International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

An Evaluation of Projection Based Multiplicative Data Perturbation for KNN Classification

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

Institute of Computer Science, Vikram University, Ujjain, Madhya Pradesh, India

Abstract: ***Random projections have recently emerged as a powerful method for dimensionality reduction. In random projection (RP), the original high-dimensional data is projected onto a lower-dimensional subspace using a random matrix whose columns have unit lengths. In this method the data is projected on to a random subspace, which preserves the approximate Euclidean distances between all pairs of points after the projection. In this research paper we give experimental results on using RP as a dimensionality reduction tool and analysis Projection Based Multiplicative data perturbation for KNN Classification as a tool for privacy-preserving data mining.***

Keywords: Random Projection, KNN Classification.

1. Introduction

In many applications of data mining, the high dimensionality of the data restricts the choice of data processing methods. Such application areas include the analysis of market basket data, text documents, image data and so on; in these cases the dimensionality is large due to either a wealth of alternative products, a large vocabulary, or the use of large image windows, respectively. RP has been found to be a computationally efficient, yet sufficiently accurate method for dimensionality reduction of high-dimensional data sets. This approach is fundamentally based on the Johnson-Lindenstrauss lemma [1], which notes that any set of m points in n-dimensional Euclidean space can be embedded into an O (Inm/ϵ2) dimensional space such that the pair wise distance of any two points is maintained with a high probability. Therefore, by projecting the data onto a lower dimensional random space, we can dramatically change its original form while preserving much of its distance-related characteristics. This research paper presents extensive theoretical analysis and experimental results on the accuracy and privacy of the random projection-based data perturbation technique.

1.1 Definition and Fundamental Properties

Random projection refers to the technique of projecting a set of data points from a high dimensional space to a randomly chosen lower dimensional space. Mathematically, let $X \epsilon R^{n\times m}$ be m data points in n-dimensional space. The random projection method multiplies X by a random matrix $R \epsilon R^{k\times n}$, reducing the n dimensions down to just k. It is well known that random projection preserves pairwise distances in the expectation. This technique has been successfully applied to a number of applications, for example, VLSI layout [2], nearest-neighbor search [3, 4], image and text clustering [5], distributed decision tree construction [6], motifs in bio-sequences [7] discovery, high-dimensional Gaussian mixture models learning [8], half spaces and intersections of half spaces learning [9].

The following are the steps to reduce the dimensionality of the data by random projections: Suppose that we have a data set $X=\{x_1, \ldots. x_n\}$ where each data point is a p dimensional vector such that $x_i \epsilon R^p$ and we need to reduce the data to a q dimensional space such that $1 \leq q < p$.

1) Arrange the data into a p × n matrix where p is the dimensionality of the data and n is the number of data points.
2) Generate a q × p random projection matrix R* using the MATLAB randn (q, p) function.
3) Multiply the random projection matrix with the original data in order to project the data down into a random projection space.

$$X^*_{q\times n} = R^*_{q\times p} * X_{p\times n}$$

Thus we can see that transforming the data to a random projection space is a simple matrix multiplication with the guarantees of distance preservation.

2. Classification by Random Projection

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

2.1 K Nearest Neighbors Classification

KNN is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure.

Algorithm (Evaluating classification)

