

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,000

Open access books available

148,000

International authors and editors

185M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Chapter

Multiplicative Data Perturbation Using Random Rotation Method

Thanveer Jahan

Abstract

Today's applications rely on large volumes of personal data being collected and processed regularly. Many unauthorized users try to access this private data. Data perturbation methods are one among many Privacy Preserving Data Mining (PPDM) techniques. They play a key role in perturbing confidential data. The research work focuses on developing an efficient data perturbation method using multivariate dataset which can preserve privacy in a centralized environment and allow publishing data. To carry out the data perturbation on a multivariate dataset, a Multiplicative Data Perturbation (MDP) using Random Rotation method is proposed. The results revealed an efficient multiplicative data perturbation using multivariate datasets which is resilient to attacks or threats and preserves the privacy in centralized environment.

Keywords: privacy, multiplicative data perturbation, random rotation method

1. Introduction

This chapter proposes a Multiplicative Data Perturbation method. It considers multivariate datasets to perturb using a geometric data perturbation method. Then, the perturbed data will use Discrete Cosine Transformation between a pair of data values to determine Euclidean distance. This proposal is clearly elaborated in the following section.

1.1 Background

Hybrid transformations are used to maintain statistical properties of data as well as mining utilities [1–3]. The statistical properties of data are mean and variance or standard deviation without any loss of data. A feasible solution [4] is provided to optimize the data transformations by maximizing privacy of sensitive attributes. A combined technique using randomization and geometric transformation is used to protect sensitive data. A randomized technique is represented as $D = X + R$, where R is additive noise, X is original data and D is perturbed data. A geometric transformation is used as a 2D rotation data matrix represented as $D' = R(\theta) \times D$, where D is the column vector containing original co-ordinates and D' is a column vector whose co-ordinates are rotated clockwise. The above method considered only single attributes as

sensitive and rest of them as non-sensitive attributes. Data perturbation method using fuzzy logic and random rotation is proposed [5, 6].

The original data is perturbed using fuzzy based approach (M) and then random rotation perturbation is used by selecting confidential numerical attributes to get the transformed data $P = M \cdot R$, where M is the dataset transformed using fuzzy based approach and R is the random dataset generated. The distorted data P is released for clustering analysis and obtained accuracy. The approach compromises in balancing privacy and accuracy. A hybrid method using SVD and Shearing based data perturbation [7] is proposed to obtain perturbed data. The approach removes the identified attributes from the dataset. These attributes are normalized using Z-score normalization to standardize to the same. Then, the dataset is perturbed using SVD transformation. Each record of the perturbed dataset is further distorted using a Shear based data Perturbation method represented as $D' = D + (Sh_D * D)$, where Sh_D is the random noise and D is the perturbed dataset obtained after SVD transformation.

The results show higher privacy is attained on hybrid methods when compared to single data perturbation methods. A hybrid technique [7, 8] based on Walsh-Hadamard Transformation (WHT) and Rotation is proposed. The Euclidean distance preserving transformation using Walsh-Hadamard (H_n) given below to generate orthogonal matrix to preserve statistical properties of the original dataset.

$$H_n = \otimes_{i=n}^D H_2 = \frac{H_2 \otimes H_2 \cdots \otimes H_2}{n} \quad (1)$$

where H_2 is $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ is a matrix and denotes the tensor or Kronecker product.

Then, Rotation transformation is applied to preserve the distance between the data points. The perturbed data preserves distance between data records and maintain accuracy using classifiers. The method is limited to numerical attributes and can be extended to categorical attributes. A hybrid approach for data transformation is proposed by Manikandan et al. [9] to sanitize data and normalize the data using min-max normalization [10]. The approach transforms original data maintaining inter-relative distance among the data. Clustering analysis shows that the numbers of clusters in original data are similar to modified data. Another approach is used to modify the original data to preserve privacy with the help of inter-relative distance on categorical data is proposed [2].

The categorical data is converted into binary data and is transformed using geometric transformation. Then the clustering algorithm is used for analysis and the results for better data utilization as well as privacy preservation. The multiplicative noise is generated using random numbers with mean as 1 and is multiplied by the original data value. A random number with a short Gaussian distribution is calculated with mean as 0 and a small variance. Geetha Mary A et al. [11] proposed a non-additive method of perturbation by randomization and data is generated based on intervals on the level of privacy specified by a user. A random number is generated that is either added or multiplied with the data to generate a random modified data. The perturbed data is classified and measures using metrics.

The condensation approach is presented by Agrawal and Yui [12] for a multidimensional perturbation technique to provide privacy for multiple columns using covariance matrix. The approach was weak in protecting data privacy. Rotation perturbation was used for privacy preserving data classification [13]. Rotation perturbations are task specific and aim to have better balance between loss of information

and loss of privacy. Multiplicative data perturbations include three types of perturbation techniques such as: Rotation Perturbation, Projection Perturbation and Geometric Data Perturbation.

A Rotation perturbation framework was adopted in privacy preserving data classification [14]. It is defined as $G(X) = RX$ where R is randomly generated rotation matrix and X is the original data. The benefit and weakness of this method is distance preservation and is prone to distance inference attacks. These attacks are addressed [15–17]. Chen et al. [14] proposed an improved version on resilience towards attacks. Oliveria et al. [17] proposed a scaling transformation along with random rotation in privacy preserving clustering.

A Random Projection perturbation is proposed [13, 18] to project a set of points from the original multidimensional space to another randomly chosen space. This resulted with an approximate model quality. A random projection matrix is used in privacy preserving data mining to enable an individual to choose their privacy levels.

