An Efficient KNN Classification by using Combination of Additive and Multiplicative Data Perturbation for Privacy Preserving Data Mining

[1]Bhupendra Kumar Pandya, [2]Umesh kumar Singh, [3]Keerti Dixit

[123]Institute of Computer Science, Vikram University, Ujjain

Abstract:

Nowadays there is huge amount of data being collected and stored in databases everywhere across the globe. Recently, there has been a growing interest in the data mining area, where the objective is the discovery of knowledge that is correct and of high benefit for users. Data miming consists of a set of techniques that can be used to extract relevant and interesting knowledge from data. Data mining has several tasks such as association rule mining, classification and prediction, and clustering. Classification techniques are supervised learning techniques that classify data item into predefined class label. It is one of the most useful techniques in data mining to build classification models from an input data set. The used classification techniques commonly build models that are used to predict future data trends. In this research paper we analysis CAMDP (Combination of Additive and Multiplicative Data Perturbation) technique for KNN classification as a tool for privacy-preserving data mining. We can show that KNN Classification algorithm can be *efficiently* applied to the transformed data and produce *exactly the same* results as if applied to the original data.

Keyword:- CAMDP, KNN classification.

1. Introduction

The ultimate goal of data mining is to extract knowledge from massive data. Knowledge is ideally represented as human-comprehensible patterns from which end-users can gain intuitions and insights. 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].Classification is one of the most important data mining task. Data classification is a two-step process. In the first step, which is called the learning step, a model that describes a predetermined set of classes or concepts is built by analyzing a set of training database instances. Each instance is assumed to belong to a predefined class. In the second step, the model is tested using a different data set that is used to estimate the classification accuracy of the model. If the accuracy of the model is considered acceptable, the model can be used to classify future data instances for which the class label is not known. At the end, the model acts as a classifier in the decision making process.

There are several techniques that can be used for classification such as decision tree, Bayesian methods, rule based algorithms, and Neural Networks. 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.

2. 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[3][4], 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. $d(x, y) + d(y, z) \geq d(x, z)$ (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 $\boldsymbol{D}$ being a record and each row an attribute. The data owner generates a n × n noise matrix $\boldsymbol{OR}$, 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.

3. 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:

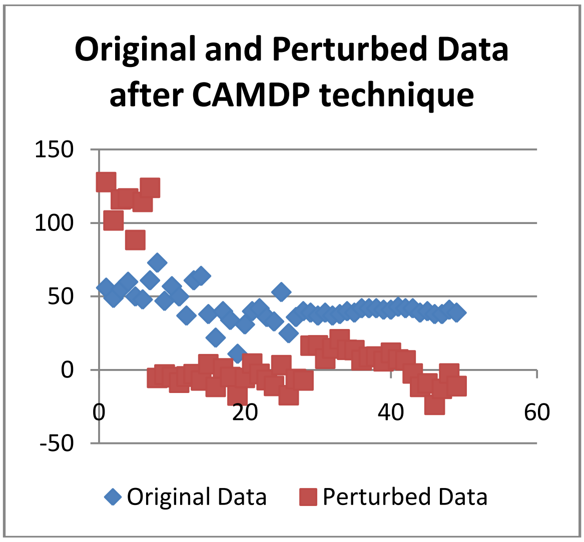


Figure 1

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

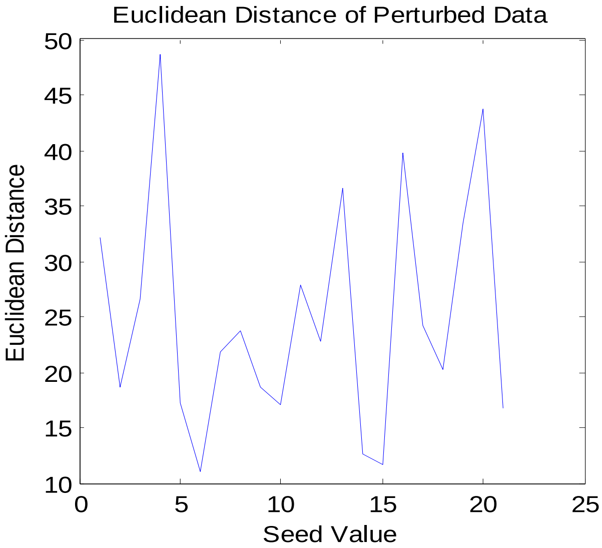
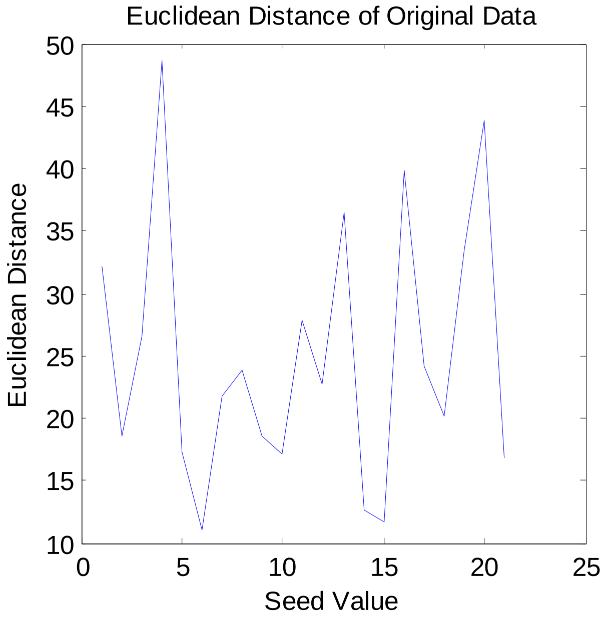


Figure 2 and 3



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.