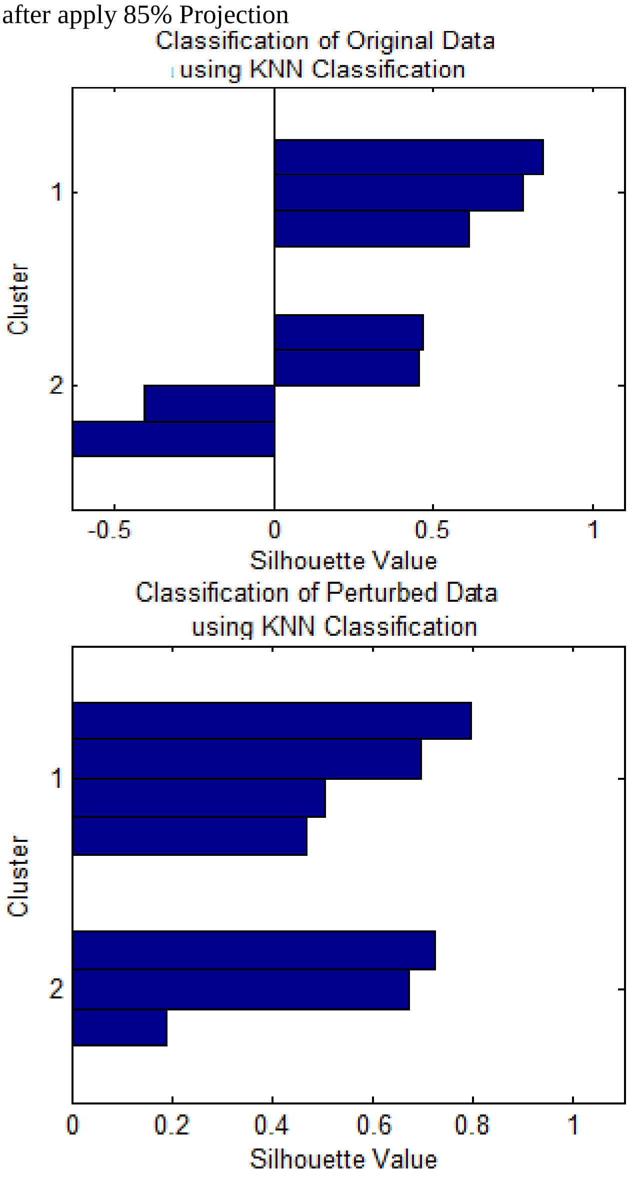
```
Requires: Dataset D, original dimensions d, projection dimensions k
1. For i = 1:100(splits)
2. Split D in to training set and test set at random
3. Train and test the data in the data space and obtain accuracy
4. Generate a (k, d) random matrix R using MATLAB randn (k, d) function.
5. Project the training and test data into the random projection space R*D.
6. Train and test the data in the reduced space and obtain accuracy
7. end for
```

3. Experimental Results

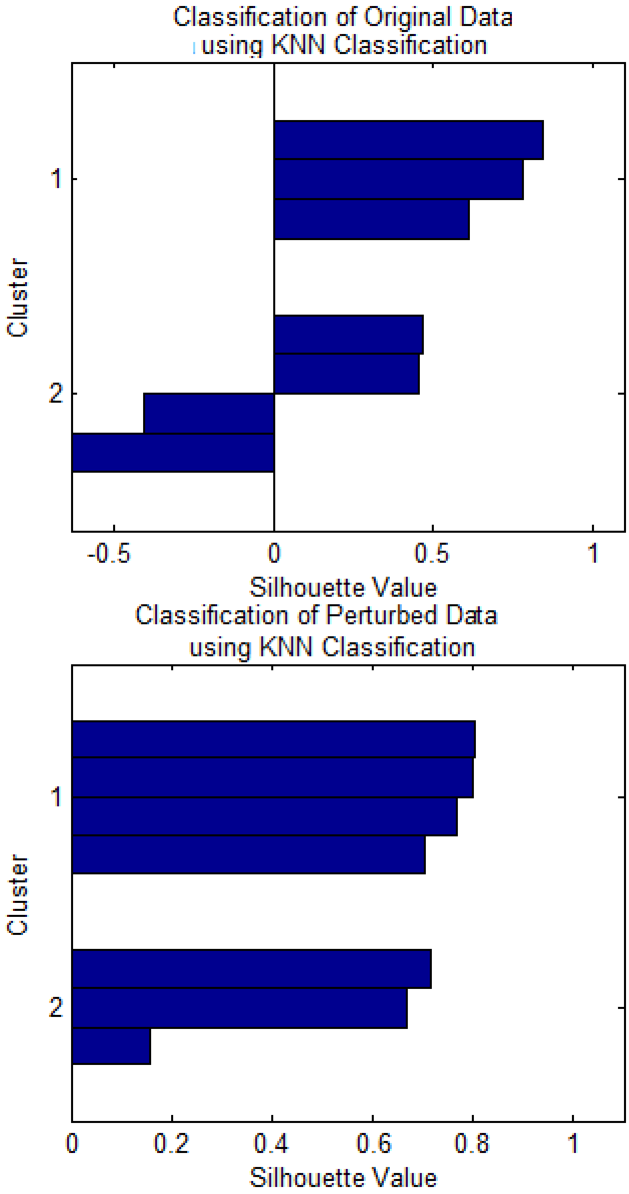
In this study we have Students result database of Vikram University, Ujjain. I randomly selected 7 rows of the data with only 7 attributes (Marks of Foundation, Marks of Mathematics, Marks of Physics, Marks of Computer Science, Marks of Physics Practical, Marks of Computer Science Practical and Marks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

