Privacy Preserving Secure Classification using Multiplicative Data Perturbation at Multilevel Trust

R.Mynavathi
Assistant Professor, Department of Information Technology
Velalar College of Engineering and Technology,Erode, TamilNadu, India

Dr.S.Malliga
Professor, Department of Computer Science and Engineering
Kongu Engineering College, Perundurai, TamilNadu, India

P.Rajendran
Assistant Professor, Department of Computer Applications
Velalar College of Engineering and Technology,Erode, TamilNadu, India

Abstract- The field of Privacy Preserving Data Mining has gained specific area of interest for researchers due to the impact of various security issues. The problem of developing specific models without accessing private information is addressed. With the voluminous growth of data, threat to individual's private information also grows. Developing useful data mining models without accessing private information but with better data mining utility has become a major concern. Many studies on data perturbation techniques for protecting sensitive data focus on adding noise to the original data. Manipulating Gaussian noise to the sensitive data has somehow balanced the privacy preservation and the utility of data mining. This paper deals with perturbing sensitive data using multiplicative Gaussian noise and builds a secure kNN classifier model that provides secured mining. We propose an efficient approach that aims to provide better secured data mining result with minimum information loss.

Keywords – Privacy preserving data mining, data perturbation, Gaussian noise, kNN classifier

I. INTRODUCTION

Data mining is concerned with extracting useful information represented in models and patterns. Infinite volumes of data streams are generated by various networks, online transactions, surveillance systems, all types of industries and other dynamic environments. The data collected from these sources needs to be analyzed for predicting future behavior and segregating anomalies. However, data owners may not be willing to share their data mostly due to privacy considerations. In recent years, Privacy preserving data mining has been extensively studied. Several techniques ranging from perturbation to secure multi party computation have been explored. Various techniques for privacy preserving data mining are discussed in [1][2][8]. Technique of random data perturbation was initiated by Agrawal and Srikant[3]. More solutions based on data perturbation can be found in [4][5][6]. In this paper, we focus primarily on the perturbation technique. These techniques are usually used in scenarios where individuals can perturb their private data with some known random noise and report the perturbed data to the data miner. Since the distribution of the added noise is known, the data miner could reconstruct the original distribution using different statistical methods and mine the reconstructed data. The work presented here addresses this issue of data privacy. The paper is organized with next section presenting the background work related to the proposed framework, Section 3 introducing the preliminary needed for the framework and Section 4 formulating the problem and providing an overview of the proposed framework. Section 5 analyses the result of applying the proposed methodology to the census data set from online repository. The last section concludes with the summary of work.

II. RELATED WORK

Charu.C.Aggarwal in [2] introduced the concept of Privacy Preserving Data Mining. Earlier research in this field can be broadly categorized under two different categories. First approach being disguising the sensitive data to some other format to preserve privacy. The second approach is based on cryptographic techniques to protect sensitive data. These cryptographic techniques include generation of cipher text based on key. There are different solutions addressing various data mining problems. Privacy preservation techniques should not only preserve the sensitive data but also enable better data utility. Different methodologies of disguising the data using perturbation schemes can be further grouped as follows

A. Modifying the data using Data Perturbation

B. Disguising the final Data Mining result

C. Perturbation of Data under distributed environments of industries

A. Disguising the data–

Techniques for modifying the data into another format include data perturbation [3], swapping the data and data randomization[9]. The sensitive attributes are modified with a new value in all these approaches. Various other techniques involving disguising the data are discussed in [7][8].

B. Disguising the Data Mining result –

Privacy preservation could also be done after the data mining process. In this method the data mining models are preserved for privacy. Classifier effectiveness, Query auditing, Inference control, Association Rule hiding are different approaches under this category. All these techniques work with data that are with minimum alteration[11][12][13].

C. Distributed Data Perturbation –

Privacy of horizontally partitioned data and vertically partitioned data can be addressed using perturbation schemes under distributed environment. Secure multiparty computation works under this distributed privacy preserving model. In distributed environment, secure multiparty computation is when two parties perform an analysis based on their private inputs, without disclosing their own output to anybody else. In horizontal partitioning the database records lie in different places. In vertical data partitioning all the values for different attributes resides in different places [14]. Different approaches needs to be applied for horizontal and vertical partitioning of data. A variety of cryptographic techniques are used for preserving horizontally and vertically partitioned data.[15].