An ideal data perturbation [19] aim with a balance tradeoff of minimizing information loss and privacy loss. However these are not balanced in the existing algorithms. Compared with the existing approaches in privacy preserving data mining, Geometric data Perturbation have significantly reduced these overcome [20].

A Geometric Data Perturbation is a sequence of random geometric transformation including multiplicative transformation (R), Translation Transformation (T) and Distance Perturbation (DP) [21, 22].

$$G(X) = R(X) + T + DP \quad (2)$$

The approach has two unique characteristics. The first characteristic is to perturb the original data with geometric rotation, translation and identify rotation invariant classifiers as given in above. The second characteristic is to build privacy model by evaluating the privacy quality of perturbation method. The privacy model generated is used to analyze the attacks, such as, Naives and ICA-based reconstruction. The quality of data perturbation approach is determined by the quality of privacy preserved. It is the difficulty level in estimating the original data from perturbed ones such estimations are named as inference attacks. The attacks are categorized into three categories such as: Naives Inference, Reconstruction based inference and distance based inference. A statistical method based inference to estimate original data from perturbed named as Naives inference attack was proposed [23]. It is represented as $O=P$, where O is the observed data and P is the perturbed data. Reconstructing the data with perturbed and released information from data is presented. Reconstruction based attacks also called as Independent Component Analysis (ICA) [24, 25]. It is represented as, $O = E^{-1} P$, where E^{-1} is the estimation of released information of data and P is the perturbed data. Identifying the images and some relevant information of data using outliers to discover the perturbation is distance based attacks. It is represented as $O = E^{-1}P$, where E^{-1} is the mapping to estimate and P is the perturbed data. The higher the inference the more the original data is protected and preserved such that attacker cannot break the perturbation. The above attacks are analyzed with a privacy model with privacy guarantee [26]. It had failed to avoid outlier attack. The existing data perturbation techniques have contradiction between data privacy metric and mining utility [27, 28]. The multiplicative data perturbations will maximize the two levels i.e. data privacy and mining utility. The multiplicative data perturbation shows challenging features to improve data privacy during mining process as well as to preserve the model specific information.

In this chapter a survey is presented on privacy preserving data mining to protect confidential data. The drawbacks of the above existing data perturbation methods have made us to resolve the issues with balanced factors, such as, data privacy and data utility. The challenges in preserving privacy using multiplicative data perturbation have been given a new direction in this research study.

2. Proposed method

The proposed Multiplicative Data Perturbation (MDP) is shown at **Figure 1** as a block diagram.

The above block diagram considers the original dataset and deals with it in two stages. In the first stage, the original dataset is perturbed using geometric data perturbation. The geometric data perturbation generates a distorted dataset. This distorted dataset is further perturbed using Discrete Cosine Transformation in the second stage to finally generate a distorted dataset. The process of generating a distorted dataset using a geometric data perturbation comprises three steps. At the first step a random dataset is created using random values as in the original dataset. This random dataset is rotated counter clockwise and then multiplied with the original dataset. The resultant dataset obtained the above step is transposed in the second step, that is, Translation Transformation. This Transposed dataset is added with an additive noise in the third step to obtain a distorted dataset. This proposal is an algorithm for multiplicative data perturbation in the next section.

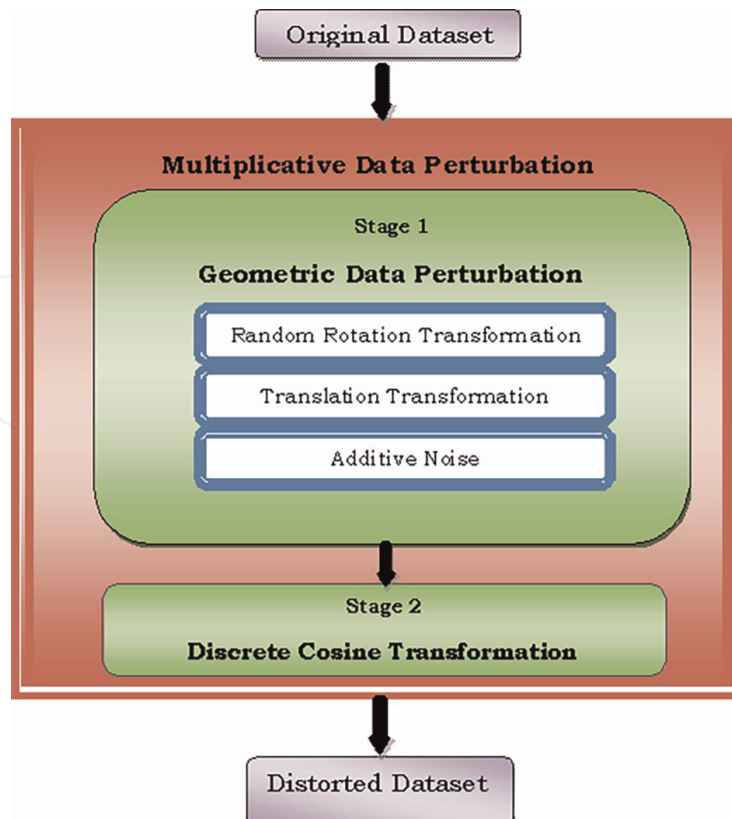


Figure 1. Block diagram for multiplicative data perturbation using random rotation method.

3. Proposed multiplicative data perturbation using random rotation algorithm

A proposal for multiplicative data perturbation is given in this section. The pseudo code of the proposed algorithm is listed below.

Algorithm:

Input: A Data Matrix $D_{p \times q}$.

