SVD based Data Transformation Methods for Privacy Preserving Clustering

M. Naga Lakshmi[1] & K Sandhya Rani[2]
[1]Research Scholar, [2]Professor
Dept of Computer Science, S.P.M.V.V,
Tirupati, Andhra Pradesh, INDIA

ABSTRACT

Nowadays privacy issues are major concern for many government and other private organizations to delve important information from large repositories of data. Privacy preserving clustering which is one of the techniques emerged to addresses the problem of extracting useful clustering patterns from distorted data without accessing the original data directly. In this paper two hybrid data transformation methods are proposed for privacy preserving clustering in centralized database environment based on Singular Value Decomposition (SVD). In hybrid method one, SVD and rotation data perturbation are used as a combination to obtain the distorted dataset. In hybrid method two, SVD and independent component analysis are used as a combination to obtain the distorted dataset. In SVD the data is analyzed in different perspectives to retain important information. Higher order statistics which contains more important information is utilized in independent component analysis. Experimental results demonstrate that the proposed methods are efficiently protects the private data of individuals and retains the important information for clustering analysis.

Keywords: Singular value decomposition, Independent component analysis, Privacy preserving clustering

1. INTRODUCTION

The advancement of information technology enables to collect huge amounts of data from various sources such as banks, telecommunications, business, web, text etc. Data mining is the process of sifting through huge databases, summarizing them and finding useful patterns. Association rules, classification, clustering and regression are some of the data mining tasks. The process of dividing the database into similar groups is known as clustering. Relative distance or relative density is taken as the similarity measure for the clustering objects. This knowledge discovery process poses privacy issues due to the disclosure of sensitive information of individuals.

To resolve the problem of privacy, a new research area called privacy preserving data mining has been evolved. The process of privacy preserving data mining is to extract useful patterns without breaching the privacy of individuals. Different techniques have been proposed for protecting the privacy of individuals such as data modification, data partitioning, data restriction and data ownership [1]. In this paper the protection of privacy is considered when clustering is performed in centralized database environment. Privacy preserving clustering is the process of protecting sensitive attribute values in the databases while preserving the general features for clustering analysis. In this paper Singular Value Decomposition (SVD) based two hybrid data transformation methods are proposed for privacy preserving clustering in centralized database environment. The related works are discussed in the following section.

2. LITERATURE SURVEY

Privacy threats for the people have been increased about the abuse of genetic information is addressed in [2], such as denying health insurance, employment, education and loans. Usage of genetic information for the healthcare setting is based on the clinician’s ethical and social responsibility. The authors in [3] described a recent survey on web users found that many of the respondents believe that participation in beneficiary programs is the cause for individual privacy. To address the privacy problem various privacy preserving data mining methods in both centralized and distributed database environment have been discussed by authors in [4]. The authors in [5] proposed a Singular Value Decomposition (SVD) strategy for privacy preservation. Further they discussed some metrics to measure the difference between distorted dataset and the original dataset and the degree of privacy protection. A SVD-based randomization approach for E-commerce which provides recommendations for collaborative filtering and also protecting privacy of individuals is described in [6]. An improved SVD based data value hiding method for privacy disclosure and the distorted dataset provides utility of the dataset without breaching privacy is presented in [7]. A privacy preserving classification algorithm using SVD for data distortion and application of SVD on structured partitions is proposed by authors in [8]. In [9], a SVD based data distortion method for privacy preserving clustering is addressed. The authors in [10] proposed privacy preserving classification hybrid method as a combination SVD and ICA. A hybrid data distortion method to preserve the confidentiality of numerical attributes in centralized database environment has been presented in [11]. The following section discusses the proposed hybrid methods for privacy preserving clustering.

3. PROPOSED METHODS

Privacy protection is an important issue when the data is shared by many users for clustering analysis. Privacy can be achieved by effective hiding of sensitive values. Many techniques have been proposed by the researchers for distorting the private data in centralized database environment. In order to improve the privacy provided by single data perturbation methods, two hybrid methods are proposed. A combination of SVD and rotation data perturbation is used in hybrid method one where as a combination of SVD and independent component analysis is used in hybrid method two.

