

Selective Data Distortion via Structural Partition and SSVD for Privacy Preservation^{*}

Jie Wang^{1,2}, Weijun Zhong¹, Shuting Xu³, Jun Zhang²

¹Department of Management Sciences and Engineering, Southeast University, Nanjing, 210096, P.R. China

²Laboratory for High Performance Scientific Computing and Computer Simulation, Department of Computer Science, University of Kentucky, Lexington, KY 40506, USA

³Department of Computer Information Systems, Virginia State University, Petersburg, VA 23806, USA

(jiewang@uky.edu, zhong.wj@public1.ptt.js.cn, sxu@vsu.edu, jzhang@cs.uky.edu)

Abstract — Accurate information extracted from datasets is required for making reasonable decisions using data mining algorithms. Privacy preservation has become one of the top priorities in the design of various data mining applications. In this paper, a novel data distortion strategy based on structural partition and sparsified Singular Value Decomposition (SSVD) technique is proposed. Three schemes, object-based partition, feature-based partition and hybrid partition, are defined to permit a tradeoff between privacy protection on centralized datasets and accuracy of data mining techniques. Some metrics to measure privacy preservation are used to examine the performance of the proposed new strategies. Data utility of the three proposed schemes is examined by a binary classification based on the support vector machine. Furthermore, the effect of different ranks of SVD and the threshold value of SSVD on data distortion and utility are also tested. Our experimental results indicate that, in comparison with standard data distortion techniques, the proposed schemes are very efficient in achieving a good tradeoff between data privacy and data utility, and it affords a feasible solution, with a significant reduction on the computational cost from SVD, to protect sensitive information and promise high accuracy in decision making.

Keywords-*component; privacy; security; SVD; SSVD; matrix partition*

I. INTRODUCTION

With the rapid growth of modern communication and data exchange technology, data mining is increasingly vital for decision makers to make a timely and accurate response from huge amounts of easily accessible information in the changing global environment. Collaborations between different parties have been considered as an essential approach for business and research success in many situations and data are commonly shared beyond the boundary of individual entities. Accurate information is required for making wise decisions using data mining algorithms. In the implementation of these new data-oriented structures there exist some important problems, such

as how an entity be entrusted with access to sensitive personal or business information, how sensitive datasets be protected from unauthorized access [1]. Without an acceptable level of privacy of sensitive information, many data mining applications would not be applicable. Therefore, privacy protection has become one of the top priorities in the design of database and data mining applications [2]. In [3], it showed that 73% of the respondents in a survey are not willing to provide their data without the protection of privacy.

Many approaches have been adopted for privacy preserving data mining. Verykios *et al.* [4] made a classification of data use based on five dimensions: data distribution, data modification, data mining algorithm, data or rule hiding and privacy preservation. The last dimension, privacy preservation, is becoming increasingly critical for future development of data mining techniques with great potential access to datasets containing personal, sensitive, or confidential information. Extracting valid data mining results while still preserving privacy of datasets is a major challenge for existing data mining algorithms. Data selective modification is required in order to keep the utility for the modified data given that the privacy is preserved [4].

Among the widely used approaches, anonymous techniques [5], with easy implementations, keep secret identities of information providers. However, the quality of the data collected with these approaches could not be guaranteed for the reason that a malicious user could provide random or falsified data and data managers cannot verify information without the identities of data providers.

Several randomization-based data distortion methods have been discussed in the literature [2]. Singular Value Decomposition (SVD) has been used for privacy issue [8]. Polat and Du suggested a SVD-based randomization approach to protecting user's privacy while providing tolerated accurate recommendations in collaborative filtering [8], although SVD is not used for the purpose of privacy preserving in their work.

^{*}Technical Report No. 449-05, Department of Computer Science, University of Kentucky, Lexington, KY, 2005. The work was supported by the Kentucky New Economy Safety and Security Consortium, and by a grant No. 02KJB630001 of Research Project Grant of JiangSu, China.

