ISSN: 0976-3104
SPECIAL ISSUE: Emerging trends in Computer Engineering and Research (ECER)

ARTICLE

PRIVACY PRESERVING IN DATA MINING USING DATA PERTURBATION AND CLASSIFICATION METHOD

Megha Dabhade[1], J Jabanjalin Hilda[2]*

[1]*School of Computational Intelligence and Engineering, Vellore Institute of Technology, Vellore, INDIA*

[2]*Faculty of school of Computational Intelligence and Engineering, Vellore Institute of Technology, Vellore, INDIA*

ABSTRACT

In today's information era, Tera bytes of data is being generated every second. It contains huge private and confidential information such as social networking, health care, finance, sensors data, criminal records etc. Data mining deals with such automatic generated data for different purposes. During such activities data gets exposed to other parties and protecting data becomes a challenge. The solution to this is provided by Privacy Preservation in Data Mining (PPDM). PPDM is the novel approach where different techniques are used to protect the privacy of data being used for Data Mining purpose. In this paper, we have done a comparative study on different data perturbation based privacy preserving methods and analyzed which one is more effective. In addition to this, Classification, the most commonly used Data Mining technique is used to develop a PPDM model. The experimental results demonstrate that how privacy is protected with respect to various privacy metrics.

KEY WORDS
PPDM, Data Mining, Classification, Data perturbation

INTRODUCTION

Recent advances in communication, computing and digital technology data is growing incredibly day by day. Such data is distributed across geographical and administrative boundaries and it demands for a powerful technique to manage and analyze such huge amount of data, called Data Mining. Data mining is the process of extracting non-trivial and potentially useful knowledge or patterns from multiple large data sources [1]

Data mining systems deals with large amount of private and confidential data from social networking, healthcare, defense activities etc. This kind of information is non-sharable, protecting such data has become an important challenge and new research stream in the field of data mining. The concept of protecting confidential and sensitive information used during different data mining activities is termed as Privacy Preservation in Data Mining (PPDM) [1] [2]. PPDM is the new techniques which not only allows us to extract useful information and provide accurate results, but also helps to prevent loss of sensitive data [3].

Received: 11 June 2017
Accepted: 7 Aug 2017
Published: 23 Sept 2017

Privacy preservation transforms dataset which contains confidential information into modified or altered form. Again, most importantly this information is hidden from unauthorized users. Privacy preserving data mining i.e. PPDM is the emerging area in data mining where efforts are being made to protect private information from unauthorized revelation. Privacy Preservation Data Mining [1] was introduced by keeping security of sensitive information as a priority in data mining process and to furnish canonical data mining process. A large portion of these security protection methodologies were proposed to safeguard private data of test dataset. Then again, protection safeguarding process which conceals data, may decrease utility of these altered dataset. At the point when their utility reduces to a specific level, the minimized data may lead to inaccurate analysis [1] .

This paper mainly focuses on how PPDM can be effectively used using Data Perturbation and Classification method. Data perturbation is widely used technique for privacy preservation. In this technique, data which is to be processed, is modified before passing to data mining process. There are different ways to modify data like data distortion, data swapping, noise addition, data hiding etc. [2] [3]. Among these method data distortion is proved to be most popular and effective method

Classification is the most commonly used data mining technique to build model. Its main objective is to build a classifier to identify class label of data based on training data [4][5]. Classifier can be represented by using decision trees, Naïve Bayes classifier, Neural Networks, SVM etc. In this paper, we mainly focus on issues related with PPDM using decision trees. Privacy preservation of individual data and accuracy of constructed classifiers examines the performance of privacy preserving technique [6][7].

In this paper, Decision tree is used as classification method. This is supervised learning. The complex decisions are further dived into smaller decisions. The complexity is controlled by the pruning technique. It can handle both numerical and categorical data. Decision trees are also easy to interpret. More precisely, ID3 algorithm is used as decision tree algorithm. For implementing such trees, ID3 is most efficient among machine learning approaches. In this approach tress are constructed based on entropy or information gain values. The original data set is divided into training data and testing data. The classifier is calculated based on training data and it is further used to predict the class for testing data [8][9].

