A Privacy-Preserving Classification Method Based on Singular Value Decomposition

Guang Li and Yadong Wang
Department of Computer Science and Engineering, Harbin Institute of Technology, China

Abstract: *With the development of data mining technologies, privacy protection has become a challenge for data mining applications in many fields. To solve this problem, many privacy-preserving data mining methods have been proposed. One important type of such methods is based on Singular Value Decomposition (SVD). The SVD-based method provides perturbed data instead of original data, and users extract original data patterns from perturbed data. The original SVD-based method perturbs all samples to the same degree. However, in reality, different users have different requirements for privacy protection, and different samples are not equally important for data mining. Thus, it is better to perturb different samples to different degrees. This paper improves the SVD-based data perturbation method so that it can perturb different samples to different degrees. In addition, we propose a new privacy-preserving classification mining method using our improved SVD-based perturbation method and sample selection. The experimental results indicate that compared with the original SVD-based method, this new proposed method is more efficient in balancing data privacy and data utility.*

Keywords: *Privacy preservation, data mining, SVD, sample selection.*

Received June 3, 2010; accepted January 3, 2011

1. Introduction

Data mining is the process of extracting patterns from data. It has become an increasingly important tool for transforming data into information. However, with the rapid development of data mining technologies, preserving data privacy poses an increasing challenge to data mining applications in many areas [9], especially the medical, financial and homeland security fields. Many people fear their private information will be misused and believe that privacy protection is important [4, 7]. In addition to social pressures, legal mechanisms exist to protect data privacy. For example, in the United States, in order to comply with the Health Insurance Portability and Accountability Act (HIPAA), individuals and organizations cannot release medical data for public use without a privacy protection guarantee. To solve this problem, Privacy-Preserving Data Mning (PPDM) methods have been studied [1, 5, 20]. They are becoming an increasingly important topic in data mining research. PPDM technology can perform data mining without accessing the details of the original data.

In the past decade, many PPDM methods have been developed. They can be divided into two main categories. The methods in the first category are based on data perturbation [2, 14, 17, 19, 21]. In these methods, the original data are not open, and users can only access perturbed data. Data mining is done on the perturbed data to extract the original data patterns. The methods in the second category are based on Secure Multi-party Computation (SMC) [3, 6, 22]. They are often used for distributed databases. They assume that there are multiple nodes, each of which has only a part of the global data set. These nodes aim to carry out data mining on the global data set, but each node does not want the other nodes to know its data. In these methods, all of the nodes exchange the information required by the mining algorithm through information exchange protocols based on SMC. These protocols allow information to be exchanged privately, without allowing any node to obtain the original data from any other node directly. One important type of methods in the first category is based on matrix decompositions, including Singular Value Decomposition (SVD) [12, 13, 18] and Nonnegative Matrix Factorization (NMF) [11]. These matrix-based global perturbation methods can provide good data mining performance while preserving privacy [15].

In essence, these methods use matrix decompositions to analyze data and to find and retain important information for data mining. Data perturbation is achieved by removing unimportant information. Because the modified data contains important information for data mining, the original data patterns can be extracted from the modified data by data mining. Because the information that is unimportant for data mining is removed, the modified data are different from the original data, and thus privacy can be preserved.

These matrix decomposition-based methods are global perturbation methods. They perturb all of the samples to the same degree. However, because different users may request different degrees of privacy protection and different samples do not hold the same importance for data mining, it is often more

appropriate to modify different samples to different degrees.

In this paper, we present an improved SVD-based perturbation method that allows different samples to be perturbed to different degrees. In addition, we propose a new privacy-preserving classification method based on the improved SVD-based perturbation method and sample selection. The experimental results indicate that as compared with the original SVD-based method, this new proposed method is more efficient in balancing data privacy and data utility. The rest of this paper is organized as follows: section 2 introduces the PPDM method based on SVD, section 3 presents our algorithm, section 4 shows the experimental results, finally, section 5 presents the conclusion.

