ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

Privacy Preserving Clustering Based on Discrete Cosine Transformation

M. Naga lakshmi[1], K Sandhya Rani[2]

Research Scholar: Dept of Computer Science, S.P.M.V.V, Tirupati, Andhra Pradesh, India[1]

Professor, Dept of Computer Science, S.P.M.V.V, Tirupati, Andhra Pradesh, India[2]

***ABSTRACT*:** The information related to an individual or an organization could be compromised when the patterns extracted from large databases through data mining technology. Privacy preserving data mining which is a new research area has been evolved in order to find the right balance between maximizing analysis results and minimizing the disclosure of private information. In this paper, a Discrete Cosine Transformation (DCT) based data distortion method is proposed for privacy preserving clustering in centralized database environment. The experimental results proved that the proposed method efficiently protects the private data of individuals and retains the important information for clustering analysis.

KEYWORDS: Discrete Cosine Transformation, Data distortion, Privacy Preservation, Clustering.

I. INTRODUCTION

Due to the degradation of hardware costs, gigantic amounts of data are being collected from business and other sources can be stored for future use. This enforces the necessity of analyzing these mountains of data and discovering important patterns. Data users are also expecting more sophisticated information from these data. Data could be large in two senses, in terms of size or in terms of dimensionality. Extraction of valuable information from these massive amounts of data is the major challenge for the organizations and makes this information available at a particular time, place and in the form needed to support the decision-making process. The extracted knowledge is helpful for business analysis and scientific computing. The increasing demand of information extraction is not supported by the simple structured query languages. This problem can be effectively addressed with data mining techniques.

Data mining is often defined as the process of discovering meaningful, new correlation patterns and trends through non-trivial extraction of implicit, previously unknown information from large amounts of data stored in repositories using pattern recognition as well as statistical and mathematical techniques [1]. The extracted knowledge is presented in the form of association rules, decision trees, clustering, and prediction. Among these techniques, clustering is an important technique, which has been widely used in various disciplines such as image processing, pattern recognition and market analysis etc.

Clustering is an unsupervised learning technique, which groups the data objects based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. Data mining techniques are useful for industrial and government organizations for various purposes such as increasing profitability and to enhance national security. Private and sensitive data available in the large databases need to be preserved and avoid the privacy leakage with the data mining system. To guard against private use of personal information, a new research area known as privacy preserving data mining is emerged to conceal sensitive information while retaining important information for data mining. In this paper, Discrete Cosine Transformation (DCT) based data distortion method is proposed for privacy preserving clustering. The related works of privacy preserving clustering is discussed in the following section.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

II. Related work

In [2], the authors discussed about different privacy preserving methods such as randomization, k-anonymization and distributed privacy preserving data mining. The evaluation and the usability of online privacy policies, how well these polices meets the requirements of the user and also suggested improvements by analyzing the available policies has been addressed in [3]. The authors in [4] proposed novel data distortion strategies via attribute partition, with different combinations of singular value decomposition (SVD), nonnegative matrix factorization (NMF), discrete wavelet transformation (DWT) to perturb sub matrices of the original datasets for privacy protection. A Fourier-related transformation based novel generalized approach to hide sensitive data values and also preserving Euclidian distances in centralized and distributed scenario is presented in [5]. A privacy preserving clustering technique using haar wavelet transformation and scaling data perturbation in centralized database environment is proposed by authors in [6]. A simple principle component analysis based transformation approach has been presented by authors in [7] and it is tested on various datasets for privacy preservation. In [8], a wavelet based data perturbation method for privacy preserving classification in distributed database environment is described. This method performs both privacy preservation and dimensionality reduction. The following section describes the proposed method.

III. Proposed Method

To achieve the dual goal of privacy preservation and valid clustering results, a discrete cosine transformation based method is proposed in centralized database environment. A discrete cosine transformation is a type of Fourier related transforms, which can converts the data from the original domain to transformed domain. Discrete cosine transformation consider a series of complex numbers x_0, x_1, ---- x_{n-1}, and generates a set of real coefficients f_0, f_1----- f_{n-1}

A. Algorithm for the Proposed Method

Discrete cosine transformation is a unitary transformation, which can preserve the Euclidian distances between original and transformed domains. The focus of DCT is preserving Euclidian distances as well as privacy preservation. Algorithm for the proposed method is described in Table1. The identifier attributes that are not relevant for data mining are removed. The dataset is normalized using z-score normalization to standardize the attributes to the same scale. The data perturbation process is in two steps. In step one, original dataset is reduced and generates a transformation matrix. In step two, distorted dataset is obtained by multiplying the transformation matrix with the original dataset.