With this data we have generated noise matrices with the help of different different projections and these resultant noise data sets are multiplied with the original data set to form the perturb data sets. We have evaluated Euclidean Distance of original and perturbed data sets with pdist() fuction of Matlab. According to the expectation the Euclidean Distance among the data records are preserved after perturbation. With the Original data we have generated 2 clusters from the knnclassifier() function of matlab. And similarly we have generated 2 clusters by using the same function with the perturbed data sets. 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 sets.

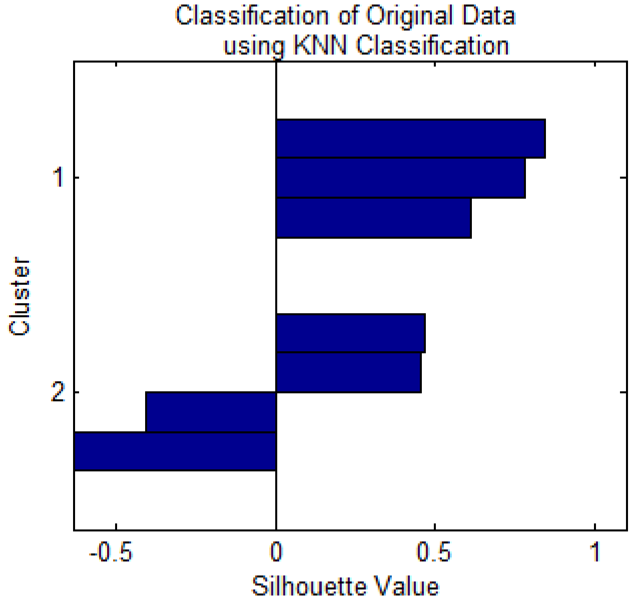
KNN Classification of Original Data and Perturbed Data after apply 85% Projection

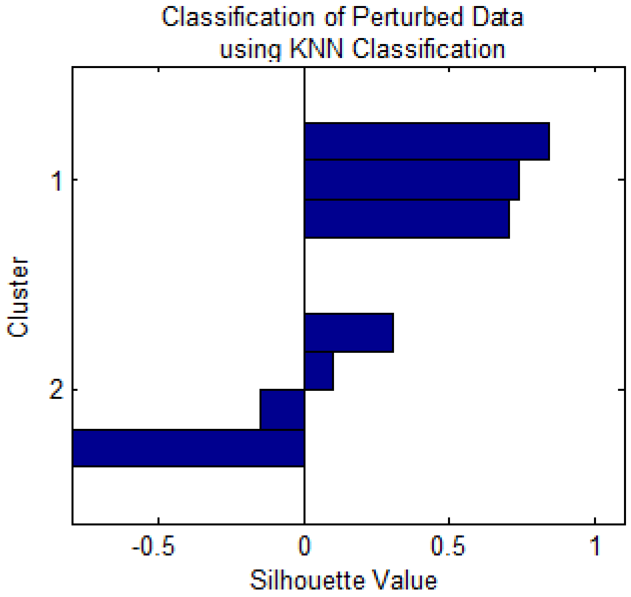


KNN Classification of Original Data and Perturbed Data after apply 70% Projection –

