IJCES International Journal of Computer Engineering Science , Volume1 Issue2, November 2011
ISSN : 2250:3439
https://sites.google.com/site/ijcesjournal
http://www.ijces.com/

Preserving Privacy in Data Mining using Data Distortion Approach

Mrs. Prachi Karandikar [#], Prof. Sachin Deshpande *

[#] *M.E. Comp,VIT , Wadala, University of Mumbai*

* *VIT Wadala ,University of Mumbai*

1. prachiv21@yahoo.co.in

2.sachin.deshpande@vit.edu.in

Abstract. Data mining, *the extraction of hidden predictive information from large databases,* is nothing but discovering hidden value in the data warehouse. Because of the increasing ability to trace and collect large amount of personal information, privacy preserving in data mining applications has become an important concern. Data distortion is one of the well known techniques for privacy preserving data mining. The objective of these data perturbation techniques is to distort the individual data values while preserving the underlying statistical distribution properties. These techniques are usually assessed in terms of both their privacy parameters as well as its associated utility measure. In this paper, we are studying the use of non-negative matrix factorization (NMF) with sparseness constraints for data distortion.

Keywords: **Data Mining, Privacy, Data distortion, NMF, Sparseness**

INTRODUCTION

Data Mining [1] ,is the extraction of hidden predictive information from large databases, is a powerful technology with great potential to help companies focus on the most important information in their data warehouses. Several data mining applications deal with privacy-sensitive data such as financial transactions, and health care records. Because of the increasing ability to trace and collect large amount of personal data, privacy preserving in data mining applications has become an important concern. There is a growing concern among citizens in protecting their privacy . Data is stored either in a centralized database or in a distributed database . According to its storage there are various privacy preserving techniques used . Among the techniques that are used for privacy preserving data mining are: Generalization , Data Sanitation , Data distortion , Blocking & Cryptography techniques . We are focusing our study on the latter approach i.e. Data distortion via data perturbation . The objective of data perturbation is to distort the individual data values while preserving the underlying statistical distribution properties. Theses data perturbation techniques are usually assessed in terms of both their privacy parameters as well as its associated utility measure. While the privacy parameters present the ability of these techniques to hide the original data values, the data utility measures assess whether the dataset keeps the performance of data mining techniques after the data distortion. Our objective is to study the use of truncated non-negative matrix factorization (NMF) with sparseness constraints for data perturbation. The rest of the paper is organized as follows. In section 2, we review the non-negative matrix factorization technique. The data distortion and the utility measures which can be used are reviewed in section 3 and section 4 respectively. Experimental results are given in section 5. Conclusion and future scope are given in section 6.

Nonnegative matrix factorization

Non negative matrix factorization (NMF) [8] refers to a class of algorithms that can be formulated as follows: Given a nonnegative $n \times r$ data matrix, V . NMF finds an approximate factorization $V \approx WH$ where W and H are both non negative matrices of size $n \times m$ and $m \times r$ respectively. The reduced rank m of the factorization is generally chosen so that $n\ r<m<nr$ and hence the product WH can be regarded as a compressed form of the data matrix V . The optimal choices of matrices W and H are defined to be those non-negative matrices that minimize the reconstruction error between V and WH . Various error functions

have been proposed. The most widely used is the squared error (Euclidean distance) function (i) and K-L Divergence function (ii) etc [2] .

$$E(W, H) = \sum_{i,j} (V_{ij} (WH)_{ij})^2 \quad . \tag{1}$$

$$D(A||B) = \sum_{ij} \left(A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij} \right) \tag{2}$$

Non negative matrix factorization requires all entries of both matrices to be non negative, i.e., the data is described by using additive components only.