Output: A Distorted Data matrices $D4, D5$.

Begin.

Step 1: Create a Random data matrix R with p rows and q column and Rotate the random data matrix as $R_{q \times p}$ //counter clock wise Rotation by 90° .

Step 2: Construct the data matrix $X_{p \times q}$ using $R_{q \times p}$ and $D_{p \times q}$ data matrices as:
 $X_{q \times q} = R_{q \times p} * D_{p \times q}$ //Multiplicative Transformation.

Step 3: Create another random data matrix $X1_{p \times q}$ with p rows, q columns with mean as 0 and standard deviation as 1.

Step 4: Construct the distorted data matrix $D4_{p \times q}$ using $X_{p \times q}$, Transpose of R and $X1_{p \times q}$ data matrices as:

$D4 = X + R^T + X1$ //Geometric data Perturbation Step 5: Call function DCT ($D4_{p \times q}; D5_{p \times q}$)//Discrete cosine transformation.

Step 6: The resultant distorted data matrix $D5_{p \times q}$ is output,

End.

Function DCT ($D4_{p \times q}; D5_{p \times q}$)//Function for Discrete Cosine Transformation.

Input: A data matrix $D4_{p \times q}$ Output: A data matrix $D5_{p \times q}$

Begin.

Step 1: Copy the data matrix $D4$ to a data matrix $D5$ //alias

Step 2: For $i = 1$ to q .

 For $k = 1$ to q .

 If $k = 1$ then

$$D4[i] = \left(\frac{1}{\sqrt{i}} * X_2(i) * (\cos(3.14 * (2 + 1)/2i)) \right)$$

 Else

$$D5[i] = \left(\frac{\sqrt{2}}{i} * X_2(i) * (\cos(3.14 * (2 + 1)/2i)) \right)$$

 End if

End For

Construct $D5$ data matrix and return as parameter.

End

The algorithm accepts the data matrix $D_{p \times q}$ with p rows and q columns as input. It creates a random data matrix R with p rows and q columns having random values as elements. This random data matrix R is rotated counter clockwise by 90° and then multiplied with data matrix $D_{p \times q}$. The data matrix that results is named as data matrix $X_{p \times q}$. Create another random data matrix $X1$ with p rows, q columns such that its mean is 0 and standard deviation is 1. Now, construct the distorted data matrix $D4$ adding the data matrices X, R^T and $X1$. This data matrix $D4$ is passed as a parameter to the called function $DCT()$. The predefined conditions are checked and data matrix $D5$

is updated. This data matrix D5 after completely updated is an output of the algorithm. The time complexity of the proposed MDP algorithm is found to be $O(n)$, where n is the dimension of the dataset.

The process of updating D5 is explained with the help of an example stated below:

Example 1.1: Consider a data matrix $D_{p \times q} = \begin{bmatrix} 4 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix}$ where $p = 2$ and $q = 3$.

At Step 1, create a random data matrix $R_{2 \times 3}$ as given below:

$R = \begin{bmatrix} -0.3034 & -0.7873 & -1.1471 \\ 0.2939 & 0.8884 & -1.0689 \end{bmatrix}$ and rotate R counter clockwise by 90° as given below:

$$R_{3 \times 2} = \begin{bmatrix} -1.1471 & -1.0689 \\ -0.7873 & 0.8884 \\ -0.3034 & 0.2939 \end{bmatrix}$$

At step 2, construct the data matrix $X = D_{2 \times 3} * R_{3 \times 2}$ is given as below:

$$X = \begin{bmatrix} -6.4664 & -2.2049 \\ -2.2378 & 0.1134 \end{bmatrix}$$

At step 3, create another random data matrix $X1$ with 2 rows and 3 columns such that the mean is 0 and the standard deviation is 1.

$$X1 = \begin{bmatrix} -6.4664 & 1.4384 & -0.7549 \\ -2.9443 & 0.3252 & 1.3703 \end{bmatrix}$$

At step 4, construct the distorted data matrix $D4 = X + R^T + X1$ as given as: $D4 =$

$$\begin{bmatrix} -3.1036 & -0.1362 & -1.3618 \\ -5.0820 & 2.1020 & 1.980. \end{bmatrix}$$

At step 5, the function call DCT (D4:D5) where

$$DCT(k) = f(k) \sum_{q=1}^q D4(q) \cos [(2k + 1)i\pi/2q] \quad k = 1, 2 \dots q; \quad i = 1 \dots p \quad (3)$$

where

$$f(k) = \begin{cases} \frac{1}{\sqrt{q}} & k = 1 \\ \frac{\sqrt{2}}{q} & 2 \leq k \leq q \end{cases}$$

Let $k = 1, q = 1, f(k) = \frac{1}{\sqrt{q}}$, then $f(1) = 1$, substituting the values in the Eq. (3)

$$Dct(1) = 1 * -3.1036 * \cos [3 * 3.14/2] = -5.7881$$

Let $k = 2, q = 1, f(2) = \frac{\sqrt{2}}{q}$, then $f(2) = 1$, substituting the above values in Eq. (3)

$$DCT(2) = 1 * -0.1362 * \cos [(2 * 2) * 3.14/2 * 2] = 1.3900$$

Similarly, the remaining data values of D4 are calculated to form a D5 data matrix as given below:

$$D5 = \begin{bmatrix} -5.7881 & 1.3900 & 0.4371 \\ 1.3989 & -1.5826 & -2.3630 \end{bmatrix}$$

The constructed data matrix D5 is the output.

