DISTANCE BASED TRANSFORMATION FOR PRIVACY PRESERVING DATA MINING USING HYBRID TRANSFORMATION

Hanumantha Rao Jalla[1] and P N Girija[2]

[1]Department of Information Technology, CBIT, Hyderabad, A.P, INDIA
hanu_it2007@yahoo.co.in
[2]School of Computer and Information Sciences, UOH, Hyderabad, A.P, INDIA
pn_girija@yahoo.com

ABSTRACT

Data mining techniques are used to retrieve the knowledge from large databases that helps the organizations to establish the business effectively in the competitive world. Sometimes, it violates privacy issues of individual customers. This paper addresses the problem of privacy issues related to the individual customers and also propose a transformation technique based on a Walsh-Hadamard transformation (WHT) and Rotation. The WHT generates an orthogonal matrix, it transfers entire data into new domain but maintain the distance between the data records these records can be reconstructed by applying statistical based techniques i.e. inverse matrix, so this problem is resolved by applying Rotation transformation. In this work, we increase the complexity to unauthorized persons for accessing original data of other organizations by applying Rotation transformation. The experimental results show that, the proposed transformation gives same classification accuracy like original data set. In this paper we compare the results with existing techniques such as Data perturbation like Simple Additive Noise (SAN) and Multiplicative Noise (MN), Discrete Cosine Transformation (DCT), Wavelet and First and Second order sum and Inner product Preservation (FISIP) transformation techniques. Based on privacy measures the paper concludes that proposed transformation technique is better to maintain the privacy of individual customers.

KEYWORDS

Privacy preserving, Walsh-Hadamard transformation, Rotation and classification

1. INTRODUCTION

Explosive growth in data storing and data processing technologies has led to the creation of huge databases that contains fruitful information. Data mining techniques are retrieving hidden patterns from the large databases. Sometimes, the organizations share their own data to third party or data miners to get useful information. So, the original data is exposed to many parties. It violates privacy issues of individual customers. Privacy infringement is an important issue in the Data Mining. People and organizations usually do not tend to provide their private data or locations to the public because of the privacy concern [1]. The researchers are intended to address this problem on the topic of Privacy Preserving Data Mining (PPDM). These methods have been developed for different purposes, such as data hiding, knowledge hiding, distributed PPDM and privacy aware knowledge sharing in different data mining tasks [2].

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 15–23, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4502

The issue of privacy protection in classification has been raised by many researchers [3, 4]. The objective of privacy preserving data classification is to build accurate classifiers without disclosing private information while the data is being mined. The performance of privacy preserving techniques should be analysed and compared in terms of both the privacy protection of individual data and the predictive accuracy of the constructed classifiers.

Recent research in the area of privacy preserving data mining has devoted much effort to determine a trade-off between privacy and the need for knowledge discovery, which is crucial in order to improve decision-making processes and other human activities. Mainly, three approaches are being adopted for privacy preserving data mining namely, heuristic based, cryptographic based and reconstruction based [5]. Heuristic based techniques are mainly adopted in centralized database scenario, whereas cryptographic based technique finds its application in distributed environment. There is a clear tradeoff between accuracy of knowledge and the privacy. That is higher the accuracy-lower the privacy and lower the accuracy-higher the privacy. Hence, privacy preserving data mining remains as an open research issue. Some data perturbation techniques which are maintaining data mining utilities may not satisfy statistical properties. However some perturbation techniques like SAN and MN may satisfy statistical properties which are lagging in privacy issues.

In this paper we suggest a Hybrid transformation technique it maintains data mining utilities and statistical properties like mean and standard deviation of the original data without information loss. Also we preserve the Euclidean distance between the data records before and after the transformation. WHT is an attractive alternative to the Fourier Transforms because it is computationally more efficient, and thus performs fast on digital computer.

This paper is organized as follows: section 2 discuss about the related work. Section 3 focus on Walsh-Hadamard Transformation, section 4 talk about usage of Rotation transformation, section 5 explains the proposed algorithm, section 6 presents experimental results and finally section 7 discuss conclusion and future scope.

