A PRIVACY-PRESERVING DATA MINING METHOD BASED ON SINGULAR VALUE DECOMPOSITION AND INDEPENDENT COMPONENT ANALYSIS

Guang Li*, Yadong Wang

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China
**Email:* hit6006@126.com

ABSTRACT

Privacy protection is indispensable in data mining, and many privacy-preserving data mining (PPDM) methods have been proposed. One such method is based on singular value decomposition (SVD), which uses SVD to find unimportant information for data mining and removes it to protect privacy. Independent component analysis (ICA) is another data analysis method. If both SVD and ICA are used, unimportant information can be extracted more comprehensively. Accordingly, this paper proposes a new PPDM method using both SVD and ICA. Experiments show that our method performs better in preserving privacy than the SVD-based methods while also maintaining data utility.

Keywords: Privacy preservation, Data mining, Singular value decomposition, Independent component analysis

1 INTRODUCTION

Data mining is the process of extracting patterns from data. It is becoming 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 fields (Cavoukian, 1998; Tavani, 1999). Many people fear that their private information will be misused and believe that privacy protection is important (Clayton, 2003; Cranor, Reagle, & Ackerman, 1999; Hall, & Rich, 2000). In addition to social pressures, legal mechanisms exist to protect data privacy. For example, in the U.S., to comply with the Health Insurance Portability and Accountability Act (HIPAA), individuals and organizations cannot reveal their medical data for public use without a privacy protection guarantee. To solve this problem, privacy-preserving data mining (PPDM) methods have been studied (Bertino, Fovino, & Provenza, 2005; Verykios, Bertino, Fovino, Provenza, Saygin, & Theodoridis, 2004). Now, PPDM technology, which can perform data mining without accessing the details of the original data, is becoming an increasingly important issue in data mining research.

In the past decade, many PPDM methods have been developed. They can be divided into two main categories, namely, methods based on data perturbation (Agrawal, & Srikant, 2000; Fung, Wang, & Yu, 2007; Liu, Kantarcioglu, & Thuraisingham, 2008; Kisilevich, Rokach, Elovici, & Shapira, 2010) and methods based on secure multi-party computation (SMC) (Emekci, Sahin, Agrawal, & Abbadi, 2007; Lindell, & Pinkas, 2002; Pinkas 2002). In the first category of methods, original data are not open, and users can only access perturbed data. Data mining is conducted on perturbed data to extract patterns from the original data. The second category of methods is often used for distributed databases. These methods assume that there are multiple nodes, each of which has only a portion of the global data set. These nodes aim to carry out data mining on the global data set, but each node does not allow the other nodes to know its data. In these methods, all of the nodes exchange the information required by the mining algorithm through protocols based on SMC. These protocols allow information to be exchanged privately, without allowing any node to directly obtain data from other nodes.

One important type of the first category is the PPDM method based on singular value decomposition (SVD) (Wang, Zhang, Xu, & Zhong, 2008; Xu, Zhang, Han, & Wang, 2006). SVD-based methods use the basic concept that the data mining and the privacy breaches require different kinds of information. Data mining only wants to derive patterns from the data, so it mainly needs general trends of data. Privacy breaches want to get the specific values of private data, thereby requiring the detailed

information of data. If only the general trends of data, which is the important information for data mining, is preserved, data mining can still be completed while also protecting privacy. The SVD-based methods perturb data using this idea. They arrange samples into a matrix, use SVD to analyze data and to identify unimportant information for data mining, and remove the unimportant information to perturb data.

SVD-based methods take two main forms. One is the basic SVD-based method, or BSVD method, and the other is the sparsified SVD-based method, or SSVD method. After SVD, every sample is turned into a sum of multiple components. The BSVD method perturbs data by removing the components corresponding to lower singular values. The SSVD method has two steps. In the first step, it uses the BSVD method to perturb data. In the second step, it implements additional perturbation of the modified data generated by the BSVD method from the first step. The SSVD method is an improvement over the BSVD method. It performs better at preserving privacy than the BSVD method and also maintains data utility.

Independent component analysis (ICA) is a data analysis method that is different from SVD. ICA and SVD analyze data from different perspectives, thereby revealing different data characteristics and identifying different unimportant information for data mining. Using SVD to analyze data is essentially equivalent to implementing principal component analysis (PCA) (Liang, Lee, Lim, Lin, Lee, & Wu, 2002; Lipovetsky, 2009; Wu, Liang, Sun, Zhou, & Lu, 2004), which considers second-order statistics; whereas ICA exploits higher-order statistics (Zhang, & Chan, 2006). If both SVD and ICA are used, we can extract unimportant information for data mining more comprehensively.

