An Efficient K-Means Clustering by using Combination of Additive and Multiplicative Data Perturbation for Privacy Preserving Data Mining

Bhupendra Kumar Pandya, Umesh kumar Singh, Keerti Dixit

Institute of Computer Science, Vikram University, Ujjain

Abstract- The collection of digital information by governments, corporations, and individuals has created tremendous opportunities for knowledge- and information-based decision making. Driven by mutual benefits, or by regulations that require certain data to be published, there is a demand for the exchange and publication of data among various parties. Data in its original form, however, typically contains sensitive information about individuals, and publishing such data will violate individual privacy. Privacy preserving data mining (PPDM) tends to transform original data, so that sensitive data are preserved. In this research paper we analysis CAMDP (Combination of Additive and Multiplicative Data Perturbation) technique for k-means clustering as a tool for privacy-preserving data mining. We can show that K-Means Clustering algorithm can be *efficiently* applied to the transformed data and produce *exactly the same* results as if applied to the original data.

Index Terms- CAMDP, K-means clustering.

I. Introduction

Data mining in recent years with the database and artificial intelligence developed a new technology that the big amount of raw data to discover the hidden, useful information and knowledge to help policy makers to find the potential between the data Associated factors found to be ignored. Data mining because of its huge business prospect are now becoming an international data library and information policy-making in the field of cutting-edge research, and caused extensive academic and industry relations note [1]. At present, data mining has been in business management, production control, electronic commerce, market analysis and scientific science and many other fields to explore a wide range of applications [2]. The face of huge amounts of data, the first task is to sort them out, cluster analysis is to classify the raw data as a reasonable way. The so-called clustering is a group of physical or abstract objects, according to the degree of similarity between them, divided into several groups, and makes the same data objects within a group of high similarity, and different groups of data objects are not similar [3][4]. Clustering of objects is as ancient as the human need for describing the salient characteristics of men and objects and identifying them with a type. Therefore, it embraces various scientific disciplines: from mathematics and statistics to biology and genetics. Since clustering is the grouping of similar instances/objects, some sort of measure that can determine whether two objects are similar or dissimilar is required. There are two main type of measures used to estimate this relation: distance measures and similarity measures. Many clustering methods use distance measures to determine the similarity or dissimilarity between any pair of objects. In this paper, we analyze 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.

II. CAMDP Technique

The CAMDP technique is a Combination of Additive and Multiplicative Data Perturbation techniques. This Method combines the strength of the translation and distance preserving method.

2.1. 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.

2.2. 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).

2.3. 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.

2.4. Data Perturbation Model

Translation and Orthogonal transformation-based data perturbation can be implemented as follows. Suppose the data owner has a private database $\boldsymbol{D}_{n\times n}$, with each column of D being

a record and each row an attribute. The data owner generates a n × n noise matrix O_R, and computes

$$\boldsymbol{D'}_{n\times n} = \boldsymbol{D}_{n\times n} * \boldsymbol{OR}_{n\times n}$$

Where $\boldsymbol{OR}_{n\times n}$ is generated by Translation and Orthogonal Transformation.

The perturbed data $\boldsymbol{D'}_{n\times n}$ 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.

III. CAMDP Algorithm

Algorithm: Privacy Preserving using 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

Comparison of Original and Perturbed Data:

