Classification Techniques for Geometric Data Perturbation in Multiplicative Data Perturbation

[1]Mr.Keyur Dodiya,[2]Shruti Yagnik
M.E. Scholar, Assistant Professor
L.J Institute of Engineering & Technology, Gujarat, India

Abstract **- It is very important to be able to find out useful information from huge amount of data. In this paper we address the privacy problem against unauthorized secondary use of information. To do so, we introduce a family of geometric data transformation methods (GDTMs) which ensure that the mining process will not violate privacy up to a certain degree of security. We focus primarily on privacy preserving data classification methods. Our proposed methods distort only sensitive numerical attributes to meet privacy requirements, while preserving general features for classification analysis. Our experiments demonstrate that our methods are effective and provide acceptable values in practice for balancing privacy and accuracy. This paper focuses on Geometric Data Perturbation to analyse large data sets.**

Keywords **- Data mining; Privacy preserving; data perturbation; randomization; cryptography; Geometric Data Perturbation**

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]. A more recent study of web users found that 86% of respondents believe that information for participation in benefits programs is a matter of individual choice privacy [2].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 May well threaten the exercise of data mining. It is of great importance used technical security to protect the confidentiality of individual values for data mining for the development of appropriate Malthus.attacks and only good for very few specific data mining models. The condensation approach (Aggarwal and Yu, 2004) 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 k-anonymization (LeFevre, DeWitt and Ramakrishnan, 2006) is only designed for general-purpose utility preservation and may result in low-quality data mining.

There is much research on privacy-preserving data mining (PPDM) [3] malfunctioning. randomization and secure multi-party system based calculations. More recently, there has been much research on anonymity Including k-anonymity and l-diversity. As a result, we now have numerous privacy and anonymity 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 [4]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 [5].

Classification- is the process of finding a model that describe and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. If-then rules, a decision tree and neural network are models in which classification can be represented[6].

Clustering- is a technique used to place data elements into related groups without advance knowledge of the group definition. K-means clustering and expectation maximization (EM) clustering are popular clustering techniques.[7]. The organization is collecting data for analysing organization's policy, customer's behaviour and getting improve its strategies. Collection of data is used for different data mining purpose like statistical analysis, decision point of view, knowledge gathering etc. During this time, we have to also protect the sensitive data from the researcher. So, before releasing dataset, sensitive data will be hidden from unauthorized researchers. This issue can be solved with the use of privacy preserving in data mining. Privacy preserving data mining has become increasingly popular because it allows sharing of privacy sensitive data for analysis purposes [8].

Privacy preserving in data mining (PPDM) is used for hiding sensitive knowledge. This sensitive knowledge is concentrate on data heuristic approach, data modification approach and data cryptography approach. This approaches are vary from researcher to researcher because some researcher may think only certain attribute value should private and some may think whole data column should be private. Privacy preserving in data mining is classified into:[9]

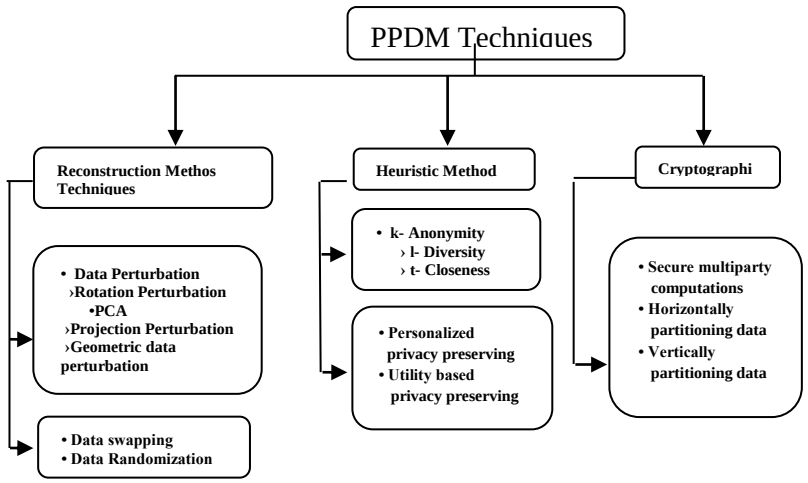


Fig. 1 PPDM techniques

II. 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.

$$G(X) = RX + \Psi + \Delta \text{ [10].}$$

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 [11].