2. The SVD-Based PPDM Method

Let A be a matrix with dimensions $n \times m$ representing the original data. The rows of the matrix correspond to data objects, and the columns correspond to attributes. The SVD of matrix A is:

$$A = USV^T \tag{1}$$

Where U is an $n \times n$ orthonormal matrix, and S is an $n \times m$ diagonal matrix with the number of nonzero diagonal entries equal to $Rank(A)$. $Rank(A)$ is the rank of matrix A. In S, all nonzero diagonal entries are in descending order. V^T is an $m \times m$ orthonormal matrix. Let $0 \le k \le Rank(A)$. In the original SVD-based PPDM method [18], the perturbed data A_k is defined as follows:

$$A_k = U_k S_k V_k^T \tag{2}$$

Where U_k is an $n \times k$ matrix that contains the first k columns of U, S_k is a $k \times k$ diagonal matrix that contains the largest k nonzero diagonal entries of S, V_k^T is a $k \times m$ matrix that contains the first k rows of V^T. In this method, utility and privacy are regulated by k. With increasing values of k, more information from the original data is retained. Increasing the value of k improves utility at a cost to privacy. Shuting *et al.* [18] also presents a PPDM method based on sparsified SVD (SSVD). In this method, the perturbed data $\overline{A_k}$ is defined as follows:

$$\overline{A_k} = \overline{U_k} S_k \overline{V_k^T} \tag{3}$$

Where $\overline{U_k}$ and $\overline{V_k^T}$ are obtained from U_k and V_k^T by setting the small entries in them equal to zero. For example, given a threshold value e, if the abstract value of u_{ij} or v_{ij} is smaller than e, let u_{ij} or v_{ij} be zero. Note that u_{ij} and v_{ij} are the U_k and V_k^T's entries in the i-th row and j-th column, respectively. Compared to the original SVD-based perturbation method, the SSVD-based perturbation method performs better with respect to preserving privacy.

Considering that some objects and attributes are not private and do not need be protected, Jie *et al.* [12, 13] divide the original matrix into two or four submatrices; the SSVD-based method is used on only one submatrix, and no perturbations are performed on the other submatrix.

3. The New Algorithm

The currently available PPDM method based on SVD perturbs every sample to the same degree. In the SVD-based method, every sample is perturbed with the same parameter k. In the SSVD-based method, the threshold value e does not consider sample differences also. In Jie *et al.* [12, 13], all of the perturbed samples are still modified to the same degree.

However, in reality, different users have different requirements regarding privacy protection, and different samples do not hold the same level of importance for data mining. So, it is often more appropriate to perturb different samples to different degrees. Based on this idea, in this paper, we analyze the SVD-based perturbation method and improve it so that it can be used to perturb different samples to different degrees. We assume the i-th row and j-th column entry of the original sample matrix A is a_{ij}; the i-th row and j-th column entry of the original SVD-based perturbation method's perturbed sample matrix A_k is b_{ij}; after SVD, $A=USV^T$; and the i-th row and j-th column entry of U, S, and V^T are u_{ij}, s_{ij} and v_{ij}, respectively. Considering that $s_{ij}=0$ ($i \ne j$) and $s_{ii}=0$ ($i > Rank(A)$), we obtain:

$$a_{ij} = \sum_{p=1}^{Rank(A)} u_{ip} s_{pp} v_{pj} \tag{4}$$

$$b_{ij} = \sum_{p=1}^{k} u_{ip} s_{pp} v_{pj} \tag{5}$$

That means, in essence, that after SVD, every entry is turned into a sum of $Rank(A)$ items, and the original SVD-based perturbation abandons the tail $Rank(A)–k$ items for every entry. So, the perturbed data matrix A_k can also be calculated as:

$$A_k = \hat{U} S V^T \tag{6}$$

Where $\hat{U}$ is calculated by setting all u_{ij} ($j > k$) to zero. It is easy to find that the i-th row of U only affects the i-th row of A and A_k. This means the i-th row of U only influences the i-th sample. So, in this new perturbed data calculation method, we can perturb different samples to different degrees. If we want to perturb the i-th sample using the original SVD-based method with parameter k_i, we need only calculate $\hat{U}$ by setting all u_{ij} ($j > k_i$) to zero. According to the same principle, the SSVD-based method can also be improved to perturb different samples to different degrees.