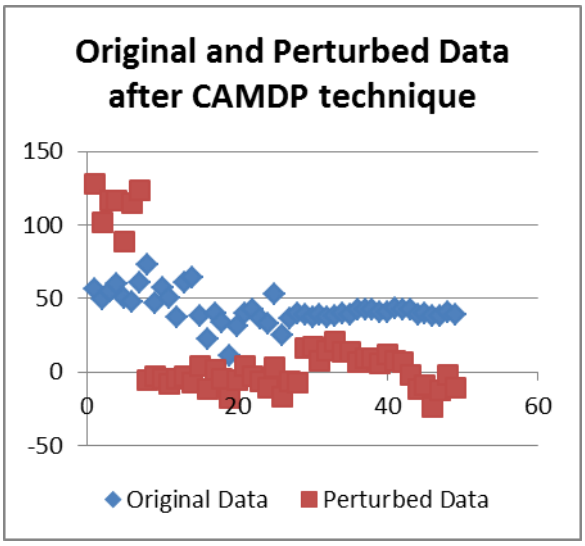

Figure 1

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

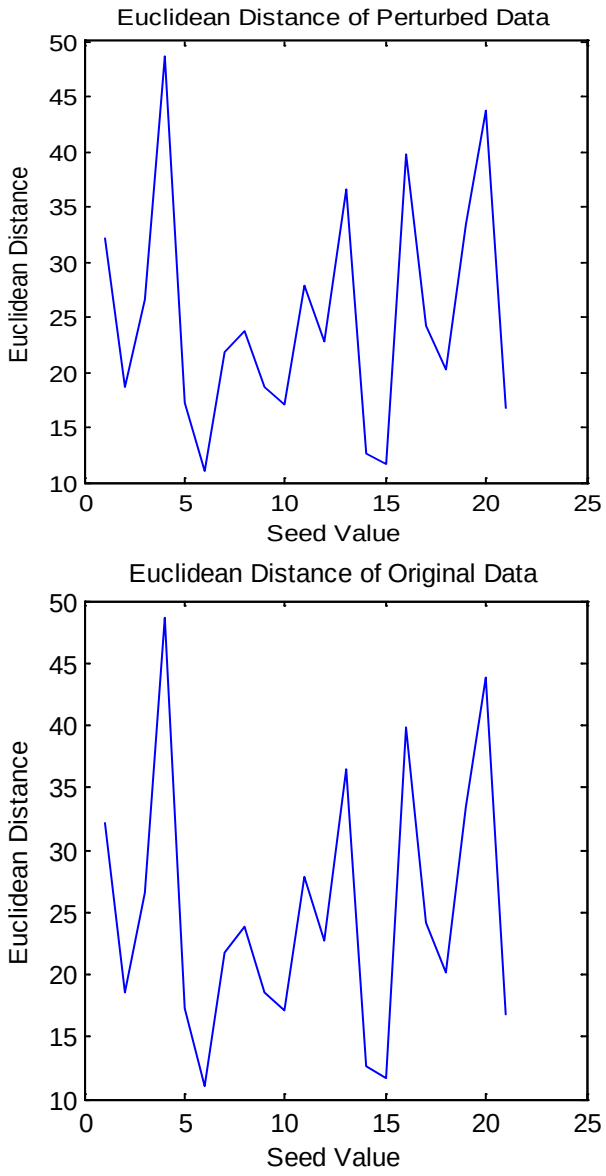

Figure 2 and 3

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 plotted graph 1 that shows the difference between Original and Perturbed Data. We have evaluated Euclidean Distance of original and perturbed data with pdist() fuction of Matlab. We have plotted the graph 2 and 3 which shows the comparison between Euclidean Distances of original data and perturbed data after applying CAMDP technique.

IV. 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.

V. Specific Partitional Clustering Techniques: K-means

The K-means algorithm discovers K (non-overlapping) clusters by finding K centroids ("central" points) and then assigning each point to the cluster associated with its nearest centroid. (A cluster centroid is typically the mean or median of the points in its cluster and "nearness" is defined by a distance or similarity function.) Ideally the centroids are chosen to minimize the total "error," where the error for each point is given by a function that measures the discrepancy between a point and its cluster centroid, e.g., the squared distance. Note that a measure of cluster "goodness" is the error contributed by that cluster. For squared error and Euclidean distance, it can be shown [5-7] that a gradient descent approach to minimizing the squared error yields the following basic K-means algorithm.

Basic K-means Algorithm for finding K clusters.

1. Select K points as the initial centroids.
2. Assign all points to the closest centroid.
3. Recompute the centroid of each cluster.
4. Repeat steps 2 and 3 until the centroids don't change (or change very little).

K-means has a number of variations, depending on the method for selecting the initial centroids, the choice for the measure of similarity, and the way that the centroid is computed. The common practice, at least for Euclidean data, is to use the mean as the centroid and to select the initial centroids randomly.

VI. Experimental Result Based on the K-means Clustering

We have taken the original data which is result set of students. With this data we have generated 3 clusters from the kmeans() function of matlab. And similarly we have generated 3 clusters by using the same function with the perturbed data. We have used silhouette function for plotting graph of the clustered data generated by the original data and also for plotting graph of the clustered data generated by perturbed data.

