Hybrid SVD BASED DATA TRANSFORMATION METHODS FOR PRIVACY PRESERVING

Syed Md. Tarique Ahmad, Shameemul Haque, Prince Shoeb Khan

Abstract— At the present time privacy issues are main 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 a hybrid data transformation method is proposed for privacy preserving clustering in centralized database environment. The proposed hybrid method takes the advantage of two existing techniques such as Singular Value Decomposition (SVD) and shearing based data perturbation. Experimental results demonstrate that the proposed method efficiently protects the private data of individuals and retains the important information for clustering analysis.

Index Terms— Singular value decomposition (SVD), Shearing data perturbation, Privacy preserving clustering.

1 INTRODUCTION

DATA collection has increased rapidly with the evolution of internet and information technology. In order to enhance their business, organizations are using data mining for extracting new patterns and relationships. The process of sifting through huge databases, for extracting useful, hidden patterns is called as data mining. The techniques of data mining are widely used in business and scientific communities such as medical, healthcare, insurance, banking, marketing etc. Association rules, classification, clustering, regression are some of the data mining tasks. Cluster analysis partitions data into several categories or useful groups (clusters) based on the similarity in the data. It is an unsupervised learning method, which is used for the exploration of inter relationships among a collection of patterns, by organizing them into homogeneous clusters. Relative distance or relative density between the objects is taken as the similarity measure for the clustering objects. Clustering is performed based on the principle of maximizing the intracluster similarity and minimizing the intercluster similarity.

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]. Data mining is providing numerous benefits, there is a negative impact with data mining is the risk of privacy invasion. This problem is addressed by a new branch of data mining which takes the privacy issues under consideration is known as privacy preserving data mining. The goals of privacy preserving data mining are

A. Protection of privacy in data release
B. Privacy is protected among multiple collaborating parties
C. Protecting the sensitive knowledge patterns extracted with data mining tools.

Privacy preserving clustering methods can extract valid clustering patterns without breaching the privacy of individuals. Different approaches have been developed to effectively shield the sensitive information contained in databases such as access control, perturbation techniques, anonymity, and secure multi-party computation. In this paper a hybrid data transformation method is proposed for privacy preserving clustering, which is a combination of Singular Value Decomposition (SVD) and shearing based data perturbation.

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]. Information disclosure occurs due to the legitimate access to the data. To prevent the information disclosure, various privacy preserving data mining mechanism are used, which are different form conventional data security and accesscontrol mechanisms are discussed in [5]. Authors in [6] presented about the increasing privacy concerns of the people have been increased due to the misuse of genetic information. DNA is considered as extremely sensitive because an individual can be uniquely identified with DNA. The authors also discussed mechanisms to prevent the misuse of DNA and how to use this genomic information ap-

- *Syed Md. Tarique Ahmad is currently pursuing Ph.d in Computer Science in Pacific University Udaipur Rajasthan, India, E-mail: tariquemca@gmail.com*
- *Shameemul Haque is currently working with Dept. of Computer Science, King Khalid University, Abha,Saudi Arabia, E-mail: shameem32123@gmail.com*
- *Prince Shoeb Khan is currently working with Dept. of Computer Science, King Khalid University, Abha,Saudi Arabia, E-mail: princeshoebkhan @gmail.com*

propriately for healthcare setting. A Singular Value Decomposition (SVD) strategy for privacy preservation, some metrics to measure the difference between distorted dataset and the original dataset , the degree of privacy protection are also presented in [7]. Authors in [8] described a SVD-based randomization approach and tolerable accurate recommendations for collaborative filtering in order to protecting privacy of individual. 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 [9], by authors. In [10], a SVD based data distortion method is proposed for privacy preserving clustering. The authors in [11] described a hybrid method for privacy preserving classification which is a combination SVD and ICA. A shearing based data transformation method is introduced in [12], which will ensure that mining data will not violate privacy to certain amount of security.

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 protection, a hybrid method is proposed which is a combination of SVD and shearing based data transformation.

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 * m matrix A of rank r into a product of three matrices, such that.

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

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

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

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.