4. Implementation

The proposed algorithm that was discussed in the previous section is implemented in MatLab. Its source code is included. The details of implementation are furnished in this section.

The implementation utilizes the built in functions available in MatLab such as load(), size(), randn(), rot90(), dct() and normrnd(). First, a load() built-in function is used to read a data into a data matrix D. The size() function is employed to retrieve the number of rows and columns. The function randn() is used to generate a random matrix R where the size is similar to data matrix D. The data matrix R is rotated using built in function available, namely rot90(). Then, to form a data matrix X, the data matrix R is multiplied by D data matrix. Next, normrnd() is called to generate a data matrix X1 having the mean as 0, the standard deviation as 1 and the size as similar to data matrix D. The distorted data matrix D4 is constructed by adding three data matrices, X1, R^T and X2. Finally, the function DCT() is employed on distorted data matrix D4 to obtain the resultant distorted data matrix D5.

5. Experimentation

The Experimentation was conducted using desktop computer system loaded with windows XP Operating system, MatLab and Tanagra data mining tool. The experimental details are elaborated in this section. The experimentation begins with the original dataset D is given as input to the proposed MDP algorithm to obtain the distorted dataset D4 and D5. Then, the original dataset D and distorted datasets D4 and D5 are uploaded into Tanagra data mining tool after appending a class attribute. These uploaded datasets are classified using classification utility available within Tanagra data mining tool. The results of classification are analyzed thereafter.

Similarly the datasets are clustered using clustering utilities available in them. The results of clustering are also analyzed and furnished at Section 6.6 under Results and Analysis. Unified column privacy metric to analyze possibility of attacks is also discussed in this section. But, their calculation is shown in section Results and Analysis. The datasets of Credit Approval, Haber-Man, Tic-Tac-toe and Diabetes are used in this experimentation. The details of Credit Approval dataset used in this experiment is furnished here and the rest of the datasets are furnished.

A Real Time Multivariate dataset, namely, Credit Approval, is downloaded from website UCI Machine Learning Repository. The details are shown at **Table 1**. Therefore the original dataset used in the experimentation is a Credit Approval dataset. It

Dataset	Size	Description
Credit Approval	690 rows & 15 columns	It consists of information of customers details concerned with credit card applications

Table 1.
Details of credit approval dataset.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
1	22.08	11.46	2	4	4	1.585	0	0	0	1	2	100	1213
0	22.67	7	2	8	4	0.165	0	0	0	0	2	160	1
0	29.58	1.75	1	4	4	1.25	0	0	0	1	2	280	1
0	21.67	11.5	1	5	3	0	1	1	11	1	2	0	1
1	20.17	8.17	2	6	4	1.96	1	1	14	0	2	60	159

Table 2.
A credit approval original dataset D.

comprises 690 rows/tuples and 15 columns/attributes including one target/class attribute.

A sample list of the original dataset D with 5 rows and 14 attributes is shown at **Table 2**.

The process in the experiment is explained as below:

First, a dataset named creditapproval.txt is loaded into X data matrix with the help of load() method. Next, the size() method on X data matrix determines the number of rows p as 690 and the number of columns q as 14. The data matrix is now named $D_p \times q$. Then, a built-in function randn(p, q) is used to create a random data matrix R. The random data matrix R is rotated with the help of built-in function rot90(). The data matrix X is constructed using data matrix R multiplied by data matrix D. The built in function normrnd(0,1, p, q) is used to create another random data matrix X1 with p rows, q columns, such that its mean is 0 and standard deviation is 1. Construct the distorted data matrix D4 by adding three data matrices X, R^T (transpose of R), X1. The distorted data matrix D4 is given as parameter to function DCT(D4) and it returns the final distorted data matrix D5 as output. When the above process is executed in experimentation it outputs a distorted datasets D4 and D5.

6. Results and analysis

The distorted datasets D4 and D5 together with the original dataset D, respectively are appended with a class attribute, YES or NO. The original dataset D after appending with a class attribute is shown at **Table 3**.

Similarly the distorted datasets D4 and D5 are also appended with a class attribute and furnished at section 6.6 as part of Results and Analysis. The above mentioned datasets D, D4 and D5 are uploaded into Tanagra data mining tool. First, classification utility is used on the dataset D and distorted datasets D4, D5. It divides the attributes into two categories, non-class attributes and class attribute. These two categories can be two inputs to the classifier chosen from the available ones.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	Class
1	22.08	11.46	2	4	4	1.585	0	0	0	1	2	100	1213	NO
0	22.67	7	2	8	4	0.165	0	0	0	0	2	160	1	NO
0	29.58	1.75	1	4	4	1.25	0	0	0	1	2	280	1	NO
0	21.67	11.5	1	5	3	0	1	1	11	1	2	0	1	YES
1	20.17	8.17	2	6	4	1.96	1	1	14	0	2	60	159	YES

Table 3.
 A credit approval original dataset with class attribute.

Suppose we select SVM (Support Vector Machine) as classifier, then, it classifies the datasets D, D4 and D5 based on class attribute into either credit card either approved or rejected. Such results are furnished at Section 6.6 under Results and Analysis. Similarly, the experimentation is repeated with Iterative Dichotomizer 3 (ID3), (Successor of ID3) C4.5, KNN (k-Nearest Neighbor) and MLP (Multi Layer Perceptron) classifiers.

The results of those experiments are furnished at Section 6.6. A Clustering utility available in Tanagra data mining tool is used to cluster the original dataset D and distorted datasets D4 and D5. Non- class attributes are considered and given as input to k-mean clustering method. As a result, categories of clusters are formed.