Based on this idea, this paper presents a PPDM method using both SVD and ICA. Our method follows the SSVD method's framework and also has two steps. First, we perturb data using the BSVD method. Then, we implements additional perturbation by using ICA. Experiments show that our method maintains good data utility and better preserves privacy than SVD-based methods.

The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 presents our algorithm. Section 4 shows the experimental results, and Section 5 is a conclusion.

2 RELATED WORK

2.1 PPDM methods based on SVD

Let A be a matrix with dimensions $n \times m$ representing the original data. The rows of A correspond to data objects while the columns correspond to the attributes. The full SVD of A is in Eq. (1), where U is an $n \times n$ orthonormal matrix. S is an $n \times m$ diagonal matrix, where the number of nonzero diagonal entries is $Rank(A)$, and all nonzero diagonal entries are in descending order. V^T is an $m \times m$ orthonormal matrix. $Rank(A)$ is the rank of A.

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

In the BSVD method, there is a parameter k such that $0 \leq k \leq Rank(A)$. The perturbed data A_k is defined in Eq. (2). In Eq. (2), 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. Larger values of k result in better data utility but worse privacy protection.

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

In the SSVD method, there are two parameters, k and d, where $0 \leq k \leq Rank(A)$ and $d > 0$. Its perturbed data $\overline{A_k}$ is defined in Eq. (3).

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

In Eq. (3), $\overline{U_k}$ and $\overline{V_k^T}$ are obtained from U_k and V_k^T. If the abstract value of either u_{ij} or v_{ij} is smaller than d, let u_{ij} or v_{ij} be zero, where u_{ij} and v_{ij} are the entries of U_k and V_k^T in the i-th row and j-th column, respectively. In the SSVD method, larger values of k and smaller values of d result in better data utility but worse privacy protection.

2.2 ICA

ICA is a linear transformation that translates data into a sum of independent components. This paper uses noiseless ICA and assumes that observation signals and source signals have the same number. That means that after implementing ICA, the $n \times m$ original data matrix A is BW, as in Eq. (4). B is an $n \times m$ matrix representing ICA coefficients. Every column of B has an independent distribution. W is an $m \times m$ mixing matrix. Prior to implementing ICA, the data should be centered and whitened (Chen, 2007; Hyvarinen, Karhunen, & Oja, 2001).

$$A = BW \tag{4}$$

ICA coefficients can be considered sparse coding of the original data. In sparse coding, the elements with larger abstract values contain more important information. The elements with lower abstract values are less important and are thus often identified as noise (Hyvarinen, Karhunen, & Oja, 2001). In our method, we adopt this assumption. We consider the elements in B with large abstract values to represent general trends of data and contain important information for data mining and the elements in B with small abstract values not important for data mining.

3 THE PROPOSED ALGORITHM

Our new method takes the same framework as the SSVD method. It has two parameters, k and d, and has two steps. In the first step, the BSVD method with parameter k is used to perturb data. Then, we implement ICA on the perturbed data from the first step and set the ICA coefficients that have abstract values less than d be zero. The details on our algorithm are shown in Figure 1. Prior to implementing ICA, the data should be centered and whitened (Chen, 2007; Hyvarinen, Karhunen, & Oja, 2001).

The experiments showed that our method performs better in preserving privacy than the SSVD method while also maintaining data utility. This is because the unimportant information for data mining extracted by SVD is enriched in the components with lower singular values. In the SSVD method, after removing these components, data are still perturbed by using the result from SVD. Usually at this time, there is not a substantial amount of unimportant information that is still preserved. So, there is not much room for the SSVD method to improve on the BSVD method. In our new method, after removing these components with lower singular values, ICA is implemented to analyze data to find additional unimportant information for data mining. Thus, our method has more room to improve on the BSVD method than the SSVD method.

The PPDM algorithm based on SVD and ICA

Input: Original data A, which is an $n \times m$ matrix, and parameters k and d. The rows of A correspond to data objects, and the columns correspond to attributes.

Output: Perturbed data A_M

Begin

Do SVD for A: $A = USV^T$

Calculate $A_k = U_k S_k V^T_k$, as in the BSVD method.

Calculate the objects' average $c = (A_k(1,:) + A_k(2,:) + \ldots + A_k(n,:))/n$.

