

Data Distortion for Privacy Protection in a Terrorist Analysis System

Shuting Xu, Jun Zhang, Dianwei Han, and Jie Wang

Department of Computer Science, University of Kentucky,
Lexington KY 40506-0046, USA
jzhang@cs.uky.edu

Abstract. Data distortion is a critical component to preserve privacy in security-related data mining applications, such as in data mining-based terrorist analysis systems. We propose a sparsified Singular Value Decomposition (SVD) method for data distortion. We also put forth a few metrics to measure the difference between the distorted dataset and the original dataset. Our experimental results using synthetic and real world datasets show that the sparsified SVD method works well in preserving privacy as well as maintaining utility of the datasets.

1 Introduction

The use of data mining technologies in counterterrorism and homeland security has been flourishing since the U.S. Government encouraged the use of such technologies. However, recent privacy criticism from libertarians on DARPA's Terrorism Information Awareness Program led to the defunding of DARPA's Information Awareness Office. Thus, it is necessary that data mining technologies designed for counterterrorism and security purpose have sufficient privacy awareness to protect the privacy of innocent people.

In this paper, we will discuss several data distortion methods that can be used in protecting privacy in some terrorist analysis systems. We propose a sparsified Singular Value Decomposition (SVD) method for data distortion. There are some publications about using SVD-related methods in counterterrorism data mining techniques, such as in detecting local correlation [6], social network analysis, novel information discovery and information extraction, etc. However, to the best of our knowledge, there has been no work on using SVD-related methods in data distortion. We also propose some metrics to measure the difference between the distorted dataset and the original dataset.

2 Analysis System and Data Distortion

2.1 A Simplified Model Terrorist Analysis System

A simplified model terrorist analysis system can be consisted of two parts, the data manipulation part and the data analysis part. Only the data owner or

authorized users can manipulate the original data. After the data distortion process, the original dataset is transformed into a completely different data matrix and is provided to the analysts. All actions in the data analysis part are operated on the distorted data matrix. For example, the analysts can apply data mining techniques such as classification, relationship analysis, or clustering, on the distorted data. As the data analysts have no access to the original database without the authorization of the data owner, the privacy contained in the original data is protected. k -anonymity protection [7] and its variance have been used in similar scenarios, but they do not work for data distortion.

2.2 Data Distortion

Data distortion is one of the most important parts in the proposed model terrorist analysis system. We will review two of the commonly used random value data distortion methods, as well as propose a class of SVD-based methods for data distortion in this section.

Uniformly Distributed Noise. In this method, the original data matrix A is added by a uniformly distributed noise matrix N_u [2]. N_u is of the same size as A , and its elements are random numbers chosen from the continuous uniform distribution on the interval from C_1 to C_2 .

Normally Distributed Noise. Similarly as the previous method, here the original data matrix A is added by a normally distributed noise matrix N_n , which has the same size as A [2]. The elements of N_n are random numbers chosen from the normal distribution with parameters mean μ and standard deviation σ .

SVD. Singular Value Decomposition (SVD) is a popular method in data mining and information retrieval [3]. It is usually used to reduce the dimensionality of the original dataset A . Here we use it as a data distortion method.

Let A be a sparse 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 singular value decomposition of the matrix A is $A = U\Sigma V^T$, where U is an $n \times n$ orthonormal matrix, $\Sigma = \text{diag}[\sigma_1, \sigma_2, \dots, \sigma_s]$ ($s = \min\{m, n\}$) is an $n \times m$ diagonal matrix whose nonnegative diagonal entries are in a descending order, and V^T is an $m \times m$ orthonormal matrix. We define $A_k = U_k \Sigma_k V_k^T$, where U_k contains the first k columns of U , Σ_k contains the first k nonzero diagonals of Σ , and V_k^T contains the first k rows of V^T . It has been proven that A_k is the best k dimensional approximation of A in the sense of the Frobenius norm. A_k can be seen as a distorted copy of A , and it may keep the utility of A as it can faithfully represent the original data. We define $E_k = A - A_k$.

Sparsified SVD. We propose a data distortion method based on SVD: a sparsified SVD. After reducing the rank of the SVD matrices, we set some small size entries, which are smaller than a certain threshold ϵ , in U_k and V_k^T , to zero. We refer to this operation as the dropping operation [5]. For example, given a threshold value ϵ , we drop u_{ij} in U_k if $|u_{ij}| < \epsilon$. Similarly, an element v_{ij} in V_k^T is also dropped if $|v_{ij}| < \epsilon$. Let \overline{U}_k denote U_k with dropped elements and \overline{V}_k^T

denote V_k^T with dropped elements, we can represent the distorted data matrix \bar{A}_k , with $\bar{A}_k = \bar{U}_k \Sigma_k \bar{V}_k^T$. The sparsified SVD method is equivalent to further distorting the dataset A_k . Denote $E_\epsilon = A_k - \bar{A}_k$, we have $A = \bar{A}_k + E_k + E_\epsilon$. The data provided to the analysts is \bar{A}_k which is twice distorted in the sparsified SVD method.