3.2 Shearing based Data Perturbation

Shearing based transformation is a linear transformation thatdisplaces each point in a fixed direction. In this method, noise term is applied to each confidential numerical attribute. So that each confidential numerical data is modified using shearing based transformation.

$$X' = X + (Sh_x \times X) \quad (3)$$

In the above formula, random noise Sh_x is generated and multiplied with X for transforming the original data.

3.3 Algorithm for Proposed Method

Hybrid data perturbation methods are providing good privacy protection when compared to the single data perturbation methods. In order to effectively extract useful patterns from huge amounts of data, a hybrid method is proposed by combining the existing techniques SVD and shearing based transformation. Table1 describes algorithm for the proposed hybrid method. 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 distorted using SVD transformation. In step two, shearing based data transformation is applied to the each record to obtain final distorted dataset. Table 1 displays the algorithm for proposed hybrid method.

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*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. For each record X′ in D′
   Randomly generate noise term Shx
   Apply Shearing based transformation to obtain final distorted dataset
   $X' = X + (Sh_x * X)$
7. End For
8. Release the distorted matrix D′ for clustering analysis.

Table 1: Algorithm for Data Transformation

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

4 Implementation Of Proposed Methods

The experimental evaluation of the proposed method is carried out by considering two measures: (1) Degree of privacy and (2) Clustering quality. Experiments are conducted on three real life datasets from UCI [14].They are Credit-g dataset with 5 numerical attributes and 1000 instances, Wabalone dataset with 5 attributes and 4177 instances, Sare dataset with 3 attributes and 306 instances. As the first experiment, Singular value decomposition (SVD) is applied on three datasets to get the misclassification error and in the second experiment shearing based transformation is applied. In the third experiment hybrid method is applied on three datasets to get privacy values.

4.1 Privacy Degree

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 [15]. This measure is given by var (X − Y). Where X represents a single original attribute and Y is the distorted attribute.

$$S = Var (X-Y) / Var (X) \quad (3)$$

The higher S values indicate that privacy protection is high. The privacy protected by the data transformation method is shown in the table 2.

Data Distortion Methods	Credit-g	Wabalone	Sare
SVD	2.0187	2.6018	1.1331
SHEAR	9	9	9
Hybrid Method (SVD & SHEAR)	15.919	19.4113	13.5327

Table 2: Privacy values of the transformed Datasets.

The privacy values of the perturbation methods SVD, shearing based data perturbation and proposed hybrid method are shown in Table 2 and it clearly shows that the proposed hybrid 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 the following Figure.

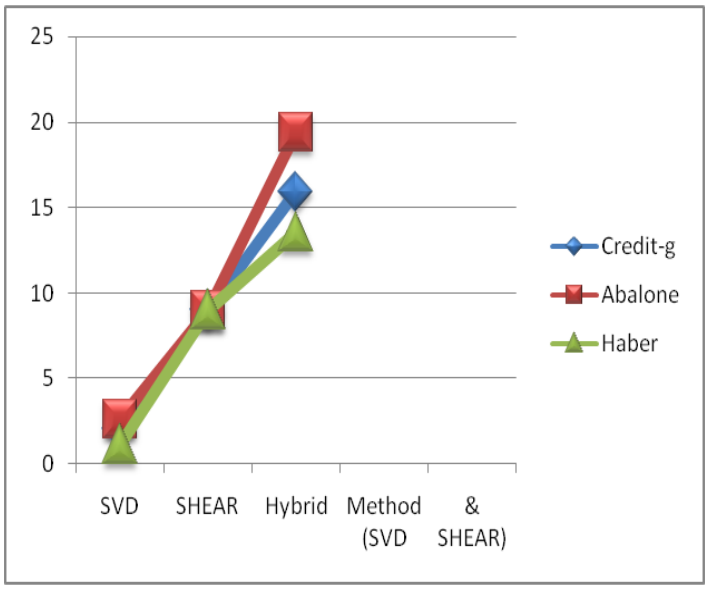

Fig. 1: Privacy Values of the Data Transformation Methods

From the results shown in Figure 1, it is proved that the proposed hybrid method gives higher privacy values for all the three datasets. Hence the privacy preservation of the proposed hybrid method is better than the single data perturbation methods SVD and shearing based data perturbation.