C is a matrix in which every row is c.

Do orthonormal diagonalization for $(A_k - C)^T(A_k - C)$, get $(A_k - C)^T(A_k - C) = PQP^T$. where Q is a diagonal matrix and P is an orthonormal matrix. In Q, all diagonal entries are in descending order of there abstract values.

Let r equal to the rank of Q. Q_r is the diagonal submatrix of Q containing all the nonzero diagonal entries. P_r is the matrix containing the first r columns of P.

Calculate $A_k = (A_k - C)P_r Q_r^{-1/2}$.

Do ICA for A_k, and get $A_k = BW$.

$A_M = B_M W$. Assuming m_{ij} and b_{ij} are, respectively, the B_M and B's entries in the i-th row and j-th column, if $|b_{ij}| < d$, let $m_{ij} = 0$, else, let $m_{ij} = b_{ij}$.

$A_M = A_M Q_r^{1/2} P_r^T + C$

End

Figure 1. The new PPDM algorithm

4 EXPERIMENTS

4.1 Utility measures

The data utility measures assess the performance of data mining techniques on a data set after data distortion (e.g., whether the data mining techniques can maintain accuracy on distorted data). To measure data utility, this paper uses three types of classifiers: the j48 decision tree in the Waikato Environment for Knowledge Analysis (WEKA) (Witten, & Frank, 2005), the nearest neighbour classifier, and the Support Vector Machine (SVM). We assume that the classification accuracy of classifiers trained on the perturbed data and the original data are R_p and R_o and then define $r = (R_o - R_p)/R_o$. Obviously, smaller values of r show better utility. Assuming r_j, r_k, and r_s are the respective r values for the j48 decision tree, the nearest neighbour classifier, and the SVM, we use $max(r) = max\{ r_j, r_k, r_s \}$ as the utility measure. In this paper, if $max(r) \leq 0.02$, then we deem that the data utility was maintained.

4.2 Privacy measures

The privacy measures assess whether the PPDM algorithm protects privacy. We use the privacy measures that are often used by the matrix decomposition-based PPDM methods (Wang, Zhang, Xu, & Zhong, 2008; Xu, Zhang, Han, & Wang, 2006). 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 (Wang, Zhang, Xu, & Zhong, 2008; Xu, Zhang, Han, & Wang, 2006).

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

$$VD = \|A - MA\|_F / \|A\|_F \tag{5}$$

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

$$RP = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left|Rank_j^i - MRank_j^i\right|}{nm} \tag{6}$$

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

$$RK = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} Rk_j^i}{nm}, \text{ where } \quad Rk_j^i = \begin{cases} 1 & Rank_j^i = MRank_j^i \\ 0 & otherwise \end{cases} \tag{7}$$

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 separately, CP is defined as

$$CP = \frac{\sum_{i=1}^{m}\left|RankV_i - MRankV_i\right|}{m} \tag{8}$$

As in the case of RK, we define CK to measure the percentage of attributes that keep their ranks of average value after the distortion. Therefore, it is calculated as

$$CK = \frac{\sum_{i=1}^{m} Ck_i}{m}, \text{ where } \quad Ck_i = \begin{cases} 1 & RankV_i = MRankV_i \\ 0 & otherwise \end{cases} \tag{9}$$

To summarize, VD is the relative value difference in the Frobenius norm. RP, RK, CP, and CK all measure the rank difference of data elements. Simply put, if privacy is protected better, VD, RP, and CP will have larger values, and RK and CK will have smaller values.

4.3 Databases

Two real-life databases were used in the experiments: the original Breast Cancer Wisconsin Data Set (WBC) and the Pima Indians Diabetes Data Set (PID). They are both from the University of California at Irvine's Machine Learning Repository (Frank & Asuncion 2010; Mangasarian & Wolberg, 1990). The WBC database has 9 attributes and 699 samples. There are 16 samples that have missing values, and there are some repeated samples. We used only complete samples and deleted the repeated samples. As a result, there were 449 usable samples in the WBC database. The PID database has 8 attributes and 768 samples. In our experiments, for both databases, 20% of the samples were selected randomly as testing samples, and the other 80% of the samples were used as training samples.

4.4 Experimental results

All the experiments were repeated 50 times and averaged to obtain our experimental data. We used the BSVD method, the SSVD method, and our proposed approach to perturb data. These methods' parameter values were selected by experiments to guarantee data utility and optimize privacy measures. Then, we compared different perturbation methods by comparing the privacy measures of databases perturbed by those methods with their appropriate parameter values.

