology, CSIT – 2011,p.609

[3] W.T. Chembian1, Dr. J.Janet, "A Survey on Privacy Preserving Data Mining Approaches and Techniques",in Proceedings of the Int. Conf. on Information Science and Applications ICISA 2010,6 February 2010, Chennai, India

[4] Xiaolin Zhang and Hongjing Bi; "Research on Privacy Preserving Classification Data Mining Based on Random Perturbation"; International Conference on Information, Networking and Automation (ICINA); 978-1-4244-8106-4

[5] Ching-Ming Chao, Po-Zung Chen and Chu-Hao Sun ;"Privacy-Preserving Classification of Data Streams"; Tamkang Journal of Science and Engineering; Vol. 12, No. 3, pp. 321_330 (2009)

[6] Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino,Loredana Parasiliti Provenza, Yucel Saygin, Yannis Theodoridis "State-of-the-art in Privacy Preserving Data Mining"

[7] Haisheng Li;" Study of Privacy Preserving Data Mining" ;Third International Symposium on Intelligent Information Technology and Security Informatics; 978-0-7695-4020-7