A unified column metric, Root Mean Square Error (RSME) is used to evaluate inference attacks. It is calculated using Eq. (3) as given below:

$$RSME(r) = \sqrt{\frac{1}{q} \sum_{i=1}^q (D - P)^2} \quad (4)$$

where $D = d_1, d_2 \dots d_q$ are the original dataset values, $P = p_1, p_2 \dots p_q$ are the perturbed dataset values and q is number of columns.

Then, privacy $(D, P) = \frac{4\sigma}{2r} = \frac{r}{2}$ (if standard deviation $\sigma = 1$). The attacks used are:

Naives inference is calculated as given in Eq. (4), where D is the original data and $P = E$ (E is estimated or Random dataset).

Reconstruction inference is calculated as given in Eq. (4), where D is the original dataset and the Perturbed dataset

$$P = E^{-1} * P. \quad (5)$$

Distance based inference is calculated as given in Eq. (5), where D is the original dataset and $P = P'$ (P' is mapped set of points of Perturbed dataset P).

The calculations of these metrics are furnished at Section 7 under Results and Analysis.

7. Results and analysis

The results obtained in the above experiment are presented in this section. The original dataset D is given as input to the proposed MDP and output distorted dataset $D4$ and $D5$ are presented below at **Table 4** and **Table 5**, respectively.

When SVM classifier is used on D, D4 and D5 datasets, the following observations are made and the same are presented at **Table 6**.

In the above **Table 4**, the first column presents the original dataset D and the distorted datasets D4 and D5. The number of tuples in the datasets considered for experimenting can be seen in the second column. The third column displays the number of training tuples classified for credit card approved as YES. The number of support vectors available is furnished in the fourth column. Fifth column reveals the error rate of SVM classifier. The computation time is tabulated at last column.

Similarly, when ID3 and C4.5 classifiers are used on D, D4 and D5 datasets the results are tabulated at **Tables 7 and 8**.

In the above **Tables 7 and 8**, the first column presents the dataset D and distorted dataset D4 and D5. The number of tuples in the datasets considered for experimenting can be seen in the second column. The third column displays the number of training tuples belonging to credit card approved as YES. A tree having number of nodes and leaves is furnished in the fourth column. Fifth column reveals the error rate of the ID3 and C4.5 classifiers, respectively. The computation time is tabulated at last column.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
7.2	15.69	14.65	5.3	18.7	9.8	2.57	-11.5	1.97	21.4	10.18	8.44	94.11	1.22
2.7	31.43	4.313	0.7	14.2	4.0	4.99	1.4	-6.69	4.77	-8.25	9.66	157.6	3.98
1.3	30.06	-13.5	17.8	3.13	22.0	-6.14	-4.23	5.28	-6.39	2.14	6.61	259.3	-2.37
9.7	26.01	9.224	-4.4	7.82	-10.3	-7.82	16.13	-10.2	1.78	13.12	-2.16	7.82	7.38
2.1	2.033	-9.22	-0.5	22.2	6.95	-0.10	2.54	-6.28	12.6	0.05	5.964	57.78	159.9

Table 4.
A credit approval distorted dataset D4.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
26.9	822.	111.6	43.7	19.6	108.	46.0	30.18	7.99	73.31	24.07	57.16	4.85	2.67
8.8	-1.6	10.90	-6.3	11.2	-4.44	6.46	6.85	-0.08	-0.08	-7.90	-5.33	-67.91	594.2
-4.8	-12.	19.15	-12.	12.4	-13.1	6.02	3.03	-10.18	-10.1	-2.99	-12.4	-229.7	-1.56
-2.3	2.86	5.801	4.62	3.55	5.80	5.89	-1.11	-11.90	-11.9	2.75	15.73	-12.28	1.87
22.9	-8.8	-11.61	-1.2	1.79	4.63	-8.72	8.00	-3.50	-3.50	2.75	-8.24	49.84	-1.27

Table 5.
A credit approval distorted dataset D5.

Dataset	Total Number of Tuples	Number of Training Tuples Classified as Approved (YES)	Number of Support Vectors	Error Rate	Computation Time (ms)
Original (D)	690	589	392	0.14	1562 ms
Distorted (D4)	690	582	621	0.446	2172 ms
Distorted (D5)	690	587	632	0.315	1969 ms

Table 6.
A credit approval dataset classified using SVM.

Dataset	Total Number of Tuples	Number of Training Tuples Classified as Approved(YES)	Tree having number of nodes and leaves	Error Rate	Computation Time(ms)
Original (D)	690	584	7 node,4 leaves	0.1464	16 ms
Distorted(D4)	690	476	3 node, 2 leaves	0.3101	31 ms
Distorted(D5)	690	580	1 node, 1 leaf	0.4464	16 ms

Table 7.
 A credit approval dataset classified using ID3.

Dataset	Total Number of Tuples	Number of Training Tuples classified as Approved(YES)	Tree having number of nodes and leaves	Error Rate	Computation Time(ms)
Original (D)	690	644	67 node, 34 leaves	0.066	47 ms
Distorted(D4)	690	621	137 nodes, 69 leaves	0.101	172 ms
Distorted(D5)	690	634	157 nodes, 79 leaves	0.1246	234 ms

Table 8.
 A credit approval dataset classified using C4.5.

Dataset	Total number of tuples	Number of Training Tuples Classified as Approved (YES)	Neighbors	Error Rate	Computation Time(ms)
Original (D)	690	537	5	0.2217	313 ms
Distorted(D4)	690	485	5	0.297	422 ms
Distorted(D5)	690	673	5	0.3145	391 ms

Table 9.
 A credit approval dataset classified using KNN.