2. RELATED WORK

PPDM techniques are mostly divided into two categories such as random perturbation and cryptographic techniques. A number of proposed privacy techniques exist based on perturbation. Agarwal and Srikanth [3], build classifier from the perturbed training data, later in 2001 a distortion-based approach for preserving the privacy was introduced by Agrawal and Aggarwal [6]. Reconstruction-based techniques for binary and categorical data are available in the literatures [7, 8]. M.Z Islam and L.Brankovic [9] proposed an algorithm known as DETECTIVE. In their work they addressed the perturbation is used for the categorical attributes based on clusters.

Cryptographic techniques are applied in distributed environment. Secure Multiparty Computation (SMC) is the well known technique in this category. In SMC two or more parties compute secure sum on their inputs and transfer to the other party without disclosing the original data [10, 11 and 12].

Jie Wang and Jun Zhang [13] addressed a frame work based on matrix factorization in the context of PPDM [13], they have used Singular Value Decomposition (SVD) and Non negative Matrix Factorization (NMF) methods. The framework focuses the accuracy and privacy issues in classification.

Recently, Euclidean distance preserving transformation techniques are used such as Fourier related transforms (DCT), wavelet transforms and linear transforms which are discussed in [14, 15 and 16].

In this paper, we also present a Euclidean distance preserving transformation technique using Walsh-Hadamard (WHT) and Rotation Transformation. WHT generates an orthogonal matrix, it preserves the Euclidean distance after transformation and as well as preserves statistical properties of the original data then we apply Rotation Transformation, it also preserve distance between data points. Hybrid Transformation technique preserves individual privacy of the customers.

3. WALSH-HADAMARD TRANSFORMATION (WHT)

Definition: The Hadmard transform H_n is a $2^n \times 2^n$ matrix, the Hadamard matrix (scaled by normalization factor), that transforms 2^n real numbers X_n into 2^n real numbers X_k.

The Walsh-Hadamard transform of a signal x of size N=2^n, is the matrix vector product $x.H_n$. Where

$$H_N = {}_{i=1}^{n}\otimes H_2 = \underbrace{H_2 \otimes H_2 \otimes \dots \otimes H_2}_{n}$$

The matrix $H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ and $\otimes$ denotes the tensor or kronecker product. The tensor product of two matrices is obtained by replacing each entry of first matrix by that element multiplied by the second matrix. For example

$$H_4 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

The Walsh-Hadamard transformation generates an orthogonal matrix(H_n), it preserves Euclidean distance between the data points.

Definition: Matrices A for which, $A^T.A = I$ are called orthogonal matrices. They have the property that the transpose of A is also the inverse: $A^T = A^{-1}$ and$(A^T)^{-1} = A$.

Theorem 1: Suppose that $T: R^n \to R^n$ is a linear transformation with matrix A, then the linear transformation T preserves scalar products and therefore distance between points/vectors if and only if the associated matrix A is orthogonal.

Proof: Suppose that the scalar product of two vectors $u, v \in R^n$ is preserved by the linear transformation. Recall that a scalar product is the same as the matrix product of one vector as a row matrix by the other vector is a column matrix: $u.v = u^T.v$ then

$$\begin{aligned}(Au).(Av) &= (Au)^T(Av) = u^T A^T Av \\ &= u^T(A^T A)v\end{aligned}$$

Hence, if the scalar product is preserved then $(Au).(Av) = u^T(A^T A)v = u.v$ which shows that the product $A^T A$ must disappear from $u^T(A^T A)v$. This certainly happens if $A^T A = I_n$, where I_n the identity matrix, for then $u^T(A^T A)v = u^T I_n v = u^T v$, as required. It is also intuitively clear, at least, that this happen only if $A^T A = I_n$. Since distance is defined in terms of the scalar product, it follows that distance is also preserved. ■

Theorem 2: Suppose that $T: R^n \to R^n$ is a linear transformation with matrix A, then the linear transformation T preserves angles between the vectors (may or may not preserve distance) if the associated matrix B is a scalar multiple of an orthogonal matrix. *i.e. B=kA,* where $A^T A = I_n$ and $k \in R$.

Proof: this theorem proof is similar to the above theorem. ■

4. ROTATION TRANSFORMATION

In Cartesian co-ordinate system Translation, rotation and reflection are Isometric Transformations. Using Translation transformation failed to protect privacy of individual customers [17]. In this paper we are using Rotation transformation to hide underlying data values with combination of WHT. The purpose of Rotation Transformation is, increase complexity to unauthorized people while accessing the data for their personal use.