4. 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.

5. Classification

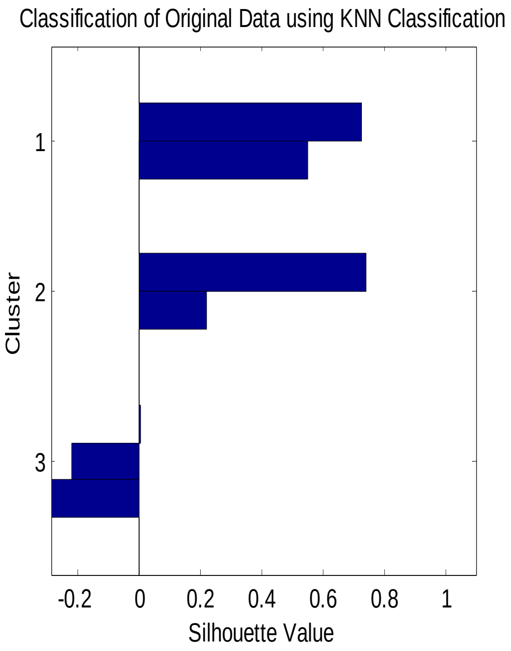
Classification is the process of building a classifier from a set of pre-classified (labelled) records. It discovers a pattern (model) that explains the relationship between the class and the non-class attributes [5]. A classifier is then used to assign (predict) a class attribute value to new unlabeled records. Classifiers also help to analyze the data sets better. They are expressed in different ways such as decision trees, sets of rules.

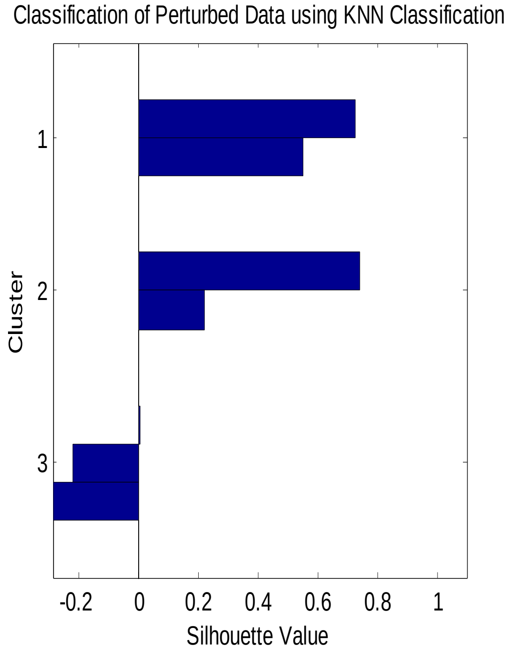
One of the techniques for building decision trees is based on information gain [5,6]. This technique first calculates the entropy (uncertainty) in estimating the class attribute values in the whole data set. It then divides the whole data set into two parts, based on an attribute, where each part contains a subset of values of the attribute and the other part contains the set of remaining values of the attribute. The attribute value that sits in the border of the two sets of values is also known as the splitting point.

Due to the division of the data set, the uncertainty in estimating the class attribute value changes which depends on the distribution of the class values. For example, let us assume that the domain size of the class attribute of a data set is two. In an extreme case, if all records belonging to one division have one class value and all other records belonging to the other division have the other class value then the uncertainty gets reduced to zero resulting in the maximum information gain. The decision tree building algorithm picks the best splitting point, among all possible splitting points of all non-class attributes, that reduces the uncertainty the most. The best splitting attribute is the root node of a decision tree and the best splitting point is the label on the edges. The same approach is applied again on each division of the data set and this process continues until the termination condition is met, resulting in a decision tree classifier.

6. Experiment result based on the KNN (k-nearest neighbour) classification

We have taken the original data which is result set of students and we have taken the subset of the original dataset and named it training. And we formed the group corresponding to training dataset. We have classified the original dataset in these groups on the basis of the training dataset with the knnclassify() function of matlab. And same classification applied on the perturbed data using the same function. We have used silhouette function for plotting graph of the classified data generated by the original data and also for plotting graph of the classified data generated by perturbed data.





As shown in the above graph the classification of original data and perturbed data remains same.

7. Discussion

It is proved by the experimental result that we get the same result after applying classification to the perturbed data as after applying classification to the original data. Hence we can say that data perturbed by this technique can be used in classification techniques.

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

8. 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 neighbour classification.

The tremendous popularity of KNN Classification 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 classification techniques.

9. 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] 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.

[4] 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.

[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, "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.