3.1 Singular Value Decomposition

Among the different methods in data mining, Singular Value Decomposition (SVD) is one of the familiar methods [12]. The dimensionality of the original dataset can be re reduced by SVD and also used as a data distortion method. The original dataset A is represented as n × m matrix. The data objects are represented as rows and attributes are represented as columns. The singular value decomposition is a more general method that factors any n × m matrix A of rank r into a product of three matrices, such that

$$A = UWV^T \qquad (1)$$

From the above formula, U is an n × n orthonormal matrix, W is an n × m diagonal matrix whose nonnegative diagonal entries (the singular values) are in descending order, and V^T is an m × m orthonormal matrix. Because of the arrangement of singular values in the matrix W the SVD transformation has the property that maximum variation in the objects are taken in the first dimension and most of remaining variations are captured in second dimension, and so on. The rank-k approximation of A_k to the matrix A can be defined as

$$A_k = U_k V_k W_k \qquad (2)$$

From the above formula, U_k contains the first k columns of U, W_k contains the first nonzero singular values, and V_k^T contains the first k rows of V^T. With k being usually small, the dimensionality of the dataset has been reduced dramatically from min (m, n) to k (assuming all attributes are linearly independent). The various steps in SVD bases data transformation to obtain distorted database are given in the following section.

3.1.1 Algorithm for SVD based Data Transformation.

1. Original dataset D consists of m rows and n columns.
2. Distorted Dataset D′ consists of m rows and n columns.
3. Suppress all identifier attributes from the given matrix $D_{m \times n}$.
4. Apply SVD on the matrix D to obtain decomposed matrices U, W, V^T.
5. Compute the distorted matrix $D' = UWV^T$
6. Release the distorted matrix D′ for clustering analysis.

3.2 Independent Component Analysis (ICA)

ICA is a statistical method for transforming a complex dataset into independent subparts. The ICA model represents the observed dataset X as a linear combination of mixing matrix and random matrix [13].

$$X = AS \qquad (3)$$

In the above formula S is an m × n random matrix whose values are assumed as independent and A is an n × n mixing matrix. Before applying ICA algorithm, the data should be centered and whitened. Centering is the process of subtracting its mean from variables and converted into zero mean variables. Whiting is a preprocessing strategy that transforms the components of data matrix into uncorrelated and the variance equal to unity. ICA coefficients of the matrix are the sparse coding representation of the original data. The important information is retained in the elements with larger abstract values in the sparse coding. The elements with lower abstract values are considered as noise.

The independent components are latent variables, meaning that they cannot be directly observed and also the mixing matrix is assumed to be unknown. Only X, the observed matrix is known and estimate both A and S using it. Then, after estimating the mixing matrix A, compute its inverse, say W in order to obtain the independent components using the following formula.

$$S = WX \qquad (4)$$

3.3 Rotation Data Perturbation

Rotation Data Perturbation (RDP) is one of the geometric transformations [14] which leaves the metric properties unaltered. In RDP, the noise term is a rotation angle θ and rotation matrix $R_o(\theta)$ is obtained from the rotation angle, which is used to rotate the k pairs of attributes from data matrix D. If number of attributes in D is odd, then the last attribute is paired with an already selected attribute randomly. Each attribute is taken once, when the number of attributes is even. $D' = R_o(\theta) \times D$, where D is the column vector containing the original coordinates and D′ is a column vector whose coordinates are rotated coordinates and $R_o(\theta)$ is a 2 × 2 matrix.

$$R_o(\theta) = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \sin\theta \end{bmatrix} \qquad (5)$$

3.4 Hybrid Methods

To enhance the performance of the SVD based single data perturbation, two hybrid data transformation methods are proposed for privacy preserving clustering. In hybrid method one, SVD and rotation data perturbation are used as a combination to obtain the distorted dataset. In hybrid method two, SVD and independent component analysis are used as a combination to obtain the distorted dataset.

