Volume 5, Issue 3, March 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Efficient Privacy Preserving Method For Classification in Data Mining System

Santosh Kumar Bhandare
Asst. Professor, Department of Computer Science and Engineering
Shri Dadaji Institute of Technology & Science
Khandwa (M.P.) India

Abstract – The privacy protection of data is very important issue in data mining system. Data mining system contain large amount of confidential and sensitive data. These confidential data cannot be share to everyone. Therefore privacy protection of data is required in data mining system for avoiding privacy leakage of data. Data perturbation is one of the well known methods for privacy protection in data mining system. In data perturbation method individual data value are distorted before applying data mining application. In this paper we present Tanh (Tan Hyperbolic) normalization transformation based data perturbation method for privacy protection of data in data mining system. The privacy protection is measured by privacy parameters and the utility measure shows the performance of data mining methods after data distortion. We performed experiment on real life datasets and the experimental results show that Tanh normalization transformation based data perturbation method is very effective to protect confidential information and also maintain the performance of data mining technique after data distortion.

Keywords: Privacy preserving, Tanh normalization, data perturbation, classification, data mining

I. INTRODUCTION

Data mining [**1**] is the process to explorer pattern from huge amount of data using data mining tool such as classification. Data mining system contain large amount of confidential and sensitive data such as financial, criminal and healthcare records. This confidential information cannot be share to everyone so privacy protection of data is required for avoiding privacy leakage. The problem of privacy protection is very important issue in data mining system. There are a lot of data mining application deal with privacy and security concern. There are a lot of research has been done in privacy preserving data mining based on secure multiparty computations, randomization, perturbation and Anonymity including K-anonymity and l-diversity. In this paper we discuss the Tanh Normalization transformation based data perturbation technique in which confidential numerical attributes are distorted for privacy protection in classification analysis. Data perturbation is one of the best techniques for privacy preserving data mining system. In data perturbation the individual data values are distorted before applying data mining application. The privacy parameters [**2**] are used for measurement of the degree of privacy protection. These parameters also show the capability of this technique to concealing the original data and the data utility measures show the performance of data mining technique after data distortion. In this paper we present the Tanh normalization transformation based data distortion method for privacy protection. We perform experiment on four real life datasets and the experimental results show that the proposed method is very effective to concealing the confidential information and also preserve the performance of data mining technique after data distortion.

II. RELATED WORK

There has been a lot of privacy preserving data mining literatures. These literatures can divide into two categories. In the first category, methods modify the data mining algorithms so that without knowing the exact values of data, they allow data mining operations on distributed dataset. In the second category, methods are modifying the values of the datasets to protect privacy of data values. In this category there are several research has been done in data distortion or data perturbation are as follow:

In the year 1982, A.C.Yao [8] firstly proposed the secure two-party computation problem, which is extended in the year 1998, [19] by O. Goldreich, to the secure multi-party computation (SMC). The secure multi-party computation protocol is based on the cryptograph secure model. This secure multi-party computation protocol can compute arbitrary function in distributed networks where each participant holds his inputs, while the participants do not trust each other, nor the channels by which they communicate with each other. However, the participants can correctly get the result of the function from their local inputs, while keeping their local data as private as possible. Each party provides his input that will keep private.

In the year 2005, Chen et al. [23] they proposed a rotation based perturbation method. The proposed method maintains zero loss of accuracy for many classifiers. Experimental results show that the rotation perturbation can greatly improve the privacy quality without sacrificing accuracy.

In the year 2006, Wang et al. [11] they proposed a new data distortion method based on Structural Partition and SSVD for Privacy Preserving data mining, they used object-based partition, feature-based partition and hybrid partition.

Singular Value Decomposition (SVD) is a popular method in data mining and information retrieval. In this proposed method structural matrix partition is used here to divide the original matrix into several sub matrices. Then perform SSVD on one selected sub matrix. Three kinds of matrix partition are proposed here, which are denoted by:

i. P1 (Object-based partition)
ii. P2 (Feature-based partition)
iii. P3 (Hybrid partition)

The performance of the proposed new strategies is measured by some metrics. Data utility is examined by a binary classification based on the support vector machine. The experimental results show that feature-based partition is a feasible and efficient solution for privacy-preserving data mining.