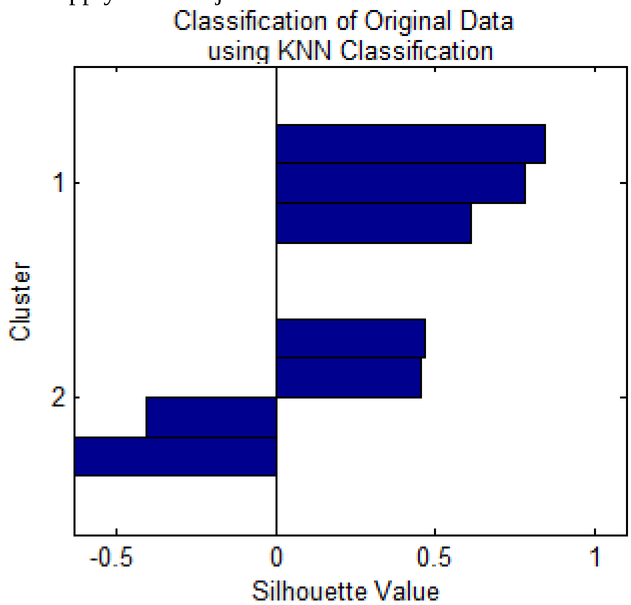


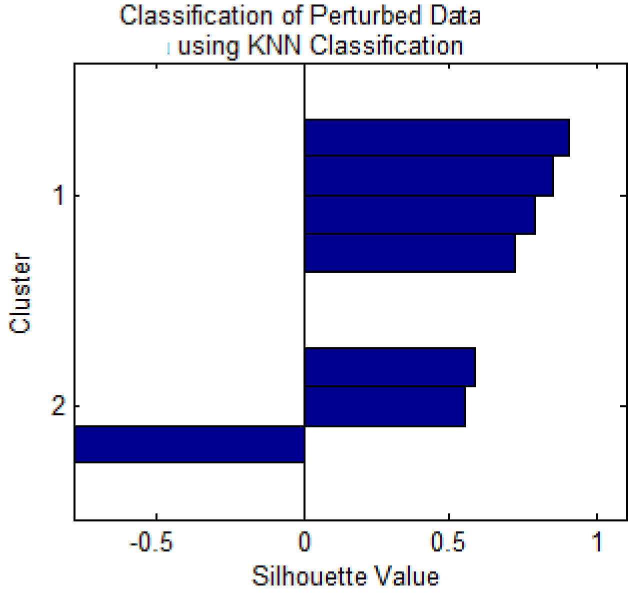
KNN Classification of Original Data and Perturbed Data after apply 60% Projection -





KNN Classification of Original Data and Perturbed Data after apply 45% Projection -





4. Discussion

It is proved by the experimental result that we get the expected 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 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.

5. Conclusion

In this research paper, we have analyzed the effectiveness of Projection based perturbation and we considered the use of this technique 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 expected result, e.g., KNN classification, with little loss of accuracy.

The tremendous popularity of KNN classification algorithm has brought to life many other extensions and modifications. Euclidean distance is an important factor in KNN 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 classification techniques.

Reference

[1] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, “On the privacy preserving properties of random data perturbation techniques,” in Proceedings of the IEEE International Conference on Data Mining, Melbourne, FL, November 2003.

[2] Z. Huang, W. Du, and B. Chen, “Deriving private information from randomized data,” in Proceedings of the 2005 ACM SIGMOD Conference, Baltimroe, MD, June 2005, pp. 37–48.

[3] S. Guo and X. Wu, “On the use of spectral filtering for privacy preserving data mining,” in Proceedings of the 21st ACM Symposium on Applied Computing, Dijon, France, April 2006, pp. 622–626.

[4] N. R. Adam and J. C. Worthmann, “Security-control methods for statistical databases: a comparative study,” ACM Computing Surveys (CSUR), vol. 21,no. 4, pp. 515–556, 1989.

[5] G. T. Duncan and S.Mukherjee, “Optimal disclosure limitation strategy in statistical databases: Dterring tracker attacks through additive noise,” Journal of The American Statistical Association, vol. 95, no. 451, pp. 720–729, 2000.

[6] R. Gopal, R. Garfinkel, and P. Goes, “Confidentiality via camouflage: The cvc approach to disclosure limitation when answering queries to databases,” Operations Research, vol. 50, no. 3, pp. 501–516, 2002.

[7] D. Agrawal and C. C. Aggarwal, “On the design and quantification of privacy preserving data mining algorithms,” in Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles

of Database Systems, Santa Barbara, CA, 2001, pp. 247–255.

[8] S. Guo, X. Wu, and Y. Li, "On the lower bound of reconstruction error for spectral filtering based privacy preserving data mining," in Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'06), Berlin, Germany, 2006.

[9] K. Muralidhar, R. Parsa, and R. Sarathy, "A general additive data perturbation method for database security," Management Science, vol. 45, no. 10, pp. 1399–1415, 1999.