3.4.1 Hybrid Method-1(SVD & RDP)

Hybrid method-1(SVD & RDP) is proposed by taking the advantage of two existing techniques SVD and rotation data perturbation to optimize the privacy provided by single data perturbation method. The details of the proposed hybrid method-1(SVD & RDP) are described in the following algorithm. The given input dataset is preprocessed by removing unnecessary attributes for data mining and normalized using z-score normalization. The dataset is decomposed using SVD data perturbation into three matrices U, W, V^T. Matrix V^T is the input for the rotation data perturbation and distorted into $V^{T'}{}'$. The final distorted dataset is calculated as a product of the matrices $U, W, V^{T'}$

3.4.2 Algorithm for Hybrid Method-1 (SVD & RDP)

1. Original dataset D consists of m records and n attributes.
2. Distorted Dataset D′ consists of m records and n attributes.
3. Identifier and other non numerical attributes are suppressed.

4. Apply SVD on the matrix D to obtain decomposed matrices U, W, V^T
5. Calculate $k = n/2$ if n is even else

   $k = (n + 1)/2$
6. For each k pairs of attributes in V^T
7. For each pair of attributes A_i, A_j from step 6 where $1 \leq i \leq n$ and $1 \leq j \leq n$
8. Compute $V^{T'}(A'_i, A'_j) = R_o(\theta) \times V^T(A_i, A_j)$ for different values of θ and identify the range that gives higher privacy vales
9. Select an angle θ from the selected range that gives highest privacy preservation to compute the noise term $R_o(\theta)$
10. Using this $R_o(\theta)$ Compute

    $V^{T'}(A'_i, A'_j) = R_o(\theta) \times V^T(A_i, A_j)$
11. Calculate the transformed matrix

    $D' = U_k W_k V_k^{T'}$
12. Release the distorted dataset D′ for clustering analysis

3.4.3 Hybrid Method-2(SVD & ICA)

The security provided by the single data perturbation SVD can be enhanced by the proposed hybrid method-2(SVD & ICA). This method is developed as the hybridization of SVD and independent component analysis. The following section describes the algorithm for proposed hybrid method-2(SVD & ICA).

3.4.4 Algorithm for Hybrid Method-2 (SVD & ICA)

1. Original dataset D consists of m records and n attributes.
2. Distorted Dataset D′ consists of m records and n attributes.
3. Identifier and other non numerical attributes are suppressed.
4. Apply SVD on the matrix D to obtain decomposed matrices U, W, V^T
5. ICA model express the input matrix V^T as a linear combination of matrices A and S as

   $V^T = AS$
6. Apply preprocessing techniques centering and whitening on V^T
7. Estimate the unknown mixing matrix A from V^T
8. Compute matrix $W = A^{-1}$ and obtain independent components by the formula

   $$S = WV^T$$
9. The matrix S consists of independent components is considered as distorted matrix $V^{T''}$
10. Calculate the final distorted dataset

    $D' = UW V^{T'}$
11. Release the distorted dataset D′ for clustering analysis

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

4. IMPLEMENTATION OF THE PROPOSED HYBRID METHODS

The proposed method is evaluated by conducting the experiments on three real life datasets from UCI [14]. The datasets used for the experiments are Iris with 4 attributes and 150 instances, Brest cancer with 10 attributes and 683 instances, and Credit-g with 5 numerical attributes and 1000 instances. The performance of the data distortion method is measures based on two factors, I) Utility measures and II) Privacy measures. The well-known k-means clustering algorithm is used to measure the clustering quality.

The utility of the dataset is measured based on the misclassification error [15]. When the dataset is transformed, the clusters in the original dataset should be equal to the clusters in the distorted dataset. WEKA (Waikato Environment for Knowledge Analysis) software [16] is used to test clustering accuracy of the original and modified data base. The misclassification error, denoted by M_E, is measured as follows.

$$M_E = \frac{1}{N}\sum_{i=1}^{k}(|Cluster_i(D)| - |Cluster_i(D')|) \quad (6)$$

In the above formula
N - Number of points in the original dataset.
K - Number of clusters.
$Cluster_i$ (T) - Number data points of the i^{th} cluster of the original data set.
$Cluster_i$ (T′) - 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. In order to measure the effectiveness of the proposed hybrid methods for privacy preserving clustering in centralized database environment, experiments are conducted on three real life datasets. These datasets are transformed by applying the hybrid method-1(SVD & RDP) algorithm as described in the section 3.4.2 and hybrid method-2(SVD & ICA) algorithm as described in the section 3.4.4. For all the three datasets misclassification error and privacy values are calculated for SVD, hybrid method-1(SVD & RDP), hybrid method-2(SVD & ICA). These values are illustrated in the following Table.

