Asian Journal of Convergence in Technology *Volume III, Issue III*
ISSN No.:2350-1146, I.F-2.71

Data Privacy Preserving using Perturbation Technique

Prof. V.S. Mahalle
Department of Computer Science and Engineering
Shri Sant Gajanan Maharaj College of Engineering
Shegaon, India
vsmahalle@gmail.com

Pankaj Jogi
Department of Computer Science and Engineering
Shri Sant Gajanan Maharaj College of Engineering
Shegaon, India
Pankajjogi716@gmail.com

Shubham Purankar
Department of Computer Science and Engineering
Shri Sant Gajanan Maharaj College of Engineering
Shegaon, India
spurankar93@gmail.com

Samiksha Pinge
Department of Computer Science and Engineering
Shri Sant Gajanan Maharaj College of Engineering
Shegaon, India
samiksha.pinge15@gmail.com

Urvashi Ingale
Department of Computer Science and Engineering
Shri Sant Gajanan Maharaj College of Engineering
Shegaon, India
urvashi.ingle@gmail.com

***Abstract*— Data Mining mainly consist of the discovery of structures, associations and the events in the data. In order to analyze the data related to sector like healthcare, privacy of data is to be maintained. In order to maintain the privacy of data, a perturbation technique is applied on original dataset and a new dataset is formed which is different from original dataset. Data mining can be performed on this perturbed dataset for various surveys and analysis.**

In this paper, perturbation technique algorithm is explained step by step in order to preserve the privacy of data.

***Keywords*— Privacy preservation, Geometric data perturbation, Random perturbation, Rotation perturbation .**

I. INTRODUCTION

In early years, amount of data was relatively small and could be store by means of computers and storage devices like hard drive, etc.

In recent years, increase in storage capacity of information devices causes increase in storing personal information about patients, customers, banking and individuals. Information is a very important asset of individuals in everyday life.

Over the years, amount of data has generated increases with a high rate which led to raise the necessity of using database management system (DBMS) and cloud storage. This information can be used for different destructive purpose like attack for individuals' identity theft, cyber terrorism, external attacks and fraud banking or ATM transactions. In order to alleviate this, we are using privacy preservation technique.

To provide the privacy to individual's data we are using various privacy preservation techniques which uses different approaches to preserve the privacy of sensitive data.

Some of the well-known approaches of privacy preservation implemented by different authors are as follow, data perturbation technique (In this technique, we alter the original data with modified one), knowledge hiding (In this technique, we hide the sensitive association rules), multiparty computation (In this technique, we build a model over the distributed database independent of knowing other inputs).

Data perturbation technique is one of effective and efficient approach, hence has many applications in medical healthcare data protection as it has highest probability of an attack being take place on such kind of subjects. Due to this reason, it is

implemented as a more solid application when it comes to the field of EHR security.

The author Sativa Lohiya and Lata Ragha has prescribed a discrete formula, which is: A (max) - A (min)/n = length.

Where, A is continuous attribute and n represent number of discrete, and length is the value of the discrete interval. But these technique does not reconstruct the original data values; it only reconstructs the distribution.

The author Jahan, G. Narsimha and C.V. Guru Rao [14] introduced a new approach in data perturbation. This approach based on singular value decomposition (SVD) and sparsified singular value distribution (SSVD) technique along with feature of selection to reduce the feature space. SSVD is efficient approach in keeping data utility, while SVD also works better than other standard data distortion methods which add noise to the data to make it perturbed.

II. MATERIALS AND METHODS

We are using geometric perturbation in this paper for privacy preservation of sensitive attributes of dataset D.

We are chosen two different datasets for implementation of this algorithm which are Bank management dataset and Adult dataset. The mentioned dataset can be obtained from UCI machine learning repository, each of datasets has number of attributes including numeric and non-numeric. We have implemented the algorithm on numeric attributes only.