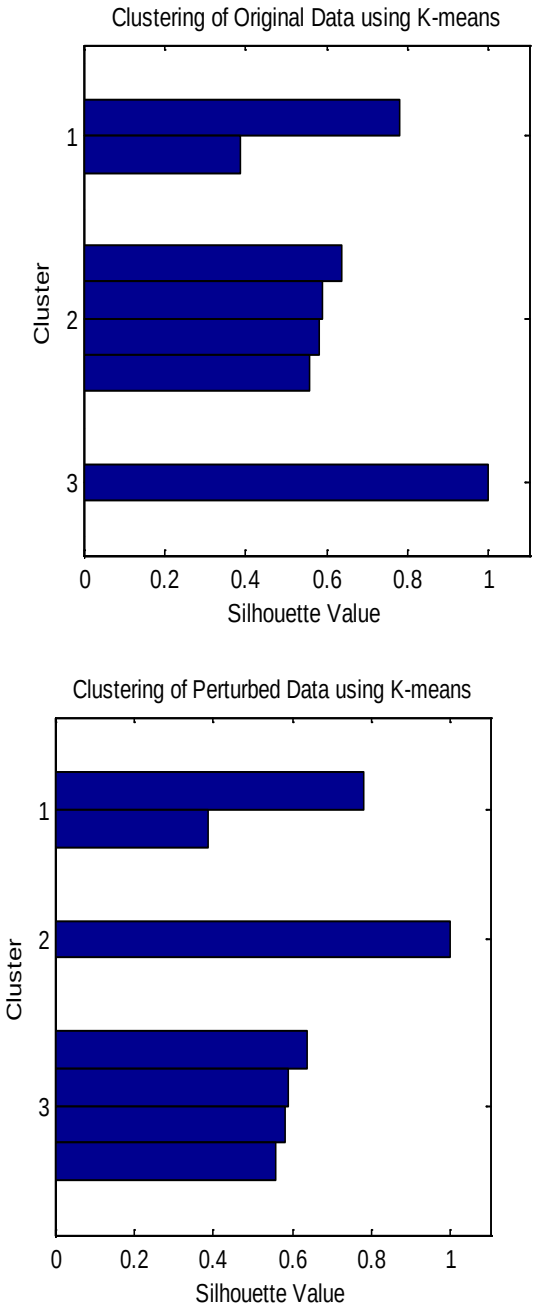

Figure 4 and 5

As depicted from the above graph it is clear that the data which shares same cluster (cluster 2) after applying the clustering on original data shares same cluster (cluster 3) after applying clustering on the perturbed data.

VII. Discussion

It is proved by the experimental result that we get the same result after applying clustering to the perturbed data as after applying clustering to the original data. Hence we can say that data perturbed by CAMDP technique can be used in clustering techniques and we can work with high dimensional data and large datasets. So we can use the perturbed data in various data mining applications like marketing, organization, land use, insurance, city planning etc.

VIII. Conclusion

In this research paper, we have analyzed the effectiveness of CAMDP technique CAMDP technique includes the linear combination of Distance Preserving perturbation and translation perturbation. This technique allows many interesting data mining algorithms to be applied directly to the perturbed data and produce an error-free result, *e.g.*, K-means clustering and K-nearest neighbor classification.

The tremendous popularity of K-means algorithm has brought to life many other extensions and modifications. Euclidean distance is an important factor in k-means clustering. In CAMDP technique the Euclidean distance is preserved after perturbation. Hence the data perturbed by this technique can be used in various clustering techniques.

References

[1] Strehl A, Ghosh J. Relationship-based clustering and visualization for high-dimensional data mining[J].INFORMS J COMPUT, 2003, 15(2):208-230.

[2] Milenova B.L. , Campos M.M.O-Cluster: scalable clustering of large high dimensional data sets[C].IEEE International Conference on Data Mining, 2002, 290-297.

[3] Daniel B.A. , Ping Chen Using Self-Similarity to Cluster Large Data Sets[J].Data Mining and Knowledge Discovery, 2003, 7(2):123-152.

[4] Wei Chi-Ping , Lee Yen-Hsien , Hsu Che-Ming. Empirical comparison of fast Partitioning-based clustering algorithms for large data sets[J].Expert Systems with Applications, 2003, 24(4):351-363.

[5] J. Han and M. Kamber. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, San Diego, CA 92101-4495,USA, 2001.

[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, "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.

Authors

First Author – Bhupendra Kumar Pandya, Institute of Computer Science, Vikram University, Ujjain, Email: bhupendra20pandya@yahoo.co.in
Second Author – Umesh kumar Singh, Institute of Computer Science, Vikram University, Ujjain, Email: umeshsingh@rediffmail.com
Third Author – Keerti Dixit, Institute of Computer Science, Vikram University, Ujjain, Email: keerti_dixit2007@yahoo.co.in