When KNN classifier is used on D, D4 and D5 datasets the following observations are made and presented at **Table 9**.

In the above **Table 9**, the first column presents the original dataset D and distorted datasets D4 and D5. The number of tuples in the

datasets considered for experimenting can be seen in the second column. The third column displays the number of training tuples classified as credit card approved as YES for KNN classifier. The fourth column displays the number of neighbors. The fifth column reveals the error rate of KNN classifier. The computation time is tabulated in the last column.

Similarly, the results are tabulated at **Table 10** when MLP classifier is used on D, D4 and D5 datasets.

In the above **Table 10**, the first column presents the original dataset D and the distorted datasets D4 and D5. The number of tuples in the datasets considered for experimenting can be seen in the second column. The third column displays the number of tuples classified for credit card approved as YES. The maximum number of

Dataset	Total Number of Tuples	Number of tuples Classified as Approved (YES)	Max Iteration	Train Error Rate	Computation Time(ms)
Original (D)	690	620	100	0.0924	578 ms
Distorted(D1)	690	552	100	0.168	562 ms
Distorted(D2)	690	589	100	0.347	625 ms

Table 10.
A credit approval dataset classified using MLP.

iteration for MLP classifier is furnished in the fourth column. The fifth column reveals the training error rate of KNN classifier. The computation time is tabulated in the last column. Based on the results presented above the accuracy of classification of datasets is presented at **Table 11**. The accuracy is the percentage of tuples that were correctly classified by a classifier.

The above **Table 11** presents the accuracy of the classifiers for Credit Approval, Haber Man, Tic-Tac-Toe and Diabetes datasets. The first column presents the dataset D, the distorted datasets D4 and D5. The second column presents the accuracy of classification obtained on Credit Approval dataset using SVM, ID3, C4.5, KNN and MLP classifiers. The third column presents the accuracy of classification obtained on Haber Man dataset using SVM, ID3, C4.5, KNN and MLP classifiers. The fourth column presents the accuracy of classification obtained on Tic-Tac-Toe dataset using SVM, ID3, C4.5, KNN and MLP classifiers. The fifth column presents the accuracy of classification obtained on Diabetes dataset using SVM, ID3, C4.5, KNN and MLP classifiers.

It is observed that accuracy of C4.5, KNN and MLP classifiers are better than the accuracy of the other classifiers for distorted dataset D5 compared to distorted dataset D4.

The above **Table 12** presents the comparison of accuracy. The first column presents the distorted dataset D4 and D5. The second column presents the accuracy obtained on the proposed MDP using Credit approval, Tic-Tac-Toe and diabetes datasets for SVM and KNN classifiers. The third column presents the accuracy for the existing geometric data perturbation methods using Credit approval, Tic-Tac-Toe and Diabetes datasets for SVM and KNN classifiers. It is observed that the accuracy on the datasets using our proposed MDP was found better than the accuracy of the Existing Geometric data perturbation. Moreover, their accuracy was found only on SVM and KNN classifiers for Credit Approval, Tic-Tac-Toe, and Diabetes datasets only.

The proposed MDP has given good accuracy for distorted dataset D5 compared to distorted dataset D4, whereas the literature does not show any accuracy for distorted data D5

The results of k-means clustering are shown below at **Table 13**, when $k = 2$ (form two clusters).

In the above **Table 13**, the first column presents the dataset D, D4, and D5. The number of objects in the dataset considered for the experiment can be seen in the second column. The third column displays the number of objects belonging to cluster1. The fourth column reveals the number of objects belonging to cluster 2. The computational time is presented in the last column. Based on the results presented above the misclassification error rate of datasets is presented at **Table 14**.

Dataset	Credit Approval					Haber Man					Tic-Tac-Toe					Diabetes				
	SVM	ID3	C4.5	KNN	MLP	SVM	ID3	C4.5	KNN	MLP	SVM	ID3	C4.5	KNN	MLP	SVM	ID3	C4.5	KNN	MLP
Original (D)	85	84	93	98	89	75	86	88	89	90	97	89	87	98	97	89	86	82	89	90
Distorted (D4)	86	70	90	88	80	68	78	79	75	76	98	73	77	96	89	80	83	79	79	85
Distorted (D5)	88	84	91	97	87	72	85	85	89	90	99	88	87	98	97	83	85	81	89	90

Table 11.
 Accuracy of classifiers (%).

Dataset	Proposed Multiplicative Data Perturbation (MDP)						Existing Geometric Data Perturbation Method					
	Credit Approval		Tic-Tac-Toe		Diabetes		Credit Approval		Tic-Tac-Toe		Diabetes	
	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN
Distorted (D4)	86	88	98.7	98	80	79	86.5	82.9	98	99.5	77	73.5
Distorted (D5)	88	97	99	98.5	83.4	89	—	—	—	—	—	—

Table 12.
Comparison of accuracy.

Dataset	Number of Objects	Number of Objects in Cluster 1	Number of Objects in Cluster 2	Computation time (ms)
Original (D)	690	259	431	94 ms
Distorted (D4)	690	391	299	109 ms
Distorted (D5)	690	336	354	125 ms

Table 13.
Clustering on credit approval dataset for $k = 2$.

Dataset	PROPOSED MULTIPLICATIVE DATA PERTURBATION (MDP)			
	Credit Approval	Haber Man	Tic-Tac-Toe	Diabetes
Distorted (D4)	0.389	0.189	0.035	0.03
Distorted (D5)	0.22	0.100	0.031	0.02

Table 14.
Comparison of misclassification error-rate.