4.2 Quality of Clustering

The perturbed dataset is compared with the original dataset to measure the clustering quality of the proposed hybrid method. After transforming the data, clusters in the original dataset should be equal to those ones in the distorted dataset.

K-means clustering algorithm is applied to the original dataset as well as the transformed dataset. Waikato Environment for Knowledge Analysis (WEKA) software is used to test clustering accuracy of the original and modified data set. The misclassification error, denoted by M_E, is measured as follows:

Where

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

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 experimental are conducted 10 times and M_E value is taken as an average of 10 experiments. The computed M_E values for SVD, shearing based data perturbation methods for all the three datasets are shown in the following table.

Data Distortion Methods	Credit-g	Wabalone	Sare
SVD	0.1808	0.08099	0.1229
SHEAR	0	0	0
Hybrid Method (SVD & SHEAR)	0.1808	0.08099	0.1229

Table 3: M_E Values of the transformed Datasets.

From above table, it is illustrated that the ME values for SVD, hybr d method are similar and for shearing based data transformation is 0. The clustering quality provided by SVD and hybrid method is similar and shearing based data transformation is providing higher clustering quality. But when comparing the Table 2 and Table 3, it clearly reveals that the proposed hybrid method is providing higher privacy values than the single data perturbation methods SVD and shearing based data transformation. The pictorial representation of the ME values are shown in the following figure.

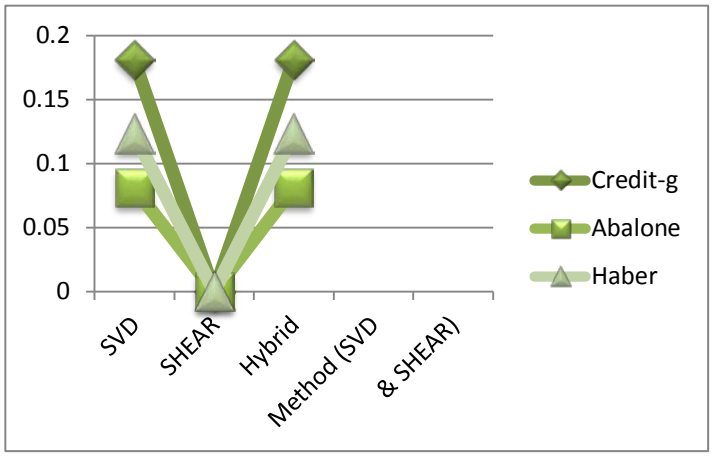

Fig. 2: M_E of the data transformation methods

The misclassification error values of single data perturba-

tion methods SVD, shearing based data transformation and hybrid methods are depicted in figure 2, reveals that the misclassification error values of SVD, hybrid methods are similar and shearing based data transformation is 0. Even though shearing based data transformation gives lower misclassification error values, many researchers pointed out that this multiplicative noise added by shearing based data transformation method can be easily filtered out using logarithmic transformation and other attack methods. Hence an intruder can get back the original dataset. Though the ME values which indicates the clustering quality for the SVD, hybrid method are similar and for shearing based data transformation is 0, the proposed hybrid method is giving higher privacy values than single data perturbation methods SVD and shearing based data transformation.

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. In this paper a hybrid method is proposed which is a combination of SVD and shearing based data transformation. The proposed hybrid method is implemented on three real life datasets from UCI for clustering analysis. The experimental results proved that, the proposed hybrid method gives higher privacy preservation and retaining the important information when compared to the single data perturbation methods SVD and shearing based transformation.

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] Liu Yu, Dap eng L, et al, "Survey of research on anonymization technology in data publication", Computer Application, pp. 2361-2364.

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

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

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

[9] 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.

[10] 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.

[11] 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, and 16 February 2011.

[12] Manikandan.G, Sudhan.R, Vaishnavi.B, "Privacy Preserving Clustering By Shearing Based Data Transformation". In Proceedings of International Conference on Computing and Control Engineering (ICCCE 2012), 12 & 13 April, 2012.

[13] 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.

[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.