The SVD sparsification concept was firstly developed by Gao and Zhang in [7] for reducing the storage cost and enhancing the performance of SVD in text retrieval applications. Xu *et al.* applied SVD and sparsified SVD methods in data distortion in a terrorist analysis system [8]. The disadvantage of SVD-based approaches is that SVD incurs a significant computational cost for large scale datasets during the matrix decomposition phase. Moreover, in many applications with large databases, not the entire database is needed for data distortion. In most cases, only part of the collection that contains sensitive information needs to be protected in a database.

Based on our previous work in [8], we present a few novel strategies to partially distort the original data matrix in order to maintain the advantage of data privacy and data usability of SVD, but obtain a significant reduction on computational cost.

The basic idea underlying the proposed strategies is to perform a dimension or rank reduction and conduct sparsification operation on one selected part of the original dataset. Three different matrix structural partition strategies are used to partition the original data into several submatrices. Sparsified SVD (SSVD) strategy is then applied on the selected submatrix to distort partial information in the subset. The distorted submatrix is combined with the original undistorted part of the matrix to form the new dataset for data mining applications. Then data mining algorithms are used on the distorted matrix. As an example, a binary classification based on SVM algorithm is used to examine the classification accuracy of the proposed distortion strategies. The effect of the rank of SVD and the threshold value of SSVD, κ and ε , on distortion level and classification accuracy are tested. We also use some metrics to measure the degree of privacy preservation.

Our experiments show that compared to other more standard data distortion techniques, the proposed selective SSVD-based structural partition approach is very efficient in maintaining both data privacy and data utility and it can be applied to protect the privacy of enterprise information and promise a high accuracy of decision making.

II. BACKGROUND AND RELATED WORK

A. Some Assumptions

In order to preserve data privacy, we assume that no one except the data owner or authorized users have the right to access the original data. The data analysts only see the distorted dataset matrix, not the original matrix and or any part of it that may have privacy concerns. The analysts run the data mining algorithms to recommend decisions based on the distorted dataset matrix, while they are not able to know the original dataset unless appropriate permission is granted. Therefore, sensitive information with high level privacy is protected from unauthorized access. Simultaneously data utility should be assured in order to control the degradation of accuracy of the data mining algorithms due to data distortion.

B. Singular Value Decomposition

Singular Value Decomposition (SVD) is a popular method in data mining and information retrieval [10], since it has a mathematical feature to find a rank- κ approximation of a

matrix with minimal change to that matrix for a given value of κ [11]. It is usually used to reduce the dimensionality of the original dataset. Here we use it as a data distortion method [8].

Let A be a matrix of dimension $n \times m$ representing the original dataset. The rows of the matrix correspond to data objects and the columns to attributes. The SVD of the matrix A can be written as

$$A = U \Sigma V^T \quad (1)$$

where U is an $n \times n$ orthonormal matrix having the left singular vectors of A as its columns, Σ is an $n \times m$ diagonal matrix whose nonnegative diagonal entries are the singular values in a descending order,

$$\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_s), s = \min\{m, n\} \quad (2)$$

V^T is an $m \times m$ orthonormal matrix having the right singular vectors of A as its columns [11]. These three matrices reflect a breakdown of original relationship into linearly independent vectors.

Due to the arrangement of the singular values in the matrix Σ (in a descending order), the SVD transformation has the property that the maximal variation among the objects is captured in the first dimension, as $\sigma_1 > \sigma_i$, for $i > 2$. Similarly much of the remaining variations are captured in the second dimension, and so on. Thus, a transformed matrix with a much lower dimension can be constructed to represent the original matrix faithfully.

We can create a rank- κ approximation A_k to the matrix A by defining,

$$A_k = U_k \Sigma_k V_k^T \quad (3)$$

where U_k contains the first κ columns of U , Σ_k contains the κ largest nonzero singular values of A , and V_k^T contains the first κ rows of V^T . It has been proven that the distance between A and its rank- κ approximation is minimized by the approximation A_k in the sense of the Frobenius norm [11,13].