NMF with Sparseness Constraint

Several measures for sparseness have been proposed. The sparseness of a vector *X* of dimension *n* is given by [8]:

$$S_x = \frac{\sqrt{n} - (\sum |x_1|) / \sqrt{\sum x_1^2}}{\sqrt{n} - 1} \tag{1}$$

Usually, most of NMF algorithms produce a sparse representation of the data. Such a representation encodes much of the data using few active components. However, the sparseness given by these techniques can be considered as a side-effect rather than a controlled parameter, i.e., one cannot in any way control the degree to which the representation is sparse. Our aim is to constrain NMF to find solutions with desired degrees of sparseness. The sparseness constraint can be imposed on either W or H or on both of them. For example, a doctor analyzing a dataset that describes disease patterns, might assume that most diseases are rare (hence sparse) but that each disease can cause a large number of symptoms. Assuming that symptoms make up the rows of her matrix and the columns denote different individuals, in this case it is the coefficients which should be sparse and the basis vectors unconstrained. We have studied the projected gradient descent algorithm for NMF with sparseness constraints proposed in [7] .

Truncation on NMF with Sparseness Constraint

In order to control the degree of achievable data distortion, the elements in the sparsified *H* matrix with values less than a specified truncation threshold $\sum$ are truncated to zero.
Thus the overall data distortion can be summarized as follows:
(i) Perform sparsified NMF with sparse constraint h_s on *H* to obtain H_{Sh} (ii) Truncate the elements in H_{Sh} that are less than $\sum$ to obtain $H_{Sh,\sum}$. The perturbed dataset is given by $W H_{Sh,C}$. Thus the new dataset is basically distorted twice by our proposed algorithm that has three parameters: the reduced rank *m* , the sparseness parameter h_s and the truncation threshold C.

DATA DISTORTION MEASURES

Throughout this work, we adopt the same set of privacy parameters proposed in [5]. The value difference (VD) parameter is used as a measure for value difference after the data distortion algorithm is applied to the original data matrix. Let V and $_V$ denote the original and distorted data matrices respectively. Then, VD is given by :

$$VD = \| V - V \| / \| V \| , \quad (1)$$

where $\| . \|$ denotes the Frobenius norm of the enclosed argument.
After a data distortion, the order of the value of the data elements also changes. Several metrics are used to measure the position difference of the data elements. For a dataset $V$with n data object and m attributes, let i , $Rank\ j$ denote the rank (in ascending order) of the j^{th} element in attribute i.
Similarly, let $Rank^i_j$ denote the rank of the corresponding distorted element. The RP parameter is used to measure the position difference. It indicates the average change of rank for all attributes after distortion and is given by

$$RP = \frac{1}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} \left| Rank^i_j - \overline{Rank^i_j} \right| \quad (2)$$

RK represents the percentage of elements that keeps their rank in each column after distortion and is given by

$$RK = \frac{1}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} Rk^i_j \quad (3)$$

where $Rk^i_j = 1$. If an element keeps its position in the order of values, otherwise $Rk^i_j = 0$.
Similarly, the CP parameter is used to measure how the rank of the average value of each attributes varies after the data distortion. In particular, CP defines the change of rank of the average value of the attributes and is given by

$$CP = \frac{1}{m} \sum_{i=1}^{m} \left| RankVV_i - \overline{RankVV_i} \right| \quad (4)$$

where $RankVV_i$ and $RankVV_i$ denote the rank of the average value of the i^{th} attribute before and after the data distortion, respectively. Similar to RK, CK is used to measure the percentage of the attributes that keep their ranks of average value after distortion. From the data privacy perspective, a good data distortion algorithm should result in a high values for the RP and CP parameters and low values for the RK and CK parameters.

UTILITY MEASURE

The data utility measures assess whether the dataset keeps the performance of data mining techniques after the data distortion. The accuracy of a simple K-nearest neighborhood (KNN) [9] can be used as data utility measure.