TABLE 1: ALGORITHM FOR PROPOSED METHOD

Input: Original dataset $D_{m \times n}$

Output: Distorted Dataset $D'_{m \times n}$

Begin

a. Identifier and other non numerical attributes are suppressed
.

b. Reduce the matrix of size n and generates the transformation matrix using Discrete Cosine Transformation,

$$T = w(k) \sum_{n=1}^{N} D(n) \cos \frac{\pi(2n-1)(k-1)}{2N}$$

For K= 1,2----- N,

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

Where $w(k) = \begin{cases} \frac{1}{\sqrt{N}}, & k = 1 \\ \sqrt{\frac{2}{N}}, & 2 < k < N \end{cases}$

c. Calculate the distorted matrix D′ = D * T

End

The implementation details of the proposed method are explained in the following section.

IV. Implementation Of The Proposed Method

The experimental evaluation of the proposed method is performed by considering two measures: (a) Degree of privacy and (b) Clustering quality. Experiments are conducted on three real life datasets from UCI [9]. They are Haber man dataset with 3 attributes and 306 instances, Liver dataset with 5 attributes and 345 instances, Credit-g dataset with 5 numerical attributes and 1000 instances. As the first experiment, DCT matrix transformation is applied on three datasets, second experiment, Principal Component Analysis (PCA) is applied on three datasets for comparison and misclassification error, overall F-measure values and privacy values are computed.

A. Degree of Privacy

The privacy provided by the data transformation method is measured as the amount of sensitive information hidden successfully. It is the variance between the actual and the perturbed values [10]. This measure is given by Var (X − Y) where X represents a single original attribute and Y is the distorted attribute.

$$S = Var\ (X - Y)/Var(X)$$

The higher S values indicate that privacy protection is high. The privacy protected by the data transformation method is shown in the following table.

Table 2: Privacy Values of the Transformed Datasets.

	Haber man	Liver	Credit-g
DCT Matrix Transformation	8.356	4.222	6.428
PCA	1.6267	1.3756	2.0337

The privacy values of the perturbation methods PCA and proposed DCT matrix transformation method are shown in Table 2 and it clearly shows that the proposed method is providing higher level of privacy guarantee. The results of the experiments are presented in terms of graph for all the three datasets in Figure 1.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

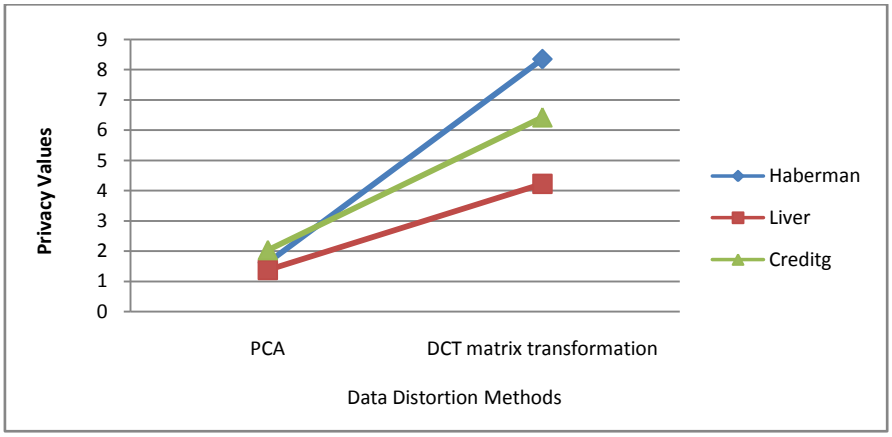

FIGURE 1: PRIVACY VALUES OF THE DATA TRANSFORMATION METHODS

From the results shown in Figure 1, it is confirm that the proposed method gives higher privacy values for all the three datasets. Hence the privacy preservation of the proposed method is providing higher privacy values when compared to data perturbation method PCA.

B. Clustering Quality

The perturbed dataset is compared with the original dataset to measure the clustering quality of the proposed method. It is measured based on the misclassification error and overall F-measure. When the dataset is transformed, the clusters in the original dataset should be equal to the clusters in the distorted dataset. K-means clustering algorithm is applied to the original dataset as well as the transformed dataset. WEKA (Waikato Environment for Knowledge Analysis) software is used to test clustering accuracy of the original and modified data set.

