Protecting the Sensitivity using Data Distortion

N. Maheswari
Associate Professor,
School of Computing Science & Engineering
VIT University,Chennai, India

M.Revathi
Assistant Professor,
Department of Computer Science and Engineering
Hindusthan College of Engineering and Technology
Coimbatore, India

Abstract— **Protecting the sensitive data publishing has attracted considerable research interest in recent years. Data privacy is the most acclaimed problem when publishing individual data. It ensures individual data publishing without disclosing sensitive data. It brings out a new branch of data mining, known as Privacy Preserving Data Mining (PPDM). Privacy-Preserving is a major concern in the application of data mining techniques to datasets containing personal, sensitive, or confidential information. Data distortion is a critical component to preserve privacy in security-related data mining applications; we propose a QR Decomposition method for data distortion. We focus primarily on privacy preserving data clustering. As the distorted data occupies small amount of storage space, the memory requirement becomes low. Finally, we evaluate the effectiveness of the method in terms of misclassification error rate. The error rate is 1.7% at the maximum. Our experiments on several data sets reveal the efficiency and effectiveness of the method and the classification error rate varies as a result of security. However, the method has much less computational cost, especially when new data items are inserted dynamically.**

Keywords*- privacy preserving; QR Decomposition; k-means clustering; data distortion; data mining*

I. INTRODUCTION *(HEADING 1)*

Data mining technologies have enabled organizations to extract useful knowledge from data in order to better understand and serve their customers and, thus, gain competitive advantages. In the information age, people are becoming more aware of the need to protect their private electronic data from falling into the wrong hands. Privacy is becoming an increasingly important issue in many data-mining applications that deal with health care, security, financial, behavioral, and other types of sensitive data. It is particularly becoming important in counterterrorism and homeland defense-related applications. These applications may require creating profiles, constructing social network models, and detecting terrorist communications among others from privacy sensitive data. In the cases of inter-corporation and security data mining applications, data mining algorithms may be applied to datasets containing sensitive or private information.

To address these concerns, researchers in the data mining community have proposed various solutions. Examples of these studies include a method for building a decision tree classifier from data where confidential values have been perturbed [2] a framework for mining association rules from data that have been randomized [5] and algorithms for hiding sensitive rules [18] among others. This stream of research often approaches the problem from a data miner's standpoint, focusing on how to develop algorithms for mining the data that are perturbed due to privacy concerns. Our study, however, approaches the issue from the standpoint of an organization that owns data. While some recent work deals with distributed databases owned by multiple parties, we consider situations where all data are owned by a single organization. We focus on how to protect individual privacy when the organization releases the data to a third party for performing data mining.

Privacy-preserving data mining techniques have been developed to address these concerns. The general goal of the privacy-preserving data mining techniques is defined as to hide sensitive individual data values from the outside world or from unauthorized persons, and simultaneously preserve the underlying data patterns and semantics so that a valid and efficient decision model based on the distorted data can be constructed. In the best scenarios, this new decision model should be equivalent to or even better than the model using the original data from the viewpoint of decision accuracy. There are currently at least two broad classes of approaches to achieving this goal. The first class of approaches attempts to distort the original data values so that the data miners (analysts) have no means (or greatly reduced ability) to derive the original values of the data. The second is to modify the data mining algorithms so that they allow data mining operations on distributed datasets without knowing the exact values of the data or without direct accessing the original datasets. This article only discusses the first class of approaches.

The rest of the paper is organized as follows. Section 2 gives the view of the previous works. Section 3 presents the proposed method for data distortion and clustering. Section 4 shows the experimental results of the performance of the algorithm. Concluding remarks and future work are described in Section 5.

II. RELATED WORK

Matrix decomposition is a key component in many data mining and computer vision tasks. Data perturbation approaches can be grouped into two main categories: the probability distribution approach and the value distortion approach. The probability distribution approach replaces the data with another sample from the same (or estimated) distribution [12],[13] or by the distribution itself, and the value distortion approach perturbs data elements or attributes directly by either additive noise, multiplicative noise, or some other randomization procedures. In this paper, we mainly focus on the value distortion approach.