In the year 2007, Xu et al. [12] they proposed the Fast Fourier Transform (FFT) based data perturbation method for privacy preserving data mining i.e. privacy protection of data is achieved by the Fast Fourier Transform. The dataset is distorted or perturbed by using Fast Fourier Transform for privacy protection of data values. They performed experiment on real life dataset to test the performance of the SVD based and FFT based data distortion methods. They compare the Fast Fourier Transform with Singular Value Decomposition based distortion method. The experimental results show that Fast Fourier Transform based data distortion technique is similar to Singular Value Decomposition based distortion technique, but the computational time used by the FFT based technique is much less than the SVD based technique. So the Fast Fourier Transform based data distortion technique is very effective.

In the year 2007, Wang et al. [13] they proposed several efficient and flexible techniques to address accuracy issue, in privacy preserving data mining through matrix factorization. They compare accuracy maintenance after data distortion by different methods using the support vector machine classification. Nonnegative matrix factorization and singular value decomposition are effective and promising methods for privacy preserving data mining. The experimental results indicate that for centralized datasets with numerical attributes, matrix factorization-based distortion strategies achieve a satisfactory performance. NMF-based distortion method is a better choice among the matrix approximation methods

In the year 2008, Dr. K. Duraiswamy et al. [14] they proposed Sensitive Rule Hiding (SRH), to hide the sensitive rules that contain sensitive data, so that sensitive rules containing specified sensitive data on the right hand side of the rule cannot be inferred through association rule mining. In this paper they discussed privacy preserving in data mining, and they proposed a method for hiding sensitive rules. This proposed method is able to hide the sensitive rule sets automatically.

In the year 2008, Liu et al. [15] they proposed the wavelet transformation for data distortion or data perturbation to preserve the privacy of data. At the same time, privacy preserving strategy based on wavelet perturbation; keep the data privacy and data statistical properties and data mining utilities. They perform experiment on real life dataset. The experimental results show that proposed method keep the distance before and after data perturbation and it also preserve the basic statistical properties of original data while maximizing the data utilities. Therefore the experimental results show that this method is very effective and promising. They also used multi-basis wavelet transformation to enhance data perturbation and compare the single basis wavelet perturbation method with the multi-basis wavelet perturbation method.

In the year 2009, Yidong Li et al., [17] Data Swapping is one of the best data perturbation method. The data swapping method is effective, if the data swapping methods preserves data privacy as well as data utility. In this paper they investigate the possibility of using data swapping with equi-width partitioning for private data publication, which is used in data perturbation due to the difficulty of preserving data protection. According to experimental result they show that, Equi-Width Swapping (EWS) may achieve a similar performance in privacy protection to that of Equi-Depth Swapping (EDS) if the number of partitions is sufficiently large i.e. number of partitioned is greater than or equal to the square root of P, where P is the size of dataset. The experimental results show that CASTLE is effective and efficient with respect to the output data quality.

In the year 2009, Yingjie Wu et al., [10], also proposed k-Anonymity privacy preserving for re-publication of incremental datasets. The k-Anonymity based privacy preserving is very popular approach for privacy protection. In this paper, they presented an effective approach for republication of incremental datasets. They analyze some possible generalizations in the anonymization for incremental updates. In the k-anonymity model, privacy is assured by ensuring that any record in the released data is indistinguishable from at least $k-1$other records with respect to a set of attributes called quasi-identifier. There are a lot of k-anonymization algorithms have been developed, most of the existing methods assume that the dataset is fixed. In this paper they perform experiment on real life dataset, and the experimental result show that the proposed approach is effective.

