A Robust Privacy Preservation by Combination of Additive and Multiplicative Data Perturbation for Privacy Preserving Data Mining

Bhupendra Kumar Pandya
Institute of Computer Science
Vikram University, Ujjain

Umesh Kumar Singh
Institute of Computer Science
Vikram University, Ujjain

Keerti Dixit
Institute of Computer Science
Vikram University, Ujjain

ABSTRACT
Most of our daily activities are now routinely recorded and analysed by a variety of governmental and commercial organizations for the purpose of security and business related applications. From telephone calls to credit card purchases, from internet surfing to medical prescription refills, we generate data with almost every action we take. These data sets need to be analyzed for identifying patterns which can be used to predict future behaviour. However, data owners may not be willing to share the real values of their data due to privacy reason. Hence, some amount of privacy preservation needs to be done on data before it is released. Privacy preserving data mining (PPDM) tends to transform original data, so that sensitive data are preserved. In this paper we have proposed a new method CAMDP (Combination of Additive and Multiplicative Data Perturbation) for privacy preserving in data mining.

Keyword
Additive and Multiplicative Data Perturbation

1. INTRODUCTION
Large volumes of detailed personal data are regularly collected and analyzed by applications using data mining. Such data include shopping, criminal records, medical history, credit records, among others [1]. On the one hand, such data is an important asset to business organizations and market-basket analysis both to decision making processes and to provide social benefits, such as medical research, crime reduction, national security, etc.[2]. On the other hand, analyzing such data opens new threats to privacy and autonomy of the individual if not done properly. We present a new data perturbation technique for privacy preserving outsourced data mining. A data perturbation procedure can be simply described as follows. Before the data owners publish their data, they change the data in certain way to disguise the sensitive information while preserving the particular data property that is critical for building meaningful data mining models. Perturbation techniques have to handle the intrinsic trade-off between preserving data privacy and preserving data utility, as perturbing data usually reduces data utility. Several perturbation techniques have been proposed for mining purpose recently, but these two factors are not satisfactorily balanced. For example, Traditional Multiplicative Data Perturbation approach [3,4] distort each data element independently, therefore Euclidean Distance among data records are usually not preserved, and the perturbed data can not be used for many data mining applications. This perturbation schemes are equivalent to additive perturbation after the logarithmic transformation. Due to the large volume of research in deriving private information from the additive noise perturbed data, the security of this perturbation scheme is questionable. The Distance Preserving data perturbation approach [5,6,7] is quite useful as it allows many interesting data mining algorithms to be applied directly to the perturbed data and produce an error-free result but cannot effectively protect data privacy. The attacker can get the original data after applying attack techniques. Hence the privacy of original data is vulnerable. The random projection perturbation approach [8,9,10] project the data onto a lower dimensional random space and can dramatically change its original form while preserving much of its distance related characteristics but with little loss of accuracy.

In this research paper, we propose a new multidimensional data perturbation technique: CAMDP (Combination of Additive and Multiplicative Data Perturbation) for Privacy Preserving Data Mining that can be applied for several categories of popular data mining models with better utility preservation and privacy preservation.

2. PROBLEM DESCRIPTION
The initial idea of it was to extend traditional data mining techniques to work with the perturbed data to mask sensitive information. The key issue is to get accurate data mining results using perturb data. The solutions are often tightly coupled with the data mining algorithms under consideration. The goal is to transform a given data set D into perturbed version D' that satisfies a given privacy requirement and loss minimum information for the intended data analysis task. In this paper data perturbation algorithms have been proposed for data set perturbation.

3. PROPOSED METHOD
Instead of applying one method alone if we apply the combination of additive and multiplicative methods, then it will be more difficult for an attacker to get back the original data. To achieve this goal here CAMDP(Combination of Additive and Multiplicative Data Perturbation) Method is proposed. This Method combines the strength of the translation and distance preserving method.

3.1 Translation based Perturbation
In TBP method, the observations of confidential attributes are perturbed using an additive noise perturbation. Here we apply the noise term applied for each confidential attribute which is constant and value can be either positive or negative.

3.2 Distance Based Perturbation

To define the distance preserving transformation, let us start with the definition of metric space. In mathematics, a metric space is a set S with a global distance function (the metric d) that, for every two points x, y in S, gives the distance between them as a nonnegative real number d(x, y). Usually, we denote a metric space by a 2-tuple (S, d). A metric space must also satisfy

1. d(x, y) = 0 iff x = y (identity),
2. d(x, y) = d(y, x) (symmetry),
3. d(x, y) + d(y, z) ≥ d(x, z) (triangle inequality).