The input to a data mining algorithm in many cases can be represented by a vector-space model, where a collection of records or objects is encoded as an nm× object-attribute matrix [6]. For example, the set of vocabulary (words or terms) in a dictionary can be the items forming the rows of the matrix, and the occurrence frequencies of all terms in a document are listed in a column of the matrix. A collection of documents thus forms a term-document matrix commonly used in information retrieval. In the context of privacy-preserving data mining, each column of the data matrix can contain the attributes of a person, such as the person’s name, income, social security number, address, telephone number, medical records, etc. Datasets of interest often lead to a very high dimensional matrix representation [1]. It is observable that many real-world datasets have nonnegative values for attributes. In fact, many of the existing data distortion methods inevitably fall into the context of matrix computation. For instance, having the longest history in privacy protection area and by adding random noise to the data, additive noise method can be viewed as a random matrix and therefore its properties can be understood by studying the properties of random matrices [11],[14].

Matrix decomposition in numerical linear algebra typically serves the purpose of finding a computationally convenient means to obtain a solution to a linear system. In the context of data mining, the main purpose of matrix decomposition is to obtain some form of simplified low-rank approximation to the original dataset for understanding the structure of the data, particularly the relationship within the objects and within the attributes and how the objects relate to the attributes [9]. The study of matrix decomposition techniques in data mining, particularly in text mining, is not new, but the application of these techniques as data distortion methods in privacy-preserving data mining is a recent interest [19], [20]. A unique characteristic of the matrix decomposition techniques, a compact representation with reduced-rank while preserving dominant data patterns, stimulates researchers’ interest in utilizing them to achieve a win-win task both on high degree privacy-preserving and high level data mining accuracy.

Data distortion is one of the most important parts in many privacy-preserving data mining tasks. The desired distortion methods must preserve data privacy, and at the same time, must keep the utility of the data after the distortion [15],[16]. The classical data distortion methods are based on the random value perturbation [2]. SMC research focuses on protocol development for protecting privacy among the involved parties or computation efficiency [22]; however, centralized processing of samples and storage privacy is out of the scope of SMC.

It has been used to reduce the dimensionality of, (and remove the noise in the noisy), datasets in practice [3]. The work in [21] addresses publication of anonymized transactional data. However, only the reidentification of individual transactions is prevented (i.e., the equivalent of the k-anonymity paradigm), whereas the privacy threat of associating sensitive items with quasi-identifiers is not addressed. In [23] the authors proposed two categories of novel anonymization methods for sparse high-dimensional data. The first category is based on approximate nearest-neighbor (NN) search in high-dimensional spaces, which is efficiently performed through locality-sensitive hashing (LSH). In the second category, two data transformations such as reduction to a band matrix and Gray encoding-based sorting that capture the correlation in the underlying image data have proposed.

The work presented here differs from the related work in some aspects, as follows: First, the aim is to address the problem of privacy preservation in clustering analysis. To the best knowledge, this problem has not been considered so far with QR data distortion along with clustering techniques. Second, the impact of the proposed method in the original database by quantifying how much information is preserved after transforming a database has been studied. The aim is not only on protecting individual data records, but also on providing accurate data for clustering analysis.

III. Proposed Work

A. QR Decomposition Clustering

The problem of dimension reduction has recently received broad attention in areas such as databases, data mining, machine learning, and information retrieval [10]. Efficient storage and retrieval of high dimensional data is one of the central issues in database and data mining research.

The model for the proposed method consists of two parts, the data manipulation part and the data clustering part. As illustrated in Figure 1, we assume that only the data owner or authorized access (authorized users) can manipulate the original data. After the data distortion process, the original dataset is transformed into a completely different data matrix. The k-means clustering [8] data mining technique can be applied on the distorted data.

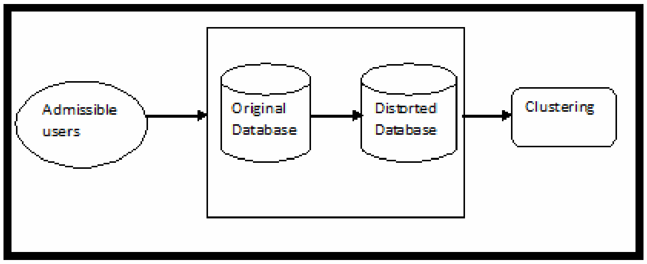


Figure 1. QR Decomposiiton Model

1) QR Decomposition

QR Decomposition (QRD) is a popular method in data mining and information retrieval. It is usually used to reduce the dimensionality of the original dataset A. QRD is useful for solving least squares’ problems and simultaneous equations. Here we use it as a data distortion method. Let A be a matrix of dimension m × n representing the original dataset. The rows of the matrix correspond to data objects and the columns to attributes.