III. PRELIMINARIES

A. Gaussian Noise –

The paper focuses on multiplying Gaussian noise to the sensitive data. The added noises are assumed to be jointly Gaussian. Jointly Gaussian property is satisfied if and only if each of the individual Gaussian noise is a linear combination of multiple independent Gaussian random variables. Gaussian vector has a probability density function as follows

$$\text{pdf}(Y) = \left(\frac{1}{\sqrt{(2\pi)^L det(K_G)}}\right)\left(e^{-(g-\mu_G)^T . K_G^{-1} . (g-\mu_G)/2}\right) \quad (1)$$

B. Data perturbation with Gaussian noise

The multiplicative data perturbation is a technique for privacy preserving data mining in which noise is added and multiplied with the data in order to mask the attribute values of records. The noise added is sufficiently large so that individual record values cannot be recovered. Multiplicative data perturbation approach is widely used and accepted method to perturb data. The methodology involves adding some random noise, D to the original data X as follows

$$Y = R*X + T + D \quad (2)$$

Here X, Y and Z are assumed to be N dimensional vectors. R is random rotation process, T is random Translation process. The original data X follows a distribution with mean μ_x and covariance K_x. The noise D is assumed to be independent of X and is jointly Gaussian with zero mean and covariance matrix Kz.

$$K_Y = K_X + K_Z \quad (3)$$

C. kNN classfier

The K-Nearest neighbor algorithm is the simplest of all machine learning algorithms. In this, an object is classified by a majority vote of its neighbors. The object is consequently assigned to the class that is most common among its kNN, where k is a positive integer that is typically small. If k=1, then the object is simply assigned to the class of its nearest neighbor. K Nearest Neighbor is one of the most fundamental and simple classification methods and should be one of the first choices for a classification study where there is little or no prior knowledge about the distribution of the data. K Nearest neighbor classification was developed from the need to perform discriminant analysis when reliable parametric estimates of probability densities are unknown are difficult to determine. Classification by kNN is done by measuring the distance of the instance with the known instances. The algorithm is efficient to compute more than one class labels for the unknown instance.

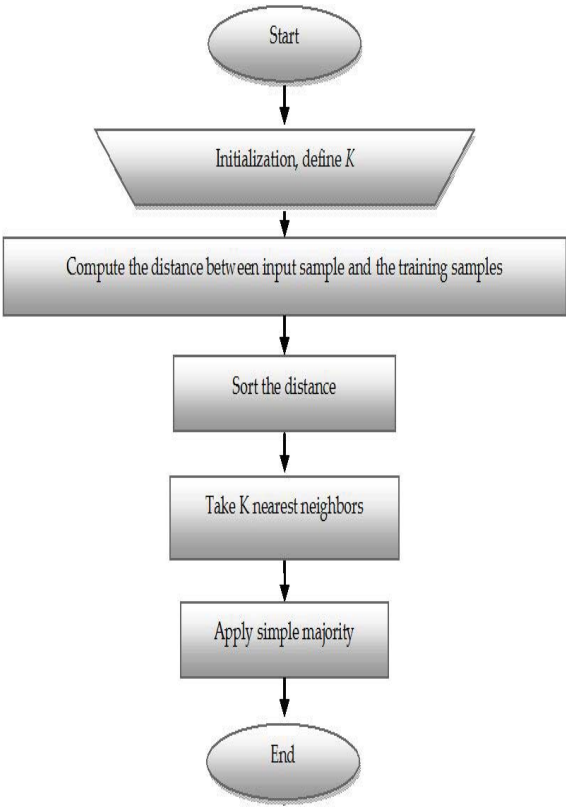

Figure 1. kNN Algorithm

IV. PROBLEM FORMULATION AND PROPOSED SOLUTION

A. Problem Setting

The proposed framework focuses on solving the classification problem over perturbed data. A secure kNN classifier over the perturbed data is developed. Perturbation is done by adding varying amount of noise to the data. The perturbed data is used to solve the classification problem. It is assumed that the data records corresponding to the k-nearest neighbors and the output class label are not known.

B. The Privacy model

The privacy model is based on introducing noise to the sensitive data without significantly changing the distribution of the original data. Information loss and privacy preservation is always a trade off. The extent to which we perturb the original data can dramatically affect the data mining results. Hence the privacy model should be generated based on both utility and security. In the proposed scheme, the data owner determines the level of trust that is to be imposed and generates perturbed copies of the data. The trust levels are predefined and noises are generated accordingly.

Algorithm 1: Generation of perturbed data
Input : Original sensitive data, covariance matrix of original data, trust level (which is assumed to be one)
Output : Perturbed data

- With the given trust level (σ^2_z) and covariance matrix of original data, generate covariance matrix of the noise to be added.
- Generate noise vector with the covariance matrix according to the probability density function for Gaussian noise
- Generate perturbed data Y = R*X + T + D, where R is random rotation, T is random translation and D is the noise generated from the above steps.
- Output Y

C. Secured kNN classifier model
From the perturbed data generated by the privacy model, classifier model is built.

Algorithm 2 : Generation of secure kNN classifier

Input **:** Perturbed data
Output : Classifier model

- Determine the parameter K which specified the number of nearest neighbors
- Calculate the distance between the query instance and all the training samples
- Sort the distance and determine the nearest neighbors based on k^{th} minimum
- Gather the category of the nearest neighbors
- Use majority of the category of nearest neighbors as the prediction value of the instance.