In the year 2010, Wang et al. [19] in this paper they proposed a novel algorithm for hiding sensitive association rules in data warehouses. A data warehouse is made by multiple dimension tables and a fact table as in a star schema. The proposed algorithm can effectively hide multi-relational association rules, because the proposed method reduce the confidence of sensitive association rule and without constructing the whole joined table. In this proposed method for hiding association rule on multiple tables, they address two issues are as follow:

i. How to calculate supports of itemsets efficiently
ii. How to reduce the confidence of an association rule by minimal modification of dimension tables

They also studied the association rule hiding problem in multi-relational databases. In this proposed method they used mining-then-joining based algorithm to deal with multi-table association rule hiding.

In the year 2010, Jianneng Cao et al. [20] they proposed Continuously Anonymizing STreaming data via adaptive cLustEring (CASTLE), a cluster-based method k-anonymize data streams and, at the same time, ensures the freshness of the anonymized data by satisfying specified delay constraints. The proposed method improves the performance without compromising the privacy protection .They also explains how the CASTLE can be easily extended to handle l-diversity. The problem occurred by *k*-anonymized data stream is also investigated in this paper. The experimental results show that CASTLE is effective and efficient with respect to the output data quality.

In the year 2011, Deivanai et al. [21] they proposed suppression based new method for achieving k-anonymity.

In this technique, they performed efficient multi-dimensional suppression i.e. values are suppressed only on certain records depending on other attribute values, without the need for manually-produced domain hierarchy trees. In this method they identify attributes that have less influence on the classification of the data records and suppress them if needed. The method was evaluated on several datasets to evaluate its accuracy as compared to other k-anonymity based methods.

In the year 2011, Tzung-Pei Hong et al. [22] they proposed a lattice-based approach for modifying original databases for hiding sensitive data. In this method they built the lattice structure which is based on the relation of sensitive item sets. The bottom-up deletion strategies is used by this approach to gradually reduce the frequency of sensitive item sets in the hiding process. They conducted experiment and the experimental results show the performance of the proposed approach.

In the year 2012, Jahan et al. [9] propose data distortion methods such as SVD (singular value decomposition) and SSVD (sparsified singular value decomposition) technique with feature selection to reduce feature space. There are various privacy metrics have been proposed which measure the difference between original dataset and distorted dataset and degree of privacy protection. They performed experiment on a real world dataset and the experimental results shows a feasible solution using sparsified singular value decomposition with a feature selection.

In the year 2013, Sara Hajian et al.,[18] they handle discrimination protection in data mining and propose new techniques applicable for direct or indirect discrimination protection individually or both at the same time. We discuss how to clean training data sets and outsourced data sets in such a way that direct and/or indirect discriminatory decision rules are converted to legitimate (nondiscriminatory) classification rules. We also propose new metrics to evaluate the utility of the proposed approaches and we compare these approaches. The experimental evaluations demonstrate that the proposed techniques are effective at removing direct and/or indirect discrimination biases in the original data set while preserving data quality.

In the year 2014, Deepti Mittal et al., [16] used a secure k-means data mining method that assume the data is distributed among different hosts preserving the privacy of the data in cloud environment. This approach maintains correctness and validity of existing k-means to generate the final results as well as in the distributed environment.

III. ASSUMPTIONS

The object-attribute relationship of real life data sets are encode into vector – space format [3]. In this format a 2-dimentional is used to share the dataset. Row of the matrix indicates individual object and each column represent a particular attribute of these objects. In this matrix, we assume that every element is fixed, discrete and numerical. Any missing element is not allowed.

IV. DATA DISTORTION MEASURES

We used the same set of privacy parameters proposed in [**2**]. The privacy measures depends only on the original matrix A and its distorted matrix Ā

A. Value Difference (VD)

After a data matrix is distorted by data distortion method, the value of its elements changes. The value difference of the datasets is defined by the relative value difference in the Frobenius norm. On the other hand VD is the ratio of the Frobenius norm of the difference of "A" and "Ā" to the Frobenius norm of A.

$$VD = \frac{\|M - \bar{M}\|}{\|M\|} \qquad (1)$$

Where || denotes the Frobenius norm of the enclosed argument

B. Position Difference

The order of the value of the data element changes after data distortion. We use several metrics to measure the position difference of the data element.