3.3 Generation of Orthogonal Matrix

Many matrix decompositions involve orthogonal matrices, such as QR decomposition, SVD, spectral decomposition and polar decomposition. To generate a uniformly distributed random orthogonal matrix, we usually fill a matrix with independent Gaussian random entries, then use QR decomposition.

3.4 Data Perturbation Model

Translation and Orthogonal transformation-based data perturbation can be implemented as follows. Suppose the data owner has a private database Dn×n, with each column of D being a record and each row an attribute. The data owner generates a n × n noise matrix OR , and computes

$$D'_{n\times n} = D_{n\times n} * OR_{n\times n}$$

Where $OR_{n\times n}$ is generated by Translation and Orthogonal Transformation.

The perturbed data $D'_{n\times n}$ is then released for future usage. Next we describe the privacy application scenarios where orthogonal transformation can be used to hide the data while allowing important patterns to be discovered without error.

This technique has a nice property that it preserves vector inner product and distance in Euclidean space. Therefore, any data mining algorithms that rely on inner product or Euclidean distance as a similarity criteria are invariant to this transformation. Put in other words, many data mining algorithms can be applied to the transformed data and produce exactly the same results as if applied to the original data, e.g., KNN classifier, perception learning, support vector machine, distance-based clustering and outlier detection.

4. PROPOSED ALGORITHM

Algorithm: Privacy Preserving using proposed CAMDP Technique.

Input: Original Data ***D***.

Intermediate Result: Noise Matrix.

Output: Perturbed data stream ***D'.***

Steps:

1. **Given** input data $\boldsymbol{D_{n\times n}}$
2. Generate an Orthonal Matrix $\boldsymbol{O_{n\times n}}$ from the Original Data $\boldsymbol{D_{n\times n.}}$
3. Create Translation Matrix $\boldsymbol{T_{n\times n}}$.
4. Creat Matrix $\boldsymbol{OT_{n\times n}}$ by adding the Translation Matrix $\boldsymbol{T_{n\times n}}$ and Orthonal Matrix $\boldsymbol{O_{n\times n.}}$
5. Generate an Orthonal Matrix(noise matrix) $\boldsymbol{OR_{n\times n}}$ from the Matrix $\boldsymbol{OT_{n\times n.}}$
6. Create Perturbed Dataset $\boldsymbol{D'_{n\times n}}$ by multiplying Original Data $\boldsymbol{D_{n\times n}}$ and Noise Matrix $\boldsymbol{OR_{n\times n}}$.
7. Release Perturbed Data for Data Miner.
8. Stop

5. FLOW CHART

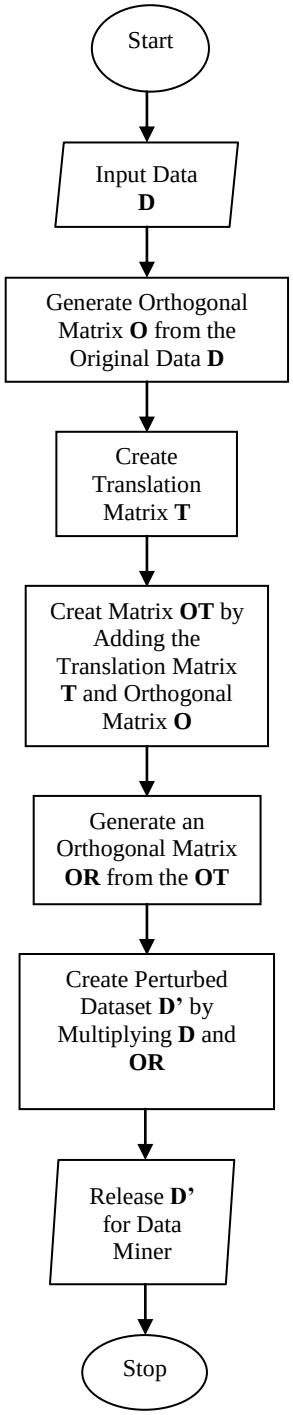


Table 1: Original Data

Table 1: Original Data

Foun dation	Maths	Physics	Com. Sc.	Phy. Prac.	Com. Sc. Prac	Project
56	73	38	42	39	42	42
49	47	22	36	37	42	39
55	57	40	33	39	42	40
60	50	34	53	37	41	38
50	37	11	25	38	41	38
48	61	31	36	40	43	41
61	64	40	40	39	42	39