1) Misclassification Error (M_E)

The misclassification error, denoted by M_E, is measured as follows:

$$M_E = \frac{1}{N}\sum_{i=1}^{k}(|Cluster_i(D)| - |Cluster_i(D')|)$$

In the above formula
N - Number of points in the original dataset.
K - Number of clusters.
$Cluster_i$ (D) - Number data points of the i^{th} cluster of the original data set.
$Cluster_i$ (D′) - Number of data points of the i^{th} cluster of the transformed dataset.
Higher M_E values indicates lower clustering quality where as lower M_E values indicate the higher utilization of the data. The experiments are conducted 10 times and M_E value is taken as an average of 10 experiments. The misclassification error values of the PCA and proposed method are shown in the following Table.

TABLE 3: MISCLASSIFICATION ERROR VALUES OF THE DISTORTED DATASETS

	Haber man	Liver	Credit-g
DCT Matrix Transformation	0.1274	0.1287	0.1591
PCA	0.116	0.09618	0.1956

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

When the misclassification error values of PCA and proposed method which are illustrated in Table 3 are compared, it clearly shows that the proposed method yields good clustering results for all the three datasets. The results of misclassification error and privacy values proved that the proposed method DCT Matrix Transformation method is effective and provide practically acceptable values for balancing privacy and accuracy. The misclassification error values of the data distortion methods for three datasets are shown in the following figure.

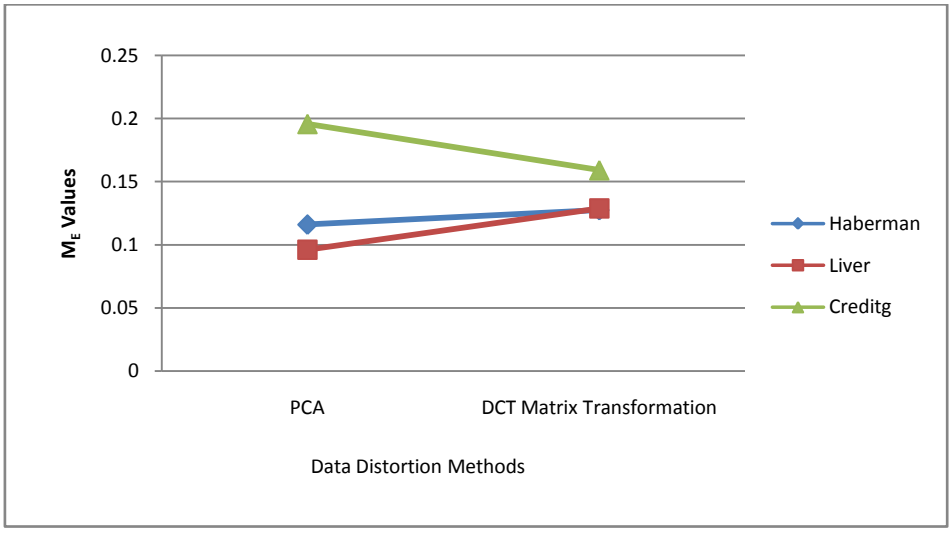

FIGURE 2: MISCLASSIFICATION ERROR VALUES OF THE DATA TRANSFORMATION METHODS

Misclassification error values indicate the utility of the data for data mining analysis. The misclassification error values of the two experiments PCA and DCT matrix transformation are depicted in Figure 2. It clearly reveals that, the proposed method yields comparable misclassification error values for all the three datasets.

2) Overall F-measure (OF)

Clustering quality can be measured based on the overall F-measure (OF) [5]. The higher OF values indicates higher clustering quality. Let C_1, C_2, ….., C_n are the clusters according to the original dataset. Let C_1', C_2', ….., C_n' are the clusters of the distorted dataset. The F-measure of a cluster generated from the original dataset C_i and a cluster generated from the distorted dataset C_j' is defined as follows.

$$F_{ij} = 2\frac{P_{ij} R_{ij}}{P_{ij} + R_{ij}}$$

$$\text{Precision } P_{ij} \text{ equals: } \frac{C_i \cap C_j'}{C_j'}$$

$$\text{Recall } R_{ij} \text{ equals: } \frac{C_i \cap C_j'}{C_i}$$

The F-measure of a cluster C_i is given by

$$F_i = \max_j F_{ij}$$

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

The overall F-measure can be calculated by using the following formula

$$F = \sum_{i=1}^{n} \frac{C_i}{N} F_i$$

In the above formula N is the total number of records in the dataset. Table 4 displays overall F-measure values on average for 2 to 5 clusters.

TABLE 4: OVERALL F-MEASURE (OF) VALUES OF THE TRANSFORMED DATASETS

	Haber man	Liver	Credit-g
DCT Matrix Transformation	0.899	0.857	0.813
PCA	0.917	0.928	0.79

The overall F-measure values illustrated in Table 4 reveals that the clustering quality of proposed method provides practically acceptable values when compared to the data perturbation method PCA for the two datasets. The following figure depicts the overall F-measure values of PCA and proposed DCT matrix transformation methods.