Figure 2 shows the data utility measures for the BSVD method with different k values. In this experiment, the smallest value of k was one, and the largest value of k was equal to the number of attributes of original data. The parameter k increased by steps of one. As we previously showed, larger values of k result in better data utility but worse privacy protection, so k should be as large as possible while maintaining data utility. In Figure 2, it can be seen that the appropriate value of k is seven for WBC and six for PID in the BSVD method.

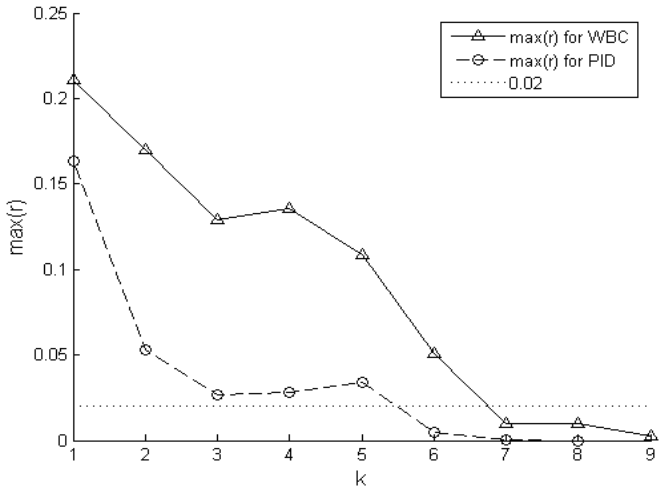

Figure 2. The data utility measures for databases perturbed by the BSVD method with different values of k.

The SSVD method and our new method have the same framework. They both have two parameters, k and d, and have two steps. In the first step, they use the BSVD method with the parameter k to perturb data. In the second step, they implement additional perturbation with parameter d on the modified data generated by the BSVD method from the first step. We used a greedy strategy to select the appropriate parameter values for them. We let the value of k in them both be equal to the appropriate value of k in the BSVD method to maximize the perturbation in its first step while maintaining data utility. Then we selected the largest possible value of d to guarantee data utility. Figures 3 and 4 show the utility measures for databases perturbed by the SSVD method and our new method with the appropriate value of k and different values of d. In these experiments, the parameter d was decided by another parameter e, which is the rate at which elements tended toward zero. Larger values of e correspond to larger values of d. In these experiments, the smallest value of e was 0.05, and the largest value of e was 0.95.

The parameter *e* was increased by steps of 0.05. In Figures 3 and 4, it can be seen that *e* should be 0.45 for the WBC and should be 0.15 for the PID in the SSVD method; and *e* should be 0.75 for the WBC and should be 0.8 for the PID in our new method.

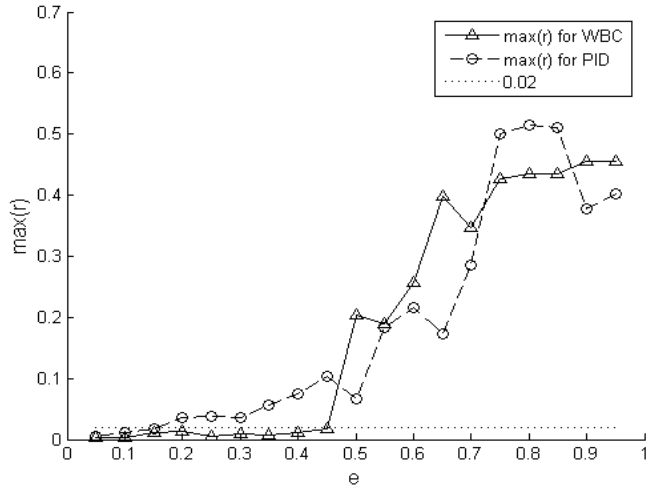

Figure 3. The data utility measures for databases perturbed by the SSVD method with different values of *e* and the appropriate value of *k*.

