ISSN: 0975-766X
CODEN: IJPTFI
*Available Online through* *Research Article*
**www.ijptonline.com**

# PRIVACY PRESERVING CLINICAL DATA OF PATIENTS USING DATA PERTURBATION

**Shamika Mukane [1], Maheswari N[2*]**
[1]School of computing science and engineering, VIT University, Chennai.
[2]School of computing science and engineering, VIT University, Chennai.
*Email: maheswari.n@vit.ac.in*

*Received on 09-08-2016* *Accepted on 05-09-2016*

**Abstract**

**Aim:** In the world of bigdata, where the data is generating every second in the huge volume, protecting confidential data is an issue. In the process of data mining, many times the sensitive data is used for analysis. To solve this problem data is perturbated and further given to mining process. In this way the privacy of the patients's original clinical data is preserved and data mining can be done on modified data obtained from data perturbation. **Method:** This paper discusses various techniques such as SVD, PCA, QR, Hilbert and Watershed for data distortion and saved to preserve privacy. The preserved data is used for analysis using hierarchical clustering. **Results and Discussion**: Experiment clearly shows that the method used for data perturbation is an effective and values shows privacy is preserved. Experimental result shows the watershed method can used for data perturbation with respect to data privacy, calculated using various privacy metrics.

**Keywords:** Privacy Preservation, data perturbation, clinical, watershed, clustering.

## Introduction

The main goal of data mining is extracting hidden patterns from the given data. Privacy Preserving is a concern in many fields where data mining can be applied, such as financial, medical, online shopping, homeland security, and etc. Variety of data is collected from various sources, which is used for mining. Pattern evaluation is done using this collected data to enhance the decision making power of the organization. On the other side, data used for mining includes private, sensitive and confidential information of the individual. Hence the privacy of confidential data should be preserved. The challenge of preserving privacy can be solved using data perturbation. Data perturbation can protect the sensitive attributes in the dataset, since distorted data is given for data mining and original data is preserved. This paper mainly consider, the case in which sensitive numerical attributes are perturbated in order to meet privacy in clustering analysis, mainly hierarchical

clustering. The motivation behind clustering is grouping similar type of objects together. The main goal is to increase the intra-cluster similarity and reduce inter-cluster dissimilarity. Clustering can be used for outlier analysis. Application of clustering are fraud identification, spatial data analysis, pattern recognition and many more. Central idea behind privacy preserving of the data is distribution of data in various location. As technology is changing rapidly, data is not stored in centralized location, whereas it is stored in distributed environment. Hence vulnerability of data increases. To preserve the privacy in the clustering, it is important to have best data perturbation method which protects the data without affecting the significance of data mining. Paper discusses about the watershed transformation technique for data perturbation which distorts the sensitive, private and confidential data from the patients's clinical dataset, so as to preserve the privacy. Watershed transformation is independent of clustering, hence after its application the data mining technique can be used to extract knowledge from given perturbated data.

As data is stored in distributed environment, many users have fear of losing their private information or they think it can be misused by the third party users. Hence it is necessary to preserve the privacy while mining the data. Privacy preserving in data mining is upcoming field, which leads to protect the original data of the user and guarantees to keep the data secure.

In this paper, it further describes the related works done in the field of data perturbation. Later the basic concept of the data perturbation and introduction of the various techniques used for perturbation of data is discussed. Further hierarchical clustering is described in brief. Various privacy measures are explained in detail along with that Cophenetic distance calculation. Finally the experimental results of various perturbation techniques is shown. Comparison is done by the calculation of privacy measures and Cophenetic distance between the clusters.

**Related Works**

The geometric data transformation methods [1] will help in preserving privacy after perturbation in cluster analysis. The clustering techniques used are partition-based (kmeans) and hierarchical. The article gives the quantifying privacy metrics to measure the privacy preserved after data perturbation. The data[2], is preserved using singular value decomposition method. In this method existing SVD method is modified and Sparsified SVD is propose which gives better results in terms of privacy. The experimental results are compared with SVD and new SSVD. Result shows that SSVD is better than SVD in balancing data utility and data privacy.