***Corresponding Author**
Email:
jabanjalin.hilda@vit.ac.in
Tel.: +91 7598193077

IIOAB JOURNAL
ISSN: 0976-3104

RELATED WORK

The privacy preservation in data mining can be done in two ways. In the first approach, data mining algorithms are modified without any knowledge of data. While in second approach, methods modify the datasets values to keep the privacy of data safe. In this paper, we are concentrating on second approach, many researches has been carried out in data perturbation. Some of them are as follows:

Santhosh et al. were the first to propose the idea of Data swapping in the year 1982. This is transformation technique where dataset is modified by altering the values of attributes of datasets from selected records. The data swapping is proved to be a distinguished data perturbation technique for privacy preservation [10].

Naga et al. in 2006 have proposed Singular Value Decomposition (SVD) approach which based on data distortion approach. They have used real world datasets for their experiment and proposed further innovation called sparsified SVD [11] [12]. The experimental results showed that this new sparsified SVD is efficient in preserving privacy and it also maintains the performance of the datasets.

Aldeen et al. in the year 2012, have proposed data distortion strategies viz. SVD and sparsified SVD along with feature selection. Their main objective was to reduce feature space in features. There are different kinds of privacy metrics which assess the utility of data. These measures calculate the performance of data mining processes by finding the difference between original dataset and distorted dataset and what is the degree of privacy protection. The real-world dataset was used for experiment and the results demonstrated a feasible solution with the use of sparsified SVD than only SVD. [13-16]

Mohammed et al In the year 2011, a study was carried out on intuitionistic i.e. a proof of contradiction method, for fuzzy clustering and application of fuzzy k-member clustering to protect privacy concretely in pattern recognition. K-member clustering is k-anonymity clustering technique in which data samples are summarized so that every sample is different from at least (k – 1) other samples. To improve quality of data summarization with k-anonymity, a fuzzy variation of k-member clustering was proposed. A secure framework was proposed for handling both vertically and horizontally distributed data in case of fuzzy co-clustering[17]

Kake et al. proposed Fuzzy based PPDM in the year 2016 where fuzzy-based mapping techniques were compared in terms of their privacy-preserving feature and their ability to employ exactly same relationship with other fields. [18] A fuzzy c-regression method was used to generate synthetic data on which statistical computations were done by third party. In fuzzy clustering approach. This method effective because it collects records into clusters where each record is not recognizable from others after within-cluster merging. Hence lossless data anonymization can be achieved.

METHOD

Classification

In this paper Decision tree is used as Classification method. In the field machine learning and statistics, the decision tree algorithm is called as “Predictive modelling technique” which build a simple tree to construct the pattern of classification data. Decision is being popular because it has ability to handle both numerical and categorical data [8] . As well they are easy to interpret. It is inverted directed tree having root at the top and has peculiarity that any complex decision making process can be converted into smaller and simple decision.

Data perturbation

In this paper, we are using data perturbation method for modifying data. Data perturbation has an important aspect in preserving the privacy of data. Perturbation is deviation of system from normal state to some other but consistent state . After perturbation, the original data set is modified and further given for the analysis process. Data perturbation can be done in several, however, the most widely used techniques are: probability distribution and data distortion. Data perturbation is easy and effective technique for preserving confidential data [6] [9]. A no. of methods has been proposed for privacy preserving in data mining. This paper mainly focusses on five major methods used for data perturbation. Those are as follows:

a) ***Noise addition:*** In this method, the origin data matrix is added by uniformly distributed noise matrix. The noise matrix is of same size as original. The elements of noise matrix are the randomly generated numbers collected from continuous uniform distribution.

b) ***PCA:*** The principle Component analysis is mainly used for dimensionality reduction. In PCA orthogonal transformation is used, so to transform the original data of co-related samples into the set of linearly uncorrelated samples. This sample is known as Principle Components. PCA becomes sensitive when

ISSN: 0976-3104

original variables show relative scaling in their values [9]. It is a widely-used tool in exploratory data analysis and developing predictive models for decision making. There are two major ways for performing PCA one by Eigen-value decomposition of a data matrix or singular value decomposition of a data matrix.

c) ***SVD:*** SVD i.e. Singular Value Decomposition is frequently used method for data perturbation. It is usually used for the dimensionality reduction of the original data set. In this paper, it is used for data perturbation method [6] [17]. Let say, A be the original matrix of order m x n. The row(n) in matrix represents the data whereas column(m) represents the attributes. The SVD of the matrix M is:

$$M = U \sum VT$$

Where U is an orthogonal matrix of order m x n, $\sum$ is an m x n diagonal matrix whose diagonal elements are positive and VT represents an n x n orthonormal matrix

d) ***QR:*** It is primarily used for the decomposition of a matrix. Modified matrix is a product of orthogonal matrix (Q) and upper triangular matrix (R). This can be represented as follows:

$$M = QR$$

If M is a complex square matrix, then there is a decomposition M = QR where Q is a unit matrix (i.e. Q*Q = I). If A has m linearly independent columns, then first m columns of Q form an orthonormal basis for the column space of A. In other words, the first k columns of Q form an orthonormal basis for any $1 \leq k \leq n$. In short any column k of A depends only on the first k columns of Q which is amenable for the triangular form of R [9] [19].