The qr function performs the orthogonal-triangular decomposition of a matrix. This factorization is useful for both square and rectangular matrices. It expresses the matrix as the product of a real orthonormal or complex unitary matrix and an upper triangular matrix.

[Q, R] = qr (A) produces an upper triangular matrix R of the same dimension as A and a unitary matrix Q so that A = Q * R. For sparse matrices, Q is often nearly full. If [m, n] = size (A), then Q is m-by-m and R is m-by-n. Matrix R might be a sparse matrix. So that it reduces the size of the memory.

Any factorization A = QR, for which the matrices Q € CP×M and R € CM×M satisfy the following conditions, a QR decomposition of A with QR factors Q and R:

1) the nonzero columns of Q are orthonormal

2) R is upper triangular with real-valued nonnegative entries on its main diagonal

3) R = QHA

When used for privacy-preserving purpose, the distorted dataset R can provide protection for data privacy, at the same time; it keeps the utility of the original data as it can faithfully represent the original data structure.

2) Proposed Algorithm

Input : Data Matrix M, No. of clusters K

Output : Distorted Data matrix M', Clusters

Step 1 : Find the confidential numerical attributes (a i) i = 1, 2…n in M.

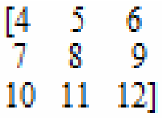
Step 2 : Form the matrix B. B = [a 1, a 2 , ………..a n]

Step 3 : Apply QR to the matrix B.

QR= QR (B)

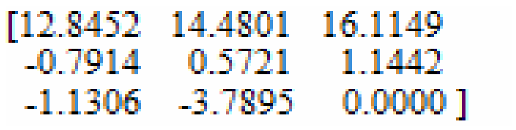
Step 4 : Update the resultant decomposed matrix in M, gives M'

Step 5 : Generate clusters for confidential numerical attributes in M'.



$$\begin{matrix}[4 & 5 & 6\\ 7 & 8 & 9\\ 10 & 11 & 12]\end{matrix}$$

Figure 2. Original Matrix



$$\begin{matrix}[12.8452 & 14.4801 & 16.1149\\ -0.7914 & 0.5721 & 1.1442\\ -1.1306 & -3.7895 & 0.0000\ ]\end{matrix}$$

Figure 3. Distorted Matrix

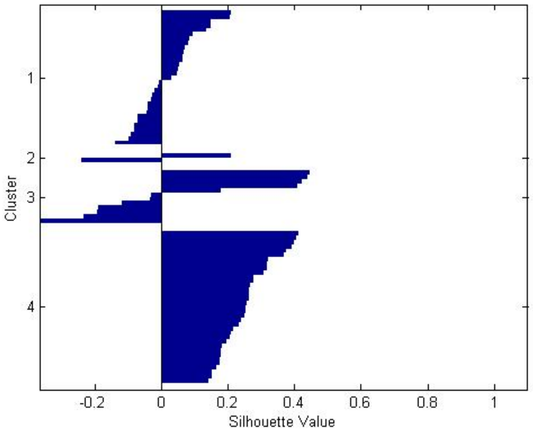


Figure 4. . Cluster representations before applying QR

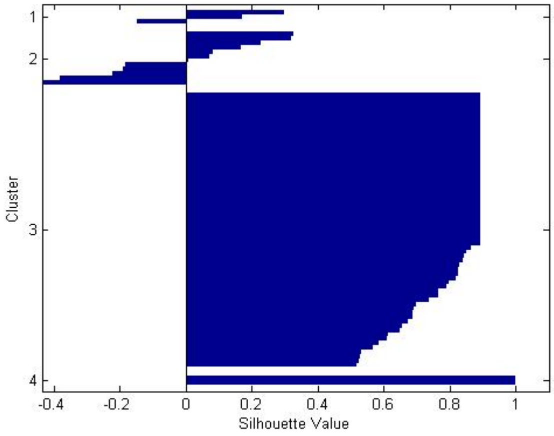


Figure 5. Cluster representations after applying QR

Figure 2: A sample data matrix; Figure 3: A distorted data matrix using QR decomposition corresponding to the original sample matrix; Figure 4: The representation of the clusters before applying QR when K=4 for the spectf data set; Figure 5: The representation of the clusters after applying QR when K=4 for the spectf data set.

To illustrate how the proposed method works, let us consider the sample matrix in Figure 2. To do so, we apply our proposed method. Figure 3 shows the distorted matrix. To get an idea of how well-separated the resulting clusters are, we can make a silhouette plot. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and ranges from -1 to +1.