Table 2: Noise Matrix generated by CAMDP technique

0.34	-0.31	0.32	-0.4	0.41	0.14	-0.59
0.54	0.04	0.37	0.1	0.37	-0.2	0.64
0.39	0.03	0.44	0	-0.8	0.01	-0.12
0.34	-0.1	-0.2	0.8	0.11	-0.2	-0.37
0.34	-0.55	-0.6	-0.3	-0.2	-0.3	0.17
0.37	0.77	-0.3	-0.3	0.04	-0.2	-0.2
0.3	0.02	-0.3	0.1	-0	0.89	0.15

Table 3: Perturbed Data

128	-5.42	3.93	-2.6	16.6	6.67	-2.29
102	-3.13	-12	-6.8	16.8	8.88	-11.4
116	-4.89	1.17	-11	7.76	9.18	-9.16
117	-8.54	-4.7	3.3	14.7	6.07	-23.7
88.5	-4.41	-17	-17	20.9	11.7	-12.8
114	-2.84	-5.4	-6.2	13.7	7.25	-2.36
124	-7.17	4.42	-7.1	13.6	6.62	-11

Comparison of Original and Perturbed Data

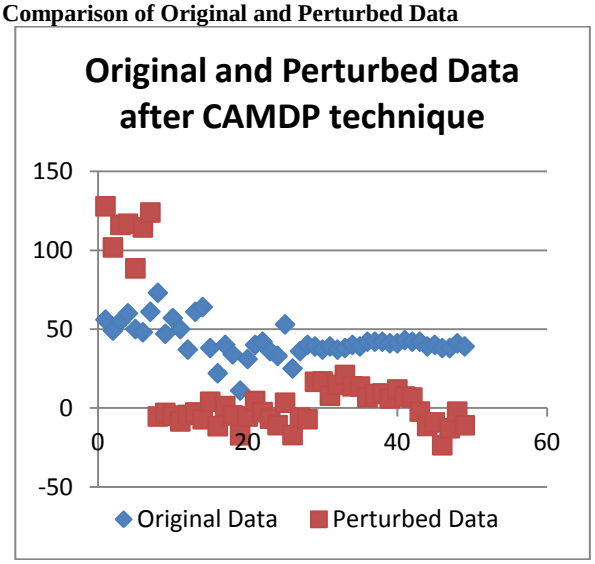


Figure 1

Euclidean Distance of Original Data

32	18.6	26.5	48.7	17.2	11	21.8
24	18.6	17.1	27.9	22.8	37	12.6
12	39.79	24.2	20.2	33.4	44	16.8

Euclidean Distance of Perturbed Data

32	18.6	26.5	48.7	17.2	11	21.8
24	18.6	17.1	27.9	22.8	37	12.6
12	39.79	24.2	20.2	33.4	44	16.8

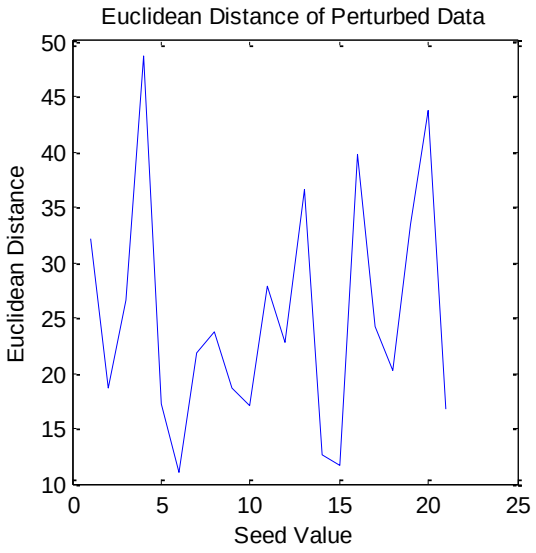
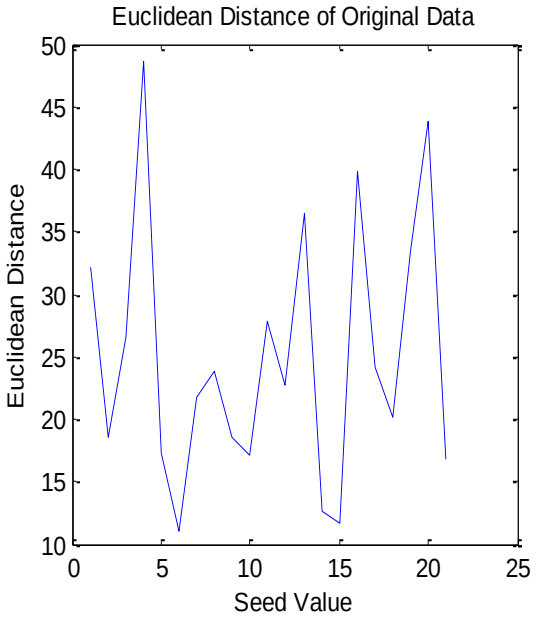