In this paper, we have come up with approach where we are using different data perturbation methods to protect or preserve the privacy of data used for data mining processes. The following figure shows the functional workflow for PPDM. The classification used to verify the performance of the original dataset after perturbation. The privacy measures are calculated for each of data perturbation methods described above. [Fig: 1] depicts the functional workflow of our implementation.

Workflow

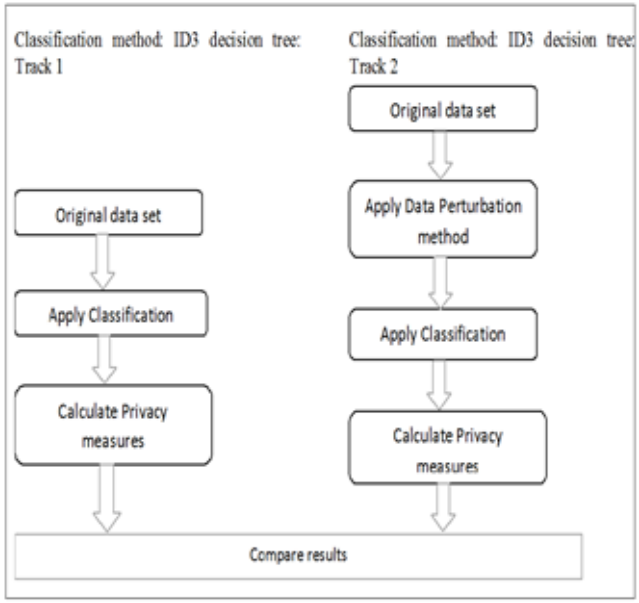

Fig. 1: Functional workflow.

The entire workflow of this paper is divided into main three modules viz:

a) ***Classification module:*** In this module, data set is being mined with one of the classification algorithm. In this paper, we have used "decision tree" as classification method. As well, accuracy is also calculated which being compared with perturbed data is set accuracy.

b) ***Perturbation module:*** In this module, the same data set is being perturbed using one of the perturbation method. We have used Noise addition as perturbation method for one of the column of dataset. We have also checked accuracy after perturbation as it will depict the distortion of data due to perturbation.

c) ***PPDM module:*** This is the important module where actual privacy parameters are being calculated which shows the results how perturbation method is useful for preserving the privacy of data. This module is applied over above two modules i.e. classification module and perturbation module.

IIOAB JOURNAL
ISSN: 0976-3104

PRIVACY MEASURES

In this paper, we have used privacy measure that are usually used by the PPDM methods, based on matrix decomposition . Privacy is said to be protected if VD, RP and CP have larger value and RK and CK will have smaller value.

a) ***Value Difference (VD):*** After applying perturbation on the data samples, the data gets modified. The modified changes are the value difference (VD) [1] between the original data and perturbed data. It is given by the relative value difference in Forbenius norm. The value difference is the ratio Forbenius norm on original data (A) and the perturbed data (PA) to the original data (A).

$$VD = |A-MA|_F / |A|_F$$

b) ***Position Difference (RP):*** After Data Perturbation on the dataset, the relative position of the data sample is modified also. There are many metrics to measure the positional difference of the data samples [1].