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 [12].

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 [11].

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

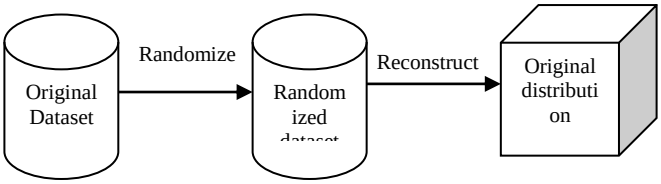
Figure 2 (a) Translation Matrix (b) Rotation Matrix

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.

This perturbation technique is a combination of Rotation, Translation and Noise addition perturbation techniques. The additional components ψ and Δ are used to address the weakness of rotation perturbation while still preserving the data quality for classification modeling. Concretely, the random translation matrix addresses the attack to rotation center and adds additional difficulty to ICA-based attacks and the noise addition addresses the distance-inference attack [13].

If the matrix $X_{d\times n}$ indicates original dataset with *d* columns and *N* records, $R_{d\times d}$ be a orthonormal random matrix, ψ be a translation random matrix and $\Delta_{d\times n}$ be a random noise matrix, where each element is Independently and Identically Distributed (iid) variable like Gaussian distribution N(0,σ2), the geometrical perturbation will be defined as following [14]:

$$G(X) = RX + \psi + \Delta \qquad (3)$$



III. ALGORITHM

Algorithm: Geometric Data Perturbation G(X)

For each attribute of **G(X),** let **R** be random rotation **X** be a original dataset, **T** be a translation and D be a Gaussian noise then the value of attribute **G(X)** is calculated using following formula[15].

$$G(X) = R*X + T + D$$

Step-1 The data stream D is taken from large dataset in a proper data with sensitive attribute
Step-2 Provide this data to the Rotation Transformation with original dataset. R x D
Step-3Then find out mean of original dataset and compute with each attribute
Step-4 Apply Gaussian Noise for data preprocessing with the help of Gaussian noise. We cannot easily identify original data without any formula

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where, μ=Mean, σ=Variance
Step-5 We get perturbed data D‘
Step-6Apply Weka classification algorithm for or D & perturbed data D‘
Step-7 Classification algorithm like J248/ Naïve Bayesian algorithm to provide accuracy & privacy for our perturbed dataset D’

IV. RESULTS AND DISCUSSION

Series of experiments were performed over define the classification accuracy. Our evaluation approach focused on the overall quality of generated classifier after dataset perturbation[16].

Experiment was based on following steps

- Setup each dataset as WEKA framework.
- Define Data to evaluate measures and classification membership matrix.
- Modified all the instances in WEKA framework by applying our proposed data perturbation method to protect the sensitive attribute value.
- J248/ Naïve Bayesian classification algorithm is used to find the Correctly Classified for our performance evaluation. Our selection was influenced by (a) J248/ Naïve Bayesian classification is one of the best known Classification Algorithm and is scalable.
- Compare how closely each classified value in the perturbed dataset matches its corresponding classified value in the original dataset. We expressed the quality of the generated classifier by computing the Kappa Statistics and Mean Absolute Error[17].

Experiments were performed to measure accuracy while protecting sensitive data. We here presents two different results, one is corresponding to classified accuracy in terms of membership matrix which was manually derived from classified result.

	Adult - Age				Adult - Education Num			
	NB		J48		NB		J48	
	original	Perturb	original	Perturb	original	Perturb	original	Perturb
correctly classified instances	0.8342	0.8318	0.8621	0.8573	0.8342	0.8291	0.8621	0.8562
incorrectly classified instances	0.1657	0.1681	0.1378	0.1426	0.1657	0.1708	0.1378	0.1437
Time taken	0.2	0.23	4.52	4.18	0.2	0.22	4.84	5.04
Kappa statistic	0.4993	0.4905	0.6004	0.5805	0.4993	0.475	0.6004	0.5721
Mean Absolute Error	0.1735	0.1759	0.1942	0.2009	0.1735	0.1771	0.1942	0.2031
Root Mean Squared Error	0.3723	0.3756	0.3196	0.3246	0.3723	0.3152	0.3196	0.3297
Relative Absolute Error	0.4745	0.4809	0.5309	0.5495	0.4745	0.4844	0.5309	0.5553
Root Relative Squared Error	0.8706	0.8783	0.7474	0.7592	0.8706	0.8772	0.7474	0.7711