Rank reduction was first proposed by Deerwest *et al.* as a method for removing the noise of a text collection [14]. We define $E_k = A - A_k$ as the noise in the original matrix A . Hence, using A_k instead of A may yield better mining accuracy. Simultaneously due to the difference between A and A_k , the distorted data A_k can preserve privacy, as it is difficult to figure out the values of A from those of A_k without the knowledge of E_k . Hence, A_k can be seen as both a distorted copy of A and a faithful representation of the original data.

A good choice of the rank of SVD could capture the main structure of a data collection and ignore the irrelevant noise. Large ranks will still retain some noise, but too small ranks will lose the structure and meaning of the original collection. How to choose the rank that provides optimal performance of data mining algorithms for any given datasets remains an open question and is normally decided via empirical tests [12]. In this paper, we conduct some experiments and take a close look at the effect of rank of SVD on the accuracy of a binary classification.

SVD computation incurs a significant computational cost for large scale data matrices. It is indicated that the cost of computing the SVD of a sparse matrix A using a Lanczos-type procedure could be expressed as [12]:

$$\text{Total cost} = I \times \text{cost}(A^T A x) + k \times \text{cost}(A x) \quad (4)$$

where I is the the number of iterations required by a Lanczos-type procedure to approximate the eigensystem of $A^T A$, x is a vector and k is the number of computed singular values and their corresponding number of non-zero entries in the sparse matrix A . The dominant computational cost of the Lanczos method is related to the number and complexity of matrix multiplication by A and A^T .

C. Sparsified Single Value Decomposition (SSVD)

Three SVD sparsification strategies, which are single threshold strategy (STS), column threshold strategy (CTS) and exponential threshold strategy (ETS), have been proposed by Gao and Zhang for reducing the storage cost and enhancing the performance of SVD in the area of information retrieval [7]. In this paper, the simplest one, STS is used to perform sparsification on SVD matrix A_k .

The basic idea of STS-based sparsification is that, given a certain threshold value ε , for any u_{ij} in U_k , if $|u_{ij}| < \varepsilon$, then we set $u_{ij} = 0$. The same operation is conducted on V_k^T .

Let \bar{U}_k and \bar{V}_k denote the new matrices created after performing STS on U_k and V_k^T respectively, and the new version of the distorted matrix A_k is

$$\bar{A}_k = \bar{U}_k \Sigma_k \bar{V}_k \quad (5)$$

STS is applied here to further distorting the A_k after the rank reduction by SVD. Obviously the degree of perturbation of \bar{A}_k is larger than A_k and the protection on data privacy is improved.

III. STRUCTURAL PARTITION STRATEGIES

Instead of conducting SVD and STS on the whole data matrix, structural matrix partition is used here to divide the

original matrix into several submatrices, and we perform SSVD on one selected submatrix.

Three kinds of matrix partition are proposed here, which are denoted by P1, P2, and P3, respectively.

1) Object-based partition (denoted by P1)

Let us partition A as

$$A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \quad (6)$$

The whole dataset is divided into two groups, A_1 and A_2 . We perform sparsified SVD on A_1 to get $B_1 = \text{SSVD}(A_1)$. Then, the partially distorted dataset is

$$A^* = \begin{bmatrix} B_1 \\ A_2 \end{bmatrix} \quad (7)$$

Here, all feature values of the first group are distorted

2) Feature-based partition (denoted by P2):

Let

$$A = [A_1 \quad A_2] \quad (8)$$

A_1 contains the first part of feature items and A_2 the second part. We perform sparsified SVD on A_1 to get $B_1 = \text{SSVD}(A_1)$. Then the new distorted matrix is

$$A^* = [B_1 \quad A_2] \quad (9)$$

In this case, only part of feature values are distorted by SSVD.

