ISSN 22773061

PRIVACY PRESERVING CLUSTERING BASED ON SINGULAR VALUE DECOMPOSITION AND GEOMETRIC DATA PERTURBATION

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

[1]Research Scholar: Dept of Computer Science

S.P.M.V.V, Tirupati, Andhra Pradesh, INDIA

nagalaxmi.mada@gmail.com

[2]Professor: Dept of Computer Science, S.P.M.V.V,

Tirupati, Andhra Pradesh, INDIA

sandhyaranikasireddy@yahoo.co.in

ABSTRACT

Privacy preservation is a major concern when the application of data mining techniques to large repositories of data consists of personal, sensitive and confidential information. Singular Value Decomposition (SVD) is a matrix factorization method, which can produces perturbed data by efficiently removing unnecessary information for data mining. In this paper two hybrid methods are proposed which takes the advantage of existing techniques SVD and geometric data transformations in order to provide better privacy preservation. Reflection data perturbation and scaling data perturbation are familiar geometric data transformation methods which retains the statistical properties in the dataset. In hybrid method one, SVD and scaling data perturbation are used as a combination to obtain the distorted dataset. In hybrid method two, SVD and reflection data perturbation methods are used as a combination to obtain the distorted dataset. The experimental results demonstrated that the proposed hybrid methods are providing higher utility without breaching privacy.

Keywords: Privacy preservation, clustering, Singular value decomposition, Geometric data perturbation.

Council for Innovative Research

Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY

Vol 10, No 3

editor@cirworld.com
www.cirworld.com, member.cirworld.com

ISSN 22773061

1. INTRODUCTION

With the advancement in hardware technology large amounts of data is collected and stored by various companies and organizations. Data mining is a useful process for analyzing and extracting information from the large amounts of data. Association rule mining, classification and clustering are the some of the data mining tasks. Clustering is a tool that segments the dataset into meaningful clusters and similar objects are placed in the same clusters and dissimilar objects are placed in different clusters. A recent survey on web users reveal that, 86% of the users believe that, participation in the survey results to violate the privacy rights of individual [1]. The storing and sharing of data and the application of data mining techniques on these data are useful for decision making process on one hand, on the other hand privacy violation occurs when the extracted data mining patterns contain delicate personal information.

Privacy preserving data mining has been developed to avoid the disclosure of sensitive information, maintain confidentiality of the organizations and also preserve privacy of individuals. This paper proposes a SVD based hybrid methods for privacy preserving clustering in centralized database environment. In hybrid method one, the dataset is perturbed using SVD and Scaling data perturbation method, In hybrid method two, SVD and reflection data perturbation methods are used to protect the sensitive attribute values. These hybrid methods preserve privacy and data utility of clustering. The related works of privacy preserving clustering is discussed in the following section.

2. LITERATURE SURVEY

The authors in [2] addressed the statistical inference problem in online query processing systems and a framework for evaluating and comparing different controls. Data perturbation techniques for privacy preserving data mining have been discussed by authors in [3]. Authors in [4] [5] presented geometric data transformation methods for privacy preserving clustering in centralized database environment. Privacy preserving classification using singular value decomposition and application of SVD on structural partitions discussed in [6]. The authors in [7] proposed sparsified SVD method for data distortion and a simplified model for terrorist analysis is proposed. In [8], a SVD based data distortion method for privacy preserving clustering has been addressed by authors. A hybrid data distortion method using isometric transformations such as translation, rotation and reflection transformation methods to preserve the confidentiality of numerical attributes in centralized data has been presented by authors in [10]. A Double Reflecting data perturbation and rotation data perturbation based hybrid data transformation to preserve the privacy of confidential numerical attributes is introduced in [11]. The proposed hybrid methods are explained in the following section.

3. SVD BASED DATA DISTORTION

Various single data perturbation techniques are existing for preserving the privacy of individuals. To enhance the privacy provided by the single data perturbation methods such as SVD, scaling data perturbation, reflection data perturbation, two hybrid data perturbation methods are proposed in this paper.

3.1 Singular Value Decomposition

Singular Value Decomposition (SVD) is a matrix factorization method [13] which is used to reduce the dimensionality of the datasets and can be used as a data distortion method. Let A be a matrix of dimension n × m representing the original dataset. The rows of the matrix correspond to data objects and the columns to attributes. The singular value decomposition is a more general method that factors any n x m matrix A of rank r into a product of three matrices, such that

$$A = UWV^T$$

Where U is an n × n orthonormal matrix, W is an n x m diagonal matrix whose nonnegative diagonal entries (the singular values) are in descending order, and V^T is an m x 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 captured in the first dimension and much 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

$$A_k = U_k W_k V_k^T$$

Where 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 to obtain distorted database are given in Table 1.