Datasets	**Description**
Bank Management	Total Instances: 22,180 Attributes: 9
Adult Dataset	Total Instances: 18, 481 Attributes: 6

(Fig.1. list of datasets)

The attributes of Adult dataset are mentioned in tabular form as:

Attribute	Data type
Age	Numeric
Fnlwgt	Numeric
Work class	Text
Education	Text
Education num	Numeric
Marital Status	Text
Occupation	Text
Relationship	Text
Race	Text
Sex	Text
Capital gain	Numeric
Hours per week	Numeric
Native country	Text

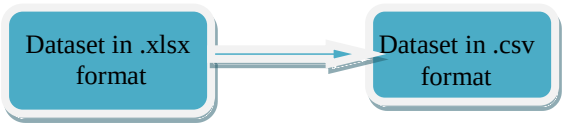
(Fig.2. Adult dataset)

A. Preprocessing of Data

We are using Adult dataset which is obtained from UCI machine learning repository as available as Excel file for implementation of this algorithm. We required the file to be in comma separated value (.cvs) file format.

The file from .xlsx format can be easily converted into .csv format using save as a file type

.



B. How to select sensitive attributes?

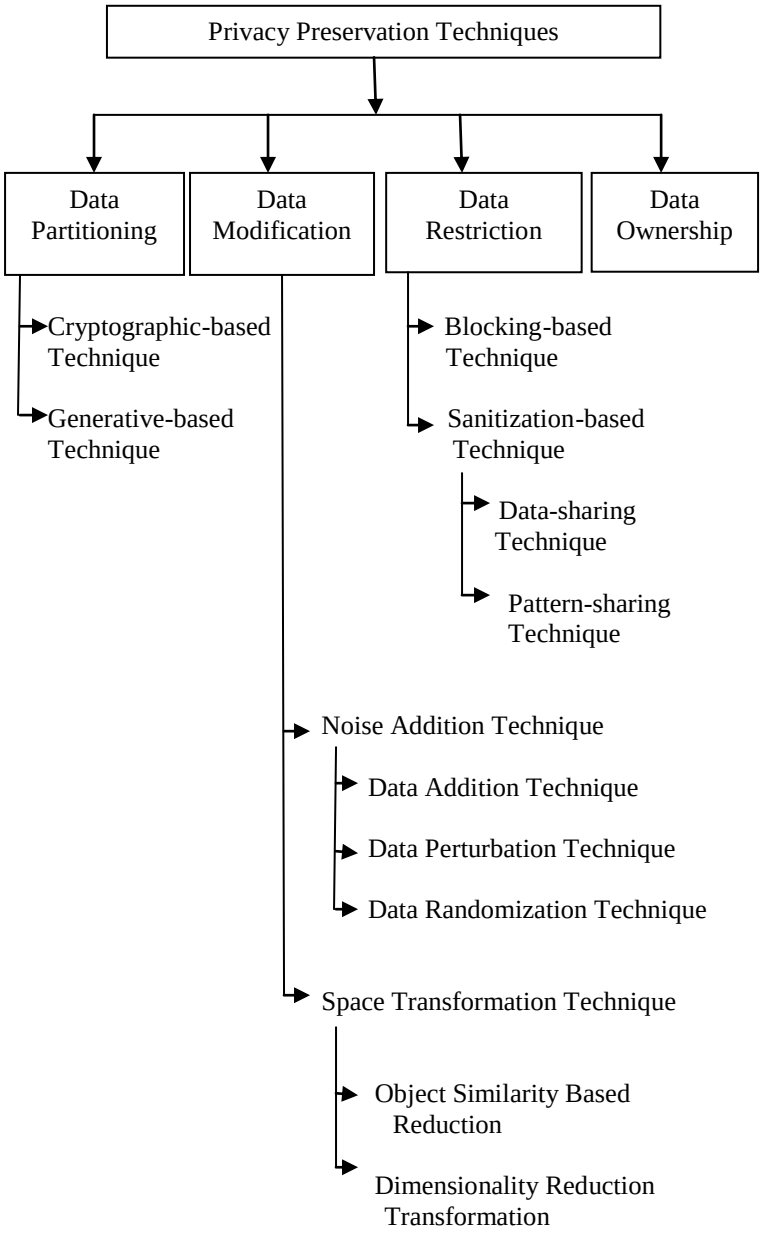
We have chosen sensitive attribute in order to implement geometric perturbation technique, as we have seen earlier this technique can be applied to numeric dataset and only one attribute at a time. Preserving privacy of both attribute important in both the cases as age (Adult Dataset) reveals the individual information while the accno (Bank Management System Dataset) creates possible threat of fraud bank transaction, hence are altered.