The above **Table 14** presents the misclassification error rate. The first column presents the distorted dataset D4 and D5. The second column presents the error rate obtained on the proposed MDP using Credit Approval, Haber Man, Tic-Tac-Toe and Diabetes datasets.

In the privacy metric mentioned in Section 1.5 in Eq. 1.2, the detailed calculation of privacy quality to analyze attacks is shown below:

Consider the data matrix $D = \begin{bmatrix} 4 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix}$ the corresponding distorted data matrix using the proposed MDP is given below:

$P = \begin{bmatrix} -5.7881 & 1.3900 & 0.4371 \\ 1.3989 & -1.5826 & -2.3630 \end{bmatrix}$, E is the estimated values (Random) as given below:

$E = \begin{bmatrix} -3.5441 & 1.3900 & 0.3211 \\ 0.9321 & 2.4567 & -6.7860 \end{bmatrix}$ and calculating $D'' = R^{-1} * P$ is given below

$D'' = \begin{bmatrix} -2.1461 & 0.2800 & 0.3211 \\ 1.8421 & 4.6767 & 4.6130 \end{bmatrix}$ and calculating P' is given below

Attacks	Proposed MDP Method				Existing Geometric Data Perturbation Method		
	Credit Approval	Haber Man	Tic-Tac-Toe	Diabetes	Credit Approval	Tic-Tac-Toe	Diabetes
Naives	1.743	1.129	1.564	1.512	1.345	1.234	1.456
Reconstruction	1.467	1.841	1.489	1.893	1.287	1.450	1.921
Distance	1.527	1.980	1.901	1.452	1.556	1.784	1.356

Table 15.
 Analysis on attacks.

$$P' = \begin{bmatrix} -1.9261 & 0.6800 & 1.3211 \\ 3.6821 & 1.6821 & -4.5920 \end{bmatrix}$$

Then, substitute the above data matrices in eq. 1.2 to analyze the following attacks:
 Naives-based Inference Attack: The RMSE is calculated by substituting the data matrices D and E. The result for RMSE r, obtained is as given below:

$$r = \sqrt{\frac{1}{3} \sum_{i=1}^2 ((D) - (E))^2} = 1.9221, \text{ Privacy } (D, P) = r/2 = 0.6796$$

Reconstruction -based Inference Attack: The RSME r is calculated by substituting the data matrices D and D". The result r obtained is as given below:

$$r = \sqrt{\frac{1}{3} \sum_{i=1}^2 ((D) - (D''))^2} = 1.6794, \text{ Privacy } (D, D'') = r/2 = 0.839$$

Distance -based Attack: The RSME r is calculated by substituting the data matrices D and P'. The result r obtained is as given below:

$$r = \sqrt{\frac{1}{3} \sum_{i=1}^2 ((D) - (P'))^2} = 1.70261, \text{ Privacy } (D, P') = 0.851$$

Similarly the RMSE r is calculated for the original D and distorted datasets D4 and D5 and the results are furnished at **Table 15** as shown below.

In the above **Table 15**, the first column presents the Naives based, Reconstruction based and Distance -based attacks. The second column displays RMSE (Root Mean Square Error) r is calculated for the proposed MDP method on Credit Approval, Haber Man, Tic-Tac- Toe and Diabetes datasets. The third column reveals the RMSE calculated for existing hybrid methods on Credit Approval and Diabetes datasets. It is observed that the RMSE r for proposed MDP method on distance -based attack is high compared to RMSE for the existing geometric data perturbation methods. The metric for the proposed MDP shows better quality in preserving the confidential data and provides high uncertainty to reconstruct the original data.

8. Conclusion

A Multiplicative Data Perturbation algorithm by combining a Geometric Data Perturbation method and Discrete Cosine Transformation is proposed in this chapter.

The proposed MDP is successfully implemented using different multivariate datasets mentioned above.

The experiments on those datasets resulted to classify accurately and create accurate number of clusters. Based on the result analysis, it is resolved that our proposed MDP algorithm is efficient to preserve confidential data during perturbation and ensures privacy while being resilient against possible of attacks the proposed methods considered a univariate datasets ex: Terrorist. A multivariate dataset is considered and a multiplicative data perturbation (MDP) was explored to effectively perturb the data in a centralized environment. This method has resulted in perturbing the data effectively and be resilient towards attacks or threats while preserving the privacy.

The research studies can explore the privacy issues on a Big Data as a future scope of research work in the following directions:

Improving Data Analytic techniques –Gather all data, filter them out with certain constraints and use to take confident decision.

Algorithms for Data Visualization- In order to visualize the required information from a pool of random data, powerful algorithms are crucial for accurate results.

In future scope includes, research can include many various methods explore many methods. These latest methods can show various results.