In the classification problem, the samples at the category edge have more information with respect to classification and are more important than the samples in the center of the category. As such, the center samples should be perturbed to a higher degree, and the edge samples should be perturbed to a lower degree. Based on this idea, this paper proposes a privacy-preserving classification method using both our improved SVD-based perturbation method and the Weighted Condensed Nearest Neighbor (WCNN) sample selection algorithm [8]. The WCNN algorithm is an improvement on the Condensed Nearest Neighbor (CNN) algorithm, which is a classical sample selection algorithm based on the nearest neighbor rule. The WCNN algorithm tries to identify category edge samples by the nearest neighbor rule and allows users to specify the number of selected samples.

Our new privacy-preserving classification method has two steps. First, the WCNN algorithm is used to select important samples. We specify the rate r for important samples. Then, perturbation is performed using our improved SVD-based method. There are two parameter values k_1 and k_2 for the SVD-based perturbation method, with $k_1>k_2$. When perturbing the important samples, let $k=k_1$, and when perturbing the other samples, let $k=k_2$.

The WCNN algorithm has two steps. First, the CNN algorithm is applied enough times to select a sufficient number of samples. Then, some of these selected samples are deleted according to their weights. Only the preserved samples are output as the selected samples. There are two parameters in the WCNN algorithm. One is the number of samples selected in the first step, and the other is the number of samples preserved and output in the second step. In our algorithm, we only attach the rate r to important samples. If the total number of samples is N, the WCNN algorithm should output $N\times r$ samples as important samples. We calculate r' as:

$$r' = \begin{cases} r + 0.15 & r \le 0.3 \\ r + 0.1 & 0.3 < r < 0.7 \\ r + 0.05 & r \ge 0.7 \end{cases} \tag{7}$$

Then, in the first step of the WCNN algorithm, we make the number of selected samples as close to $N\times r$' as possible.

4. Experiments

4.1. Utility Measures

Data utility measures assess whether a data set maintains the performance of data mining techniques after data distortion, e.g., whether the original data patterns can be extracted from the perturbed data. In this paper, we chose to examine the accuracy of three kinds of classifiers as data utility measures: the J48 decision tree, NaiveBayes and IB1 in the Waikato Environment for Knowledge Analysis (WEKA) [10]. If the classifier trained on the perturbed data has a similar accuracy as that trained on the original data, the perturbation method can be said to maintain good data utility.

4.2. Privacy Measures

We used the privacy measures that are often used by the matrix decomposition-based PPDM methods [11, 12, 13, 18]. We assume the original data are A, and the modified data are MA.A and MA are both $n\times m$ matrices. There are five privacy measures: VD, RP, RK, CP and CK. The first measure, VD, is the ratio of the Frobenius norm of the difference of MA from A to the Frobenius norm of A. It is calculated as:

$$VD = \|A - MA\|_F \,/\, \|A\|_F \tag{8}$$

The Frobenius norm of an $n\times m$ matrix A, with i-th row and j-th column entry denoted by a_{ij}, is calculated as:

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2} \tag{9}$$

If $Rank_j^i$ and $MRank_j^i$ denote the rank in ascending order of the j-th element for the i-th attribute in A and MA, respectively, the second measure, RP, is defined as:

$$RP = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left|Rank_j^i - MRank_j^i\right|}{nm} \tag{10}$$

The third measure, RK, represents the percentage of elements that maintain their ranks in each column after distortion. RK is computed as:

$$RK = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} Rk_j^i}{nm} \tag{11}$$

Where $Rk_j^i = \begin{cases} 1 & Rank_j^i = MRank_j^i \\ 0 & otherwise \end{cases}$.

The fourth measure, CP, is used to define the change in rank of the average value of the attributes. If $RankV_i$ and $MRankV_i$ are ranks in ascending order of the average value of the i-th attribute in A and MA, respectively, CP is defined as:

$$CP = \frac{\sum_{i=1}^{m}\left|RankV_i - MRankV_i\right|}{m} \tag{12}$$

The fifth measure, CK, represents the percentage of attributes that maintain their ranks of average values after distortion. It is calculated as:

$$CK = \frac{\sum_{i=1}^{m} Ck_i}{m} \tag{13}$$

Where $Ck_i = \begin{cases} 1 & RankV_i = MRankV_i \\ 0 & otherwise \end{cases}$.