Table 1: Algorithm 1: SVD based Data Transformation.

Input : Dataset D containing m rows and n columns.
Output: Distorted Dataset D′ containing m rows and n columns.
Begin

Step 1: Suppress all identifier attributes from the given matrix $D_{mxn.}$

Step 2: Apply SVD on the matrix D to obtain decomposed matrices U, W, V^T.

Step 3: Compute the distorted matrix $D' = UWV^T$

Step 4: Release the distorted dataset D′ for clustering analysis.

End

3.2 Geometric Data Transformation Methods

A geometric data transformation method of dimension d is an ordered pair, defined as GDTM = (V, f) where: a) $V \subseteq R^d$ is a representative vector subspace of data points to be transformed: b) f is a geometrical transformation function, f: $R^d \rightarrow R^d$. For geometric data transformation methods, the inputs are the vectors V, composed of confidential numerical attributes and the uniform noise vector N, while the output is the transformed vector space V′. The geometric data transformation methods are translation data perturbation, scaling data perturbation, rotation data perturbation, and reflection data perturbation. Among these scaling data perturbation and reflection data perturbation are adopted in the proposed hybrid methods to transform the original dataset to protect the privacy of individuals and maintaining the similarity between the data objects.

3.2.1 Scaling Data Perturbation (SDP): In scaling data perturbation method, the noise term is applied to each confidential numerical attribute. A positive or negative constant is multiplied to all values of a selected attribute. The data is transformed with scaling data perturbation to obtain the distorted dataset and this is shared for clustering analysis.

3.2.2 ReFlection Data Perturbation (RFDP): In reflection data perturbation method, the noise term is nothing but a rotation angle which is applied to the confidential numeric attributes. k pairs of attributes from data matrix D are selected. 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.

Single data distortion methods are not providing good privacy protection. In many cases original data can be extracted from the perturbed data. The metric properties remain unaltered after the transformations are called isometric transformations. The SVD data transformation retains the general trends in the data and also protecting privacy. In order to provide better privacy preservation, two hybrid methods are proposed in this paper which takes the advantages of existing techniques SVD and geometric data perturbation. In hybrid method one, SVD and scaling data perturbation are used as a combination to obtain the distorted dataset. In hybrid method two, SVD and reflection data perturbation are used as a combination to obtain the distorted dataset. The following section presents the hybrid method-1(SVD & SDP).

3.3 Hybrid Method-1 Based On SVD & SDP

The main aim of developing hybrid technique is to efficiently hide sensitive data from outside world and simultaneously extracts the useful patterns in the dataset. In case one a hybrid method is proposed by combining two existing techniques singular value decomposition and scaling data perturbation. Table 2 shows the algorithm for proposed hybrid method-1 based on SVD and SDP.

Table 2: Algorithm 2: Algorithm for Hybrid Method-1 (SVD & SDP)

Input : Dataset D containing m rows and n columns.

Output: Distorted Dataset D′ containing m rows and n columns.

Begin

Step 1: Suppress all identifier attributes from the given matrix $D_{mxn.}$

Step 2: Apply SVD on the matrix D to obtain decomposed matrices U, W, V^T.

Step 3: Compute the distorted matrix $D' = UWV^T$

Step 4: For each confidential numerical attribute A_j in D′, where $1 \le j \le n$ do

1. Select the noise term e_j for the attribute A_j
2. For each a_{ij} an instance of A_j where $1 \le i \le m$ do

$a_{ij} \leftarrow a_{ij} * e_j$

End For

End For

Step 5: Release the distorted dataset D″ for clustering analysis.

End

This data perturbation process of algorithm 2 consists of two steps. In step one, using SVD data perturbation the given input dataset is transformed and which is used as input to the scaling data perturbation to obtain the final distorted in step two achieve higher privacy preservation. The final distorted dataset is released for clustering analysis. The algorithm for case two which is based on SVD and reflection data perturbation methods is discussed in the following section.

3.4 Hybrid Method-2 Based On SVD & RFDP

To provide higher privacy protection hybrid data perturbation method is proposed as a combination of singular value decomposition and reflection data perturbation methods. Table 3 shows the algorithm for the proposed hybrid method-2.

The given input dataset is perturbed in two steps. In step one, the given dataset is transformed using SVD data perturbation, which is used as input to the reflection data perturbation to obtain the final distorted dataset in step two.

Table 3: Algorithm 3: Algorithm for Hybrid Method-2 (SVD & RFDP).