Definition: let Transformation $T: R^n \rightarrow R^n$ be a transformation in the n-dimensional space. T is said to bean isometric transformation if it preserves distances satisfying the constraint

$|T(U) - T(V) = |U - V|$ for U, V $\in R^n$

T: $\begin{pmatrix} x' \\ y' \end{pmatrix} = T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$

In this work we choose transform angle θ based on Variance of attributes before and after Transformation.

Var(X) = Var$(x_1, x_2, x_3, \ldots., x_M) = \frac{1}{M} \times \sum_{i=1}^{M}(x_i - \overline{x})^2$

Where $\overline{x}$ is arithmetic mean of$x_1, x_2, x_3, \ldots., x_M$.

We are following guidelines in [17] for choosing the transform angle θ, is calculated as follows $p_0 = min\left(var\left(A_i - A_i^{'}\right), var(A_j - A_j^{'})\right)$ Where A_i and $A_i^{'}$ are original and transformed data respectively and $\theta = p_0 * pl$.

5. EXPERIMENTAL WORK

Assume that, we represent a dataset as a matrix format. A row indicates an object and a column indicates an attribute. If the number of columns is less than2^n, here n=0, 1, 2, 3… Then we are adding the columns to its nearest value of 2^n. All the added columns are padding with zeros. Every element is discrete and numerical missing element is not allowed.

Algorithm:

Input: Dataset D, privacy_level Pl;

Output: Modified Dataset$D^{'}$;

1. Pre-Process the data if no. of columns less than N (N=2^n, n=0, 1, 2, 3….)
2. Generate Walsh NxN Matrix, N= number of columns.
3. Obtain the modified dataset by multiply original dataset with Walsh matrix.
4. Divide modified dataset into N/2 pairs.
5. For each pair apply rotation transformation.
6. Based on privacy level we choose optimal transform degree value.
7. Obtain the modified dataset by multiplying with Rotation matrix

6. EXPERIMENTAL RESULTS

We conducted experiments on two real life datasets Iris and Australian Credit dataset obtained from UCI Machine learning Repository [18]. The dataset properties are as follows.

Table 1. Dataset Description

Dataset name	No. of Records	No. of attributes	No. of classes
Iris	150	4	3
Australia credit	690	14	2

The Iris consists of flower dataset. It contains features of three types of flowers (classes) like Iris *Setosa*, Iris *Versicolor* and Iris *Virginica*. The four attributes are Sepal Length (SL), Petal Length (PL), Sepal Width (SW) and Petal Width (PW).

The Australian credit is banking dataset. It consists of two types of classes, *good* and *bad*. It consists of 690 instances with 14 attributes. Out of these 14 attributes, 6 attributes are numerical and 8 attributes are categorical. In this work we consider only numerical attributes. Two extra columns are added to dataset and those columns are appended with zeros.

We are using KNN (K-Nearest Neighbor) as a classifier in WEKA Tool [19]. KNN classifier is well known classifier. That works based on the distance between records. In the experiments parameter k is set with the values 3, 5, and 7, this transformation preserves distance between the records before and after transformation. Original data takes a matrix format, row is treated as an object and column is treated as an attribute.

Table 2. Original Dataset

SL	PL	SW	PW
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2

Table 3. Modified Dataset

SL	PL	SW	PW
-1.69866	8.109623	-2.54931	6.198623
-1.13457	7.34827	-1.98522	5.437271
-1.54761	7.465414	-2.29962	5.653045
-1.36575	7.382185	-2.31502	5.372555

Original and modified values are shown in Table 2 and 3 respectively. Distance between first and remaining records in original dataset are 0.5385, 0.5099 and 0.6480. In modified dataset the distances are 1.0770, 1.0198 and 1.2961. WHT transformation is worked based on Theorem 2. Distance matrix of modified dataset is an integer multiple of original dataset distance matrix. Due to this reason the distance between data records is not modified, next we apply rotation with some angle of degree which is based on privacy level of customer then the modified dataset is obtained. In this work the privacy level is set to 0.6. K-NN classification algorithm works well on modified dataset without information loss.

Accuracy of K-NN classifier on IRIS dataset is compared between existing perturbations methods like SAN, MN etc., and also with our proposed method is Hybrid Transformation , which is comparatively also better than the other distance preserving transformation methods given in Table 4.