Silhouette S(i) = (min(b(i,:),2) - a(i)) ./ max(a(i),min(b(i,:)))

where a(i) is the average distance from the ith point to the other points in its cluster, and b(i,k) is the average distance from the ith point to points in another cluster k.

Figure 4 and Figure 5 gives the silhouette plot for the spectf data set. The clusters before and after distortion can be seen in Figure 4 and Figure 5. From the silhouette plot of Figure 5, we can see that most points in clusters 3 and 4 have a large silhouette value, greater than 0.8, indicating that those points are well-separated from neighboring clusters. However, some clusters also contain a few points with low silhouette values, indicating that they are nearby to points from other clusters.

IV. Experimental Results

In this section, we present the results of our performance evaluation. We start by describing the methodology that we used. Then we study the effectiveness of our method under K-means clustering method.

A. Methodology

All the experiments were conducted on a PC, Pentium IV with 512 MB of RAM running a windows operating system. The MATLAB package is used to execute the proposed method We used four different synthetic datasets (uci repository) such as glass data set with 214 objects and 11 attributes, wine with 178 objects and 14 attributes, spectf with 80 objects and 45 attributes, haberman with 306 objects and 4 attributes as shown in Table 1. For the dataset, we analyzed a specific number of clusters as 3 and 4. The effectiveness is measured in terms of the proportion of the points that are grouped in the same clusters after we apply a transformation on the data. We refer to such points as legitimate ones.

TABLE I. Data objects and Clusters=3

Data Objects (points)	**Before Distortion K=3 (clusters)**				**After Distortion K=3 (clusters)**			
	Database				***Database***			
	Glass	***Wine***	***Spectf***	***Haberman***	***Glass***	***Wine***	***Spectf***	***Haberman***
	113	47	66	1	1	174	67	1
	99	62	2	304	212	2	2	304
	2	69	12	1	1	2	11	1

TABLE II. Data objects and Clusters=4

Data Objects (points)	**Before Distortion K=4 (clusters)**				**After Distortion K=4 (clusters)**			
	Database				***Database***			
	Glass	***Wine***	***Spectf***	***Haberman***	***Glass***	***Wine***	***Spectf***	***Haberman***
	54	39	31	29	1	172	3	1
	54	57	2	77	1	2	12	303
	53	23	12	114	207	2	63	1
	53	59	35	86	5	2	2	1

In Table 1 and 2 Data Objects denotes the objects. K denotes the no. of clusters to group the objects. In Table 2, before distortion the wine data set objects 39, 57, 23 and 59 are grouped in cluster 1, 2, 3 and 4. After distortion, objects 172, 2, 2 and 2 are grouped in cluster 1, 2, 3 and 4. Similarly the other 3 data sets are clustered before and after distortion.

B. Measuring Effectiveness

The effectiveness is measured in terms of the number of legitimate points grouped in the original and the distorted databases. After transforming the data, the clusters in the original databases should be equal to those ones in the distorted database. However, this is not always the case, and we have some potential problems after data transformation: a noise data point end-up clustered, a point from a cluster becomes a noise point, or a point from a cluster migrates to a different cluster. Since the K-means clustering method we used, do not consider noise points, we concentrate only on the third case.

We call this problem Misclassification Error [8] and it is measured in terms of the percentage of legitimate data points that are not well-classified in the distorted database. Ideally, the misclassification error should be 0%. The misclassification error, denoted by ME, is measured as follows:

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

where N represents the number of points in the original dataset, k is the number of clusters under analysis, and ｜ Clusteri(X) ｜represents the number of legitimate data points of the i th cluster in the database X.

TABLE III. Results of Misclassification

Data Objects	**Misclassification Ratio for clusters, k=3**	**Misclassification Ratio for clusters, k=4**
214	1.0	1.4
178	1.4	1.4
80	0.0	1.7
306	0.0	1.4

As can be seen in Table 3, the proposed technique yielded the result as 0% in two various data sets and yielded 1% and 1.4% in other two sets of data for three number of clusters. For four numbers of clusters the three datasets yields 1.4% and one data set yields 1.9% of misclassification. The increase of percentage shows the variation in the number of objects stored in the clusters before and after distortion of data sets. Thus the method yielded very good results when we compare the cluster analysis of the original and the distorted datasets for k=3 than k=4. These results suggest that our technique perform well for comprising the infeasible goal of having both complete privacy and complete accuracy for clustering analysis.