The SVD sparsification concept was proposed by Gao and Zhang in [5], among other strategies, for reducing the storage cost and enhancing the performance of SVD in text retrieval applications.

3 Data Distortion Measures

We propose some data distortion measures assessing the degree of data distortion which only depend on the original matrix A and its distorted counterpart, \bar{A} .

3.1 Value Difference

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 and \bar{A} to the Frobenius norm of A : $VD = \|A - \bar{A}\|_F / \|A\|_F$.

3.2 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.

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 ascending order of its value changes. Assume dataset A has n data objects and m attributes. $Rank_j^i$ denotes the rank of the j th element in attribute i , and \overline{Rank}_j^i denotes the rank of the distorted element A_{ji} . Then RP is defined as: $RP = (\sum_{i=1}^m \sum_{j=1}^n |Rank_j^i - \overline{Rank}_j^i|) / (m * n)$. If two elements have the same value, we define the element with the lower row index to have the higher rank. RK represents the percentage of elements that keep their ranks of value in each column after the distortion. It is computed as: $RK = (\sum_{i=1}^m \sum_{j=1}^n Rk_j^i) / (m * n)$. If an element keeps its position in the order of values, $Rk_j^i = 1$, otherwise, $Rk_j^i = 0$.

One may infer the content of an attribute 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: $CP = (\sum_{i=1}^m |RAV_i - \overline{RAV}_i|) / m$, where RAV_i is the rank of the average value of attribute i , while \overline{RAV}_i denotes its rank after the distortion. Similarly as RK , we define CK to measure the percentage of the attributes that keep their ranks of average value after the distortion. So it is calculated as: $CK = (\sum_{i=1}^m Ck^i) / m$. If $RAV_i = \overline{RAV}_i$, $Ck^i = 1$. Otherwise, $Ck^i = 0$.

The higher the value of RP and CP , and the lower the value of RK and CK , the more the data is distorted, and hence the more the privacy is preserved. Some privacy metrics have been proposed in literature [1, 4]. We will relate the data distortion measures to the privacy metrics in our later work.

4 Utility Measure

The data utility measures assess whether a dataset keeps the performance of data mining techniques after the data distortion, e.g., whether the distorted data can maintain the accuracy of classification, clustering, etc. In this paper, we choose the accuracy in Support Vector Machine (SVM) classification as the data utility measure. SVM is based on structural risk minimization theory [9]. In SVM classification, the goal is to find a hyperplane that separates the examples with maximum margin. Given l examples $(x_1, y_1), \dots, (x_l, y_l)$, with $x_i \in R^n$ and $y_i \in \{-1, 1\}$ for all i , SVM classification can be stated as a quadratic programming problem: minimize $\frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i$, subject to (1) $y_i(\langle w, x_i \rangle + b) \leq 1 - \xi_i$, (2) $\xi_i \geq 0$, (3) $C > 0$, where C is a user-selected regularization parameter, and ξ_i is a slack variable accounting for errors. After solving the quadratic programming problem, we can get the following decision function: $f(x) = \sum_{i=1}^l \alpha_i y_i \langle x_i, x \rangle + b$, where $0 \leq \alpha_i \leq C$.

5 Experiments and Results

We conduct some experiments to test the performance of the data distortion methods: SVD, sparsified SVD (SSVD), adding uniformly distributed noise (UD) and adding normally distributed (ND) noise.

5.1 Synthetic Dataset

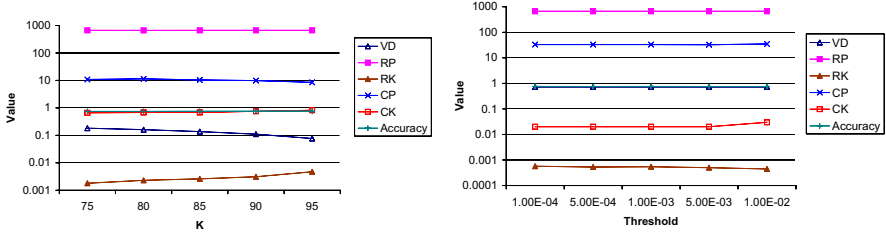
First, we compare the performance of the four data distortion methods using a synthetic dataset. The dataset is a 2000 by 100 matrix (Org), whose entries are randomly generated numbers within the interval [1,10] obeying a uniform distribution. We classify the dataset into two classes using a randomly chosen rule. The uniformly distributed noise is generated from the interval [0, 0.8]. The normally distributed noise is generated with $\mu = 0$ and $\sigma = 0.46$. The parameters of UD and ND are chosen so that the VD value of UD, ND, and SVD is approximately equal. For SVD and SSVD, the rank k is chosen to be 95, and in SSVD, the dropping threshold value ϵ is 10^{-3} .