The terrorist attack related data is discussed [3] and Sparsified SVD method is applied for data perturbation. The experimental results are compared using various privacy metrics like VD, RP, RK, CP and CK. The result shows the SSVD method is efficient in preserving privacy and maintaining utility of the data. Geometric data perturbation [8] approach is used for perturbing the multiparty data and to secure data from attackers. This article has 3 unique key points: first is geometric data perturbation and other two are data utility and guarantee of privacy. The proposed approach has scalability with respect to number of attributes participating. The multiparty cluster information [9] is preserved using principle component analysis. The result shows the efficiency of the proposed techniques after applying k-means clustering on multiparty data. Result shows that how PCA gives better privacy and performance than the geometric data transformation. A method [11] for ensuring partial disclosure while allowing a miner to explore detailed data is discussed. In this approach, one first builds a local decision tree over true data, and then swaps values amongst records in a leaf node of the tree to generate randomized training data. The swapping is performed over the confidential attribute only, where the confidential attribute is the class label. This approach deals with a trade-off statistical precision against security level, i.e., the closer to the root, the higher the security but lower the precision.

**Basic Concepts**

**Data Perturbation**

Perturbation can be defined as a deviation of a system, moving object, or process from its regular or normal state or path, caused by an outside influence. Data perturbation is important aspect for preserving that confidential data. After perturbation, the original data set is modified and further given for the analysis process. There are two major categories of data perturbation: probability distribution and fixed data (data distortion). Data perturbation is easy and effective technique for preserving confidential data.

**Methods**

**Singular Value Decomposition (SVD)** is frequently used method for data perturbation. It is usually used for the dimensionality reduction of the original data set. Here it is used for data perturbation method.

Let A is the original matrix of order *n x m*. The row in matrix represents the observations whereas column represents the attributes. The SVD of the matrix A is: [2]

$$A = U \sum V^T$$

Where U is an orthogonal matrix of order *n x n*, $\sum$ is an *n x m* diagonal matrix whose diagonal elements are non-negative and $V^T$ is an *m x m* orthonormal matrix.

**Principle Component Analysis (PCA)** is mainly used for the dimensionality reduction. In PCA orthogonal transformation is used, so as to transform the original data samples of co-related samples into the set of linearly uncorrelated samples. This sample is known as Principle Components. PCA is sensitive to the relative scaling of the original variables. PCA is mostly used as a tool in exploratory data analysis and for making predictive models. PCA can be done by eigenvalue decomposition of a data covariance (or correlation) matrix or singular value decomposition of a data matrix, usually after mean centering (and normalizing or using Z-scores) the data matrix for each attribute.

**QR decomposition** is used for the decomposition of a matrix. Modified matrix is a product of orthogonal matrix (Q) and upper triangular matrix (R).

$$A = QR$$

If instead *A* is a complex square matrix, then there is a decomposition $A = QR$ where *Q* is a unitary matrix (so $Q^*Q = I$).If *A* has *n* linearly independent columns, then the first *n* columns of *Q* form an orthonormal basis for the column space of *A*.

More generally, the first *k* columns of *Q* form an orthonormal basis for the span of the first *k* columns of *A* for any $1 \leq k \leq n$.[1] The fact that any column *k* of *A* only depends on the first *k* columns of *Q* is responsible for the triangular form of *R*.

**Watershed** is a basin-like landform defined by highpoints and ridgelines that descend into lower elevations and stream valleys [6]. In image processing watershed is used for segmentation purpose. Watershed uses the region based approach and searches for the pixel and region similarities.