Simply put, if privacy is better protected, then VD, RP and CP will have larger values, and RK and CK will

have smaller values. For example, if the original data and the modified data are:

$$A=\begin{bmatrix}2 & 6\\ 4 & 8\end{bmatrix} \text{ and } MA=\begin{bmatrix}6 & 4\\ 8 & 2\end{bmatrix}$$

then

$$A-MA=\begin{bmatrix}-4 & 2\\ -4 & 6\end{bmatrix}$$

We can derive $\|A-MA\|_F=6\sqrt{2}$ and $\|A\|_F=2\sqrt{30}$, and so $VD \approx 0.77$. Also, $Rank_1^1=1$, $Rank_2^1=2$, $Rank_1^2=1$, $Rank_2^2=2$, $MRank_1^1=1$, $MRank_2^1=2$, $MRank_1^2=2$, and $MRank_2^2=1$, and so $RP=0.5$ and $RK=0.5$. In addition, $RankV_1=1$, $RankV_2=2$, $MRankV_1=2$, and $MRankV_2=1$, and so $CP=1$ and $CK=0$.

4.3. Databases

We used two real-life databases for our experiments. They are the Wisconsin Breast Cancer (WBC) original data set [16] and the Wisconsin Diagnostic Breast Cancer (WDBC) data set. They are both from the University of California at Irvine's Machine Learning Repository. The WBC database has 9 attributes and 699 samples. There are 16 samples in it that have missing values, and it also has some repeat samples. We only used complete samples, and we deleted the repeat samples. As a result, there are 449 samples in the final WBC database that we used. The WDBC database has 30 attributes and 569 samples. In our experiments, 20% of the samples in both databases were selected randomly as testing samples, and the other 80% of the samples were used as training samples.

4.4. Experimental Results

For every training sample set, we used three methods to perform the perturbation. The first method was our proposed privacy-preserving classification method. In the WCNN algorithm, 30% of the training samples were selected as important samples. In our improved SVD-based perturbation method, we set k_1=8 and k_2=1 for the WBC data, and we set k_1=15 and k_2=1 for the WDBC data. The second method was the original SVD-based perturbation method with $k=k_1$. This means that k=8 for the WBC data and k=15 for the WDBC data. The final method was again the original SVD-based perturbation method, but this time, we set $k=k_2$. This means that for both the WBC and WDBC data, k=1. The experiments were repeated three times and averaged to obtain our experimental results.

Table 1 shows the privacy measures for all three perturbation methods. It is easy to see that the SVD-based perturbation method with $k=k_2$ had the best performance for privacy preservation of these three perturbation methods. Our method performed better in terms of preserving privacy than the SVD-based perturbation method with $k=k_1$.

Table 1. The privacy measures of all three perturbation methods.

Data	Perturbation Method	VD	RP	RK	CP	CK
WBC	SVD (k = 8)	0.07	62.58	0.02	0.22	0.78
	SVD (k = 1)	0.38	90.77	0.01	1.11	0.44
	Our Method	0.31	77.20	0.01	1.04	0.44
WDBC	SVD (k = 15)	0	27.83	0.38	0	1
	SVD (k = 1)	0.09	98.02	0.01	0.36	0.69
	Our Method	0.07	90.95	0.01	0.2	0.8

All three sets of perturbed training samples and the original training samples were used to train the J48 decision tree, NaiveBayes and IB1 classifiers by using WEKA. The accuracies of these classifiers for the testing samples were used as data utility measures. Figures 1 and 2 show the data utility measures for the WBC and WDBC data sets, respectively. It can be determined that for both the WBC and WDBC data sets, our method and the original SVD-based method with $k=k_1$ maintain good data utility. For these two perturbation methods, all three classifiers trained on the perturbed data have a similar accuracy to that trained on the original data. We also determined that the original SVD-based method with $k=k_2$ cannot maintain good data utility. For the WDBC data set, all three classification algorithms showed very low-accuracy classifiers for the perturbed data. For the WBC data set, all three kinds of classifiers trained on the perturbed data had lower accuracies than those trained on the original data, and this phenomenon is especially obvious for the J48 decision tree.