3) Hybrid partition (denoted by P3)

Let the partition be

$$A = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix} \quad (10)$$

We perform sparsified SVD on A_1 to get $B_1 = \text{SSVD}(A_1)$. Then, the selectively distorted matrix is

$$A^* = \begin{bmatrix} B_1 & A_2 \\ A_3 & A_4 \end{bmatrix} \quad (11)$$

Here, a part of the feature values for a part of the objects are selected for distortion.

As we discussed in Section II, the dominant computational cost of SVD is related to the number and the complexity of sparse matrix multiplication by A and A^T . Computing SVD

on part of the original matrix would result in a reduction on the computational cost and an improvement on the efficiency of data mining algorithms by removing unnecessary data distortion. This is because the matrix multiplication is now performed with respect to the submatrix A_1 , not to the full matrix A .

The level of distortion and data utility are dependent on the partition scheme in use. Depending on specific goals of various applications, one of the above three schemes can be chosen. The analysis of the performance of the above proposed strategies will be done in the next sections.

IV. DATA DISTORTION MEASURES

The privacy protection measure should indicate how closely the original value of an item can be estimated from the distorted data [9]. Some privacy metrics have been proposed in the literature [8,15,16]. Some data distortion measures defined in [8] are used here to assess the level of data distortion which only depends on the original matrix A and its distorted counterpart A^* .

A. Value Difference (VD)

After a data matrix is distorted, the value of its elements changes. The value difference (VD) of the datasets is represented by the relative value difference in the Frobenius norm. Thus VD is the ratio of the Frobenius norm of the difference of A^* from A to the Frobenius norm of A :

$$VD = \frac{\|A - A^*\|_F}{\|A\|_F} \quad (12)$$

B. Position Difference

After a data distortion, the order of the value of the data elements changes, too. We use several metrics to measure the position difference of the data elements.

1) Rank Position (RP)

RP is used to denote the average change of rank for all the attributes. After the elements of an attribute are distorted, the rank of each element in an ascending order of its value changes. Assume that dataset A has n data objects and m attributes. $Rank_j^i$ denotes the rank of the j th element in attribute i , and $(Rank_j^i)^*$ denotes the rank of the distorted element A_{ji} . Then RP is defined as:

$$RP = \frac{\sum_{i=1}^m \sum_{j=1}^n |Rank_j^i - (Rank_j^i)^*|}{m \times n} \quad (13)$$

If two elements have the same value, we define the element with the lower row index to have the higher rank.

2) Rank Maintenance (RM)

RK represents the percentage of elements that keep their ranks of value in each column after the distortion. It is computed as:

$$RK = \frac{\sum_{i=1}^m \sum_{j=1}^n Rk_j^i}{m \times n} \quad (14)$$

where Rk_j^i means whether an element keeps its position in the order of values:

$$Rk_j^i = \begin{cases} 1, & \text{if } Rank_j^i = (Rank_j^i)^* \\ 0, & \text{otherwise} \end{cases} \quad (15)$$

If an element keeps its position in the order of values, $Rk_j^i = 1$, otherwise, $Rk_j^i = 0$.

3) Change of Rank of Features (CP)

One may infer the content of one feature from its relative value difference compared with the other attributes. Thus it is desirable that the order of the average value of each attribute varies after the data distortion. Here we use the metric CP to define the change of rank of the average value of the attributes:

$$CP = \frac{\sum_{i=1}^m |RAV_i - RAV_i^*|}{m} \quad (16)$$

where RAV_i is the rank of the average value of attribute i , while RAV_i^* denotes its rank after the distortion.

4) Maintenance of Rank of Features (CK)

CK is defined to measure the percentage of the features that keep their ranks of average value after the distortion. So it is calculated as:

$$CK = \frac{\sum_{i=1}^m Ck^i}{m} \quad (17)$$

where Ck^i is computed as:

$$Ck^i = \begin{cases} 1, & \text{if } RAV_i = RAV_i^* \\ 0, & \text{otherwise} \end{cases} \quad (18)$$

The value of RP and CP is proportional to the level of the distortion. On the contrary, the value of RK and CK is inversely related to the level of distortion.