In graphs, watershed lines may be defined on the nodes, on the edges, or hybrid lines on both nodes and edges. Watersheds may also be defined in the continuous domain.[1] There are also many different algorithms to compute watersheds.

$$T[n] = \{(s,t) \mid g(s,t) < n\}$$

g(*s*,*t*) is intensity. n= min+1 to n = max +1. And let T[n]=0, others 1

$$C[n] = \bigcup_{i=1}^{R} C_n(M_i) \qquad C_n(M_i)$$

**Hilbert** is mainly used for signal processing. It is an analytical representation of the signal. In this paper Hilbert transform is used as the method of data perturbation. The Hilbert transform gives the transformation in complex number.

$$G_H(f) = H(f)G(f)$$

$$g_H(\tau) = h(x) * g(x), \quad \text{where } h(x) = \frac{1}{\pi x}$$

$$H(f) = -j\,\mathrm{sgn}(f)$$

**Hierarchical Clustering**

Hierarchical clustering group data over a variety of scales by creating a cluster tree or *dendrogram*. The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. This allows to decide the level or scale of clustering that is most appropriate for the application. The dendrogram function plots the cluster tree. The hierarchical, binary cluster tree created is most easily understood when viewed graphically using dendrogram. In hierarchical clustering, we can change the number of clusters anytime during in the process if we want.

*Agglomerative* is bottom-up approach. Initially each object is consider as individual cluster, later it merges the objects according to the similarity between them. It continues till all objects belong to one cluster. Various linkage methods are used to calculate the distance between the objects such as single-link, complete link, average-link and centroid. It produces ordering of objects, which is informative for displaying the data.

**Privacy Measures**

Value Difference: After applying perturbation on the data samples, the data is modified. The modified changes is the value difference (VD) between the original data and perturbated data. It is given by the relative value difference in Forbenius norm. The value difference is the ratio Forbenius norm of difference in original data (A) and the perturbated data (PA) to the original data (A). [3]

$$VD = ||A\text{-}PA||_f / ||A||_f$$

Position Difference: After Data Perturbation on the dataset, the relative position of the data sample is modified also. There are several metrics to measure the positional difference of the data samples [3]

*RP:* It is use to represent the average change of the order for every attributes in data sample. After the data of an attribute is perturbated, the order of each data is changed. Let us say Original data A has n observation and m attributes. $Order^{i}_{j}$ denotes the ascending order of the perturbated sample $A_{ij}$. The RP is defined as [3]

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

*RK:* RK gives percentage of the sample that keeps their orders of value in each attribute after the perturbation. It is calculated as: [3]

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

Where, $RK^{i}_{j}$ gives whether or not a sample keeps its position in the order of the value:

$$Rk_j^i = \begin{cases} 1, & \text{if } Ord_j^i = \overline{Ord}_j^i \\ 0, & \text{otherwise.} \end{cases}$$

*CP:* The content of an attribute can be inferred from its relative value difference compared with the other attributes. Thus it is necessary to know the order of the average value of the attribute varies after the data perturbation. The CP metric can be used to define the change of the average value of attributes: [3]

$$CP = (\sum_{i=1}^{m}|(OrdAV_i - \overline{OrdAV}_i|)/m$$

Where $OrdAV_{i,}$ is the ascending order of the attribute i, while $\overline{OrdAV}_i$ represents its ascending order after the perturbation.

*CK:* Similar to RK, CK can be defined to measure the percentage of the attributes that keep their orders of average value after the perturbation. Hence it is given as: [3]

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

Where $CK^{i}$ is calculated as:

$$Ck^i = \begin{cases} 1, & \text{if } OrdAV_i = \overline{OrdAV}_i \\ 0, & \text{otherwise.} \end{cases}$$

The value of RP and CP should be highest and the value of RK and CK should be Low, to preserve more privacy.

**Cophenetic Distance:**

In a hierarchical cluster tree, any two objects in the original data set are eventually linked together at some level. The height of the link represents the distance between the two clusters that contain those two objects. This height is known as the *cophenetic distance* between the two objects. One way to measure how well the cluster tree is generated is to compare the cophenetic distances with the original distance data. If the clustering is valid, the linking of objects in the cluster tree should have a strong correlation with the distances between objects in the distance vector. The cophenet function compares these two sets of values and computes their correlation, returning a value called the *cophenetic correlation coefficient*. The closer the value of the cophenetic correlation coefficient is to 1, the more accurately the clustering solution reflects your data.