Table 4. Accuracy of K-NN Classifier on IRIS

Method	Accuracy (%)		
	K=3	K=5	K=7
Original	95.33	95.33	95.33
SAN	95.33	95.33	95.33
MN	95.33	95.33	95.33
DCT	95.33	93.33	94.00
FISIP	96.00	95.33	96.67
Hybrid	96.67	95.33	95.33

Table 5. Accuracy of K-NN Classifier on Australian Credit

Method	Accuracy (%)		
	K=3	K=5	K=7
Original	73.33	72.60	72.31
SAN	73.33	72.60	72.31
MN	73.33	72.60	72.31
DCT	66.23	66.52	68.98
FISIP	66.56	67.39	69.42
Hybrid	67.82	68.98	67.68

Accuracy of KNN classifier on Australian Credit dataset is compared with existing methods shown in Table 5.

6.1 Privacy Measures

Privacy measures are adopted from [20].

6.1.1 Value difference

After a dataset is modified, the values of its elements are changed. The Value Difference (VD) of the dataset is represented by the relative value difference in Frobenius form. VD is the ratio of the Frobenius norm of the difference of D and $|D'|$ to the Frobenius form of D.
$VD = \|D - |D'|\|_F / \|D\|_F$.

6.1.2 Position Difference

After a data modification, the relative order of the value of the attribute changes, too. We use RP represents the average change of order for all the attributes. After data modification, the order of each value changes. Assume dataset D has *n* data objects and *m* attributes. Ord_j^i Denotes the ascending order of the j^{th} value in i^{th} Attribute, and $\overline{Ord_j^i}$ denotes the ascending order of the modified valueD_{ij}. Then RP is defined as

$$RP = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} \left| Ord_j^i - \overline{Ord_j^i} \right| \right) / (m * n)$$

RK denoted as percentage of elements keep their order in modified data.

$$RK = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} RK_j^i \right) / (m * n)$$

Where RK_j^i represents whether or not an element keeps its position in the order of values.

$$RK_j^i = \begin{cases} 1, if\ ord_j^i\ = \overline{ord_j^i} \\ 0, \quad otherwise \end{cases}$$

The metric CP is used to define the change of order of average value of attribute.

$$CP = \left(\sum_{i=1}^{m} \left| \left(ordAV_i - \overline{ordAV_i} \right) \right| \right) / m$$

Where $ordAV_i$ is the ascending order of the average value of attribute i, while $\overline{ordAV_i}$ denotes its ascending order after modification.

CK is to measure the percentage of the attributes that keep their orders of average value after distortion.

$$CK = \left(\sum_{i=1}^{m} CK^i \right) / m$$

Where CK^i is calculated as

$$CK^i = \begin{cases} 1, if\ ordAV_i = \overline{ordAV_i} \\ 0, othrwise \end{cases}$$

The higher the value of RP and CP and the lower the value of RK and CK, the more privacy is preserved [18]. We calculate above data distortion measures on both modified datasets, results are shown in Table 6. Our transformation technique is compared with existing distance preserving transformation techniques such as FISIP and wavelet transformations. Privacy measures of IRIS dataset using Wavelet Transformations are taken from [15]. Based on these values, we say that our proposed transformation preserves distance as well as knowledge without loss.

We adopted the data distortion metrics used in [19] to measure the degree of data perturbed. According to their definitions, we know that a larger RP and CP, and smaller RK and CK value indicates more the original data matrix is distorted. Which implies the data distortion method is better in preserving Privacy. Data distortion measures on Iris dataset are showed

Table 6. Privacy Measures

Data (Method)	VD	RP	RK	CP	CK
IRIS (Hybrid)	-0.4390	47.028	0.50	0	1
AUS (Hybrid)	-0.5428	259.86	9.66e-4	0	1
IRIS (FISIP)	-0.032	42.0883	0.0033	0	1
IRIS (Wavelet)	0.91276	29.6266	0.015	1.0	0.25

7. CONCLUSION AND FUTURE WORK

Some data mining algorithms works based on statistical properties based on that we propose a Hybrid transformation for PPDM. It preserves distance between data records so, knowledge should be same. It modifies the data but maintains Accuracy of classifier as original data without information loss. Our proposed transformation is applicable only to numerical attributes. It can be extended to categorical attributes.

REFERENCES

[1] D. Lin, E. Bertino, R. Cheng, and S. Prabhakar,(2008) "Position transformation: a location privacy protection method for moving objects" ,Proc. of Int'l Workshop on Security and Privacyin GIS and LBS, pp. 62–71.