Table 1: Misclassification Error Values of Transformed Datasets

	Iris	Brest cancer	Credit-g
SVD	0.07473	0.0625	0.1808
HybridMethod-1 (SVD &RDP)	0.06133	0.0588	0.1732
HybridMethod-2 (SVD &ICA)	0.0553	0.0441	0.1805

The data transformation methods displayed in Table 1 are compared, it clearly indicates that the proposed hybrid methods provide lower misclassification error values and are better than the single data perturbation SVD method.

The effectiveness of the proposed data transformation method is depicted in the following figure.

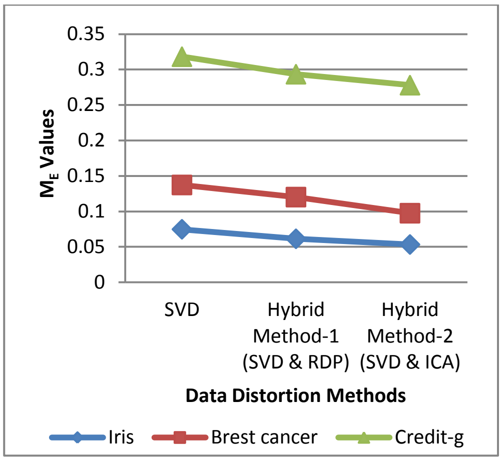

Figure 1: Comparison of Misclassification Error Values of Hybrid Methods

Figure 1 depicts the misclassification error values of the data distortion methods SVD, hybrid method-1(SVD &RDP) and hybrid method-2(SVD &ICA) for all the three data sets. The misclassification error values of the hybrid methods are lower compared to the SVD method. Lower misclassification error indicates the higher utilization of the data. Hence the hybrid data distortion methods provide higher data utilization than the single data perturbation method SVD.

The privacy provided by the data transformation method is measured as the amount of sensitive information hidden successfully. This measure is nothing but the variance between the actual and the perturbed values [15]. This measure is computed as Var (X − Y) where X represents a single original attribute and Y is the distorted attribute

$$S = Var\ (X - Y)/Var(X) \qquad (7)$$

The higher S values indicate that privacy protection of data transformation method is high. The privacy values of three data transformation methods are shown in the Table 2.

Table 2: Privacy Values of Transformed Datasets

	Iris	Brest cancer	Credit-g
SVD	0.1089	2.344	2.0187
Hybrid Method-1 (SVD&RDP)	1.014	2.404	2.7325
Hybrid Method-2 (SVD&ICA)	6.144	11.027	5.4137

From the above table, it can be easily understood that, the hybrid methods are giving good results when compared to single data perturbation method SVD. The following Figure 2 depicts the privacy values of SVD and two hybrid data transformation methods.

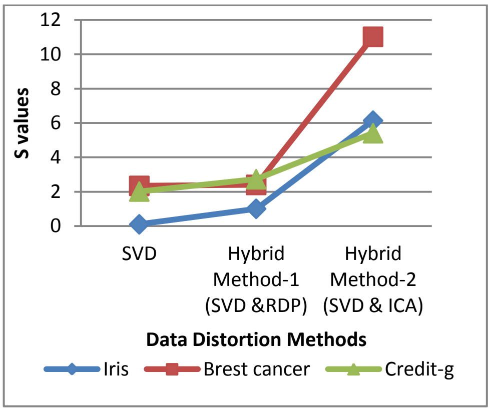

Figure 2: Comparison of Privacy Values of Hybrid Methods

The graphical representation of privacy values of the SVD, hybrid method-1(SVD & RDP) and hybrid method-2(SVD & ICA) reveals that, hybrid method-1(SVD & RDP) gives higher privacy values than the single data perturbation method SVD and hybrid method-2(SVD &ICA) yields the higher privacy values among the three methods. It can be understood that sensitive information can be hidden more securely with hybrid method-2(SVD & ICA).