V. EXPERIMENTS AND RESULTS

We conducted our experiments on census bureau database found at www.census.gov. The dataset contains 16,000 tuples with 15 attributes. We took the Age and Income attributes and ran our algorithm. For each attribute X (age and income), the perturbed data is generated by adding Gaussian noise according to the perturbation magnitude which is determined by the data owner. We then execute our random query over the perturbed data using secured classifier model. The following figures show the utility of perturbed copies using kNN classifier model. It is identified that the perturbed copy has the same utility as that by the original scheme

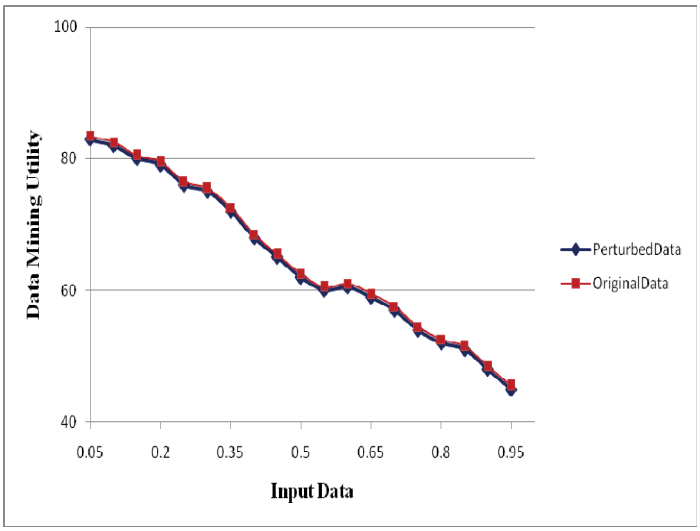

Figure 2. Data Utility comparision

VI.CONCLUSION

User's privacy is protected by different privacy preserving schemes. In this paper, we implement the scope of geometric data perturbation to the database environment. We propose a novel privacy preserving kNN classifier model over the perturbed data. The model preserves the privacy and hides the data access patterns. We assume that the noise added to the perturbed model is at single trust level. We prove that the mining utility of the perturbed data is similar to that of the original data. There are many interesting directions that are worth exploring. The generation of noise could be done at different levels based on different privacy requirements. Various other classification models can be analyzed with the perturbed model. Studying the different kinds of attacks over these models is an interesting future direction.

REFERENCES

[1] Verykios VS, Bertino K, "State-of-the-Art in Privacy Preserving Data Mining", ACM SIGMOD, Record.33,pp 50-57

[2] Charu C.Aggarwal, Philip S.Yu, "Privacy – Preserving Data Mining : Models and Algorithms"

[3] Agrawal R & Srikant R, 2000, "Privacy-preserving data mining", In proceeding of the ACM SIGMOD conference on Management of Data, ACM press, pp 439-450.

[4] Agrawal D & Aggarwal C, 2001, "On the design and quantification of privacy preserving data mining algorithms", In proceedings of the 20th ACM SIGACT SIGMOD SIGART symposium on Principles of Database systems, pp 247-254.

[5] Polat H & Wenliang D, 2003, "Privacy preserving collaborative filtering using randomized perturbation techniques", In proceedings of the 3rd IEEE international conference on Data Mining, pp 625-628.

[6] Keke Chen, Ling Liu, 2005, "Privacy preserving Data Classification with Rotation Perturbation", Technical Report.

[7] Mynavathi.R , Sowmiya.N, Vanitha.D, 2014, "Survey of Various Techniques to Provide Multilevel Trust in Privacy Preserving Data Mining", International Journal of Innovative Research in Computer and Communication Engineering, Vol.2, Pg.118-123.

[8] Mynavathi.R, Rajendran.P, Dr.Malliga.S, 2015, "An insight into recent developments towards Privacy Preserving techniques for Secure Data Mining", International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10,Pg.28406-28410.

[9] Hillol Kargupta and Souptik Datta, Qi Wang and Krishnamoorthy Sivakumar, 2003,"On the Privacy Preserving Properties of Random Data Perturbation Techniques",IEEE International Conference on Data Mining, Pg.99-106.

[10] Xinjun Qi , Mingkui Zong, 2012,"An Overview of Privacy Preserving Data Mining", ScienceDirect, Procedia Environmental Sciences 12,Vol.12 ,Pg.1341 – 1347.

[11] Tamir Tassa, 2014,"Secure Mining of Association Rules in Horizontally Distributed Databases", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, Pg. 970 - 983.

[12] Vaidya.J and Clifton.C.W, 2002,"Privacy Preserving Association Rule Mining in Vertically Partitioned Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, Pg.639-644.

[13] Yehuda Lindell and Benny Pinkas,2002, "Secure Multiparty Computation for Privacy-Preserving Data Mining ",The Journal of Privacy and Confidentiality,Pg.59-98.

[14] Vaidya.J and Clifton.C.W,2003, "Privacy-Preserving K-Means Clustering over Vertically Partitioned Data," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,Pg.206-215.

[15] Xun Yi, Mohammed Golam Kaosar, Russell Paulet, and Elisa Bertino, Fellow, IEEE, 2013,"Single-Database Private Information Retrieval from Fully Homomorphic Encryption", IEEE Transactions On Knowledge And Data Engineering, Vol. 25,Pg. 1125 – 1134.