V. DATA UTILITY MEASURE

Data utility measures indicate the accuracy of data mining algorithms on distorted data after the conduction of certain perturbation. In this paper, Support Vector Machine (SVM)

classification is chosen as the data utility measure. SVM is based on structural risk minimization theory [17]. In SVM classification, the goal is to find a hyperplane that separates the examples with maximum margin.

VI. EXPERIMENTS AND RESULTS

The objective of the experiments is to examine the performance of the proposed new data distortion strategies. Three kinds of matrix partition schemes are compared on both privacy preservation and data utility. And the effect of different rank of SVD and threshold value of SSVD, κ and ε , on accuracy of data mining algorithms are tested respectively. A synthetic dataset is used in our experiments.

A. Synthetic Dataset

A synthetic dataset (org), a 2000 by 100 matrix, is constructed with entries that are randomly generated numbers within the interval [1,10] obeying a uniform distribution. We classify all the objects into two classes using a random rule:

$$classlab = \begin{cases} 1, & |\sin(org(i,1)) - org(i,88)| \times |\cos(org(i,45))| \times org(i,78) > 15 \\ -1, & otherwise \end{cases}$$

SVM classification is used to learn from the synthetic dataset and construct the classifier. The classification results are obtained by a five-fold cross validation.

- Experiment 1

The five distortion strategies, uniformly distributed noise (UD), normally distributed noise (ND), SVD, SSVD, SSVD with matrix partition, are implemented on the same dataset to compare the performance [8].

Table 1 shows the comparison among these five data distortion methods under certain parameters. The normally distributed noise is generated with $\mu = 0$ and $\sigma = 0.46$, see [8] for the meaning of these two parameters. The uniformly distributed noise is generated from the interval [0, 0.8]. The rank κ in SVD is 20, and threshold value in SSVD is 1e-3.

The matrix partition schemes described in the previous section are applied in the experiment, represented by SSVD[P1], SSVD[P2] and SSVD[P3] in Table I. The part of the distorted submatrix is 2000 by 50 for P1, 1000 by 100 for P2, and 1000 by 50 for P3, respectively.

Based on the comparison results in Table I, a conclusion can be made that compared to randomization-based data distortion methods, UD and ND, SVD-based strategies achieve a higher level of distortion and can provide better protection on privacy. Roughly speaking, sparsified SVD is better than SVD on most of the five metrics. The CK value for SSVD-based methods is 0, which means all the features change their rank in average value after performing certain data transformations.

TABLE I. COMPARISON OF DIFFERENT DISTORTION STRATEGIES

| Dataset | Level of Distortion | | | | | Accuracy (%) |
|----------|---------------------|----------|--------|-------|------|--------------|
| | VD | RP | RK | CP | CK | |
| Org | - | - | - | - | - | 76.15 |
| UD | 0.0760 | 664.0489 | 0.0062 | 0 | 1 | 76.20 |
| ND | 0.0758 | 665.1643 | 0.0043 | 0 | 1 | 75.80 |
| SVD | 0.3665 | 666.9214 | 0.0007 | 21.28 | 0.39 | 76.60 |
| SSVD | 0.7464 | 664.0129 | 0.0005 | 36.42 | 0 | 76.50 |
| SSVD[P1] | 0.5059 | 667.5759 | 0.0011 | 34.02 | 0.02 | 66.75 |
| SSVD[P2] | 0.4866 | 332.7783 | 0.5002 | 35.48 | 0 | 77.35 |
| SSVD[P3] | 0.3655 | 333.8874 | 0.5007 | 34.44 | 0 | 76.70 |

Among the three proposed matrix partition strategies, SSVD[P1] performs best with the same level of distortion as SSVD. SSVD[P2] and SSVD[P3] is comparable on distortion level with the largest RK value and the lowest RP value. All these three methods have great variance on feature values.