V. Conclusion and Future Work

We have presented a method to distort only confidential numerical attributes to meet privacy requirements, while preserving general features for clustering analysis. We have presented a QR Decomposition method for data distortion to achieve privacy-preserving in data mining applications. We focus primarily 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 various data distortion techniques with other data mining algorithms.

REFERENCES

[1] D. Achlioptas.”Random matrices in data analysis*”, Proceedings of the 15th European Conference on Machine Learning*, pp. 1-8, 2004, Pisa, Italy.

[2] R. Agrawal, and R. Srikant. “Privacy-preserving data mining*”, Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 439-450, 2000, Dallas, TX.

[3] M.W. Berry, Z. Drmac, and E.R,Jessup,(1999). “Matrix, vector space, and information retrieval”, *SIAM Review*, 41, 335-362..

[4] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Zhu.(2003).”Tools for privacy preserving distributed data mining*”. ACM SIGKDD Explorations*, 4(2), 1-7.

[5] A. Evfimievski, R. Srikant ,R.Agarwal and J. Gehrke, (2002).”Privacy Preserving Mining of Association Rules.” Proc. Eighth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, pp. 217-228,2002.

[6] W. Frankes, and R. Baeza-Yates, *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall,Englewood Cliffs, NJ,1992.

[7] J. Gao and J. Zhang, *Sparsification strategies in latent semantic indexing*. Proceedings of the 2003 Text Mining Workshop, pp. 93-103,2003, San Francisco, CA.

[8] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, CA,2001.

[9] L.Hubert, J. Meulman and W. Heiser, *Two purposes for matrix factorization: a historical appraisal*. SIAM Review, 42(4), 68-82, 2000.

[10] Jieping Ye, Qi Li, Hui Xiong, Haesun Park, Ravi Janardan, Vipin Kumar, *IDR/QR: An Incremental Dimension Reduction Algorithm via QR Decomposition.* IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 9,2005.

[11] H. Kargupta, K. Sivakumar and S.Ghosh, *Dependency detection in mobimine and random matrices.* Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 250-262, 2002, Helsinki, Finland.

[12] Kun Liu, Hillol Kargupta, *Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining,* IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No.1, 2006.

[13] D.D. Lee and H.S. Seung, *Learning in parts of objects by non-negative matrix factorization*. Nature, 401, 788- 791,1999.

[14] M.L. Mahta, *Random Matrices*. 2nd edition,19991. Academic, London.

[15] A. Pascual-Montano, J.M.Carazo, K. Kochi, D. Lehmann, and P.D. Pascual-Marqui, . *Nonsmooth nonnegative matrix factorization (nsNMF*). IEEE Transactions on Pattern Analysis and Machine Intelligence, 28, 403-415, 2006.

[16] V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin, and Y. Theodoridis, *State-of-the-art in privacy preserving data mining*. ACM SIGMOD Record, 3(1), 50-57,2004.

[17] J.Wang, W.J. Zhong, and J.Zhang, *NNMF-based factorization techniques for high-accuracy privacy protection on non-negative-valued datasets*. Proceedings of the IEEE Conference on Data Mining, 2006.

[18] Xiao-Bai Li, Sumit Sarkar,. *A Tree-Based Data Perturbation Approach for Privacy-Preserving Data Mining,* IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 9, 2006.

[19] S. Xu, J. Zhang, D. Han, and J. Wang, *Singular value decomposition based data distortion strategy for privacy protection.* Knowledge and Information Systems, 10(3), 383-397,2006.

[20] S. Xu, J. Zhang, D. Han, and J.Wang, *Data distortion for privacy protection in a terrorist analysis system*. Proceedings of the 2005 IEEE International Conference on Intelligence and Security Informatics, pp. 459-464, 2005, Atlanta, GA.

[21] Y. Xu, K. Wang, A.W.-C. Fu, and P.S. Yu, *Anonymizing Transaction Databases for Publication*, Proc. SIGKDD, pp. 767- 775, 2008.

[22] Y.Kim ,Shaneck, M. *Efficient Cryptographic Primitives for Private Data Mining.* The forty third Hawaii international Conference on System Sciences, HICSS, 1-9., 2010.

[23] Yufei Tao, Gabriel Ghinita, Panos Kalnis, *Anonymous Publication of Sensitive Transactional Data* IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No.2, February 2011.