While coming to privacy preservation there are two dimensions which are as follows:

Individual Privacy Preservation: This focuses mainly on the privacy of the individuals or private entities.

Collective Privacy Preservation: As the name suggests, main area of this dimension is on the privacy of the overall organizations.

Following figure lists out few of privacy preserving techniques practiced by various organizations dealing with the sensitive data.



III. DATA PERTURBATION TECHNIQUES

In earlier years, plenty of work has been reported on perturbation techniques for purpose of privacy preservation. The perturbation technique includes different techniques like rotation perturbation, random noise addition technique, projection perturbation and k-anonymization model. Project primarily focuses on the Geometric data perturbation which can be classified as one of the random noise addition methods." [8].

Our focus is primarily on geometric data perturbation which can be classified as one of the technique from random noise addition method. Eventually user is not equally concern about the privacy of all attribute in the records. Hence user may be want to provide those contain values with some modification with help of available data perturbation technique. These values can be modified using programming language like C++, Java, .net, etc. [13].

DATA MINING PROBLEMS DO NOT NECESSARILY NEED INDIVIDUAL RECORDS BUT ONLY DISTRIBUTIONS. SINCE THE PERTURBATION DISTRIBUTIONS ARE KNOWN, IT CAN BE POSSIBLE TO RECONSTRUCT

aggregate distributions and this aggregate information can be used for the data mining algorithms [13].
Any Perturbation technique is evaluated mostly on two bases: the level of privacy preservation and the level of preserved data utilization. Though both are conflicting goals, main goal of any perturbation technique is to ensure maximum privacy preservation as well as maximum preserved data utilization.

Data Privacy: Data privacy is generally identified as a level of difficulty attacker has to face in estimating the original data from the perturbed data. For a given perturbation technique, the more difficult estimation of original values, and the higher level of privacy that technique provides. Geometric data perturbation provides moderate level of data privacy but is more efficient compared to other algorithms [8].

Data Utility: The level of data utility refers to the amount of critical information preserved after perturbation.

"More specifically, the sensitive information is to be model or task oriented. For example, decision tree and k-nearest Neighbor classifier for classification modeling utilizes different sets of information about the datasets. Decision tree construction primarily concerns the related columns of distributions; the k-nearest neighbor relies on the distance relationship involving all the columns. It is interesting to note that data privacy level enhancement share contradictive relationship with data utilization in most of the data perturbation techniques. Mostly perturbation algorithms aiming at maximizing privacy preservation have to bear with less data utility. This innate correlation between both factors makes it challenging for any data perturbation techniques to find a balance [8].

A. Rotation Perturbation

"It was initially proposed for privacy preservation in data clustering. It is one of the major components in Geometric data perturbation. Rotation perturbation is defined as G(X) = R*X where $X_{d\times n}$ is the original dataset and $R_{d\times d}$ is randomly

generated rotation matrix. Distance preservation is the unique benefit as well as major weakness of this method [5]. This method is vulnerable to distance-inference attacks" [6, 8].

B. Random Projection Perturbation

"It refers to the technique which projects a set of data points from the original multi-dimensional space to another randomly chosen space. Let $P_{k\times d}$ be a random projection matrix. Where, P's rows are orthonormal [2].

$$G(X) = \sqrt{\frac{d}{k}}PX$$

The above formula is applied to perturb the dataset C. The rationale of projection perturbation depends on its approximate distance preservation, which is supported by Johnson – Linden Strauss Lemma. The lemma indicates that any given dataset in Euclidean space can be embedded into another space in a way that pairwise distance of any two points is maintained with small error, resulting into model quality preservation" [8].

C. Geometric Data Perturbation

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

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

D. Multiplicative Transformation®:

This component can be rotation matrix or random projection matrix. Rotation matrix exactly preserves distances while random projection matrix only approximately preserves distances. A key feature of rotation matrix is preserving Euclidean distance. Rotation perturbation is a key component of geometric perturbation which provides primary protection to the perturbed data from naïve estimation attacks. Other components of geometric perturbation are used to protect rotation perturbation from more complicated attacks. A random project matrix $R_{k\times d}$ is defined as $R=\sqrt{\frac{d}{k}}R_0$. The Johnson-Linden-Strauss Lemma proves that random projection can approximately preserve Euclidean distances if certain conditions are satisfied. [8]

.

