iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** X **Month of publication:** October 2014

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology(IJRASET)

An analysis of Euclidean Distance preserving perturbation for Privacy Preserving Data Mining

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

Institute of Computer Science, Vikram University, Ujjain

Abstract:- Privacy preserving data mining is a novel research direction in data mining. In recent years, with the rapid development in Internet, data storage and data processing technologies, privacy preserving data mining has been drawn increasing attention. 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 as a tool for privacy-preserving data mining.

Keyword:- Distance Preserving Perturbation

I. 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 privacy preserving data mining, and points out their merits and demerits.

1.1 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.

1.1.1 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.

1.1.2 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.

1.1.3 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).

Table 1.1: Shows the original dataset.
Table 1.2: Orthogonal matrix
Table 1.3: perturbed data after apply distance preserving perturbation.

Table 1.1: Original Data

Foundation	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 1.2: Orthogonal Matrix

-0.42	0.4	0.23	-0.42	0.42	-0.5	-0.2
-0.34	-0.33	-0	-0.27	-0.26	-0.3	0.74
-0.38	0.17	0.17	0.67	-0.44	-0.4	-0.1
-0.39	-0.03	-0.9	-0.09	-0.08	0.02	-0.3
-0.31	-0.79	0.22	0.14	0.35	0.03	-0.3
-0.38	0.07	0.34	-0.38	-0.48	0.56	-0.2
-0.41	0.27	-0	0.36	0.46	0.5	0.42

Table 1.3: Perturbed Data

-125	-13.7	3.06	-16.8	-2.96	-15	22
-102	-9.57	4.08	-18.5	6.27	0.3	7.56
-115	-8.69	11.4	-10.6	-0.28	-11	11.7
-119	-5.02	-6.4	-17.4	3.53	-11	0.99
-90	-8.46	12.1	-22.4	15.4	5.4	4.11
-112	-14.9	6.04	-15	-0.25	-5	16.5
-122	-9.12	6.57	-16	-0.6	-17	13.4

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

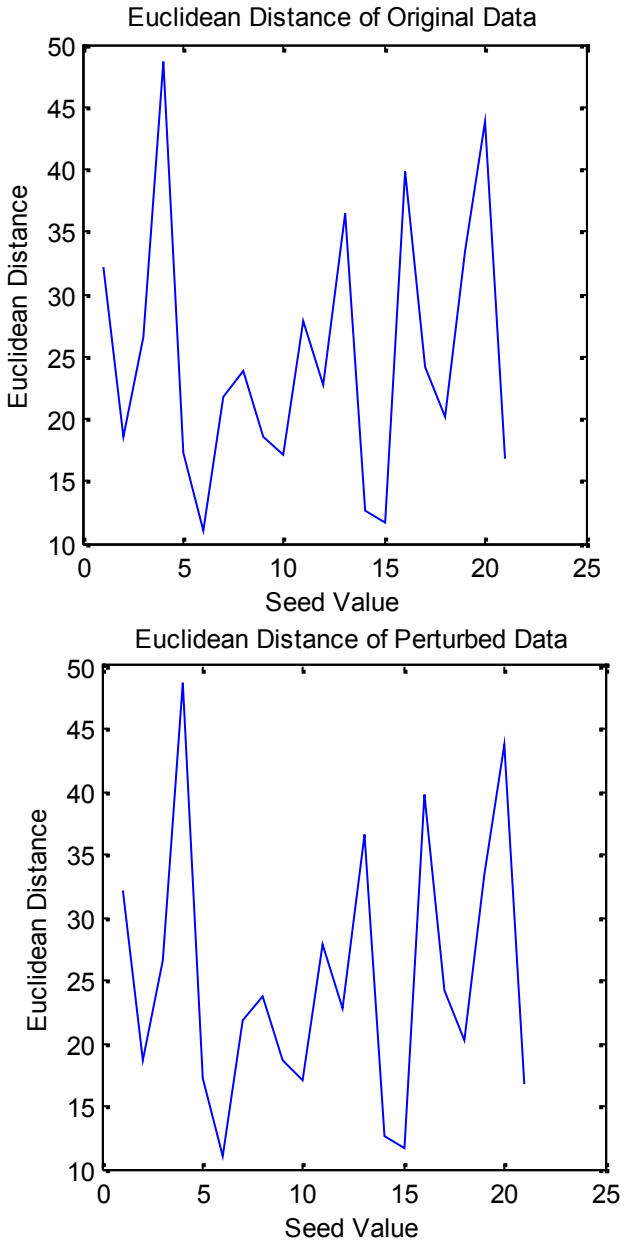


Figure 1.1

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 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.

