Performance of Euclidean Distance Preserving Perturbation for K-nearest Neighbor Classification

Bhupendra KumarPandya
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
Data Mining has many applications in the real world. One of the most important and widely found problems is that of classification. Recently, distance preserving data perturbation has gained attention because it mitigates the privacy/accuracy trade-off by guaranteeing perfect accuracy. Many important data mining algorithms can be *efficiently* applied to the transformed data and produce *exactly the same* results as if applied to the original data. *e.g.*,distance-based clustering and k-nearest neighbor classification. In this research paper we analysis Euclidean distance-preserving data perturbation for k-nearest neighbor classification as a tool for privacy-preserving data mining.

Keyword
K- nearest neighbor classification

1. INTRODUCTION
Data mining is a well-known technique for automatically and intelligently extracting information or knowledge from a large amount of data, however, it can also disclosure sensitive information about individuals compromising the individual's right to privacy [1]. A number of effective methods for privacy preserving data mining have been proposed. But most of these methods might result in information loss and side-effects in some extent, such as data utility-reduced, data mining efficiency-downgraded, etc. That is, an essential problem under the context is trade-off between the data utility and the disclosure risk. This paper provides an analysis of the Euclidean distance preserving methods for k-nearest neighbor classification as a tool for privacy preserving data mining.

2. DISTANCE PRESERVING PERTURBATION
This section offers an overview of distance preserving Perturbation: its definition, application scenarios, etc. Throughout this chapter (unless otherwise stated), all matrices and vectors discussed are assumed to have real entries. All vectors are assumed to be column vectors and M′ denotes the transpose of any matrix M. An m × n matrix M is said

to be orthogonal if $M' M = I_n$, the n × n identity matrix. If M is square, it is orthogonal if and only if $M' = M^{-1}$ [2]. The determinant of any orthogonal matrix is either +1 or −1. Let O_n denotes the set of all n × n, orthogonal matrices.

2.1 Definition and Fundamental Properties
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).

A metric space (S_1, d_1) is isometric to a metric space (S_2, d_2) if there is a bijection T: S1 → S2 that preserves distances. That is, $d_1(x, y) = d_2(T(x), T(y))$ for all x, y ϵ S1. The metric space which most closely corresponds to our intuitive understanding of space is the Euclidean space, where the distance d between two points is the length of the straight line connecting them. In this chapter, we specifically consider the Euclidean space and define d(x, y) = ||x − y||, the l^2-norm of vector x − y. A function $T : R^n \rightarrow R^n$ is distance preserving in the Euclidean space if for all x, y ϵ R^n, ||x − y|| = ||T(x) − T(y)||. Here T is also called a rigid motion. It has been shown that any distance preserving transformation is equivalent to an orthogonal transformation followed by a translation [2], . In other words, there exists M_T ϵ O_n and v_T ϵ R^n such that T equals x ϵ $R^n \rightarrow M_T x + v_T$. If T fixes the origin, T(0) = 0, then $v_T = 0$; hence, T is an orthogonal transformation. Henceforth we assume T is a distance preserving transformation which fixes the origin – an orthogonal transformation. Such transformations preserve the length (l^2-norm) of vectors: ||x|| = ||T(x)|| (i.e., given any M_T ϵ On, ||x|| = $||M_T x||$). Hence, they move x along the surface of the hyper-sphere centered at the origin with radius ||x||. From a geometric perspective, an orthogonal transformation is either a rigid rotation or a rotoinversion (a rotation followed by a reflection). This property was originally discovered by Schoute in 1891 [3]. Coxeter [4] summarized Schoute's work and proved that every orthogonal transformation can be expressed as a product of commutative rotations and reflections. To be more specific, let Q denote a rotation, R a reflection, 2q the number of conjugate imaginary eigenvalues of the orthogonal matrix M, and r the number of (-1)'s in the n − 2q real eigenvalues. The orthogonal transformation is expressible as $Q^q R^r (2q + r \leq n)$. Especially, in 2D space, det(M) = 1 corresponds to a rotation, while det(M) = −1 represents a reflection.

2.2 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. Stewart [5] replaced this with a more efficient idea that Diaconis and Shahshahani [6] later generalized as the subgroup algorithm. We refer the reader to these references for detailed treatment of this subject.

2.3 Data Perturbation Model

Orthogonal transformation-based data perturbation can be implemented as follows. Suppose the data owner has a private database $X_{n\times m}$, with each column of X being a record and each row an attribute. The data owner generates an n × n orthogonal matrix M_T , and computes

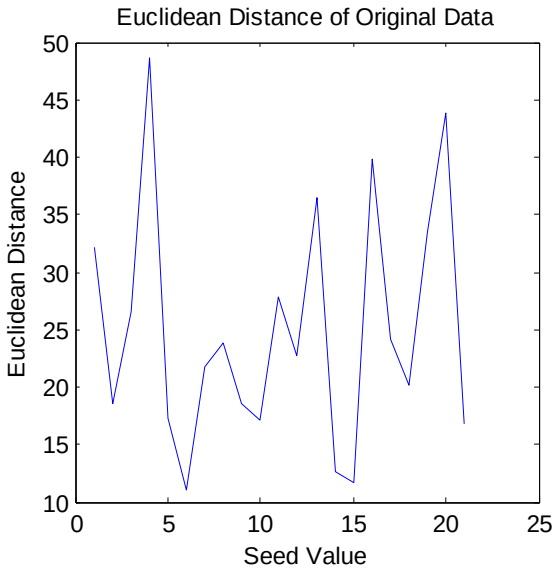
$$Y_{n\times m} = M_{Tn\times n} X_{n\times m}$$

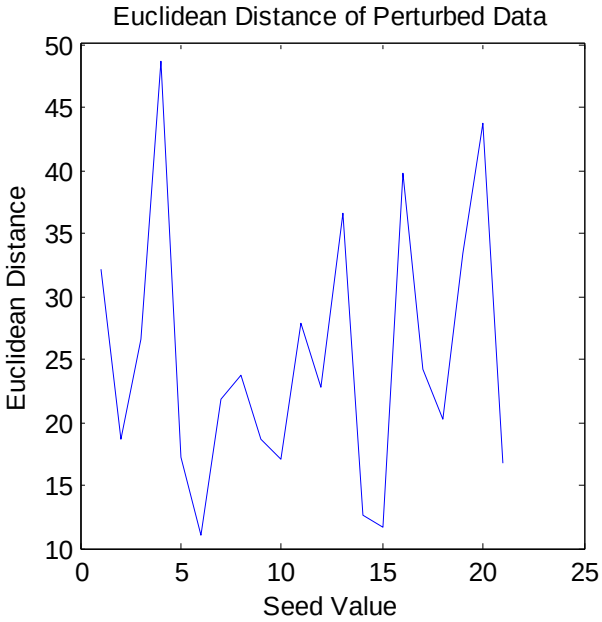
The perturbed data $Y_{n\times m}$ is then released for future usage. Next we describe the privacy application scenarios where orthogonal transformation can be used to hide the data while allowing important patterns to be discovered without error.

Orthogonal transformation 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 orthogonal 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. We refer the reader to [7] for a simple proof of rotation-invariant classifiers.

In this study we have Students result database of Vikram University, Ujjain. I randomly selected 7 rows of the data with only 7 attributes (Marks of Foundation, Marks of Mathematics, Marks of Physics, Marks of Computer Science, Marks of Physics Practical, Marks of Computer Science Practical and Marks of Job Oriented Project).

With this data we have generated a noise matrix with the help of orthogonal 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 graph 1.1 which shows the comparison between Euclidean Distances of original data and perturbed data after applying Distance Preserving Perturbation.