Figure 2 and 3



We have taken the original data which is result set of students. With this data we have generated a noise matrix with the help of CAMDP transformation and this resultant noise data set is multiplied with the original data set to form the perturb data. We have evaluated Euclidean Distance of original and perturbed data with pdist() fuction of Matlab. We have plotted the graphs 2 and 3 which shows the comparison between Euclidean Distances of original data and perturbed data after applying CAMDP Technique.

6. DISCUSSION

The above graph shows that the Euclidean Distance among the data records are preserved after perturbation. Hence the data perturbed by CAMDP technique can be used by various data mining applications such as k-means clustering, k_nearest neighbourhood classification, decision tree etc. And we get the same result as obtained with the original data.

7. CONCLUSION

The increasing ability to track and collect large amounts of data with the use of current hardware technology has lead to an interest in the development of data mining algorithms which preserve user privacy. We present a random CAMDP (Combination of Additive and Multiplicative Data Perturbation) approach for privacy preserving data Mining. CAMDP technique includes the linear combination of Distance Preserving perturbation and translation perturbation. Euclidean Distance is an important factor in K-means Clustering and KNN Classification. The experimental results have shown that the CAMDP technique can preserve the important distance related properties, thus by using this technique, data owners can share their data with data miners for various Data Mining Applications.

8. REFERENCE

[1] R. Agrawal and R. Srikant, "Privacy preserving data mining," In Proceedings of SIGMOD Conference on Management of Data, pp. 439-450, 2000.

[2] D. Agrawal and C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithm," In Proceedings of ACM SIGMOD, pp. 247-255, 2001.

[3] B. Pandya,U.K.Singh and K. Dixit, "Effectiveness of Multiplicative Data Perturbation for Perturbation for Privacy Preserving Data Mining" International Journal of Advanced Research in Computer Science, Vol. 5, No. 6 pp 112-115,2014.

[4] B. Pandya,U.K.Singh and K. Dixit, "Performance of Multiplicative Data Perturbation for Perturbation for Privacy Preserving Data Mining" International Journal for Research in Applied Science and Engineering Technology, Vol. 2, Issue VII, 2014.

[5] B. Pandya,U.K.Singh and K. Dixit, "An Analysis of Euclidean Distance Presrving Perturbation for Privacy Preserving Data Mining" International Journal for Research in Applied Science and Engineering Technology, Vol. 2, Issue X, 2014.

[6] B. Pandya,U.K.Singh and K. Dixit, "Performance of Euclidean Distance Presrving Perturbation for K-Means Clustering" International Journal of Advanced Scientific and Technical Research, Vol. 5, Issue 4, pp 282-289, 2014.

[7] B. Pandya,U.K.Singh and K. Dixit, "Performance of Euclidean Distance Presrving Perturbation for K-Nearest Neighbour Classification" International Journal of Computer Application, Vol. 105, No. 2, pp 34-36, 2014.

[8] B. Pandya,U.K.Singh and K. Dixit, "A Study of Projection Based Multiplicative Data Perturbation for Privacy Preserving Data Mining" International Journal of Application or Innovation in Engineering and Management, Vol. 3, Issue 11, pp 180-182,2014.

[9] B. Pandya,U.K.Singh and K. Dixit, "An Analysis of Projection Based Multiplicative Data Perturbation for K-Means Clustering" International Journal of Computer Science and Information Technologies, Vol. 5, Issue. 6, pp 8067-8069, 2014.

[10] B. Pandya,U.K.Singh and K. Dixit, "An Evaluation of Projection Based Multiplicative Data Perturbation for K-Nearest Neighbour Classification" International Journal of Science and Research, Vol. 3, Issue. 12, pp 681-684, 2014.