E. Translation Transformation (Ψ)

For any two points x and y in the original space, with translation the new distance will be $|(x - t) - (y - t)| = |x - y|$. Therefore, translation always preserves the distance. Only translation perturbation does not provide protection to the data, it can be simply canceled if the attacker knows that only translation perturbation is applied. Translation combined with rotation perturbation, can increase the overall resistant to the attacks.

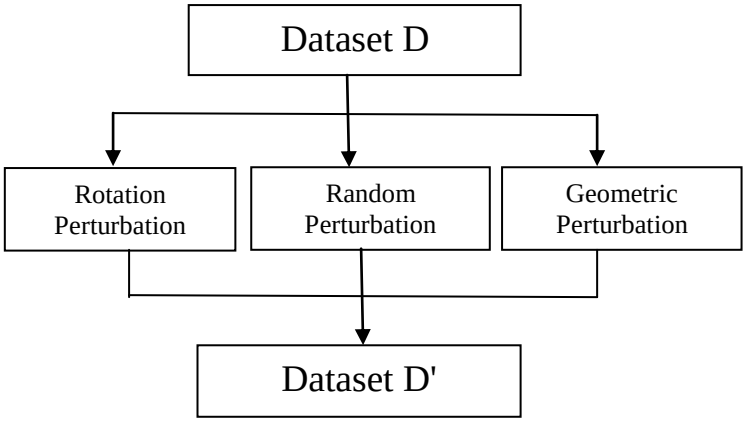
F. Distance Perturbation(Δ)

The above two components preserve the distance relationship. However, distance preserving perturbation can be under distance- inference attacks. The goal of distance perturbation is to preserve distances approximately, while effectively increasing the resistance to the distance-inference attacks. Here, distance perturbation can be noise. Solely applying noise without applying other two components will not preserve the privacy since noise intensity is low.

The major issue of distance perturbation is a tradeoff between reduction of model accuracy and gain of privacy guarantee. The data owner may opt not to use distance perturbation with the assumption that data is secure and attacker does not know about the original data. Hence, distance-inference attacks cannot happen [8].

Geometric perturbation technique can be applied to real time data as well as traditional data. Here, traditional datasets are used for experimental purpose.

Applied Framework:



The above graphical representation of the applied framework illustrates privacy preserving process using Geometric data perturbation technique.

Dataset D is altered by applying any of Multiplicative Data Perturbation techniques and modified dataset D' is obtained.

IV. ALGORITHM: GEOMETRIC DATA PERTURBATION

The idea behind using Geometric Data Perturbation algorithm is, because of its simplicity. Geometric perturbation is nothing but the enhancement to the rotation perturbation by coupling it with additional components like random translation perturbation and noise addition to the basic form of multiplicative perturbation Y = R £ X. It will be clear that by adding those additional components, Multiplicative perturbation for privacy preserving data mining geometric perturbation exhibits more robustness and provide efficiency in countering the attacks compared to normal rotation based perturbation [1,3].
For each attribute of G(X), let T be the translation, random rotation R, D be a Gaussian Noise and X be the original dataset. The value of the attribute G(X) can be found using following formula.

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

Procedure: Geometric transformation based Multiplicative data perturbation

Input: Dataset D, Sensitive attribute S.

Result: Perturbed dataset D'

Remarks: As of now, will be using traditional dataset instead of data stream and sensitive attributes will be numerical only.

Steps:

1. Given input data D with tuple size n, extract sensitive attribute $[S]_{n\times1}$.
2. Rotate $[S]_{n\times1}$ into 180^{o} clock-wise direction and generate $[Rs]_{n\times1.}$
3. Multiply element of [S] with [Rs], transformed sensitive attribute values will be
   $$[X]_{n\times1}=[S]_{n\times1} \times [Rs]_{n\times1}$$
4. Calculate translation T as mean of sensitive attribute $[S]_{n\times1}$.
5. Generate Transformation $[St]_{n\times1}$ by applying translation T to $[S]_{n\times1}$.
6. Calculate Gaussian distribution P(x) as a probability density function for Gaussian noise
   $$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$
   where, µ=Mean, Ϭ=Variance
7. Geometric data perturbation of sensitive attribute
   $$[Gs]_{n\times1}=[X]_{n\times1} + [St]_{n\times1}+ P(x).$$
8. Create perturbed dataset D' by replacing sensitive attribute $[S]_{n\times1}$ in original dataset D with $[Gs]_{n\times1}$.