Input : Dataset D containing m rows and n columns.

Output: Distorted Dataset D′ containing m rows and n columns.

Begin

Step 1: Suppress all identifier attributes from the given matrix $D_{mxn.}$

Step 2: Apply SVD on the matrix D to obtain decomposed matrices U, W, V^T.

Step 3: Compute the distorted matrix $D' = UWV^T$

Step 4: Calculate k = n/2 if n is even else k = (n+1)/2;

Step 5: For each k pairs of attributes in D′

Step 6: For each pair of attributes A_i, A_j from step 5 where $1 \le i \le n$ and $1 \le j \le n$

Compute $D''(A_i', A_j') = R_o(\theta) \times D'(A_i, A_j)$ for different values of θ and identify the range that gives higher privacy vales

Select an angle θ from the selected range that gives highest privacy preservation to compute the noise term $R_o(\theta)$

Using this $R_o(\theta)$ Compute $D''(A_i', A_j') = R_o(\theta) \times D'(A_i, A_j)$

End For

Step 7: Release the distorted dataset D″ for clustering analysis.

End

Experimental results of the proposed methods are discussed in the next section.

4. IMPLEMENTATION OF PROPOSED METHODS

The proposed methods are validated empirically by conducting experiments on three real life datasets obtained from UCI [9]. Haber man dataset with 3 attributes and 306 records, Credit-g dataset with 5 numerical attributes and 1000 records, Abalone dataset with 5 numerical attributes and 4177 instances are considered in this paper. The performance of the data distortion method is measured based on two factors I) Utility measures and II) Privacy measures. The dataset is highly utilized when the data distortion technique is giving high clustering accuracy after the data distortion. 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. After transforming the data, clusters in the original dataset should be equal to those ones in the distorted dataset. WEKA (Waikato Environment for Knowledge Analysis) software is used to test clustering accuracy of the original and modified data base. The misclassification error, denoted by M_E, is measured as follows:

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

In the above formula

N - Number of points in the original dataset.

K - Number of clusters.

$\text{Cluster}_i(D)$ - Number data points of the i^{th} cluster of the original data set.

$\text{Cluster}_i(D')$ - Number of data points of the i^{th} cluster of the transformed dataset.

The cluster quality of the distorted dataset is measured by calculating misclassification error ($M_{E)}$ values. Higher M_E values indicates lower clustering quality where as lower M_E values indicate higher clustering quality. K-means clustering algorithm is used to generate the clusters for the three original as well as distorted datasets. Each experiment is repeated 10 times

because k-means algorithm is not deterministic. M_E value is calculated on an average of 10 values. Table 4 shows the M_E values of all three datasets for SVD method, scaling data perturbation method, reflection data perturbation method and hybrid method-1 (SVD &SDP), hybrid method-2 (SVD & RFDP).

Table 4: Misclassification Error Values

	Haber man	Credit-g	Abalone
SVD	0.1229	0.1808	0.08099
SDP	0	0	0
RFDP	0.131	0.187	0.09256
Hybrid Method -1 (SVD & SDP)	0.1229	0.1808	0.08099
Hybrid Method -2 (SVD & RFDP)	0.08821	0.1	0.06271

When the misclassification error values in the Table 4 are compared, it clearly indicates that the proposed hybrid method-2 (SVD&RFDP), yields lower misclassification error rates for all the three datasets and SDP method gives the lowest misclassification error values. Even though SDP gives lower misclassification error values, many researchers pointed out that this multiplicative noise added by SDP method can be easily filtered out using logarithmic transformation and other attack methods. Hence an intruder can get back the original dataset. Among these methods SVD gives lower misclassification error compared to ReFlection Data Perturbation (RFDP) and also protects the privacy of individuals. So SVD is selected and included in hybrid methods. The following graphical representation of figure 2 depicts the effectiveness of the proposed hybrid methods.

