IJCAT - International Journal of Computing and Technology, Volume 3, Issue 6, June 2016
ISSN : 2348 - 6090
www.IJCAT.org

Privacy Preservation of Data using Factorization

[1] N. Maheswari, [2] Ankita Adhawale , [3] Amol Bhausaheb Wale

[1,2,3] School of Computing Science and Engineering
VIT University, Chennai

Abstract - Preserving the privacy of the personal information, valuable company data, important records of organizations become the difficult task day by day. Using data mining technique in many applications, there are various problems regarding security, privacy issue comes front. There are lots of techniques used to maintain privacy of data records. Transformation of data values is one of the well known methods to maintain the privacy of data records. This article presents LU-factorization method to maintain the privacy of data records. Clustering techniques had performed on the original and the distorted data sets. Performance measures have been used to evaluate the distorted data records with original data records. The experimental result shows that the LU factorization method maintains the balance between privacy and accuracy. The accuracy of the clustering has been measured and it produced acceptable results.

Keywords **- Privacy Preserving, Data Distortion, Data Mining, Clustering.**

1. Introduction

The rapid growth in the use of data mining leads to the issue of security and privacy of personal organizations, government and military information. Nowadays in this running world data gathered very fast and at a very large amount in various places such databases includes medical records, profiles of company, persons bank account detail, criminal history, defense, military secret information, government plan and policy in to business related information. Processing such type of data leads to the problem of privacy preserving in data mining. This confidential information cannot be shared with every person. There are many related research work done on privacy preserving in data mining related with data transformation, perturbation, data distortion. This article discuss about LU-Factorization method. This method transforms the matrix of datasets into the distorted form. Privacy measurement techniques are applied to know the privacy preservation degree. The performance measures between the Principal Component Analysis (PCA), Singular Value

Decomposition (SVD) and the LU factorization techniques have been analyzed.

2. Related Work

Liew et al.[5], done the work on probability distribution based data distortion. Sweeney L. et al...[7], described about k-anonymity. In this the authorized user of the data wants to share his information without knowing the identity of him. Data suppression with the data generalization is used for this purpose.

Xu et al.[2], this article uses the fast Fourier transform to distort the datasets.

Liu et al.[3], have done transformation on datasets based on wavelet transform. The perturbated data keeps the privacy of the data records as well as this also preserve the datasets utility. It can perform the data mining operations on the perturbated data.

Chieh Ming wu et al.[8],have done greedy approach to hide the value of sensitive information. This has been done without adding any extra rule to the datasets.

Peng et al. [4], uses the SVD, i.e. Singular Value Decomposition, Discrete Wavelet Transformation and the factorization of the non negative value matrix. This is the simple idea to convert the original dataset into the new distort form and then apply the data mining operations on distorted data sets.

Guang Li et al. [1], done the work on Singular Value Decomposition method. They proposed the new technique to treat all perturbated value equally in the result. Because of this methodology it improves the performance of the SVD. They have also given a new algorithm by using Weighted SVD

Santosh Kumar Bhandare [11] proposed the perturbation for the real time data based on the data mining application. Author used the Tan Hyperbolic normalization for the data transformation. He applies privacy measures to calculate the difference between the original datasets and the distorted data sets. Marina Blanton et al. [15] shows that in the presence of participants the available methods of the perturbation are not satisfy the goal of the privacy preservation

Yaping Li et al. [14] provides the solution to prevent fetching the original dataset from the data which is made by transformation. In this article user allows to do transformation of the dataset only when it is needed.

Majid Bashir Malik [13] gives the available tools and techniques of the privacy preservation in the present market and also proposed some new feature techniques.

3. Basics

3.1 Data Transformation

Data owner transforms (Fig.1) the original dataset to distorted dataset. It is one of the valuable parts of the given proposed system in this article. The given transformation techniques must keep the data utility for the further processing and doing analytical operation as well[6] as the privacy of the data sets after transformation. The basics described the SVD techniques and PCA techniques.

Fig. 1. Model of Privacy Preservation Using Data Transformation

3.2 Singular Value Decomposition

Singular Value Decomposition (SVD) is well known and widely used method of the dimensionality reduction of the data. It is frequently used [16] for transformation of the dataset. It gives the distorted matrices as define in equation (1).

Let A is the original dataset of dimension $n \times m$. Data object are given by the row and attributes are given by the column of the matrix.