A.Computation:

Generation of T (Translation Matrix):

Here, Translation matrix T = (Original value + Mean value)

Where Mean value = (Sum of original data / no. of elements)
For example,

Original dataset(x) = 1, 2, 3, 4, 5
Mean value (M) = (1+2+3+4+5)/5
= (15/5) = 3

Translation Matrix (T):

= (1+M, 2+M, 3+M, 4+M, 5+M)
= (4, 5, 6, 7, 8)

At the end, all these values of translation matrix will be added to rotated original data.

Rotation of Original data (Rotation Matrix):

Here, rotation matrix R = Rotation of original data to 180 degree
For example,
Original dataset(x) = 1, 2, 3, 4, 5
Rotation matrix (R) = 5, 4, 3, 2, 1

Generation of Noise (D):

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

The above equation is for the random Noise D. It is Gaussian Noise in this case.
Equation = [(1/ (variance * sqrt(2*PI))) * (pow (e, (-pow (x-mean), 2)/2*pow (variance, 2)))]

Explanation of variance:

Variance = sqrt (pow ((original dataset (X) - Mean), 2)/no of elements)
For example,
Step 1: pow (take data element first - mean, 2) =
pow (1-3, 2) =4

Step 2: pow (take data element second - mean, 2) =
pow ((2-3), 2) =1
Step 3: pow (take data element third - mean, 2) =
pow ((3-3), 2) =0
Step 4: pow (take data element forth - mean, 2) =
pow ((4-3), 2) =1
Step 5: pow (take data element fifth - mean, 2) =
pow ((5-3), 2) =4
Sum (4 + 1 + 0 + 1 + 4) = 10
Ans = sum/no. of element = 10/5 = 2
Variance = Sqrt (Ans) = sqrt (2) = 1.4142

V. ALGORITHM: GEOMETRIC DATA PERTURBATION

As we have seen earlier, the Rotation translation alters the original data by preserving Euclidean distance. Also, it is an important component of Geometric perturbation which protects sensitive data from naïve estimation attack by exactly preserving distance. But alone Rotational perturbation technique is not sufficient to provide protection to data as it has threats from estimation attack in which attacker can estimate the distance to get the original data.

Another important key component of Geometric perturbation is Translation transformation. It also preserves distance similar to Rotational perturbation technique. In this technique, we have calculated mean of sensitive attributes (numeric) and added it to each of sensitive value. But like Rotational perturbation it does not provide protection to sensitive attributes in dataset. One can easily have access to original data if it known that only Translation transformation technique is implemented. Hence more resistant to attack can be provided by combine implementation of both Rotational and Translational transformation.

Each of Translational and Rotational transformation preserves distance relationship among the data, however each of them has a threat from distance-inference attack. In case of distance perturbation, we need to increase the resistance to the distance-inference attack while preserving the distance. Hence third key component of Geometric perturbation technique introduced i.e. Gaussian noise, only noise without implementing other two also does not provide protection to data as intensity off noise is low.

The major issue of distance perturbation is a tradeoff between reduction of model accuracy and gain of privacy guarantee. The data owner may opt not to use distance perturbation with the assumption that data is secure and attacker does not know about the original data. Hence, distance-inference attacks cannot happen. The below table will help summarizing about Random rotation, Random projection and Geometric data perturbation.

Random Rotation	Geometric Perturbation	Random Projection
Y = R*X	Y=RX + T + D	Y=A*X
X is the original dataset for all three formulas Y is the perturbed dataset for all three formulas R is the random rotation matrix.	R is the secret rotation matrix (preserves Euclidean distances) T is the secret random translation matrix. D is the secret random noise matrix.	A is the random projection matrix.
Distances are preserved. Less secured [12].	Distances are approximately preserved [11].	Distances is not well preserved. Loss of Data [11].
Accuracy depends on the rotation matrix.	Good accuracy than any other perturbation techniques.	Worse accuracy than Geometric data perturbation.

