International Journal of Computer and Communication Technology

Volume 4 | Issue 4 Article 2

October 2013

Min Max Normalization Based Data Perturbation Method for Privacy Protection

YOGENDRA KUMAR JAIN
Head of the Department Computer Science & Engineering Samrat Ashok Technological Institute Vidisha (M. P.) 464001 India, ykjain_p@yahoo.co.in

SANTOSH KUMAR BHANDARE
Research Scholar Computer Science & Engineering Samrat Ashok Technological Institute Vidisha (M. P.) 464001 India, santosh.mits@gmail.com

Min Max Normalization Based Data Perturbation Method for Privacy Protection

YOGENDRA KUMAR JAIN
Head of the Department
Computer Science & Engineering
Samrat Ashok Technological Institute
Vidisha (M. P.) 464001 India
e-mail: ykjain_p@yahoo.co.in

SANTOSH KUMAR BHANDARE
Research Scholar
Computer Science & Engineering
Samrat Ashok Technological Institute
Vidisha (M. P.) 464001 India
e-mail: santosh.mits@gmail.com

***Abstract* - Data mining system contain large amount of private and sensitive data such as healthcare, financial and criminal records. These private and sensitive data can not be share to every one, so privacy protection of data is required in data mining system for avoiding privacy leakage of data. Data perturbation is one of the best methods for privacy preserving. We used data perturbation method for preserving privacy as well as accuracy. In this method individual data value are distorted before data mining application. In this paper we present min max normalization transformation based data perturbation. The privacy parameters are used for measurement of privacy protection and the utility measure shows the performance of data mining technique after data distortion. We performed experiment on real life dataset and the result show that min max normalization transformation based data perturbation method is effective to protect confidential information and also maintain the performance of data mining technique after data distortion.**

Keywords: Privacy preserving, normalization, data perturbation, classification, data mining

I. INTRODUCTION

Data mining [1] is the process of finding pattern from large amount of data using tool such as classification. The problem of privacy preserving is very important issue in data mining system. There are a lot of data mining application deal with privacy and security concern. Data mining system contain large amount of private and secure data i.e. financial, criminal and healthcare records. These records cannot be share to every one so privacy of data is required for avoiding privacy leakage. There are a lot of research has been done in privacy preserving data mining based on randomization, secure multiparty computations, perturbation and Anonymity including K-anonymity and l-diversity. In this paper we discuss the data perturbation technique in which confidential numerical attributes are distorted for privacy protection in classification analysis. Data perturbation is one of the best techniques for privacy preserving data mining system. In data perturbation the individual data values are distorted before data mining application. The privacy parameters [2] are used for measurement of the degree of privacy protection. These parameters also show the capability of this technique to concealing the original data. The data utility measures show the performance of data mining technique after data distortion. In this paper we proposed min max normalization transformation based data distortion method. Our experimental results show that the proposed method is effective to concealing the confidential information and also preserve the performance of data mining technique after data distortion.

II. RELATED WORK

There has been a lot of privacy preserving data mining literatures. These literatures can divide into two categories. In the first category, methods modify the data mining algorithms so that without knowing the exact values of data, they allow data mining operations on distributed dataset. In the second category, methods are modifying the values of the datasets to protect privacy of data values. In this category there are several research has been done in data distortion or data perturbation are as follow:

In the year 1985, Liew et al., they proposed data distortion method based on probability distribution [17]. This method involves three steps: (i) identification of the underlying density function, (ii) generation of a distorted series from the density function, and (iii) mapping of the distorted series onto the original series.

In the year 2000, Agrawal R. et al.,[18] they proposed an additive data perturbation method for building decision tree classifiers. Every data element is randomized by adding some noise. These random noise chosen independently by a known distribution like Gaussian distribution. The data miner rebuilds the distribution of the original data from its distorted version.

In the year 2002, Sweeney L. et al., [19] In this paper the k-Anonymity model consider the problem that a data owner wants to share a collection of person-specific data without revealing the identity of an individual. This goal is achieve by data generalization and suppression methods are used to protect the confidential information. This paper also examines the re-identification attacks.

In the year 2005, Chen et al., they proposed a rotation based perturbation method [20]. The proposed method maintains zero loss of accuracy for many classifiers. Experimental results show that the rotation perturbation can

greatly improve the privacy quality without sacrificing accuracy.