The singular value decomposition of the matrix *A* is

$$\mathbf{A = UWV^T} \qquad (1)$$

where *U* is an n x n orthonormal matrix, W=diag(σ1,σ2,….. σs) (s = min (m,n)) is an *n* x *m* diagonal matrix whose nonnegative diagonal entries (the singular values) are in a descending order, and $V^T$is an m x m orthonormal matrix. The number of nonzero diagonal entries of W is equal to the rank of the matrix A.

3.3 Principal Component Analysis

Principal Component Analysis technique is used to reduce the complexity of high dimensional data. It also gives the reduced dimension of the given datasets. PCA used to identify the pattern in data and once the patterns are identified then using pattern it is easier to reduce the dimensions of data and also by reduction there is no loss of information in the reduced datasets.

Steps for calculating PCA

1) Transform the Matrix 'X' into a Matrix 'Y' by calculating the mean of all the columns in 'X' and then subtract the mean from the original matrix i.e. from each row.

2) Then solve the covariance matrix using the equation 3.2:

$$C = \frac{Y^T Y}{N-1} \qquad (2)$$

Where,

Y^TIs the transpose of matrix Y.
N is the size of the matrix.

3) Then calculate the Eigen values and Eigen vectors of C : $Cx = \lambda x$

Where x is eigen vector and 'λ' is the eigen value.

4) The Eigen vector with the highest Eigen value is the principal component of the data.

4. Methodology

4.1 LU-Factorization Method

Data Privacy Preservation has been performed using mathematical transformation technique LU Factorization as given in equations (3),(4) and (5), where L denotes lower triangular matrix and U denotes upper triangular matrix. This method transforms the original dataset values into new transformation of data set.

Suppose that A is n*n matrix that can be reduced to an upper triangular matrix U by Gaussian elimination without using row exchanges.

Then,

$$A = LU \qquad (3)$$

Where L is lower triangular and U is upper triangular matrix as mentioned in equation 4 and 5.

- Each entry on the diagonal of L is 1.
- For each i > j, the (i, j) entry of L is the multiple m of row j that was subtracted from row i during the elimination process.
- The diagonal entries of U are the pivots from the elimination process.

The distorted data matrix is given as,

In upper triangular matrix:

$$u_{ij} = a_{ij} - (k=1, \Sigma, i-1)\ u_{kj} l_{ik} \quad (4)$$

In lower triangular matrix:

$$l_{ij} = (1/u_{ij})\ (a_{ij} - (k=1, \Sigma, j-1)\ u_{kj} l_{ik}) \quad (5)$$

Where,
i is index for row,
j is index for column,
a is original data matrix.

4.2 Data Clustering

Kmeans clustering [9] technique is one the oldest and most popular clustering technique. In Kmeans clustering clusters are formed according to value define by users, it is heuristic algorithm, there is no assurance of optimum results. It uses two features which are important feature of Kmeans algorithm, one is Euclidean distance as its metric to calculate the distance between data objects and another feature is "K" values ,which indicate how many cluster need to be form. Agglomerative clustering [9] also known as "bottom up" technique, Its name itself suggest that here cluster are individual data point at early stage and then they go on combining as one moves up to the hierarchy. Finding distances between different objects is the first step towards clustering, then linking up those objects whose distance is in close proximity, also to determine where to cut the hierarchy into different clusters is important step otherwise tree will grow to make it complex and consequently the wrong cluster analysis

5. Privacy Measures

The privacy measures have been used to analyze the performance of the data distortion methods, which are applied on the original datasets and the distorted datasets [16].

5.1. Value differences (VD)

On applying the transform technique the value of relative position in the new distorted matrix changed. This difference between the original dataset value [10] and the new transformation dataset value is called as a value difference as given in equation (6).

It is given as,

$$\text{Value difference} = \| P - Pm \|f / \| P \|f \quad (6)$$

Where,
P = original matrix dataset
Pm = modified matrix dataset
|| is gives the frobenius norms of the values in between this symbol.

5.2 Position Difference (PD)

In this measure the position of the values taken in to consideration. Position of the dataset value changes in the transformation process by analyzing the difference in changed and original position this measure can be calculated.

RP

RP shown in equation (7) gives the average change in the order of every attribute.