[2] Giannotti, F., and Pedreschi, D.(2006)" Mobility, Data Mining and Privacy", Springer, Germany.

[3] R.Agrawal and R. Srikant,(2000) "Privacy Preserving Data Mining", In Proceeding of SIGMOD Conference on Management of Data, pp 439-450

[4] Y. Lindell and B. Pinkas,(2000) "Privacy-Preserving Data Mining", In Advances in Cryptology-CRYPTO, pp36-54.

[5] Verykios, V.S., Bertino, E., Fovino, I.N., Provenza, L.P., Saygin, Y., Theodoridis, Y,(2004) "State-of-the-art in Privacy Preserving Data Mining", SIGMOD Record 33 ,pp50—57.

[6] D. Agrawal and C. C. Aggarwal,(2001) "On the Design and Quantification of Privacy Preserving Data Mining Algorithm", In Proceeding of ACM SIGMOD, pp247-255.

[7] A. Evfimieski, R. Srikant, R. Agrawal and J. Gehrke,(2002) "Privacy Preserving Mining of Association Rules", In Proceedings of the 8th ACM SIGKDD, Edmonton, Canada ,pp 217-228

[8] S. Rizi and J.R. Haritsa,(2002) "Maintaining Data Privacy in Association Rule Mining", In the proceedings of the 28th VLDB Conference, Hong kong, China ,pp 682-693.

[9] M. Z. Islam, and L. Brankovic,(2005) "DETECTIVE: A Decision Tree Based Categorical Value Clustering and Perturbation Technique in Privacy Preserving Data Mining", In Proc. of the 3rd International IEEE Conference on Industrial Informatics, Perth,Australia

[10] A. C. Yao,(1986) "How to Generate and Exchange Secrets", In Proceedings 27th IEEE Symposium on Foundations of Computer Science, pp 162-167.

[11] J. Vaidya and C. Clifton,(2002) "Privacy Preserving Association Rule Mining in Vertically Partitioned Data", In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 639-644.

[12] J. Vaidya and C. Clifton,(2003) "Privacy-Preserving k-Means Clustering over Vertically Partitioned Data", In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 206-215.

[13] Jie Wang and Jun Zhang, "Addressing Accuracy Issues in Privacy Preserving Data Mining through Matrix Factorization ".

[14] Shibnath Mukharjee , Zhiyuan Chen, Aryya Gangopadhyay,(2006)"A Privacy-preserving technique for Euclidean distance-based mining algorithms using Fourier-related transforms ",the VLDB Journal,pp (293-315).

[15] Vinod patel and yogendra kumar jain,(2009)"wave let transform based data perturbation method for privacy protection",IEEE.

[16] Jen-Wei Huang,Jun-Wei Su and Ming-Syan Chen,(2011) "FISIP: A Distance and Correlation Preserving Transformation for Privacy Preserving Data Mining"IEEE.

[17] ZHANG guo-rong,(2012)" An Effective Transformation Approach for Privacy Preserving Similarity Measurement", FSKD.

[18] http://kdd.ics.uci.edu/

[19] http://www.wekaito .ac.nz/ml/weka

[20] Shuting Xu, Jun Zhang, Dianwei Han, and Jie Wang,(2005)"Data distortion for privacy protection in a terrorist Analysis system", P. Kantor et al (Eds.): ISI 2005, LNCS 3495, pp. 459-464.

AUTHORS

P.N Girija is presently working as Professor in the School of Computers and Information Sciences, university of Hyderabad, Hyderabad. Her research areas are Speech Recognition, Speech Synthesis and Human Computer Interaction. She has published nearly eighty papers in various national and International journals and conferences. She visited School of Computer Science, Camegie Mellon University, Pittsburgh, U.S.A as a visiting Scholar during Jun-August 2004. She chaired several Sessions like COCOSDA, NTU Singapore etc, She completed sanctioned research projects from DST, AICTE, UPE etc.

HanumanthaRao Jalla completed B.Tech in computer science and Engineering from the VRSEC, Nagarjuna University, Guntur, A.P, in 2003 and M.Tech in information Technology from University of Hyderabad, A.P. Presently working an assistant professor in the Department of Information Technology, Chaitanya Bharathi Institute of Technology, Hyderabad, A.P. His research interests Privacy-Preserving Data Mining