As to data utility, the accuracy of the three new schemes are 66.75%, 77.35% and 76.7%. Naturally SSVD[P1] is worst on data mining accuracy, due to its best preservation on privacy. SSVD[P2] supplies the best data utility with the higher accuracy than the original dataset. From the above analysis, we can make a reasonable conclusion that considering a tradeoff between privacy preservation and data utility, the performance of SSVD[P2] is the best among these three matrix partition strategies.

- Experiment 2

To examine the change of data utility of the three partition schemes with increasing the rank of SVD, we conduct the experiment using the same synthetic dataset as the one in the Experiment 1, where the threshold value in SSVD is 1e-3, and in SVM, radial base function (rbf) is chosen as the kernel function and gamma vale in rbf is 0.001 [8].

Fig.1 illustrates the influence of rank of SVD on classification accuracy. P2 and P3 show the similar graphs of accuracy. The accuracy decreases with κ till κ is larger than the half of the number of features, $\kappa = 50$ in our experiment. For any of $\kappa > 50$, the accuracy of P1 and P2 is equal to that of the original dataset. The highest accuracy is obtained with the rank of 1/10 of the number of features.

P1 shows worse performance on data utility than P2 and P3 and its accuracy is lower than that of the original dataset. It also demonstrates a different trend of change. The accuracy of P1 increases with k when $\kappa < 60$ and decreases with k for $\kappa > 60$.

How to choose the rank of SVD is still unsolved and empirical tests are required. Our experiment implies one possible good choice of the rank of SVD for our distortion strategies if only considering data utility. If P1 scheme is used, 3/5 of the number of features is a good choice for k .

For P2 and P3, we can choose 1/10 of the number of features as the rank of SVD.

- Experiment 3

Here we examine the influence of the threshold value, ϵ , in the single threshold strategy (STS) used in our distortion strategies. Fig.2 illustrates classification accuracy under ϵ in the interval from 0 to 0.1. In the experiment, the rank of SVD is 40 and the same configuration of SVM as in Experiment 2 is used. With the increase of ϵ in SSVD, it exhibits no observable trend in data utility for all three distortion schemes. This implies that the sparsification parameter does not affect the classification accuracy sensitively in this study.

- Experiment 4

The previous experiments demonstrate that feature-based partition scheme can provide a high mining accuracy with an acceptable level of data distortion. The further test on this partition scheme is implemented from the viewpoint of both data distortion and data utility. In Fig.3, VD, RP and CP increase with the number of features while RK and CK decrease. It shows an intuitive result that the level of distortion increases with the number of features in A_1 .

Fig.4 exhibits a critical point with the highest accuracy when the number of column in A_1 is 70, which means A_1 contains 70 percent of the features.

The CPU time used to compute the SVD and partial SVD of the dataset on a SunBlade 150 workstation is 46.12 seconds for SSVD, 13.27 seconds for SSVD[P1], 22.95 seconds for SSVD[P2], and 5.07 seconds for SSVD[P3]. It can be seen that the cost of computing SVD is proportional to the size of the submatrix used for the partial SVD computation.

B. Summary

Some important conclusions can be drawn from these graphs:

- The overall performance of the SVD-based distortion approaches is better than randomization-based approaches.
- Most of the SVD-based approaches can obtain a higher accuracy on classification than the original data.
- For feature-based partition and hybrid partition distortion strategies, the classification accuracy decreases with the increment of the rank of SVD. This special property provides a possible way to achieve high accuracy with a significant reduction on computational cost due to the use of a small rank value.
- The overall performance of three structural partition strategies:
 1. Object-based partition has the highest distortion level with the lowest data utility level. Feature-based partition has the highest data utility level.

2. Feature-based and hybrid partitions have a quite comparable distortion level.
3. Feature-based partition should be a reasonable solution with a good compromise on the distortion and utility of data.

Of course, which partition strategy to use in a particular application is dependent on the circumstances of that applications, i.e., on the nature of the data to be distorted.