EXPERIMENTAL RESULTS

In order to test the performance of our proposed method, we conducted a series of experiments on some real world datasets. In this section, we present a sample of the results obtained when applying our technique to the original Wisconsin breast cancer downloaded from UCI machine Learning Repository [10].
For the breast cancer database, we used 569 observations and 30 attributes (with positive values) to perform our experiment. For the accuracy calculation , TANAGRA data mining tool has been used .Throughout the experiment KNN classifier has been used with K=19.

Using Squared error function –

From the Figure 1, it is clear that $m = 2$ provides the best choice with respect to the privacy parameters.

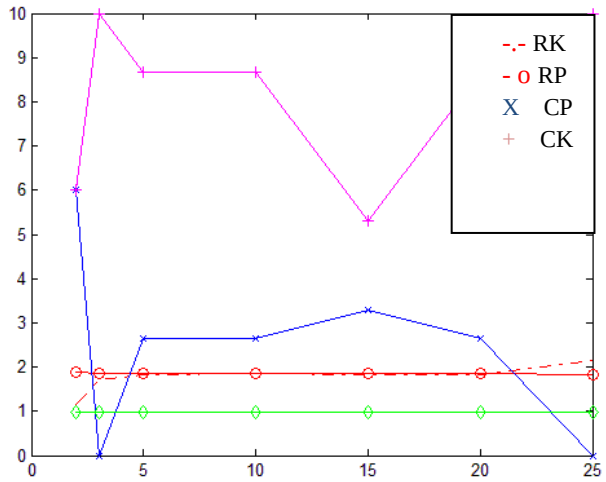

Fig. 4. Effect of the reduced rank *m* on the privacy parameters.

Table 1 shows the how the privacy parameters and accuracy vary with the sparseness constraint *Sh* .

S_h	RP	RK	CP	CK	VD	Acc
0	180.28	0.0371	0.266	0.866	0.0431	93.14 %
0.45	185.90	0.0153	0.266	0.866	0.3751	92.23 %
0.65	187.41	0.0115	0.6	0.6	0.98	91.21 %
0.75	186.43	0.015	0.26	0.86	0.99	92.61 %

TABLE I EFFECT OF THE SPARSENESS CONSTRAINT ON THE PRIVACY PARAMETERS AND ACCURACY

Table 2 shows the effect of threshold ε on the privacy parameters. From the table, it is clear that there is a trade-off between the privacy parameters .

ε	RP	RK	CP	CK	VD
0.035	184.90	0.0179	0.2667	0.8667	0.1398
0.040	195.54	0.0173	0.2667	0.8667	0.1044
0.045	196.11	0.0166	0.2667	0.8667	0.0952
0.050	197.71	0.0144	0.2667	0.8667	0.0830

TABLE II EFFECT OF THRESHOLD ε ON THE PRIVACY PARAMETERS AND ACCURACY

Using K-L Divergence error function –

From the Figure 2, it is clear that $m = 3$ provides the best choice with respect to the privacy parameters. So, we fixed $m = 2$ throughout the rest of our experiments with this dataset.