1. RP- RP parameter is used to represent the average change of rank for all attributes after data distortion. For a dataset M with q data object and p attributes. Let O_j^i is the rank (in ascending order) of the j[th] element in attribute i. Similarly $\overline{O_j^i}$ is the rank of the corresponding distorted element. Then the RP parameter is given by:

$$RP = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q}\left|O_j^i - \overline{O_j^i}\right|}{p * q} \qquad (2)$$

2. RK- RK parameter represents the percentage of elements that keeps their rank in each column after distortion. The RK parameter is given by:

$$RK = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q} Rk_j^i}{p*q} \quad (3)$$

Where $Rk_j^i = 1$ if $O_j^i = \overline{O_j^i}$ otherwise $Rk_j^i = 0$

3. CP- CP parameter is used to measure how the rank of the average value of each attributes varies after data distortion. CP represents the change of rank of the average value of the attributes. CP parameter is given by:

$$CP = \frac{1}{p}\sum_{i=1}^{p}\left|OM_i - \overline{OM_i}\right| \quad (4)$$

Where OM_i and $\overline{OM_i}$ represent the rank of the average value of i[th] attribute before and after data distortion respectively.

4. CK- Similar to RK, CK is used to measure the percentage of the attributes that keep their rank of average value after data distortion. CK parameter is given by:

$$CK = \frac{1}{p}\sum_{i=1}^{p} Ck^i \quad (5)$$

Where

$Ck^i = 1$ If $OM_i = \overline{OM_i}$

Otherwise $Ck^i = 0$

C. Utility Measure

After the conduction of certain perturbation the data utility measures indicate the accuracy of data mining algorithms on distorted data. In this paper we choose the accuracy of a J48 [5] as our data utility measure.

V. TANGENT HYPERBOLIC NORMALIZATION *(TANH)*

Tanh normalization is a data transformation method. Tanh normalization method maps the scores in the range [-1, 1] in a non linear transformation. In this method the values around the mean of the scores are transformed by a linear mapping and a compression of the data is performed for the high and low values of the scores. The tanh Normalization is based on Hampel estimators [24] and is given by

$$X_{Tanh} = \frac{1}{2}\left\{tanh\left(K\frac{X-\mu}{\sigma}\right)+1\right\} \quad (6)$$

Where μ and σ are, respectively, the mean and standard deviation estimates, of the genuine score distribution introduced by Hampel and *k* is a suitable constant. The advantage of Tanh normalization is the suppression of the effect of outliers, which is absorbed by the compression of the extreme values.

VI. THE PRIVACY RESERVING METHODOLOGY

Let A be an original data matrix of dimension P×Q. The rows of the matrix represent objects and the column of the matrix represent attributes. Now the original data matrix A whose size is P×Q, must be first transformed by Tanh normalization transformation to get transformed matrix Ā whose size has the same size P×Q as the original data matrix. The Tanh normalization transformed each element of the original data matrix A into the specific range [-1, 1]. We will use the negative number as a suitable constant k in Tanh normalization transformation. The negative value of k will change the order of each element in distorted data matrix Ā with respect to original data matrix A. Now after applying Tanh normalization transformation on original data matrix A, we have obtained a distorted data matrix Ā, which is very similar to the original data matrix A, but not identical. More importantly, Ā preserve the properties of A, so Ā can work as a distorted version of the original data matrix A.

Algorithm: Tanh Normalization Based Data Distortion Method For Privacy Preserving Data Mining.

Input: Numerical Dataset, Suitable Constant k (a negative number).

Output: Distorted Data Matrix.

Step 1: Initialized original data matrix A according to dataset, whose size is P × Q. Where each row of the matrix A, indicate individual object and each column represent a particular attribute of these objects.

Step 2: The original data matrix A must be first transformed by Tanh Normalization transformation. The Tanh Normalization transformed each element of the original data matrix A into a specific range [-1, 1].

Step 3: Now after applying Tanh Normalization transformation on original data matrix A, we get a transformed or distorted data matrix whose size is also P × Q.