A. CONCLUSION AND FITURE WORK

Geometric Transformation technique based on Multiplicative Data Perturbation approach has been applied for adding random noise to the original dataset to preserve privacy of sensitive attributes. This approach has been in direction to keep statistical relationship intact to mine useful results with perturbed data. It takes sensitive attributes as dependent attributes whereas, remaining attributes of dataset except class attribute are considered as independent attributes. Any calculations for adding tuple specific random noise is done only on dependent attributes of the dataset. Accuracy of the results from the perturbed data will be less than the accuracy of the results from the original dataset. But, it is possible to achieve main objective of preserving the privacy of the sensitive info with less accuracy loss and the loss can further be minimized.

These paper uses Geometric Data Perturbation technique for numeric data only. The algorithm can be applied on non-numeric dataset using k-anonymization techniques. The algorithm can also be expanded to check real time data using stream analysis tool. The applied algorithm is working to extract single column value only and can be extended for more than one column at a time and also this algorithm can be applied to more classification as well as clustering algorithms. One of the future goals can also be to improve efficiency of the data mining of altered dataset and make privacy preserving more effective with minimal accuracy loss.

REFERENCES

[1]. Krupali N. Vachhani, Prof. Dinesh Vaghela, "A Survey on Geometric Data Transformation for privacy preserving on data stream" International Journal of Technical Research and Applications e-ISSN: 2320-8163, Volume 3, Issue 2 (Mar-Apr 2015), PP. 257-259.

[2]. Kun Liu, HillolKargupta, Senior Member, IEEE, and Jessica Ryan, "Random Projection Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining", IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 1, january 2006, p.92.

[3]. Twinkle Ankleshwaria, Prof. J. S. Dhobi," Geometric Data Perturbation Approach for Privacy Preserving in data Stream Mining" Engineering Universe for Scientific research and Management, Impact factor 3.7, Volume 6, Issue 4, April 2014.

[4]. Hitesh Chhikaniwala [Ganpat University, India], Dr. Sanjay Garg [Nirma University, India],"Privacy Preserving Data Mining Techniques: Challenges and Issues".

[5]. Kun Liu, Hillol Kargupta, Senior Member, IEEE and Jessica Ryan, "Random Projection based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining", IEEE transaction on knowledge, Vol. 18, No. 1, January 2006.

[6]. Stanley R. M. Oliveria, Osmar R. Zaiane, "Data Perturbation by rotation for privacy preserving Clustering", Technical Report TR 04-17, August 2004.

[7]. Prashant Lahane, R K Bedi, Prasad Halgaonkar, "Data Stream Mining", International Journal of Advances in computing and Information Researches, ISSN:2277-4068, Volume 1-No. 1, January 2012.

[8]. Keke Chen [Dept. of CSE, Wright state university, Dayton], Ling Liu [CCGIT,Atlanta, GA], "Geometric Data Perturbation for Privacy Preserving Outsourced Data Mining", [Online Available:http://knoesis.wright.edu/library/download/geometric_perturbation.pdf].

[9]. Ms. Ompriya Kale, Ms. Prachi Patel, "A Survey on Privacy Preserving Data Mining Techniques", Global journal of Advanced Engineering Technologies, ISSN: 2277-6370, vol2, Issue3-2013.

[10]. Neha Gupta, Indrajeet Rajput, "Preserving privacy using data perturbation in Data Stream" International Journal of advanced Research in computer engineering & technology (IJARCET) Volume 2, No. 5, May 2013.

[11]. Yabo Xu, Ke Wang, Ada Wai-Chee Fu,Rong She, and Jian Pei, Privacy-Preserving Data Stream Classification, springer,pp.489-510(2008).

[12]. Charu C. Aggarwal and Philip S. Yu, "A Condensation Approach to Privacy Preserving Data Mining", IBM T. J. Watson Research Center, Hawthorne, NY.

[13]. R.VidyaBanu and N.Nagaveni," Preservation of Data Privacy using PCA based Transformation",2009 International Conference on Advances in Recent Technologies in Communication and Computing, in 2009 IEEE computer society,p.439.

[14]. T. Jahan, G.Narsimha and C.V Guru Rao, "Data Perturbation and Features Selection in Preserving Privacy" in proceedings of 978-1-4673-1989-8/12, IEEE 2012.

[15]. S. Lohiya and L. Ragha, "Privacy Preserving In Data Mining Using Hybrid Approach" proceedings of 2012 Fourth International Conference on Computational Intelligence and Communication Networks, IEEE 2012.