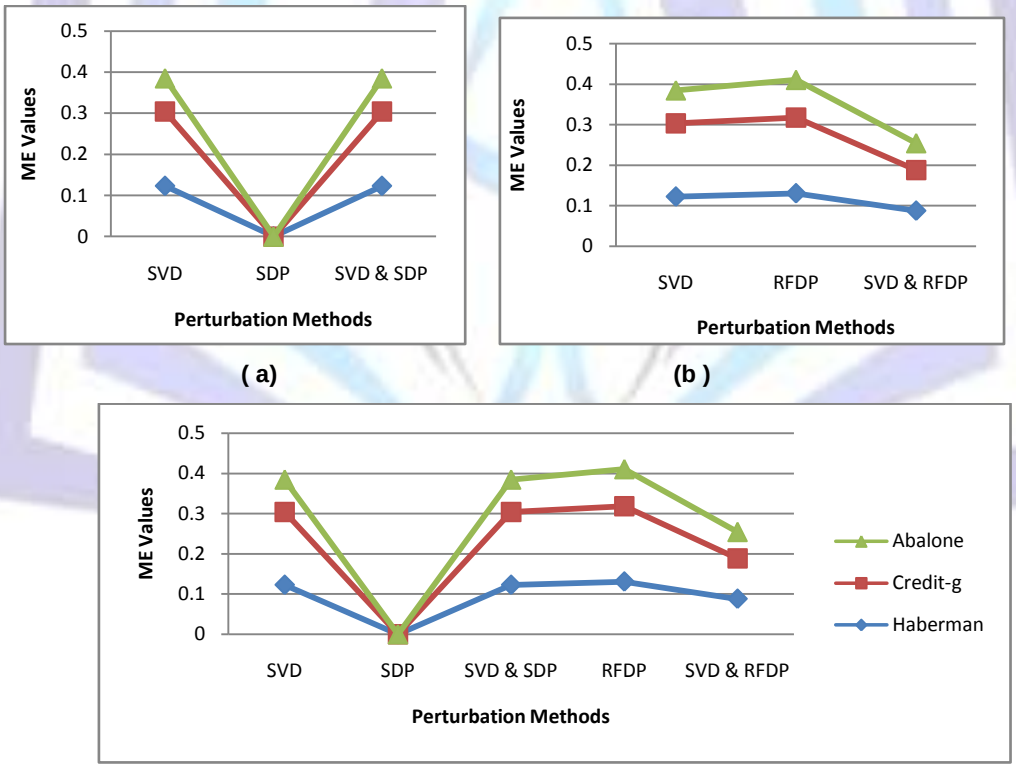


Figure 2: Misclassification Error Values

Figure 2(a) illustrates the misclassification error values related to hybrid method-1 (SVD & SDP). Among the three methods SVD, SDP, hybrid method-1(SVD&SDP), the geometric data perturbation method SDP gives the lower misclassification error.

The misclassification error values depicted in Figure 2(b) are related to hybrid method-2 (SVD & RFDP) are compared, SVD method gives lower misclassification error when compared to RFDP and hybrid method-2(SVD &RFDP) gives the lowest misclassification error among the three methods.

The misclassification error values of single data perturbation and hybrid methods are illustrated in Figure 2(c). When hybrid methods are compared, hybrid method-2(SVD &RFDP) gives the lower misclassification error values and SDP method gives the 0 misclassification error values which are lowest among all the methods.

The privacy of the perturbation technique is measured as the variance between the actual and the perturbed values [4]. This measure is given by var (X − Y) where X represents a single original attribute and Y is the distorted attribute

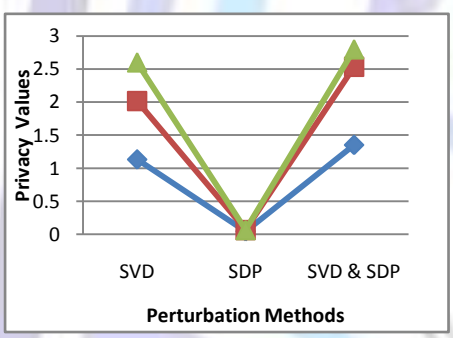
$$S = \text{var}(X - Y) / \text{var}(X).$$

The higher S values indicate that privacy protection is high. S values are computed for the three data perturbation methods SVD, scaling data perturbation, reflection data perturbation for the three datasets and hybrid methods 1 & 2 are shown in table 5.

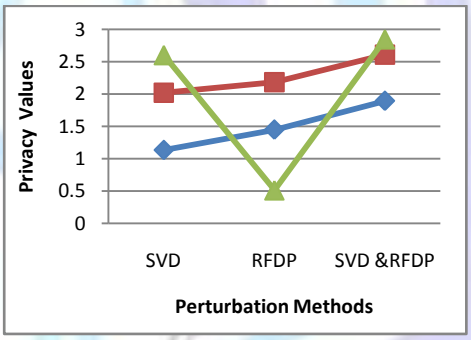
Table 5: Privacy Values of SVD and Geometric Data Perturbation.