$$RP = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} | Rank_j^i - MRank_j^i |}{nm}$$

c) ***RK:*** It exhibits the percentage of elements that keep their values in each column after distortion. It is used to represent the average change of order for every attribute in data sample. After the data of an attribute is perturbed, the order of each data is changed. Let us say original data A has n observation and m attributes. Orderij depicts the ascending order of the perturbed sample Aij. The RK is defined as:

$$RK = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} | Rk_j^i}{nm}$$

Where, RK gives whether a sample retains its position in the order of the value:

$$Rk_j^i = \begin{cases} 1 & Rank_j^i - MRank_j^i \\ 0 & otherwise \end{cases}$$

d) ***CK:*** Like RK, CK can be defined to calculate the percentage of the attributes that retain their orders of average value after the perturbation. Hence CK is given as follows:

$$CK = \frac{\sum_{i=1}^{m} Ck_i}{nm}$$

Where CK_i is calculated as follows:

$$Ck_i = \begin{cases} 1 & Rank_j^i - MRank_j^i \\ 0 & otherwise \end{cases}$$

e) ***CP:*** The values of an attribute can be inferred from its relative value difference compared with the other attributes. Hence it is necessary to know the order of average value of the attributes that varies after the data perturbation. The CP metric can be used to define the change of the average value of attributes:

$$CP = \frac{\sum_{i=1}^{m} |Rank_j^i - MRank_j^i|}{nm}$$

Where RankVi, is the ascending order of the attribute, while MRankVi represents its ascending order after the perturbation[19].

The higher value of RP and CP and lower value of RK and CK, denotes the more privacy is preserved for given dataset.

IMPLEMENTATION

The idea introduced in this paper has been implemented with RStudio. It is the platform used for R language to develop programs especially in machine learning and data mining. R language is used to demonstrate classification method (Decision tree). [Fig: 2] shows decision tree with more specification like pruning which restricts unusual growth of tree and [Fig: 3] provides probability of each class in tree for better understanding of classifier.

ISSN: 0976-3104

Dataset: Cardiotocography Data Set
(https://archive.ics.uci.edu/ml/datasets/Cardiotocography)
Number of Instances: 2126
Number of Attributes: 23

EXPERIMENTAL RESULT AND ANALYSIS

Classification results

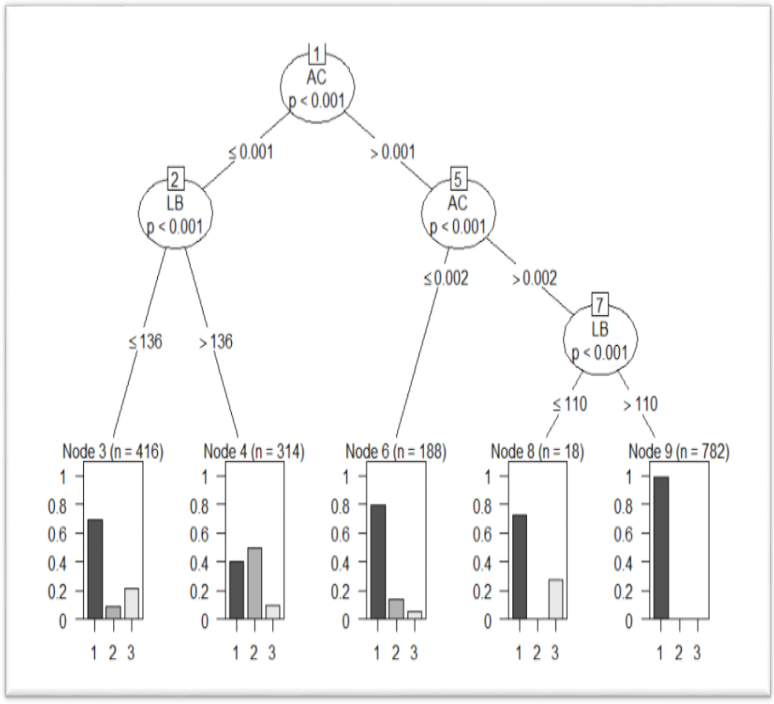

Fig. 2: Decision tree after pruning.

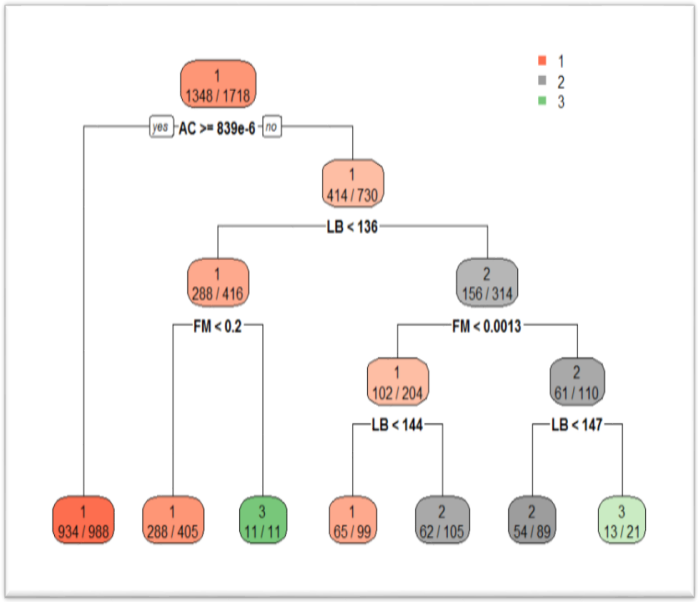

Fig. 3: Decision tree with probability of classes.