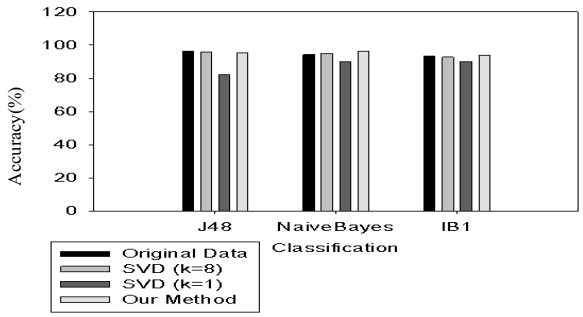

Figure 1. The accuracies of classifiers trained on the perturbed data and the original data for the WBC database.

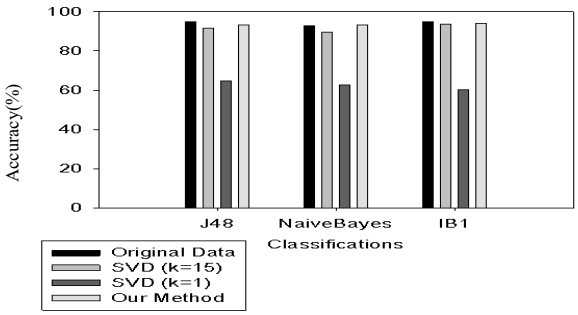

Figure 2. The accuracies of classifiers trained on the perturbed data and the original data for the WDBC database.

In conclusion, compared with the original SVD-based method, our new PPDM method is more efficient in balancing data privacy and data utility. This is because in our new method, different samples are modified to different degrees. The important samples for data mining are perturbed to a lower degree, and the unimportant samples are perturbed to a higher degree. This approach allows us to identify and then remove more information that is unimportant for data mining than the original SVD-based method.

5. Conclusions

Currently, the SVD-based perturbation method for PPDM perturbs every sample to the same degree. However, in reality, different users have different requirements for privacy protection, and different samples do not hold the same importance with respect to data mining. So, it is better to perturb different samples to different degrees. This paper improves existing SVD-based data perturbation methods by perturbing different samples to different degrees. In addition, we propose a new privacy-preserving classification method based on our improved SVD-based data perturbation method and the WCNN sample selection algorithm. The basic idea of this new privacy-preserving classification method is to divide the samples into important samples and unimportant samples and then perturb the important samples to a lower degree and perturb the unimportant samples to a higher degree. This method uses the WCNN algorithm to select important samples and then uses our improved SVD-based method to perform data perturbation. The experimental results indicate that compared with the original SVD-based method, our new method is more efficient in balancing data privacy and data utility.

Because of the WCNN sample selection algorithm, this proposed PPDM method can only be used for classification problems; but data mining is not confined to classification. In future research, we plan to study PPDM methods based on the same idea in the context of other types of data mining problems.

Acknowledgements

We thank Dr. William H. Wolberg of the University of Wisconsin Hospital at Madison for providing us with the data for experiments.

References

[1] Ashraf E., "Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous Database," *The International Arab Journal of Information Technology*, vol. 7, no. 2, pp. 152-160, 2010.

[2] Benjamin F., Ke W., and Philip Y., "Anonymizing Classification Data for Privacy Preservation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 5, pp. 711-725, 2007.

[3] Benny P., "Cryptographic Techniques for Privacy-Preserving Data Mining," *ACM SIGKDD Explorations Newsletter*, vol. 4, no. 2, pp.12-19, 2002.

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

[5] Elisa B., Igor F., and Loredana P., "A framework for Evaluating Privacy Preserving Data Mining Algorithms," *Data Mining and Knowledge Discovery*, vol. 11, no. 2, pp. 121-154, 2005.

[6] Fatih E., Agrawal D., and El-Abbadi A., "Privacy Preserving Decision Tree Learning Over Multiple Parties," *Data & Knowledge Engineering*, vol. 63, no. 2, pp. 348-361, 2007.