We can see in Table 1 that the SVD-based methods achieve a higher degree of data distortion. And SSVD is better than SVD in all the position distortion measures. The Accuracy column in Table 1 shows the percentage of the correctly classified data records. Here all the distorted methods obtain the same accuracy as using the original data.

Figure 1 illustrates the influence of the parameters in the SVD-based methods. With the increase of k in SVD, VD and CP decrease while RK , CK and

Table 1. Comparison of distortion methods for the synthetic dataset

Data	VD	RP	RK	CP	CK	Accuracy
Org	-	-	-	-	-	76%
UD	0.0760	662.8	0.0058	0	1	76%
ND	0.0763	661.6	0.0067	0	1	76%
SVD	0.0766	664.0	0.0047	8.5	0.82	76%
SSVD	0.7269	666.6	0.0005	33.2	0.02	76%

(a) The influence of k in SVD(b) The influence of threshold ϵ **Fig. 1.** The influence of the parameters in the SVD-based methods**Table 2.** Comparison of the distortion methods using a real world dataset

Data	Classification 1						Classification 2					
	VD	RP	RK	CP	CK	Accuracy	VD	RP	RK	CP	CK	Accuracy
Org	-	-	-	-	-	67%	-	-	-	-	-	67%
UD	0.0566	0	1	11.4	0.15	67%	0.0575	31.9	0.0166	9.5	0.07	66%
ND	0.0537	31.9	0.0298	12.2	0.27	66%	0.0566	34.1	0.0390	12.0	0.07	64%
SVD	0.0525	31.2	0.0251	12.2	0.12	70%	0.0525	31.2	0.0251	12.2	0.12	70%
SSVD	1.0422	37.5	0.0066	13.1	0.05	69%	1.3829	35.0	0.0090	11.5	0.02	65%

Accuracy increase. But with the increase of ϵ in SSVD ($k = 95$), there is no observable trend in data distortion or utility measures.

5.2 Real World Dataset

For a real world dataset, we download information about 100 terrorists from a terrorist analysis web site [8]. We selected 42 attributes, such as their nationality, pilot training, locations of temporary residency, meeting attendance, etc. To test the real world dataset, the uniformly distributed noise is chosen from the interval $[0, 0.09]$. The normally distributed noise is generated with $\mu = 0$ and $\sigma = 0.05$. The rank k for SVD and SSVD is chosen to be 25, and ϵ in SSVD is 10^{-3} .

In Classification 1, we classify the terrorists into two groups, those are related with Bin Laden and those are not. Here the SVD-based methods improve the accuracy a little bit. For data distortion, SSVD is the best for all the measures.

In Classification 2, the terrorists are grouped according to whether or not they have relationship with the terrorist organization Al Qaeda. Here SVD is the best for the classification accuracy, the other three methods decrease the accuracy slightly. For the data distortion measures, SSVD again works best.

6 Concluding Remarks

We proposed to use the sparsified SVD method for data distortion in a simplified model terrorist analysis system. The experimental results show that the sparsified SVD method works well in preserving privacy as well as maintaining utility of the datasets.

References

1. Agrawal, D., Aggarwal, C.C.: On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems.*, Santa Barbara, California, USA, (2001)
2. Agrawal, R., Srikant, R: Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD*, Dallas, Texas, (2000)
3. Deewester, S., Dumais, S., *et al.*: Indexing by latent semantic analysis, *J. Amer. Soc. Infor. Sci.*, **41** (1990) 391–407
4. Evfimievski, A., Gehrke, J., Srikant, R.: Limiting privacy breaches in privacy preserving data mining. In *Proceedings of PODS 2003*, San Diego, CA, June, (2003)
5. Gao, J., Zhang, J.: Sparsification strategies in latent semantic indexing. In *Proceedings of the 2003 Text Mining Workshop*, San Francisco, CA, (2003) 93–103
6. Skillicorn, D.B.: Clusters within clusters: SVD and counterterrorism. In *Proceedings of 2003 Workshop on Data Mining for Counter Terrorism and Security*, 12 pages, San Francisco, CA, May 3, (2003)
7. Sweeney, L.: k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, **10** (2002) 557–570
8. www.trackingthethreat.com
9. Vapnik, V.N.: *Statistical Learning Theory*. John Wiley & Sons, New York, (1998)