The RP parameter is given by,

$$RP = (\sum_{i=1}^{p} \sum_{j=1}^{q} Ord_j^i - \acute{Ord}_j^i \vee / (p*q) \quad (7)$$

Where,
p is data objects
q is attributes.

RK

RK value as given in equation (8) gives the percentage of attributes of the given sets which keeps its order of values in each column after applying transformation.

RK parameter is given by,

$$RP = (\sum_{i=1}^{p} \sum_{j=1}^{q} Rk_j^i \vee / (p*q) \quad (8)$$

Where Rk_j^i is the position in its order of values,

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

CP

CP parameter as mentioned in equation (9) is used to verify the change in rank of average value after transformation of the datasets.

It is given by,

$$CP = \frac{ON_i - \acute{O}N_i \vee}{\frac{1}{\alpha}\sum_{i=1}^{\alpha}} \quad (9)$$

Where,
ON_i Is the ascending order of the average value of attribute i.

$O'N_i$ Is the ascending order of the average value of attribute i after transformation.

CK

CK value as given in equation (10) is used to calculate the percentage of the attributes that keep their order of average value after data transformation.

CK is given by,

$$CK = \frac{1}{a}\sum_{i=1}^{a} Ck^i \qquad (10)$$

Where,

$Ck^i = 1$ If $ON_i = O'N_i$

Otherwise,

$Ck^i = 0$

6. Experimental Results

In this article two databases have been used. Glass Identification dataset and banknote authentication Dataset are taken from University of California at Irvine's Machine Learning Repository [12]. The Glass Identification dataset has 10 attributes and 214 samples and there are no missing values. Banknote authentication dataset has 5 attributes and 1372 samples and there are no missing values. All attributes have been considered for the analysis. Table 1 and 2 shows values of the privacy measures applied on the data distortion methods. The privacy measures of all three data distortion methods are shown in Table 1 and Table 2.

Table 1: Results for Glass Identification data

	VD	*RP*	*RK*	*CP*	*CK*
LU Factorization	1.0007	65.0867	0.0884	1.3929e+04	0
SVD	1.0008	57.9864	0.1525	1.2409e+04	0
PCA	0.9044	62.6822	0.0956	13414	1

Table 2: Results for Bank dataset

	VD	*RP*	*RK*	*CP*	*CK*
LU Factorization	1.0575	525.42	4.3732e-04	7.2092e+05	0
SVD	0.1464	450.84	5.8309e-04	6.1856e+05	0
PCA	1.2758	421.14	0.0017	57781	0

Greater the value of RP and CP preserves more privacy. In these comparisons the RP and CP values are higher for Lu-Factorization which means relative change in row position and relative change in column position is more than SVD and PCA. Lower the values of RK and CK in better for privacy preservation. RK value is very low which indicate only some elements keep their order in distorted dataset. The value of CK is zero for both LU-factorization and SVD but it is one for PCA which means there is change in order of every attribute in LU-Factorization and SVD.

The space requirement for the technique is the size of the input data. The complexity for LU factorization is $O(n^3)$. K-means has the complexity as O(n), as the number of iterations and the number of cluster are relatively small than 'n'. The computational complexity required for agglomerative clustering is $O(n^2)$. Where 'n' is the size of the input data. So the overall complexity required to perform all the process is $O(n^3)$.

7. Measuring Effectiveness

The effectiveness is calculated to identify the number of points present in both the original database and distorted database during clustering. The clusters form in the original dataset must be equal to the clusters form after transformation of the datasets. This cannot always be possible as the database has distorted. Some points migrate from one cluster to another. It will not give exact cluster with the distorted dataset. For the technique discussed in this article, K-Means Clustering and Agglomerative Clustering is used. The bank dataset has been used to analyze the accuracy of K-Means Clustering and Agglomerative Clustering. The performance of the technique can be expressed in terms of misclassification error rate[9] as given in equation (11).

$$Error\ Rate = \frac{Numbr\ Of\ Wrong\ Predictions}{Total\ Number\ of\ Predictions}$$

$$\text{Error Rate} = \frac{f_{10}+f_{01}}{f_{11}+f_{10}+f_{01}+f_{00}} \qquad (11)$$

The entry f_{ij} denotes the number of records from cluster i predicted to be of cluster j. For instance, f_{01} is the number of records from cluster 1 incorrectly predicted as cluster 2. The total number of incorrect predictions is $(f_{10} + f_{01})$. The denominator in equation (11) represents the total number of predictions performed between the clusters. For the standard result misclassification error should be 0% but for the real time experiment it is as good as near to 0%.