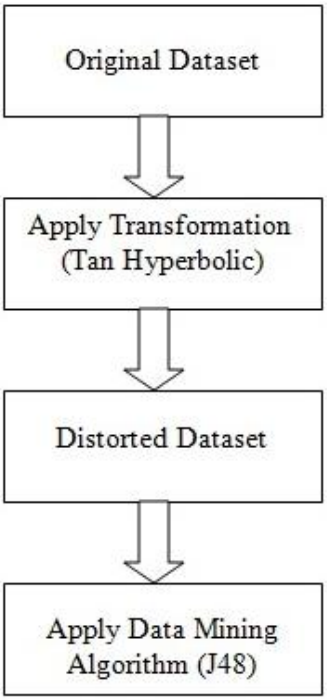

Fig 6.1: Privacy Preserving Methodology

Table I: The summary of the database

Database	Number of Instances	Number of Features	Number of Classes
Glass Identification	214	10	7
Haberman's Survival data	306	3	2
Bupa Liver Disorders	345	6	2
Iris	150	4	3

VI. EXPERIMENTAL RESULTS

We have conducted experiments to evaluate the performance of data distortion method. We choose four real-life Databases obtained from the University of California Irvine (UCI), Machine Learning Repository [6]. Datasets are the Glass Identification, Haberman's survival data, Bupa Liver Disorders and Iris Dataset. The summaries of the original database are given in Table [I] and Table [II] show the performance of proposed method. We use WEKA (Waikato Environment for Knowledge Analysis) [7] software to test the accuracy of distorted method. The privacy parameters are measured by a separate Java programme. We have constructed the classifier for J48 classification [5], and a 10-fold cross validation to obtain the classification results.

We apply our data perturbation method with K= -1 on the real life datasets mentioned in table I and get the results. All these result shown in table II.

Table II: How the privacy parameters and accuracy vary in four datasets

Data	VD	RP	RK	CP	CK	Acc in %
Glass Identification (Original)	-	-	-	-	-	96.729
Glass Identification (Distorted)	0.994660	101.25140	0.00654	5.0	0	97.6636
Haberman's survival data (Original)	-	-	-	-	-	71.8954
Haberman's survival data (Distorted)	0.985759	151.98257	0	1.33333	0.33333	71.5686
Bupa Liver Disorders (Original)	-	-	-	-	-	68.6957
Bupa Liver Disorders (Distorted)	0.989951	172.40966	0	3.0	0	68.4058
Iris (Original)	-	-	-	-	-	96
Iris (Distorted)	0.798795	74.33333	0	2.0	0	96.6667