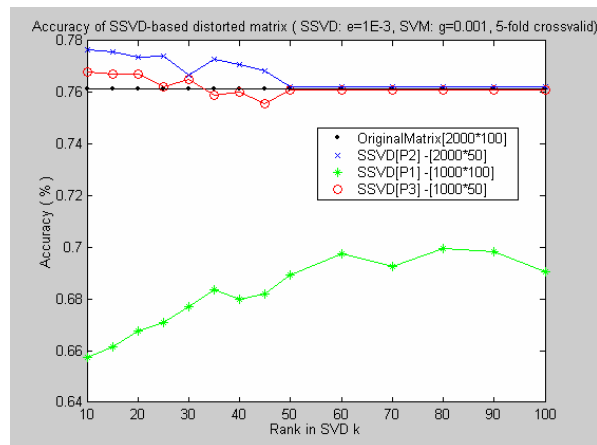


Figure 1. The effect of rank k in SVD on the accuracy

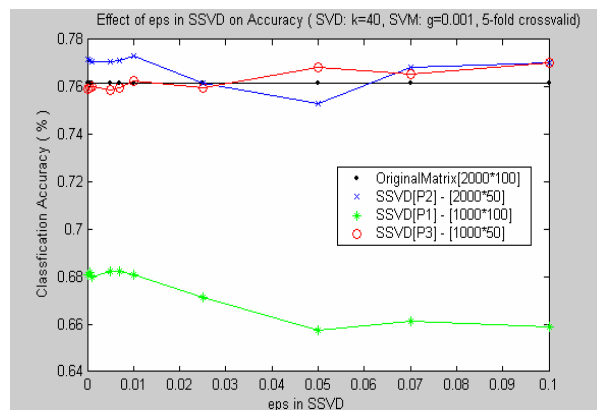


Figure 2. The effect of threshold value, ϵ , in SSVD on the accuracy

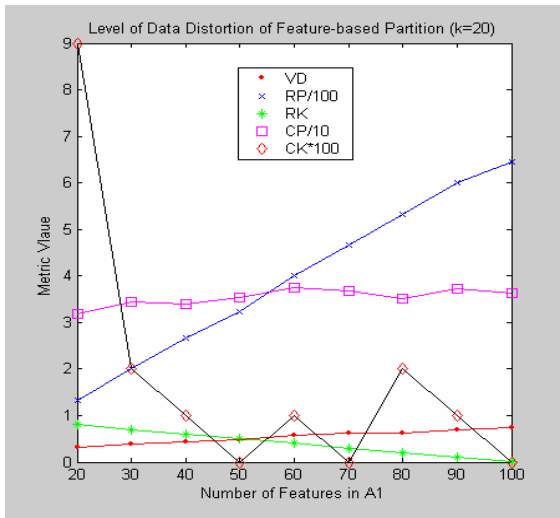


Figure 3. Level of data distortion of feature-based distortion

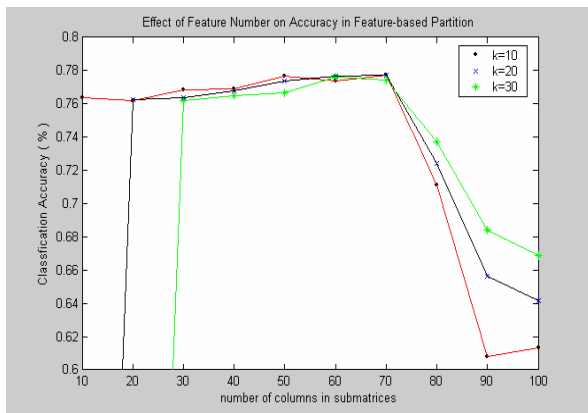


Figure 4. Effect of the number of features on accuracy of feature-based partition