IntechOpen

Author details

Thanveer Jahan
Vaagdevi College of Engineering, India

*Address all correspondence to: tanveer_j@vaadevi.edu.in

IntechOpen

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Li L, Zhang Q. A privacy preserving clustering technique using hybrid data transformation method. In: 2009 IEEE International Conference on Grey Systems and Intelligent Services (GSIS 2009). Vol. 2010. Nanjing: IEEE; 2009. pp. 1502-1506. DOI: 10.1109/GSIS.2009.5408151
- [2] Natarajan AM, Rajalaxmi RR, Uma N, Kirubhkar G. A hybrid transformation approach for privacy preserving clustering of categorical data. In: Innovations and Advanced Techniques in Computer and Information Sciences and Engineering. Dordrecht: Springer. 2007. pp. 403-408. DOI: 10.1007/978-1-4020-6268-1_72
- [3] Selva Rathnam S, Karthikeyan T. A survey on recent algorithms for privacy preserving data mining. International Journal of Computer Science and Information Technologies. 2015;6(2): 1835-1840
- [4] Patel A, Patel K. A hybrid approach in privacy preserving data mining. In: 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). Vol. 2. Ahmedabad, Gujarat, India: IEEE; 2016. p. 3
- [5] M. Naga Lakshmi and K. Sandhya Rani, "A privacy preserving clustering method based on fuzzy approach and random rotation perturbation", Publications of Problems & Application in Engineering Research-Paper, Vol. 04, Issue No. 1, pp. 174-177, 2013.
- [6] Mary AG. Fuzzy-based random perturbation for real world medical datasets. International Journal of Telemedicine and clinical Practices. 2015;1(2):111-124. DOI: 10.1504/IJTMCP.2015.069749
- [7] M. Naga Lakshmi, K Sandhya Rani, "Privacy preserving hybrid data transformation based on SVD", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8, 2013, 2278-1021
- [8] Jalla HR, Girija PN. An efficient algorithm for privacy preserving data mining using hybrid transformation. International Journal of Data Mining & Knowledge Management Process. 2014; 4(4):45-53. DOI: 10.5121/ijdkp.2014.4404
- [9] Manikandan G, Sairam N, Saranya C, Jayashree S. A hybrid privacy preserving approach in data mining. Middle- East Journal of Scientific Research. 2013; 15(4):581-585. DOI: 10.5829/idosi.mejsr.2013.15.4.1.991
- [10] Saranya C, Manikandan G. Study on normalization techniques for privacy preserving data mining. International Journal of Engineering and Technology (IJET). 2013;5(3):2701-2704
- [11] Geetha Mary AN, Iyenger NSC. Non-additive random data perturbation for real world data. Procedia Technology. 2012;4:350-354. DOI: 10.1016/j.protcy.2012.05.053
- [12] Aggarwal CC, Yu PS. A condensation approach to privacy preserving data mining. In: Proceedings of International Conference on Extending Database Technology (EDBT). Vol. 2992. Heraklion, Crete, Greece: Springer; 2004. pp. 183-199
- [13] Liu K, Kargupta H, Ryan J. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Transactions on Knowledge and Data Engineering (TKDE). 2006;18(1):92-106

- [14] Chen K, Liu L. "A Random Rotation Perturbation Based Approach to Privacy Preserving Data Classification", CC-Technical Report GIT-CC-05-12. USA: Georgia Institute of Technology; 2005
- [15] Lui K, Giannella C, Kargupta H. An Attacker's view of distance preserving maps for privacy preserving data mining. In: Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'06). Berlin, Heidelberg: Springer-Verlag; 2006
- [16] Xu H, Guo S, Chen K. Building confidential and efficient query services in the cloud with RASP data perturbation. *IEEE Transactions on Knowledge and Data Engineering*. 2014;**26**(2):322-335
- [17] Oliveira SR, Zaiane OR. Privacy preserving clustering by data transformation. *Journal of Information and Data Management (JIDM)*. 2010; **1**(1):37-51
- [18] Guo S, Wu X. Deriving private information from arbitrarily projected data. In: Proceedings of the 11th European conference on principles and practice of knowledge Discovery in databases (PKDD07). Warsaw, Poland. 2007
- [19] Balasubramaniam S, Kavitha V. A survey on data retrieval techniques in cloud computing. *Journal of Convergence Information Technology*. 2013;**8**(16):15-24
- [20] Liu J, Yifeng XU. Privacy preserving clustering by random response method of geometric transformation. Harbin, Heilong Jiang, China: IEEE. 2010: 181-188. DOI: 10.1109/ICICSE.2009.31
- [21] Balasubramaniam S, Kavitha V. Geometric data perturbation-based personal health record transactions in cloud computing. *The Scientific World Journal*. 2015;**2015**:927867, 1-927869. DOI: 10.1155/2015/927867
- [22] Chen K, Lui L. Geometric Data Perturbation for Privacy Preserving Outsourced Data Mining. London: Springer-Verlag Limited; 2010
- [23] Hyvarinen AK, Oja E. Independent Component Analysis. New York/Chichester/Weinheim/Brisbane/Singapore/Toronto: Wiley-Interscience; 2001
- [24] Brankovic L, Estivill-Castro V. Privacy issues in knowledge discovery and data mining. In: Proceedings of Australian Institute of Computer Ethic Conference (AICEC99). Melbourne, Victoria, Australia: Lecture Notes in Computer Science. 1999;**4213**:297-308. DOI:10.1007/11871637_30
- [25] Liu K, Giannella C, Kargupta H. An Attacker's view of distance preserving maps for privacy preserving data mining. In: European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD). Berlin, Germany; 2006
- [26] Li L, Zhang Q. A privacy preserving clustering technique using hybrid data transformation method. In: Grey Systems and Intelligent Services, 2009 GSIS 2009, IEEE International Conference. Nanjing, China: IEEE; 2010. DOI: 10.1109/GSIS.2009.5408151, 08
- [27] Rajesh N, Sujatha K, Kumar AALS. Survey on privacy preserving data mining techniques using recent algorithms. *International Journal of Computer Applications Foundation of Computer Science (FCS)*. 2016;**133**(7):30-33
- [28] Patel L, Gupta R. A survey of perturbation technique for privacy-preserving of data. *International Journal of Emerging Technology and Advanced Engineering Website*. 2013;**3**(6):162-166