**Results and Discussion**

The dataset used for experiment is Haberman dataset. Dataset is multivariate. It has 4 attributes and 306 observations. It has numerical data. There are no missing values in dataset. The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer. (http://archive.ics.uci.edu/ml/datasets/Haberman's+Survival).

Attributes are age of patient at the time of operation, Patient's year of operation, number of positive axillary node detected, Survival status. Table 1 shows the results of the privacy measures. Various data perturbation methods are applies on Haberman dataset. The data perturbation methods used for comparisons are principle component analysis (PCA), singular value decomposition (SVD), orthogonal transformation (QR), watershed and Hilbert. VD is highest in SVD, which shows the data is differed from the original dataset more than the other methods but comparatively watershed gives the better result in all measures. Lowest VD is the of Hilbert method, which states that the distorted values are not efficient comparatively. The RP value is highest in watershed than the others, which states that the order of each attribute is changed efficiently in this method. The CP value is highest in watershed, which states that data distortion is done in changing the order of the attributes. The lowest RK is in SVD but watershed shows slight high which is not bad result in comparison. Lowest RK value states that percentage of elements that keep their orders of value in each column. Cophenetic distance of watershed is 0.9 which is closer to 1, hence the clustering is done effectively.

**Table 1: Results of Privacy Measures.**

| | VD | RP | RK | CP | CK | Cophenetic distance |
|---|---|---|---|---|---|---|
| PCA | 0.9960 | 84.3301 | 0.0474 | 25805 | 0 | 0.6017 |
| SVD | 1.3966 | 61.4379 | 0.0033 | 18800 | 0 | 1.000 |
| QR | 1.3765 | 62.0131 | 0.0090 | 18976 | 0 | 0.9997 |
| Watershed | 1.0556 | 84.4461 | 0.0049 | 25841 | 0 | 0.9197 |
| Hilbert | 0.0000-0.1609i | 76.7533 | 0.0310 | 23487 | 0 | 0.6017 |

Hence from observation states that the suggested watershed technique gives the good results in maximum measures of the privacy. The figures 1 and 2 gives the result in the dendrogram after applying hierarchical clustering on the dataset. Figure 1 shows the dendrogram applied on the original dataset before data perturbation, whereas figure 2 shows the dendrogram applied on the perturbated dataset. The perturbation has been done using watershed technique. As observed in the Y axis of both the figures it is inferred that the data is altered. The Y axis gives the distance between each cluster and X-axis represents the objects.