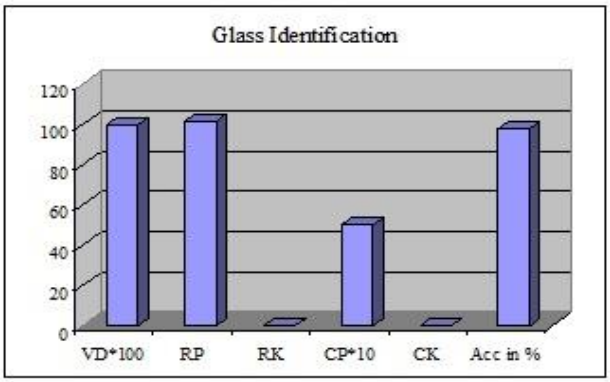

Fig 7.1: Accuracy and Privacy Parameter of Glass Identification dataset

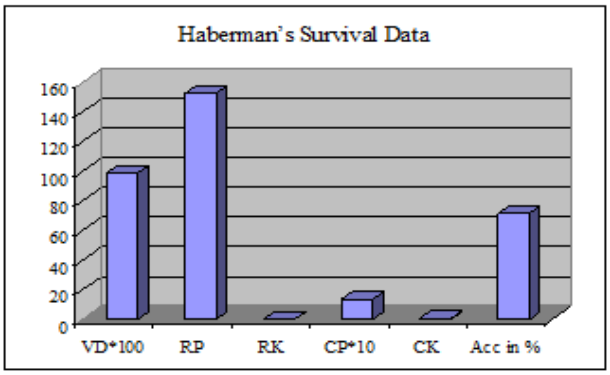

Fig 7.2: Accuracy and Privacy Parameter of Haberman's Survival dataset

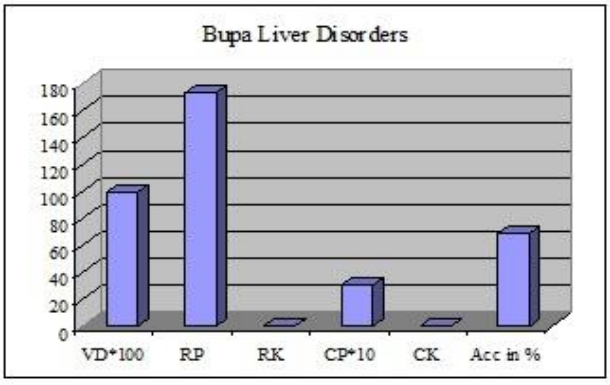

Fig 7.3: Accuracy and Privacy Parameter of Bupa Liver Disorder dataset

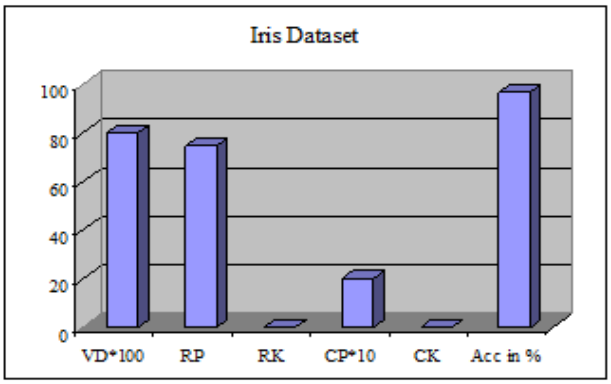

Fig 7.4: Accuracy and Privacy Parameter of Iris dataset

VII. CONCLUSION

In this paper we present Tanh (Tan Hyperbolic) Normalization transformation based data distortion method for privacy preserving data mining. In this method we distorted original dataset before applying data mining application. We performed experiment on four real life datasets and the experimental results show that Tan Hyperbolic normalization transformation based data perturbation method is very effective to protect confidential and sensitive information. The privacy parameters used in this work show the degree of privacy protection by the proposed method. In addition, the proposed method also maintains the performance of data mining technique after data distortion, it is interesting to use the other data transformation methods and compare its result with Tan Hyperbolic normalization.

REFERENCES

[1] M. Chen, J. Han, and P. Yu, "Data mining: An Overview from a database Prospective", IEEE Trans. on Knowledge and Data Engineering, vol. 8, no. 6, pp. 866-883, Dec. 1996.

[2] S. Xu, J. Zhang, D. Han, J. Wang, "Data distortion for privacy protection in a terrorist analysis system", Proceeding of the IEEE International Conference on Intelligence and Security Informatics, pp. 459-464, 2005.

[3] W. Frankes and R. Baeza-Yates, "Information Retrieval: Data Structures and Algorithms", Prentice–Hall, Englewood cliffs, NJ, 1992.

[4] J. Han, and M. Kamber, "Data Mining: Concepts and Techniques", Second edition, 2006, Morgan Kaufmann, USA.

[5] Ian H. Witten, Eibe Frank,"Data Mining Practical Machine Learning Tools and Techniques", Second Edition, 2005.

[6] UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets.html

[7] The Weka Machine Learning Workbench. http://www.cs.waikato.ac.nz/ml/weka

[8] A.C.Yao, "Protocols for secure computations", In Proc Of the 23rd Annual IEEE Symposium on Foundations of computer Science, 1982.

[9] Jahan, Narsimha and Rao, "Data perturbation and feature selection in preserving privacy" Ninth International Conference on Wireless and Optical Communications Networks (WOCN), pp 1-6, 2012.