Classification Accuracy: It is the ratio of number of correct predictions to the total number of predictions made which is multiplied by 100 (%).
Again, only accuracy is not enough to conclude any classification or perturbation method to be efficient. Some other performance parameters must have calculated.

IIOAB JOURNAL
ISSN: 0976-3104

[Table 1] shows classification results while [Table 2] shows privacy measures calculated for four different data perturbation methods.

Table 1: Classification results

	Dataset without Perturbation	Dataset with Perturbation
No of Correctly Predicted classes	1389 (training data) 323 (testing data)	1430 (training data) 324 (testing data)
No. of wrongly Predicted Classes	329 (training data) 85 (testing data)	270 (training data) 84 (testing data)
Accuracy	0.7916667	0.7941176

PPDM results

Table 2: Privacy measures calculated for data perturbation methods

Privacy Measures	VD	RP	CP	CK	RK
PCA	1.000041	44.07045	2061263	1	0
SVD	1.000001	43.77193	2047301	1	0
Simple Noise Addition	0.006210513	0.0796302	3724.464	1	0.95455
QR	1.423345	45.09783	2109316	1	0

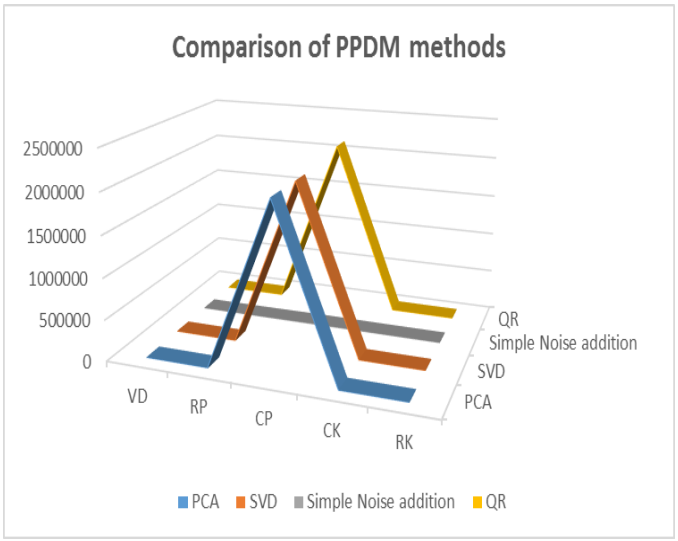

Fig. 4: Comparison between different data perturbation method.

Analysis

The experimental results show that PCA is the best among rest of three methods. Because it has higher value of RP and CP while lower values for CK and RK than other methods, which is considered as an efficient with regards to privacy measures. For simple noise addition method though it has the lowest VD, it can't be best because rest of measures are not significant for it to be best. Our focus is also on the privacy preservation classification; the experimental results clearly shows that classification is highly effective in terms of accuracy. The accuracy is getting preserved even after data perturbation hence both dataset used for experiment and data mining process is the best combination to perform out PPDM process.

CONCLUSION

This paper discussed various data perturbation methods and privacy measures have been calculated for each of the method. Hence Privacy Preservation in Data Mining i.e. PPDM proved to be novel approach to protect data. In addition to that Classification (Decision Tree) also played important role in efficient execution of Privacy Preserving methods and preserving accuracy of data. Though Data perturbation is quite obsolete method, its combination with Classification method made the approach very efficient. As

IIOAB JOURNAL
ISSN: 0976-3104

the field of Privacy is growing day by day, it important to establish a framework considering variety of privacy protecting algorithms.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

I express my sincere gratitude to Prof. Jabanjalin Hilda, Senior Assistant Professor, Department of Computational Intelligence, VIT University - Vellore, for her expert guidance and valuable support throughout project especially in Data mining. I am extremely thankful to VIT University – Vellore for providing me such great infrastructure facilities for this project as well I want to thank our Departmental Dean, HOD, all the other Staff Members and my batch mates for their encouragement throughout the course of project work. Lastly, I would like to acknowledge with gratitude, the support, and love of my family – my parents, friends without them it would not have been possible.