VII. CONCLUSIONS AND FUTURE WORK

In this paper, we studied several existing strategies in privacy-preserving data mining, including two randomization-based methods, one rank reduction method (SVD), one STS-based sparsification method (SSVD). A set of new privacy preserving strategies are proposed based on matrix partition techniques, and object-based partition, feature-based partition and hybrid partition are defined. We compare the performance of all these methods both on data privacy level and data utility level. The experimental results demonstrate that the efficiency of the proposed strategies. With consideration on the overall performance, we see that feature-based partition is a feasible and efficient solution for privacy-preserving data mining. Future work may include performance tests on real world datasets. We also plan to examine the other two sparsification strategies, CTS and ETS, proposed in [7], in data distortion applications. The determination of optimal rank of SVD and

threshold value of SSVD under different applications is also desirable in order to implement our proposed approach on real applications with little user intervention. Further investigation is worthwhile on combining our method with other privacy preservation techniques to further improve efficiency.

REFERENCES

- [1] L. Korba, "Privacy in distributed electronic commerce," in Proceedings of the 35 Hawaii International Conference on System Science, Hawaii, January 2002
- [2] Z. Yang, S. Zhong, R. N. Wright, "Privacy-preserving classification of customer data without loss of accuracy," In proceedings of the 5th SIAM International Conference on Data Mining, Newport Beach, CA, April 21-23, 2005
- [3] L. Cranor, editor. Special Issue on Internet Privacy, Comm. ACM 42(2), 1999
- [4] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, Y. Theodoridis. "State-of-the-art in privacy preserving data mining," *SIGMOD Record*, 33(1):50-57, 2004
- [5] M. K. Reiter, A. D. Rubin, "Crowds: Anonymity for Web transaction," *ACM Transaction on Information and System Security*, 1(1):66-92, 1998
- [6] S. Datta, H. Kargupta, K. Sivakumar. "Homeland defense, privacy-sensitive data mining, and random value distortion," In Proceedings of the SIAM Workshop on Data Mining for Counter Terrorism and Security, San Francisco, CA, May 2003
- [7] J. Gao, J. Zhang. "Sparsification strategies in latent semantic indexing," in Proceedings of the 2003 Text Mining Workshop, M.W. Berry and W.M. Pottenger, (ed.), pp. 93-103, San Francisco, CA, May 3, 2003
- [8] S. Xu, J. Zhang, D. Han, J. Wang, "Data distortion for privacy protection in a terrorist analysis system," In Proceedings of the 2005 IEEE International Conference on Intelligence and Security Informatics, pp. 459-464, Atlanta, GA, May 2005.
- [9] H. Polat, W. Du, "SVD-based collaborative filtering with privacy," In the 20th ACM Symposium on Applied Computing, Track on E-commerce Technologies. Santa Fe, New Mexico, USA. March 13-17, 2005
- [10] R. Agrawal, A. Evfimievski, R. Srikant. "Information sharing across private databases," in Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, pp.86-97, San Diego, CA, 2003
- [11] C. Eckart, G. Young, "The approximation of one matrix by another of low rank," *Psychometrika*, 1:211-218, 1936
- [12] M. W. Berry, Z. Drmac, E. R. Jessup, "Matrices, vector spaces, and information retrieval," *SIAM Review*, 41(2): 335-362, 1995
- [13] L. Mirsky, "Symmetric gauge functions and unitarily invariant norms", *Quart. J. Math.*, 11:50-59, 1960
- [14] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *Journal of the Society for Information Science*, 41:391-407, 1990
- [15] D. Agrawal, C.C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," In Proceedings of the 20th ACM SIGACT-SIGMODSIGART Symposium on Principles of Database Systems, Santa Barbara, California, 2001
- [16] A. Evfimievski, J. Gehrke, R. Srikant, "Limiting privacy breaches in privacy preserving data mining," In Proceedings of PODS 2003, San Diego, CA, June, 2003
- [17] V. N. Vapnik, *Statistical Learning Theory*. John Wiley & Sons, New York, 1998
- [18] T. Joachims, *Making large-scale SVM learning practical*. Advances in kernel methods-support vector learning, B.Scholkopf, C. Burges, A. Smola, (ed.), MIT-Press, 1999