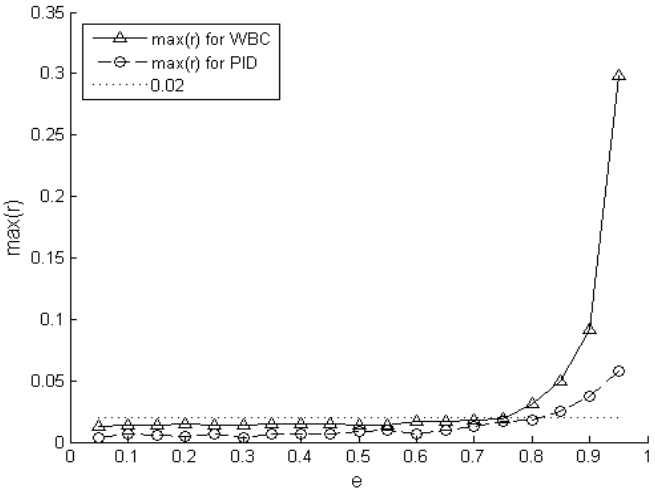

Figure 4. The data utility measures for databases perturbed by our new method with different values of *e* and the appropriate value of *k*.

It is interesting to see what happens if only ICA is applied in order for perturbation, which is equivalent to using our new method and letting the parameter *k* be equal to the number of attributes of original data. Figure 5 shows the utility measures for databases perturbed using only ICA with different values of *d*. And *d* was also decided by *e*. *e* was from 0.05 to 0.95 and increased by steps of 0.05. In Figure 5, it can be seen that *e* should be 0.6 for the WBC and should be 0.8 for the PID.

Table 1 shows the privacy measures for our new method and the SVD-based methods with the parameter values selected in the above experiments. As shown previously, these parameter values can maintain data utility and optimize privacy measures. In Table 1, it can be found that our new method, which uses both ICA and SVD, can perform better for privacy protection than the SVD-based methods and the method using only ICA. This experimental result confirmed our point of view that SVD and ICA can find different unimportant information for data mining, so when using both SVD and ICA as in our new method, we can extract more unimportant information for data mining and so can protect privacy better than using only the SVD or ICA.

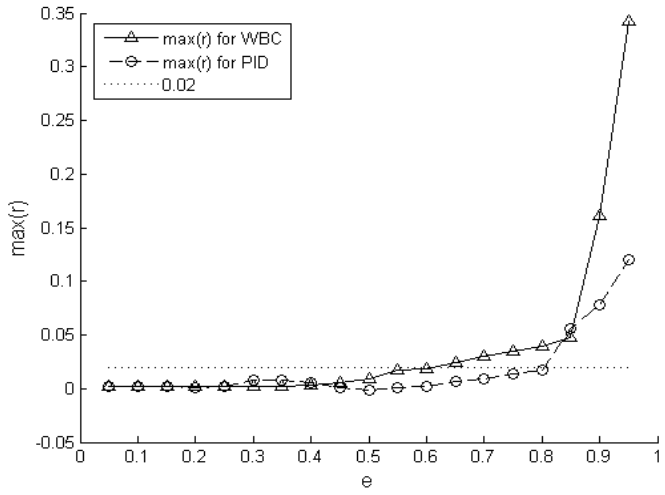

Figure 5. The data utility measures for databases perturbed only using ICA with different values of *e*.

Table 1. The privacy measures for our method and the SVD-based methods.

Database	Perturbation method	VD	RP	RK	CP	CK
WBC	BSVD	0.11	31.9	0.019	0.3	0.8
	SSVD	0.25	37.3	0.015	0.3	0.8
	Only ICA	0.19	40.1	0.014	0.1	0.9
	Both ICA and SVD	0.34	58.2	0.008	0.4	0.7
PID	BSVD	0.01	48.3	0.126	0	1
	SSVD	0.03	56.2	0.064	0	1
	Only ICA	0.25	99.1	0.013	0	1
	Both ICA and SVD	0.27	118.1	0.009	0	1

5 CONCLUSIONS

This paper proposes a new PPDM method based on both SVD and ICA. The basic idea behind this method comes from the SVD-based PPDM methods, which protect privacy by constructing a perturbed data set to replace the original data. The SVD-based methods analyze data using SVD, thereby finding and retaining only important information for data mining; data perturbation is achieved by removing unimportant information. ICA is a data analysis method different from SVD. ICA and SVD analyze data from different perspectives and identify different kinds of important information for data mining. Our new method uses both SVD and ICA to analyze data, so that more unimportant information for data mining will be found. And so our method can perform better than methods only using SVD or ICA, which was confirmed by experiments in this paper.

Our algorithm has two steps. In the first step, SVD is implemented, and the components corresponding to lower singular values are deleted. Then, ICA is implemented, and the ICA coefficients that have lower abstract values will become zero. Using experiments, we demonstrate that while maintaining data utility, our new method can protect privacy better than the methods using only SVD or ICA.