	Haber man	Credit-g	Abalone
SVD	1.1331	2.0187	2.6028
SDP	0.0433	0.066	0.066
RFDP	1.4485	2.1808	0.5138
Hybrid Method -1 (SVD & SDP)	1.3529	2.5356	2.8
Hybrid Method -2 (SVD & RFDP)	1.8935	2.6105	2.8463

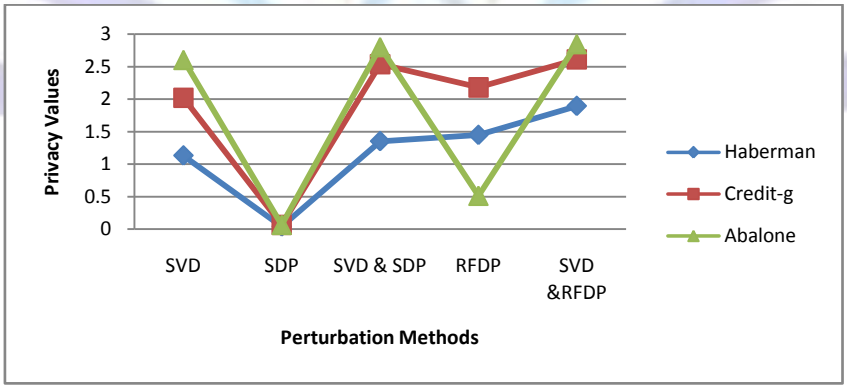
The privacy measures assess the privacy protection of data distortion methods. The graphical representation of privacy values for all the individual methods and hybrid methods is given in Figure 3. When privacy values of the perturbation methods are compared, it clearly indicates that the hybrid methods are providing higher security and can preserve the privacy.



(a)



(b)



(c)

Figure 3: Privacy Values

The privacy values depicted in Figure 3(a) clearly reveals that, the proposed hybrid method-1 (SVD & SDP), yields higher privacy values. Hence hybrid method-1(SVD & SDP) gives the higher privacy preservation than the single data perturbation methods SVD and SDP.

The privacy values illustrated in Figure 3(b) shows that, the proposed hybrid method-2(SVD & RFDP), gives higher privacy values than single data perturbation methods SVD and RFDP. Hence the hybrid method-2(SVD & RFDP) ensures the privacy preservation.

The privacy values of single data perturbation and hybrid methods 1 & 2 are illustrated in Figure 3(c). When all methods are compared, hybrid method-2(SVD & RFDP) gives the higher privacy values for all the three datasets. These results confirmed that hybrid data perturbation methods are providing higher privacy preservation.

5. CONCLUSION

Privacy preservation is essential for organizations when data of individuals is used for various purposes that contain sensitive information. It is a challenge to use the data by protecting privacy and retaining the important information for data analysis. Geometric data transformations provide better security without degrading clustering quality. In this paper hybrid methods are proposed for privacy preserving clustering as a combination of SVD and scaling data perturbation, SVD and reflection data perturbation. The proposed hybrid methods are implemented on three real life datasets from UCI for clustering analysis. The experimental results proved that, among the five data distortion methods compared hybrid method-2(SVD&RFDP) gives higher privacy preservation and retaining the important information. Hybrid method-1(SVD&SDP) also provides good privacy preservation when compared to the single data perturbation methods SVD, SDP, RFDP.

6. REFERENCES

[1]. A.F.Wstin Free bies and privacy: what net users think Technical report opinion Research corporation July 1999.Availabel from http://www.privactyexchange.org/iss/surveys/sr5990714.html

[2]. D.E.Denning A security model for the statistical database problem. ACM Transactions on the database systems.

[3]. D.Agrawal and C.Agrawal.On the design and quantification of privacy preserving data mining algorithms. In proceeding of the 20th ACM SIGMOD-SIGACT-SIGART Symposium principle database systems .pages 247-255. Santha, Barbara, California, May2001.

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

[5]. Stanley R.M.Oliveria, Osmar R. Zaiane. Achieving Privacy Preservation When Sharing Data for Clustering.

[6]. 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

[7]. S.T.Xu, J.Zhang, D.Han and J.Wang., A Singular Value Decomposition Based Data Distortion Strategy for Privacy Protection" accepted for publish and in press. Knowledge and Information Systems (KAIS) journal 2006.

[8]. 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.

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

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

[11]. Liming Li, Qishan Zhang A Privacy preserving Clustering Technique Using Hybrid Data Transformation Method, in proceedings of 2009IEEE international conference of grey systems and intelligent services.

[12]. H. T. Croft, K. J. Falconer, and R. K. Guy. Unsolved Problems in Geometry: v.2 New York: Springer-Verlag, 1991.

[13].R. Agrawal, A. Evfimievski, R. Srikant. Information sharing across private databases in Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, pp.86-97, San Diego, CA, 2003.