5. CONCLUSION

Personal information existed in huge databases should be preserved when these databases are shared for clustering analysis. In this paper, two hybrid methods are proposed to hide the sensitive numerical attributes available in the database by taking the advantage and strength of existing techniques SVD, rotation data perturbation and independent component analysis. Information that is not important for data mining can be efficiently identified by SVD. Important information can be unveiled by independent component analysis. Rotation data perturbation can retains the statistical properties of the dataset. To utilize the strength of these methods, hybrid method-1(SVD & RDP) as a combination of SVD and rotation data perturbation, hybrid method-2(SVD & ICA) which is a combination of SVD and independent component analysis are proposed in this paper. The proposed hybrid methods are successfully implemented on three real life datasets from UCI and the experimental results proved that, the proposed methods efficiently achieve the dual demand of privacy preservation of individual and correctness of knowledge discovery than single data perturbation method SVD.

6. REFERENCES

[1] S.R.M.Oliveria, Data Transformation for Privacy-Preserving Data Mining, PhD thesis, University of Alberta, 2005.

[2] Clayton W, “Ethical, Legal and Social Implications of Genomic Medicine”, New England Journal of Medicine, vol.349, no. 6, pp.562-569, 2003.

[3] A.F.Westin, Freebies and privacy: what net users think, Technical report, Opinion Research Corporation, July 1999, Available from http://www.privacyexchange.org/iss/surveys/sr990714.html

[4] E.Bernito, I Fovino, and L.Provenza, A framework for evaluating privacy preserving data mining

algorithms Data Mining and Knowledge Discovery, vol. 29, no 2, pp. 439-450, 2000.

[5] S. Xu, J. Zhang, D.Han, J.Wang, Singular value decomposition based data distortion strategy for privacy protection Knowledge and Information Systems, vol. 10, no. 3, pp. 348-361,2007.

[6] H.Polat, W. Du, SVD-based collaborating filtering with privacy, In the 20th ACM Symposium on applied computing, Track on E-commerce Technologies, Santa Fe, New Mexico, USA. March 13-17, 2005.

[7] J.Wang, J.Zhan, and J.Zhang, Towards real-time performance of data value hiding for frequent data updates, In Proceedings of the IEEE International conference on Granular Computing. IEEE Computer Society, 2008, pp.606-611.

[8] Jie W, Zhong WXu S, and Zhang J, Selective data distortion via structural partition and SSVD for privacy preservation in proceedings of International conference on Information and knowledge Engineering .pp-:114-120, CSERA press, Las Vegas, Nevada, USA, June.

[9] N.Maheswari, K.Duraiswamy, CLUST-SVD: Privacy Preserving Clustering in Singular value Decomposition World, Journal of Modeling and Simulation, Vol.4 (2008), No.4, pp 250-256.

[10] Guang Li,Yadong Wang A Privacy- Preserving Data Mining Method based on SingularValue Decomposition and Independent Component Analysis In proceeding of Data Science Journal,Volume9,16 February 2011.

[11] M.Kalitha, D.K.Bhattacharyya, M.Dutta (2008), Privacy Preserving Clustering-A Hybrid Approach ADCOM 2008.

[12] Guang Li, Yadong Wang, “A Privacy-Preserving Classification Method Based on Singular Value Decomposition”, The International Arab Journal of Information Technology, Vol. 9, No. 6, November 2012.

[13] Hyvarinen, A, Karhunen, J, & Oja, E (2001) Independent Component Analysis, Hoboken, New Jersey, US: John Wiley & Sons Inc.

[14] Machine Learning Repository http://archive .ics.uci.edu/ml/datasets.html.

[15] Stanley R.M.Oliveria, Osmar R. Zaiane, Privacy Preserving Clustering By Data Transformation. Proceedings of the 18th Brazilian Symposium on Databases, 2003.304-318.

[16] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.