[10] Yingjie Wu, Zhihui Sun, Xiaodong Wang, "Privacy Preserving k-Anonymity for Re-publication of Incremental Datasets", 2009 World Congress on Computer Science and Information Engineering, pp. 53 – 60,2009

[11] J. Wang, W. J. Zhong, J. Zhang and S.T. Xu, "Selective Data Distortion via Structural Partition and SSVD for Privacy Preservation", Proceedings of International conference on Information & Knowledge Engineering, pp. 114-120, June 2006.

[12] Shuting Xu, Shuhua Lai, "Fast Fourier transform based data perturbation method for privacy protection", Proceeding of IEEE International Conference on Intelligence and Security Informatics, pp. 221-224, 2007

[13] Jie Wang, Jun Zhang," Addressing Accuracy Issues in Privacy Preserving Data Mining through Matrix Factorization", pp. 217-220, 2007.

[14] Dr.K. Duraiswamy, Dr.D. Manjula , N. Maheswari (Corresponding Author), "A New Approach to Sensitive Rule Hiding", Journal on Computer and Information Science, CCSE 2008, vol. 1, No. 3, pp. 107-110, 2008.

[15] Lian Liu, Jie Wang, Jun Zhang, "Wavelet-Based Data Perturbation for Simultaneous Privacy-Preserving and Statistics-Preserving", Proceeding of IEEE International Conference on Data Mining Workshop, pp. 27-35, 2008

[16] Mittal, Deepti ; Kaur, Damandeep ; Aggarwal, Ashish, "Secure Data Mining in Cloud Using Homomorphic Encryption", International Conference on Cloud Computing in Emerging Markets, pp 1-7, 2014.

[17] Yidong Li, Hong Shen, "Equi-Width Data Swapping for Private Data Publication", International Conference on Parallel and Distributed Computing, Applications and Technologies, pp. 231- 238, 2009.

[18] Sara Hajian and Josep Domingo-Ferrer , "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining", Proceeding of the IEEE Transactions on Knowledge and Data Engineering, VOL. 25, NO. 7, pp. 1445 -1459, JULY 2013

[19] Shyue-Liang Wang, Tzung-Pei Hong, Yu-Chuan Tsai, Hung-Yu Kao3, "Hiding Sensitive Association Rules on Stars", IEEE International Conference on Granular Computing, pp. 505-508, 2010.

[20] Jianneng Cao, Barbara Carminati, Elena Ferrari, Kian-Lee Tan, "Continuously Anonymizing Data Streams", IEEE Transactions On Depedable And Secure Computing, pp.1, 2010.

[21] Deivanai, Nayahi and Kavitha, "A hybrid data anonymization integrated with suppression for preserving privacy in mining multi party data", International Conference on Recent Trends in Information Technology (ICRTIT), pp. 732 - 736, 2011.

[22] Tzung-Pei Hong, Chun-Wei Lin, Kuo-Tung Yang and Shyue-Liang Wang; "A lattice-based data sanitization approach", IEEE International Conference on Computational Sciences and Optimization (CSO), pp. 272 - 275, 2011.

[23] K. Chen and L. Liu, "Privacy preserving data classification with rotation perturbation", Proceeding of the 5th IEEE International Conference on Data Mining (ICDM 2005), pp. 589-592, 2005.

[24] F.R. Hampel, P.J. Rousseeuw, E.M. Ronchetti, And W.A. Stahel Robust Statistics: The Approach Based on Influence Functions, Wiley, 1986.

AUTHOR PROFILE

Mr. Santosh Kumar Bhandare is presently working as an Asst. Professor in Department of Computer Science & Information Technology at Shri Dadaji Institute of Technology & Science Khandwa (M.P.) 450001 India. The degree of B.E. secured in Computer Science & Engineering from Madhav Institute of Technology & Science Gwalior in 2006, M.Tech. in Computer Science & Engineering from Samrat Ashok Technological Institute Vidisha in 2011. Research Interest includes Data Mining, Network Security and Cloud Computing. Object Oriented Technology and Computer Graphics & Multimedia.
Mobile: +91-9827099815,
E-mail: santosh.mits@gmail.com