[7] Hall A. and Rich S., "Patients' Fear of Genetic Discrimination by Health Insurers: The Impact of Legal Protections," *Genetics in Medicine*, vol. 2, no. 4, pp. 214-221, 2000.

[8] Hao H. and Jiang R., "Training Sample Selection Method for Neural Networks Based on Nearest Neighbor Rule," *Acta Automatica Sinica*, vol. 33, no. 12, pp. 1247-1251, 2007.

[9] Herman T., "Information Privacy, Data Mining, and the Internet," *Ethics and Information Technology*, vol. 1, no. 2, pp. 137-145, 1999.

[10] Ian W. and Eibe F., *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Massachusetts, 2005.

[11] Jie W., Weijun Z., and Jun Z., "NNMF-Based Factorization Techniques for High-Accuracy Privacy Protection on Non-Negative-Valued Datasets," *in Proceedings of 6th IEEE International Conference on Data Mining-Workshops*, China, pp. 513-517, 2006.

[12] Jie W., Zhong W., Xu 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*, USA, pp. 114-120, 2006.

[13] Jie W., Zhong W., Xu S., and Zhang J., "A Novel Data Distortion Approach via Selective SSVD for Privacy Protection," *International Journal of Information and Computer Security*, vol. 2, no. 1, pp. 48-70, 2008.

[14] Li L., Murat K., and Bhavani T., "The Applicability of the Perturbation Based Privacy Preserving Data Mining for Real-World Data," *Data and Knowledge Engineering*, vol. 65, no. 1, pp. 5-21, 2008.

[15] Lian L., Jie W., and Jun Z., "Wavelet-Based Data Perturbation for Simultaneous Privacy-Preserving and Statistics-Preserving," *in*

Proceedings of IEEE International Conference on Data Mining-Workshops, Italy, pp. 27-35, 2008.

[16] Olvi M. and William W., "Cancer Diagnosis via Linear Programming," *SIAM News*, vol. 23, no. 5, pp. 1-18, 1990.

[17] Rakesh A. and Ramakrishnan S., "Privacy-Preserving Data Mining," *ACM Special Interest Group on Management of Data Record*, vol. 29, no. 2, pp. 439-450, 2000.

[18] Shuting X., Zhang J., Han D., and Wang J., "Singular Value Decomposition Based Data Distortion Strategy for Privacy Protection," *Knowledge and Information Systems*, vol. 10, no. 3, pp. 383-397, 2006.

[19] Slava K., Rokach L., Elovici Y., and Shapira B., "Efficient Multidimensional Suppression for K-Anonymity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 3, pp. 334-347, 2010.

[20] Vassilios V., Bertino E., Fovino I., Provenza L., Saygin Y., and Theodoridis Y., "State-of-the-Art in Privacy Preserving Data Mining," *ACM SIGMOD Record*, vol. 33, no. 1, pp. 50-57, 2004.

[21] Wenliang D. and Zhijun Z., "Using Randomized Response Techniques for Privacy-Preserving Data Mining," *in Proceedings of 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, USA, pp. 505-510, 2003.

[22] Yehuda L. and Benny P., "Privacy Preserving Data Mining," *Journal of Cryptology*, vol. 15, no. 3, pp. 177-206, 2002.

Guang Li is currently a PhD candidate in the Department of Computer Science and Engineering, Harbin Institute of Technology, China. He received his BSc and MSc degree in computer science and engineering from Harbin Institute of Technology, China, in 2004 and 2006. His research interests include privacy preserving data mining and bioinformatics.

Yadong Wang received his BSc degree in computer science from Heilongjiang University, China, in 1986, and the MSc degree in computer science from the Harbin Institute of Technology, China, in 1989. Currently, he is a professor of school of computer science and technology at the Harbin Institute of Technology. He has been working in research areas such as artificial intelligence, expert system, machine learning, knowledge engineering, medical information technology. His main research area is focusing on bioinformatics and computational biology since 2003, including: biological database, microarray data analysis, regulatory network, biological knowledge mining algorithm in biology medical literature.