6 ACKNOWLEDGMENT

We thank Dr. William H. Wolberg of the University of Wisconsin Hospital in Madison, WI, for providing us with the experimental data.

7 REFERENCES

Agrawal, R., & Srikant, R. (2000) Privacy-preserving data mining. *ACM SIGMOD Record* 29(2), 439-450.

Bertino, E., Fovino, I., & Provenza, L. (2005) A framework for evaluating privacy preserving data mining algorithms. *Data Mining and Knowledge Discovery* 11(2), 121-154.

Cavoukian, A. (1998) Data mining: staking a claim on your privacy. Retrieved October 25, 2010 from the World Wide Web: http://www.ipc.on.ca/images/Resources/datamine.pdf.

Chen, Y. (2007) *A Study on the Algorithm of Digital Watermarking Based on the Independent Component Analysis*, Master thesis, Xiamen University, Xiamen, Fujian, China.

Clayton E. (2003) Ethical, legal, and social implications of genomic medicine. *New England Journal of Medicine* 349 (6), 562-569.

Cranor, L., Reagle, J., & Ackerman, M. (1999) Beyond Concern: Understanding Net Users' Attitudes About Online Privacy. Retrieved October 25, 2010 from the World Wide Web: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.40.7328&rep=rep1&type=pdf.

Emekci, F., Sahin, O., Agrawal, D., & Abbadi, A. (2007) Privacy preserving decision tree learning over multiple parties. *Data & Knowledge Engineering* 63(2), 348-361.

Frank, A. & Asuncion, A. (2010) UCI Machine Learning Repository. Retrieved June 15, 2010 from the World Wide Web: http://archive.ics.uci.edu/ml/.

Fung, B., Wang, K., & Yu, P. (2007) Anonymizing classification data for privacy preservation. *IEEE TKDE* 19(5), 711-725.

Hall M. & Rich S. (2000) Patients' fear of genetic discrimination by health insurers: the impact of legal protections. *Genetics in Medicine* 2 (4), 214-221.

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

Kisilevich, S., Rokach, L., Elovici, Y., & Shapira, B. (2010) Efficient Multidimensional Suppression for K-Anonymity. *IEEE TKDE* 22(3), 334-347.

Liang Y., Lee H., Lim S., Lin W., Lee K., & Wu C. (2002) Proper Orthogonal Decomposition and Its Applications-Part I: Theory. *Journal of Sound and Vibration* 252(3), 527-544.

Lindell, Y. & Pinkas, B. (2002) Privacy preserving data mining. *Journal of Cryptology* 15(3), 177-206.

Lipovetsky, S. (2009) PCA and SVD with nonnegative loadings. *Pattern Recognition* 42(1), 68-76.

Liu, L., Kantarcioglu, M., & Thuraisingham, B. (2008) The applicability of the perturbation based privacy preserving data mining for real-world data. *Data & Knowledge Engineering* 65(1), 5-21.

Mangasarian, O. & Wolberg, W. (1990) Cancer diagnosis via linear programming. *SIAM News* 23(5), 1-18.

Pinkas, B. (2002) Cryptographic techniques for privacy-preserving data mining. *ACM SIGKDD Explorations Newsletter* 4(2), 12-19.

Tavani, H. (1999) Information privacy, data mining, and the internet. *Ethics and Information Technology* 1(2), 137-145.

Verykios, V., Bertino, E., Fovino, I., Provenza, L., Saygin, Y., & Theodoridis, Y. (2004) State-of-the-art in privacy preserving data mining. *ACM SIGMOD Record* 33(1), 50-57.

Wang, J., Zhang, J., Xu, S., & Zhong, W. (2008) A novel data distortion approach via selective SSVD for privacy protection. *International Journal of Information and Computer Security* 2(1), 48-70.

Witten, I. & Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*, Burlington, Massachusetts, US: Morgan Kaufmann.

Wu, C., Liang, Y., Sun, Y., Zhou, C., & Lu, Y. (2004) On the Equivalence of SVD and PCA. *Chinese Journal of Computers* 27(2) 286-288.

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

Zhang, K. & Chan, L. (2006) ICA by PCA Approach: Relating Higher-Order Statistics to Second-Order Moments. *Lecture Notes in Computer Science* 3889, 311-318.

(Article history: Received 4 November 2010, Accepted 31 January 2010, Available online 8 February 2011)