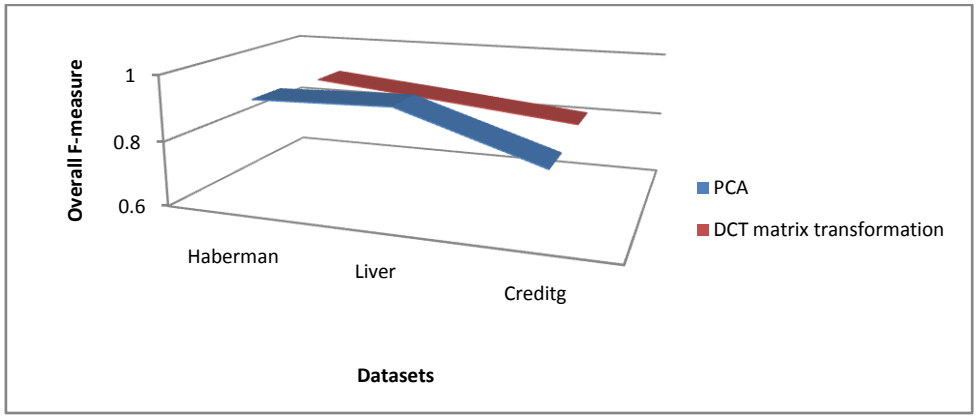

FIGURE 3: OVERALL F-MEASURE (OF) VALUES OF THE DATA TRANSFORMATION METHODS

From Figure 3, it can easily understand that the proposed method gives comparable overall F-measure values. This means the clustering quality of proposed method is good for one dataset, for other two datasets clustering quality is slightly lower when compared to data perturbation method PCA. This lower clustering quality can be effectively offset with the higher security values provided by the proposed DCT matrix transformation method.

V. CONCLUSION

Privacy and security issues are restricts the mining of large and private databases. Privacy preserving data mining is a new area, which promises the data owners for the privacy preservation of individuals as well as valid clustering results. In this paper, a discrete cosine transformation based data distortion method is proposed for privacy preserving clustering in centralized database environment. The proposed method is successfully implemented on three real life datasets from UCI for clustering analysis. The experimental results proved that the proposed method efficiently protects the private data of individuals and retains the important information for clustering analysis when compared to the principle component analysis based transformation approach.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

REFERENCES

[1]. U. M. Fayyad., G. Piatetsky-Shapiro., P. Smyth., "From Data Mining to knowledge Discovery in databases", In proceedings of American Association of Artificial Intelligence, March 1996.
[2]. C.C.Aggrawal., P.S. Yu., A General Survey of Privacy Preserving Data Mining Models and Algorithms, Volume 34 of Advances in database Systems, chapter 2, pp.11-52.Springer US, 2008.
[3]. Carlos Jensen., Colin Potts., "Privacy policies as decision making tools: an evolution of online privacy notices", Proceedings of the SIGCHI conference of Human factors in computing systems, Apr 2004, pp.471-478.
[4]. Bo Peng, Xingyu Geng., JunZhang., "Combined Data Distortion Strategies for Privacy-Preserving Data Mining", In proceedings of 3rd International Conference on Advanced Computer Theory and Engineering (1CACTE) ,2010.
[5]. S.Mukherjee., Z.Chen., A.Gangopadhyay., "A Privacy-Preserving technique for Euclidian distance-based mining algorithms using Fourier-related transforms", Proc. Of the 32nd Int'l Conference on Very Large Databases (VLDB '06),2006, pp.293-315.
[6]. Sara Hajian., Mohammad Abdollahi Azgomi., "A Privacy Preserving Clustering Technique Using Haar Wavelet Transform and Scaling Data Perturbation", In proceedings of Innovations in Information Technology, 2008.
[7]. Samir Patel., Kiran R. Amin., "Privacy preserving based on PCA transformation using data perturbation technique", Vol 4, No.05 of International Journal of Computer Science & Engineering Technology, May 2013, pp.477-484.
[8]. S.Bapna., A.Gangopadhyay., "A Wavelet-based approach to preserve privacy for classification mining", Decision Sciences Journal, 37(4), pp.623-642, 2006.
[9]. Machine Learning Repository http://archive .ics.uci.edu/ml/datasets.html.
[10]. Oliveira, S. R. M., Zaiyane, O. R., "Achieving Privacy Preservation When Sharing Data for Clustering", In Proceedings of the Workshop on Secure Data Management in a Connected World, in conjunction with VLDB'2004, pp. 67–82, 2004.