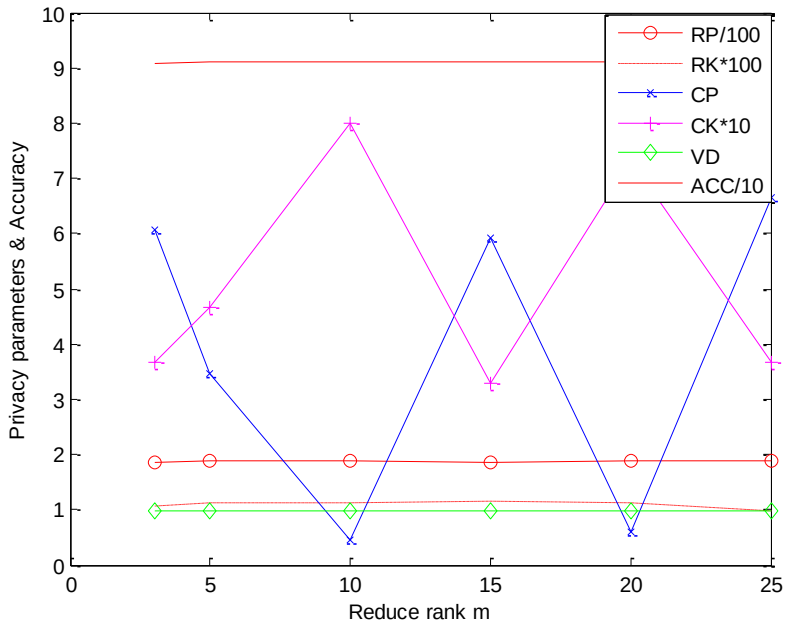

Fig. 5. Effect of the reduced rank m on the privacy parameters

Table 3 shows the how the privacy parameters and accuracy vary with the sparseness constraint S_h .

S_h	RP	RK	CP	CK	VD	Acc
0	187.51	0.0113	0	1	0.0536	92.14%
0.3	187.28	0.0103	3.2	0.53	0.999	90.56 %
0.65	185.93	0.0108	6.06	0.366	0.999	90.86 %

TABLE III EFFECT OF THE SPARSENESS CONSTRAINT ON THE PRIVACY PARAMETERS AND ACCURACY

From the results in Table 3, it is clear that S_h = 0.65 not only improves the values of the privacy parameters, but also improves the classification accuracy.

Table 4 shows the effect of threshold ε on the privacy parameters and accuracy. From the table, it is clear that there is a trade-off between the privacy parameters and the accuracy.

ε	**RP**	**RK**	**CP**	**CK**	**VD**
0.005	186.09	0.0108	6.066	0.3667	0.999
0.01	187.064	0.0108	6.066	0.3667	0.999
0.05	194.84	0.0083	6.066	0.3667	0.999
0.1	198.51	0.0064	6.066	0.3667	0.999

TABLE IV EFFECT OF THRESHOLD ε ON THE PRIVACY PARAMETERS

CONCLUSION AND FUTURE SCOPE

Non-negative matrix factorization with sparseness constraints can provide an effective data perturbation tool for privacy preserving data mining. In order to test the performance of the proposed method we would like to conduct a set of experiments on some standard real world data sets . While using the above mentioned privacy parameters , we would like to test the ability of these techniques to hide the original data values .

REFERENCES

[1] M. Chen, J. Han, and P. Yu, "*Data Mining: AnOverview from a Database Prospective*", IEEE Trans. Knowledge and Data Engineering, 8, 1996.Z.

[2] 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] Saif M. A. Kabir1, Amr M. Youssef2 and AhmedK.Elhakeem1 Concordia University, Montreal, Quebec, Canada , " On data distortion for privacy preserving data mining "

[4] Rakesh Agrawal and Ramakrishnan Srikant, "Privacy-preserving data mining," In Proceeding of the ACM SIGMOD Conference on Management of Data, pages 439–450, Dallas, Texas, May 2000. ACM Press.

[5] Shuting Xu, Jun Zhang, Dianwei Han, and Jie Wang, Data distortion for privacy protection in a terrorist Analysis system. P. Kantor et al (Eds.):ISI 2005, LNCS 3495, pp.459-464, 2005

[6] V. P. Pauca, F. Shahnaz, M. Berry and R. Plemmons. Text Mining using non-negative Matrix Factorizations, Proc. SIAM Inter. Conf. on Data Mining, Orlando, April, 2004.

[7] Patrik O. Hoyer. Non-negative Matrix Factorization with Sparseness Constraints. Journal of Machine Learning Research 5 (2004) 1457–1469

[8] D. D. Lee and H. S. Seung. Algorithms for nonnegative matrix factorization. In Advances in Neural Information Processing 13 (Proc. NIPS 2000). MIT Press, 2001.
[9] R.Duda,P.Hart,and D. Stork, "Pattern Classification," John Wiley and Sons, 2001.