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

3. CLASSIFICATION

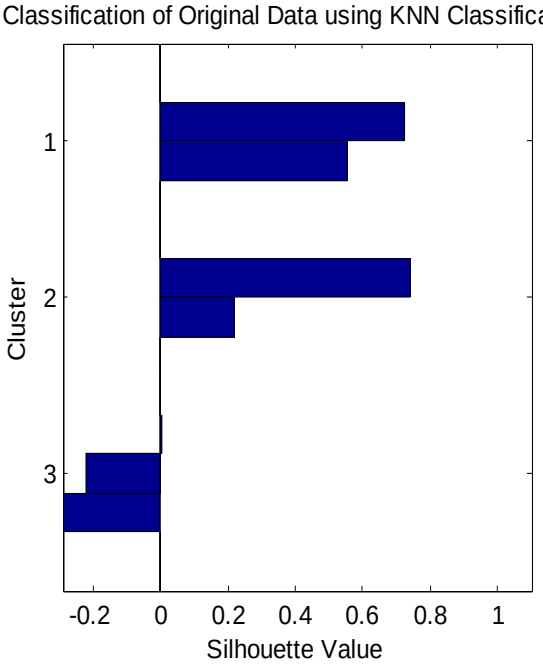
Classification is the process of building a classifier from a set of pre-classified (labelled) records. It discovers a pattern (model) that explains the relationship between the class and the non-class attributes [8]. A classifier is then used to assign (predict) a class attribute value to new unlabeled records. Classifiers also help to analyze the data sets better. They are expressed in different ways such as decision trees, sets of rules.

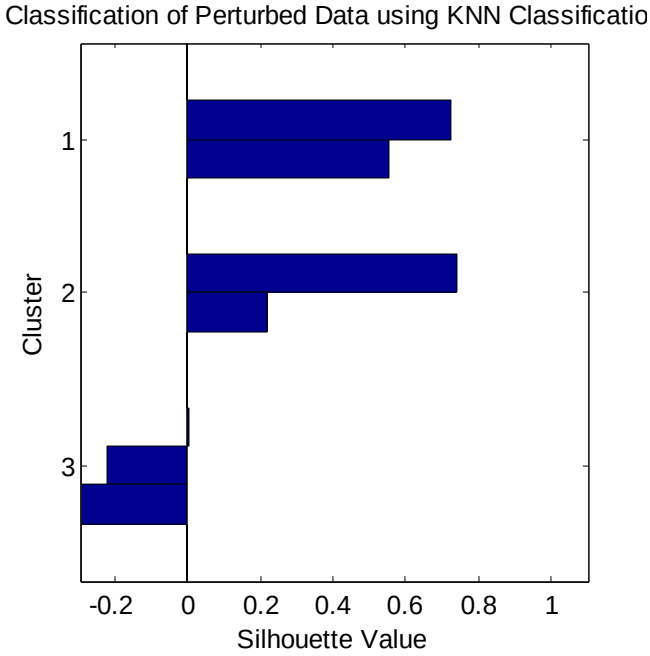
One of the techniques for building decision trees is based on information gain [8]. This technique first calculates the entropy (uncertainty) in estimating the class attribute values in the whole data set. It then divides the whole data set into two parts, based on an attribute, where each part contains 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.

4. 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.

5. 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.

6. CONCLUSION

In this research paper, we have analyzed the effectiveness of Distance preserving perturbation and we considered the use of distance-preserving maps (with origin fixed) as a data perturbation technique for privacy preserving data mining. This technique 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, *e.g.*, K-means clustering and K-nearest neighbor classification.

7. REFERENCE

[1] Han Jiawei, M. Kamber, Data Mining: Concepts and Techniques, Beijing: China Machine Press, pp.1-40,2006.

[2] M. Artin, Algebra. Prentice Hall, 1991.

[3] P. H. Schoute, "Le d´eplacement le plus g´en´eral dans l'espace `an dimensions," Annales de l'´Ecole Polytechnique de Delft, vol. 7, pp. 139–158, 1891.

[4] H. S. M. Coxeter, Regular Polytopes, 2nd ed., 1963, ch. XII, pp. 213–217.

[5] G. W. Stewart, "The efficient generation of random orthogonal matrices with an application to condition estimation," SIAM Journal of Numerical Analysis, vol. 17, no. 3, pp. 403–409, 1980.

[6] P. Diaconis and M. Shahshahani, "The subgroup algorithm for generating uniform random variables," Probability in Engineering and Information Sciences, vol. 1, pp. 15–32, 1987

[7] K. Chen and L. Liu, "Privacy preserving data classification with rotation perturbation," in Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05), Houston, TX, November 2005, pp. 589–592.

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