FINANCIAL DISCLOSURE

None

REFERENCES

[1] Samir Patel, Kiran R. Amin. [2013] Privacy Preserving Based on PCA Transformation Using Data Perturbation Technique, International Journal of Computer Science & Engineering Technology, ISSN: 2229-3345, 4(5): 477-484

[2] Inthumathi MS, Damodharan P.[2016] PPDM and Data Mining Technique Ensures Privacy and Security for Medical Text and Image Feature Extraction in E-Health Care System, International Journal of Computer Science and Information Technologies, 6 (6):5126-5129

[3] Rajesh N, Sujatha K, Arul Lawrence Selvakumar A. [2016] Survey on Privacy Preserving Data Mining Techniques using Recent Algorithms, International Journal of Computer Applications (0975 – 8887) 133(7):20

[4] Suchitra Shelke, Babita Bhagat.[2015] Techniques for Privacy Preservation in Data Mining, International Journal of Engineering Research & Technology, 4(10)

[5] Bhupendra Kumar Pandya, Umesh Kumar Singh, Keerti Dixit.[2015] A Robust Privacy Preservation by Combination of Additive and Multiplicative Data Perturbation for Privacy Preserving Data Mining, International Journal of Computer Applications (0975 – 8887) 120(1)

[6] Yousra Abdul Alsahib S. Aldeen1, Mazleena Salleh and Mohammad Abdur Razzaque, [2015] A comprehensive review on privacy preserving data mining, Springer Plus 4:694 DOI 10.1186/s40064-015-1481

[7] Tamanna Kachwala, Sweta Parmar. [2014] An Approach for Preserving Privacy in Data Mining, International Journal of Advanced Research in Computer Science and Software Engineering, 4(9)

[8] M.Syamala kumari, B.Govinda lakshmi, Privacy Preserving of Unrealised data sets using classification, IJCEA, Volume VII, Issue II, August 2014

[9] Sharmila A Harale, Bongale AK. [2014] Privacy Preservation and Restoration of Data Using Unrealized Data Sets, International Journal of Engineering Research and Applications, ISSN: 2248-9622, 4(7) :107-111

[10] Santosh Kumar Bhandare. [2013] Data Distortion Based Privacy Preserving Method for Data Mining System, International Journal of Emerging Trends & Technology in Computer Science, 2(3)

[11] Naga Lakshmi M, Sandhya Rani k.[2013] SVD based Data Transformation Methods for Privacy Preserving Clustering, ISSN: 0975 – 8887, 78 (3)

[12] Nagendra kumar.S, Aparn .R. [2013] Sensitive Attributes based Privacy Preserving in Data Mining using k-anonymity, International Journal of Computer Applications (0975 – 8887) 84(13)

[13] Aldeen YAAS, Salleh M, Razzaque MA.[2016] A comprehensive review on privacy preserving data mining, SpringerPlus, 4(1): 1-36

[14] Xinjun Qi, Mingkui Zong.[2013] An Overview of Privacy Preserving Data Mining, ICESE 2011, Procedia Environmental Sciences 12 (2012) 1341 – 1347 International Journal of Computer Applications (0975 – 8887) 8(13

[15] Guang Li, Yadong Wang.[2012] A Privacy preserving classification method based on single value decomposition, The International Arab Journal ofInformation Technology, 9(6)

[16] Mohammad Reza Keyvanpour, Somayyeh Seifi Moradi. [2011] Classification and Evaluation the Privacy Preserving Data Mining Techniques by using a Data Modification–based Framework, International Journal on Computer Science and Engineering, ISSN: 0975-3397 (2) : 862- 871

[17] Kamakshi P, Vinaya Babu A.[2010] Preserving Privacy and Sharing the Data in Distributed Environment using Cryptographic Technique on Perturbed data, Journal of Computing, 2(4)

[18] Keke Chen, Ling Liu.[2009] Privacy-preserving Multiparty Collaborative Mining with Geometric Data Perturbation, IEEE Transaction on Parallel and Distributed Computing

[19] Li Liu, Murat Kantarcioglu, Bhavani Thuraisingham. [2006]The Applicability of the Perturbation Model-based Privacy Preserving Data Mining for Real-world Data, Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)