II. DISCUSSION

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, k_nearest neighbourhood classification, decision tree etc. And we get the same result as obtained with the original data.

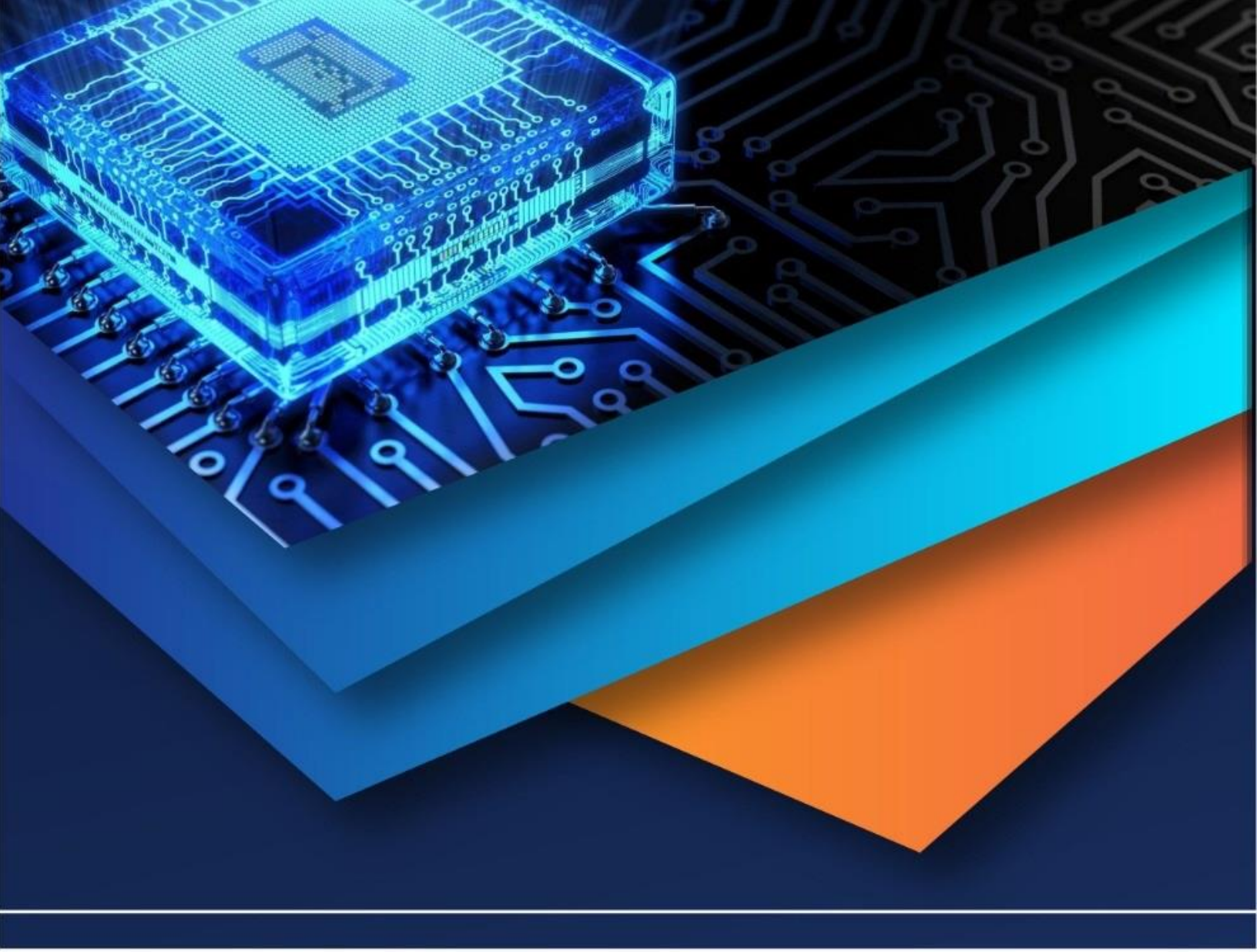
III. CONCLUSION

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.

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 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.

REFERENCES

[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.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY