JPI
Journal of Pattern Intelligence
Volume 1, Issue 1, 2011, PP-05-07
Online available at http://www.bioinfo.in/contents.php?id=49

PRIVACY PRESERVING CLASSIFICATION USING WAVELET BASED DATA DISTORTION

VINOD PATEL
Department Computer Science & Engineering, SATI, Vidisha, MP, India
*Corresponding author. E-mail: vinod.gwl@gmail.com, 9826234810

Received: January 08, 2011; Accepted: January 12, 2011

Abstract- The rapid development of modern data collection and data warehouse technologies, data mining is becoming more and more a standard practice. Data distortion is a critical component to preserve privacy in security-related data mining applications. We propose a wavelet-based method for data distortion, and compare it with the Singular Value Decomposition (SVD) based method. The experimental results show that the wavelet based method can obtain similar performance as SVD based method in preserving privacy as well as maintaining utility of the datasets, however, the computational time used by the wavelet based method is much less than the SVD based method. We conclude that the Wavelet based method is a very promising data distortion method.

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 [1].However, government access to and use of personal information in commercial databases raise concerns about the protection of privacy and due process [2]. 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. Extracting valid data mining results while still preserving privacy of datasets is a major challenge for existing data mining algorithms.

Anonymity [3] is a technique to remove identifiers (e.g. names, social security numbers, addresses, etc.) to protect privacy. However, the released data after removing identifiers may contain information that can be linked with other datasets to re-identify individuals or entities. . Data distortion is one of the important methods to avoid this kind of privacy leakage. After data distortion, on one hand, the distorted data are so different from the original data that the original data could not be derived from the distorted data. On the other hand, the distorted data must keep the main property of the original data so the data analysis algorithms could obtain similar performance as working on the original data. Many data distortion methods have been proposed in liter- ature. However, most of them are used in statistical database to maintain some statistical characteristics of datasets, such as mean, sum, variance, etc. They may not work well in keeping the performance of data mining algorithms like classification and clustering. It is shown in [4] that the SVD and sparsified SVD method is very efficient in keeping both data privacy and data utility, but SVD based methods are expensive in computational costs (time complexity is 0(n3)). We propose to Wavelet based method 0(n) which is better than SVD based methods. To the best of our knowledge, no similar Wavelet based privacy-preserving data distortion method has been reported. We have conducted experiments and the results show that the Wavelet method can obtain similar performance as SVD in preserving privacy as well as maintaining utility of the datasets, but the Wavelet method is much faster than the SVD.

Assumptions

The matrix representation (vector-space format) is one of the most popular ways to encode the object-attribute relationships in many real-life datasets. In this format, a 2-dimensional (2D) matrix is used to store the dataset in which each row of the matrix stands for an individual object and each column represents a particular attribute of these objects. Apparently, in this matrix, the privacy is a set of all confidential attributes represented by columns and all secret objects represented by rows. In such a matrix, we assume that every element is fixed, discrete, and numerical. Any missing element is not allowed.

Wavelet decomposition

In mathematical terms, a discrete wavelet transformation (DWT) is a wavelet transformation for which the input discrete samples are divided into approximation coefficients and detail coefficients, which correspond to the low frequency and high frequency decompositions of the original samples, respectively. Such wavelet decomposition process is applied recursively with high and low passing filters on the approximation coefficients of the previous level and then down-sampled. A brief transformation process is illustrated in Figure 1:

According to the above graph and introduction in [9], the 2D DWT decomposition firstly uses the high pass filters and low pass filters to process the original entire a*b matrix. Then the b columns with length a are passed to the first filter which operates on the columns horizontally These filter outputs are then downsampled by a half. In other words, a half of the column elements are thrown away. The remaining half column outputs are further decomposed using the second filter which considers the in put data as transposed a/2 rows of length b. Similarly, the second filter throws away a half of the remaining column elements. After these processes, one approximation coefficient and three detail coefficient submatrices are produced at this level. For the next level DWT decomposition process, it just recursively processes the approximation coefficients of the previous level as the new input matrix. Through the above description, it is clear that the discrete wavelet transformation only depends on the maximum decomposition level and the filters (wavelet basis). For a given wavelet basis, the maximum number of decomposition levels, n, of DWT mainly depends on the dimensions of the input signals. Although the standard 2D DWT decomposition needs the input matrix represented in 2a * 2b dimensions, where a and b are two integers, we can still deal with matrices of any dimension size as follows [9].

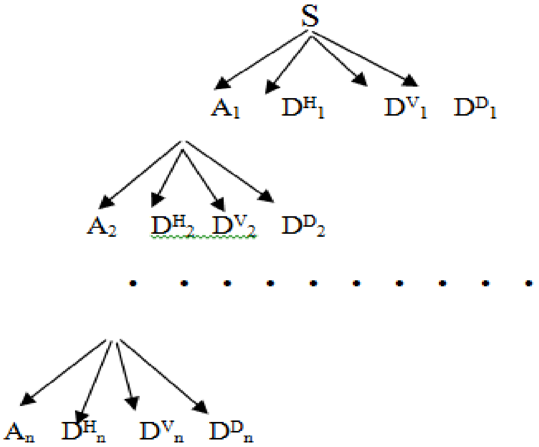

Fig. 1- The DWT decomposition schema. S is the original 2D matrix. Ai contains the i_-th level approximation coefficients; DHi , DVi and DDi are the horizontal, vertical and diagonal detail coefficients, respectively.

For any a*b dimension matrix, the DWT decomposition can process and downsample all columns through the standard DWT filters, but the rows may not be sufficiently decomposed (for simplicity, we assume that a>b). However, in the data distortion techniques, it does not matter because we can still suppress the entire detail coefficients and then reconstruct them and the approximation coefficients, to be introduced in the next section, to successfully distort the whole original data if n is large enough.

Thus, the maximum number of decomposition levels, n, of a data matrix of any dimension a*b is defined as:

n= ┌ log2 min(a,b)┐ .

Experimental results

Data Privacy Measures

We choose the five data distortion privacy measure metrics, VD, RP, RK, CP and CK, first defined in [10], and then in [3], to evaluate the proposed data distortion methods. We also develop a new measure metric, RangePer, to measure the preserved attribute values and the privacy range. The objective of these measure metrics is to evaluate the possibility of estimating the true values and range of the original data from the distorted data [5, 6, 8].

In brief, VD value is the ratio of the Frobenius norm of the difference between the original matrix S and the distorted matrix S* to the Frobenius norm of S. RP value presents the ratio of the average change of ranks for all attributes to the number of total elements of the matrix. RM denotes the percentage of elements, which keep their ranks of values in each column after the distortion. CP stands for change of ranks of the average values of the attributes. CK is defined to evaluate the percentage of the attributes that keep their ranks of average values after the distortion. The detailed definitions of these measure metrics are presented in [10].

According to their definitions, we know that a larger VD, RP, CP and RangePer value, and a smaller RK and CK value refer to a better privacy-preserving level.

Distortion Experiments

In the experiment section, we choose two real-life Databases obtained from the University of California Irvine (UCI), Machine Learning Repository [7]. They are the Wisconsin breast cancer original dataset (WBC) donated by Olvi Mangasarian, and the Wisconsin breast cancer diagnostic database (WDBC) donated by Nick Street. The summary of the two original databases is in Table 1.

Table 1-The summary of the WBC and WDBC databases

Database	Number of Instances	Number of Features	Number of Classes
WBC	699	9	2
WDBC	569	30	2

In addition to the summary, the attributes of the two databases only have numerical values and no missing value. (In the original WBC database, there are a few missing values in the sixth column. We replace these missing values by 1 if the object belongs to the malignant class and 2 if the object is in the benign class, according to the standard classification provided by the UCI Repository.) Tables 2 and 3 demonstrate our privacy preserving distortion experimental results. In the experiments we choose the SVD-based data distortion method for comparison [10, 3]. We use the simplest SVD data distortion method, i.e., no sparsification strategy is implemented.

In the SVD data distortion experiment, we choose the reduced rank k value to be 5 in WBC and 15 in WDBC. In the wavelet transformation (S) of both Tables 2 and 3,we choose the Haar basis wavelet for decomposition. The results of our experiments, especially the run Time, are averaged values of five repeated experiments, obtained from a Dell desktop workstation with a P4-2.8GHz CPU, 40G hard disk, and 256MBmemory inMatlab 6.5.0.180913a with a Linux operation system. For the results reported in Tables 2 and 3, the support vector machine (SVM light) with a five-fold cross validation [11, 12] is employed as the standard classification tool which is used to measure the data utility accuracy in our experiments. According to Tables 2 and 3, we can draw the following conclusions:

Table 2- Performance comparison of SVD and wavelet transformation on WBC.

Database	VD	RP	RK	CP	CK	Run Time (Seconds)	Accuracy
Original							96.0%
SVD	0.2080	239.4	0.006358	1.556	0.4444	0.07882	95.9%
Wavelet	0.2557	238.6	0.004769	1.333	0.5556	0.03081	96.0%

Table 3- Performance comparison between SVD and wavelet transformation on WDBC

Database	VD	RP	RK	CP	CK	Run Time (Seconds)	Accuracy
Original							85.4 %
SVD	0.000035	121.3	0.3454	0	1.0000	0.13880	85.4 %
Wavelet	0.000843	165.3	0.1083	4.800	0.4000	0.05166	85.4 %

The data accuracy level of the wavelet-based distortion methods is as good as that of the SVD and the original data The run time of the wavelet-based distortion methods is faster than that of the SVD-based method Most of the privacy preservation metrics show that the wavelet-based distortion methods can keep a Better privacy level than the standard SVD-based Method. In the three wavelet-based distortion methods their analysis accuracy and privacy preserving And run time performances are similar

Conclusion

In this paper, we propose a class of new privacy preserving data distortion methods based on wavelet transformation Through experiments, we demonstrate that the wavelet-based data distortion methods can effectively and efficiently render a balance between data utilities and data privacy beyond its remarkable fast run time in comparison with the SVD-based distortion method which has already been demonstrated as a promising privacy preserving data distortion method [10]. Further research work along this line can be carried out in terms of data distortion and data utilities. Comparing the wavelet transformation based data distortion methods with the more aggressive SVD-based data distortion methods, such as the sparsified SVD-based methods [10, 3], should be of interest.

References

[1] Taipale K.A. (2003) Colum. Sci. & Tech. Law Rev., 5:1-83.

[2] Dempsey J.X. and Rosenzweig P. (2004) Legal Memorandum #11, the Heritage Foundation.

[3] Klosgen W. (1995) Proceeding of the First International Conference on Knowledge Discovery and Data Mining (KMM-95,), 186–191, Montreal, Canada.

[4] Xu S., Zhang J., Han D., Wang J. (2006) Knowledge and information system (KAIS) Journal, 10 (3), 383-397.

[5] Agarwal D. and Agarwal C.C. (2001) Proceeding of the 20th ACM SIGACT-SIGMODSIGART Symposium on Principals of Databases System, 247-255, Santa Barbara.CA.

[6] Evfimievski A, Gehrke J. and Srikant R. (2003) *Proceedings of PODS 2003*, San Diego, CA, 211-222.

[7] Newman D.J., Hettich S., Blake C.L. and Merz C.J. (1998) *UCI repository of machine learning databases. University of California, Department of Information and Computer Science, Irvine, CA*.

[8] Polat H. and Du W. (2005) *SVD-based collaborative filtering with privacy. In the 20th ACM Symposium on Applied Computing, Track on E-commerce Technologies*, pp. 791-795, Santa Fe, NM, 2005.

[9] Weeks M. and Bayoumi M.A. (2002) *IEEE Transactions on Signal Processing*, 50(8):2050-2063.

[10] Xu S., Zhang J., Han D. and Wang J. (2005) *Proceedings of the 2005 IEEE International Conference on Intelligence and Security Informatics*, pp. 459-464.

[11] Joachim's T. (2002) *Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms. Kluwer Academic Publisher, Norwell, MA*.

[12] Joachims T. (1999) *Making large-scale SVM learning practical. In Advances in Kernel Methods – Support Vector Learning, B. Scholkopf and C. Burges and A. Smola (ed.), MIT Press, Cambridge, MA*.