Table 3. Misclassification Error for K-Means clustering

Method /Cluster	*Cluster (K=2)*	*Cluster (K=4)*
LU	0.62	0.35
PCA	1.0	0.37
SVD	1.0	0.37

Table 4. Misclassification Error for Agglomerative clustering

Method	*Agglomerative Clustering*
LU	0.01
PCA	0.02
SVD	0.02

In the result depicted in Table 3, LU Factorization has the misclassification error rate as 0.62 and 0.35 for the clusters. With compare to the SVD and PCA it shows less error rate. As per Table 4, LU Factorization has the misclassification error rate as 0.01 for the clusters. Though it does not show significant difference among other two techniques it is better than the other two techniques. Therefore LU-Factorization shows the better clustering accuracy of the data set distortion among the SVD and PCA.

8. Conclusion

This article proposes a new technique for privacy preservation using LU-Factorization method. The original datasets are transformed and the privacy measures are applied to know the percentage of privacy preservation. By applying the privacy measure to the existing system like SVD and PCA and comparing the results with proposed method, it has been concluded that LU-Factorization is better in preserving the privacy of the dataset. Privacy measures and the misclassification error results show the balance between clustering accuracy and privacy.

References

[1] Guang Li, Yadong Wang, "A New Method for Privacy-Preserving Data Mining Based on Weighted Singular Value Decomposition," JCIT: , Vol. 6, No. 3, pp. 28 ~ 34, 2011.

[2] ShutingXu, ShuhuaLai, "Fast Fourier transformation based data perturbation method for privacy protection," IEEE International Conference on Intelligence and Security Informatics, pp.221-224, 2007.

[3] Lian Liu, Jie Wang, Jun Zhang, "Wavelet-Based Data Perturbation for Simultaneous Privacy-Preserving and Statics-Preserving," Proceeding of IEEE International Conference on Data Mining Workshop, pp. 27-35, 2008.

[4] Bo PengXingyuGeng, Jun Zhang, "Combined data distortion strategies for privacy-preserving data mining," Proceeding of the IEEE International Conference on Advanced Computer Theory and Engineering (ICACTE), pp. V1-572-V1-576, 2010.

[5] C.K Lieu, U.J. Choi, and C.J. Liew, "A Data Distortion by Probability Distribution," ACM Transaction on Database System (TODS), vol.10, no. 3, pp.395-411, Sep.1985.

[6] R Agrawal and R.Srikant, "Privacy-Preserving data mining," Proceeding of the ACM SIGMOD Conference on management of data, pp. 439-450, may 2000.

[7] L. Sweency, "k-Anonymity: A Model for Protecting Privacy," International Journal on Uncertainty, Fuzziness, and Knowledge based System, vol 10, no 5, pp.557-570,2002.

[8] Chieh-Ming Wu, Yin-Fu Huang and Jian-Ying chen, "Privacy Preserving Association Rules by using Greedy Approach," World Congress on Computer Science and Information Engineering, pp. 61-65, 2009

[9] J. Han, M. Kamber. "Data Mining : Concepts and Techniques. Morgan Kaufmann Publisher, 2001.

[10] S. Xu, J. Zhang, D. Han, J.Wang, "Data distortion for privacy protection in a terrorist analysis system," IEEE International Conference on Intelligence and Security Informatics, 2005, 459-464.

[11] Santosh Kumar Bhandare, "Data Distortion Based Privacy Preserving Method for Data Mining System," International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Volume 2, Issue 3, May – June 2013

[12] https://archive.ics.uci.edu/ml/datasets.html,[Machine learning Repository]

[13] Majid Bashir Malik, M. Asger Ghazi, Rashid Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", Third International Conference on Computer and Communication Technology, 2012.

[14] Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang, "Enabling Multilevel Trust in Privacy

Preserving Data Mining", IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 9, September 2012.

[15] Marina Blanton, "Achieving Full Security in Privacy Preserving Data Mining", IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, 2011.

[16] S. Xu, J. Zhang, D. Han, and J. Wang, Singular value decomposition based data distortion strategy for privacy protection. Knowledge and Information Systems, 10(3), 383-397,2006.