Table 1: Classification on Adult Dataset

	Bank - Age				Bank - Duration			
	NB		J48		NB		J48	
	original	Perturb	original	Perturb	original	Perturb	original	Perturb
correctly classified instances	0.8807	0.8805	0.9031	0.903	0.88	0.866	0.9031	0.8924
incorrectly classified instances	0.1193	0.1195	0.0968	0.0969	0.1193	0.1339	0.0968	0.1075
Time taken	0.43	0.45	6.5	7.17	0.44	0.46	7.72	7.94
Kappa statistic	0.4391	0.4346	0.4839	0.4846	0.4391	0.3413	0.4839	0.3354
Mean Absolute Error	0.1532	0.1542	0.1269	0.1276	0.1532	0.1681	0.1269	0.157
Root Mean Squared Error	0.3088	0.3075	0.2773	0.2781	0.3088	0.3305	0.2773	0.2986
Relative Absolute Error	0.7416	0.7464	0.6142	0.6175	0.7416	0.8135	0.6142	0.7596
Root Relative Squared Error	0.9606	0.9567	0.8628	0.8653	0.9606	1.028	0.8628	0.9289

Table 2: Classification on Bank Dataset

V. CONCLUSIONS

The increasing ability to track and collect large amounts of data with the use of current hardware technology has lead 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 privacy preserving 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. We conclude that we have reached from reviewing this area, manifest that privacy issues can be effectively consider only within the limits of certain privacy preserving data mining approaches[19].

REFERENCES

[1] L.F.Cranor,J.Reagle and M.S.Ackerman. "Be-yond concern: Understanding netusers' attitudes about on line privacy".Technical Report TR99.4.3, AT&TLabs Research, April1999.

[2] A.F.Westin. "Freebies and privacy: What net users think". Technical report, Opinion ReSearch Corporation, July1999. Available from http://www.privacyexchange.org/iss/surveys/sr990714.html.

[3] Stanley R.M Oliveira, Osmar R. Zaiane, "Towards Standardization in privacy Preserving data Mining" In the proceeding of IEEE International Conference on data Mining 2006

[4] Rakesh Agarwal, Ramakrishnan Srikant, "Privacy Preserving Data Mining "In Proceedings of the 1st International Conference on Knowledge Discovery and Databases

[5]] P. Samarati, "Protecting respondent's privacy in micro data release", In IEEE Transaction on Knowledge and Data Engineering, 2001, pp.1010-1027.

[6] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining", In Proc of ACM SIGMOD, 2004, pp. 50–57.

[7] Ackerman, M. S., Cranor, L. F., and Reagle, J, "Privacy in e- commerce: examining user scenarios and privacy preferences", In Proc. EC99, 1999, pp. 1-8.

[8] Vassilios S.Verylios,E.Bertino,Igor N,"State –of-the art in Privacy preserving Data Mining",published in SIGMOD 2004 pp.121-154.

[9] L.Golab and M.T.Ozsu ,Data Stream Management issues-"A Survey Technical Report",2003.

[10] Keke Chen, Ling Liu," Geometric data perturbation for privacy preserving outsourced data mining", Springer,2010.

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

[12] Jie Liu, Yifeng XU, "Privacy Preserving Clustering by Random Response Method of GeometricTransformation", IEEE 2010

[13] S. R. M. Oliveira, O. Zaiane and Y. Saygin, Secure Association Rule Sharing. PAKDD Conference, 2004.

[14] Y. Saygin, V. Verykios and C. Clifton, Using Unknowns to prevent discovery of Association Rules. ACM SIGMOD Record, Vol 30, Issue 4, 2001.

[15] Y. Saygin, V. Verykios and A. Elmagarmid, Privacy Preserving Association Rule Mining. 12th International Workshop on Research Issues in Data Engineering, 2002.

[16] 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.

[17] S. Nabar, B. Marthi, K. Kenthapadi, N. Mishra and R. Motwani, Towards Robustness in Query Auditing. VLDB Conference, 2006.

[18] K. Kenthapadi, N. Mishra and K. Nissim, Simulatable Auditing. ACM PODSConference, 2005.

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