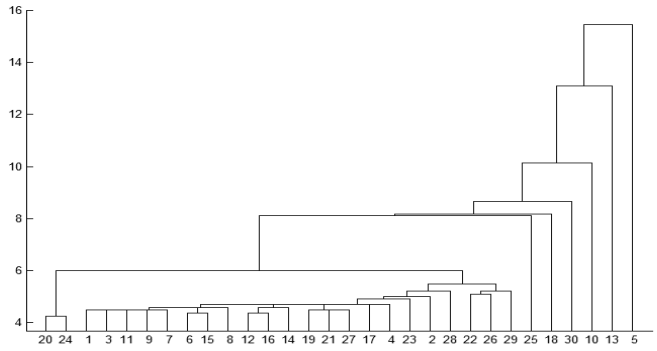


**Figure 1. Before Data Perturbation.**

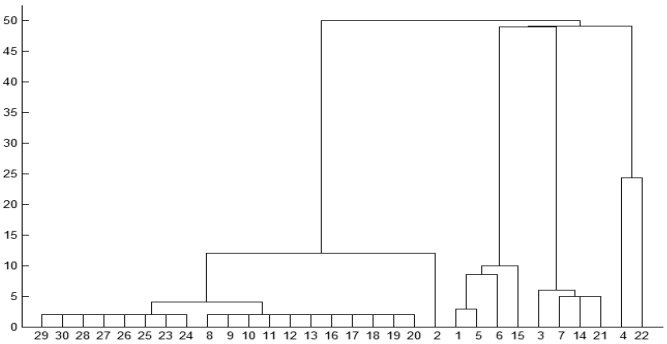


**Figure 2. After Data Perturbation.**

**Conclusion**

The paper discussed various data distortion methods. The experimental results shows that waterhsed is best in preserving data privacy. As the statistical result shows watershed gives the best result as compared to SVD, PCA, QR and Hilbert transformation. The focus is also on privacy preserving data clustering. Experimental results have demonstrated that the proposed method is highly effective for high accuracy privacy protection, in the sense that they can provide high degree of data distortion and maintain high level of data utility with respect to the data mining algorithms. In future, it is certainly of interest for the research community to experiment these data distortion technique with other data mining algorithms.

**References**

1. Stanley R.M. Oliveria, Osmar R. Zaiane, Privacy Preservation Clustering by Data Transformation, *Journal of Information and Data Management*, 2010,*Vol. 1* no. 1.
2. Gaung Li , Yadong Wang, A Privacy-Preserving Classification Method Based on Singular Value Decomposition, *International Arab Journal of Information technology*,2012, *Vol 9*, No 6.
3. Shuting Xu, Jun Zhang, Dianwei Han, Jie Wang, Data Distortion for privacy Protection in terrorist Analysis System, *Technical Report* No 432-05, Department of Computer Science, University of Kentucky, (2005)
4. Kavitha S, Raja Vadhana P, Data Privacy Preservation Using Various Perturbation Techniques, *International Journal of Innovative Research in Computer and Communication Engineering*, 2015, *Vol 3* Issue 2.
5. Samir Patel, Kiran R. Amin, Privacy Preserving Based on PCA Transformation Using Data Perturbation Technique, *International Journal of Computer Science and Engineering Technology*, 2013, *Vol 4*, No. 5.
6. Lamia Jaafar Belaid, Walid Mourou, *Image Segmantation*: A watershed Transformation Algorithm, *Image Anal Stereol*, 2009; *vol.28*: 93-102.
7. Santosh kumar Bhandare , Data Distortion Based Privacy Preserving Method for Data Mining System, *International Journal of Emerging Trends &technology in Computer Science*, 2013,Vol.2 Issue 3.
8. Keke Chen, Ling Liu, Privacy Preserving Multiparty collaborative mining with geometric data perturbation, *IEEE transaction on parallel and distributed Systems*, 2009, vol 20, issue 12.
9. S. Chidambaram; K. G Srinivasagan, A combined random noise perturbation approach for multilevel privacy preservation in data mining, *Recent Trends in Information Technology (ICRTIT)*, (2014).

10. C.Gokulnath, M.K.Priyan, E.Vishnu Balan, K.P.Rama Prabha, R.Jeyanthi, Preservation of privacy in data mining by using PCA based perturbation technique, *International conference on smart technologies and management for computing, communication, control, energy and materials*, ICSTM, (2015).
11. Priya Meshram; Sonali Bodkhe, Review on privacy preservation method by applying discrimination rules in data mining, *Pervasive Computing (ICPC),* (2015).
12. Lei Xu, Chunxiao Jiang, Yan Chen,Jian Wang,Yong Ren,A Framework for Categorizing and Applying Privacy-Preservation Techniques in Big Data Mining, *IEEE Computer* (2016) ,Vol.49, Issue 2".
13. Zahra Nazari, Dongshik Kang, M.Reza Asharif, Yulwan sung, Seiji Ogawa, A new hierarchical clustering algorithm, *International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS*), (2015).
14. Thanveer Jahan, G. Narshima, C.V. Guru Rao, Data perturbation and feature selection in preserving privacy *Ninth International Conference on Wireless and Optical Networks (WCON),*( 2012).

**Corresponding Author:**
**Maheswari N*,**
**Email:** *maheswari.n@vit.ac.in*