In the year 2006, Wang et al., has used the Non-negative matrix factorization (NNMF) for data mining [8]. In this work, they combined non-negative matrix decomposition with distortion processing. The presented method have two important aspects (i) non-negative matrix factorization (NMF) is used to provide a least square compression version of original datasets and (ii) Using iterative methods to solve the least square optimization problem is provided an attractive flexibility for data administrator. The presented result given that the careful choice of iterative parameter settings, two sparse non-negative factors can solve by some efficient algorithms. Alternating least square using projected gradients in computing NNMF converges faster than multiplicative update methods. Iterative NMF based distortion method provides good solution for data mining problem on the basis of discriminate functions.

In the year 2006, Li et al., proposed kd-tree based perturbation Method [9]. This method recursively partitions a data set into smaller subsets, so that the data records within each subset are more homogeneous after each partition. Then the confidential data in each final subset are perturbed using the subset average. This method is both efficient and effective, due to the recursive divide-and-conquer technique used. The experimental results show that the proposed method is effective.

In the year 2006, Xu et al., proposed Singular value decomposition (SVD) based data distortion strategy for privacy protection [15]. In this work they propose a sparsified Singular Value Decomposition (SVD) method for data distortion. They conducted experiment on synthetic and real world datasets and the experimental result show that the sparsified SVD method is effective in preserving privacy as well as maintaining the performance of the datasets.

In the year 2006, Wang et al., they proposed a new data distortion method based on Structural Partition and SSVD for Privacy Preservation [16].
They used object-based partition, feature-based partition and hybrid partition. The experimental results show that feature-based partition is a feasible and efficient solution for privacy-preserving data mining.

In the year 2007, Saif et al., also used non-negative matrix factorization for data perturbation [10]. They investigated the use of truncated non-negative matrix factorization (NMF) with sparseness constraints. The experimental results show that the Non-negative matrix factorization with sparseness constraints provides an efficient data perturbation tool for privacy preserving data mining. The privacy parameter used in the proposed work provides some indication on the ability of these techniques for concealing the original data values.

In the year 2007, Wang et al., they proposed several efficient and flexible techniques to address accuracy issue, in privacy preserving data mining through matrix factorization [21]. Experimental results indicate that for centralized datasets with numerical attributes, matrix factorization-based distortion strategies achieve a satisfactory performance.

In the year 2007, Xu et al., has used the Fast Fourier Transform (FFT) for data perturbation [11]. The dataset is distorted or perturbed by using Fast Fourier Transform (FFT) for privacy protection of data values.

In the year 2008, Liu et al., has used the wavelet transformation for data distortion or data perturbation to preserve the privacy of data [12]. Privacy preserving strategy based on wavelet perturbation; keep the data privacy and data statistical properties and data mining utilities at the same time. The results show that presented method keep the distance before and after data perturbation and it also preserve the basic statistical properties of original data while maximizing the data utilities.

In the year 2009, Lin et al., has presented a method for data perturbation. In this method, the data matrix is vertically partitioned into several sub-metrics and held by different owners [13]. For perturbing their individual data, each data holder can randomly and independently choose a rotation matrix. The presented results show that random rotation based method for data perturbation preserve the data privacy without affecting the accuracy.

In the year 2010, Peng et al., used combine data distortion strategies for privacy preserving data mining [14]. They designed four schemes via attribute partition, with single value decomposition (SVD), non-negative matrix factorization (NMF), discrete wavelet transformation (DWT) for distortion of submatrix of the original dataset for privacy preserving. The basic idea of the proposed strategies was to perform distortion on sub matrices of original dataset using different method. The results show that proposed method was very efficient in maintaining data privacy as well as data utility in comparison to the individual data distortion techniques such as SVD, NMF and DWT.

III. ASSUMPTIONS

The object-attribute relationship of real life data sets are encode into vector – space format [3]. In this format a 2-dimentional is used to share the dataset. Row of the matrix indicates individual object and each column represent a particular attribute of these objects. In this matrix, we

assume that every element is fixed, discrete and numerical. Any missing element is not allowed.

IV. DATA DISTORTION MEASURES

We used the same set of privacy parameters proposed in [2]. The privacy measures depends only on the original matrix M and its distorted matrix $\bar{M}$

A. Value Difference (VD)

After a data matrix is distorted by data distortion method, the value of its elements changes. The value difference of the datasets is defined by the relative value difference in the Frobenius norm. On the other hand VD is the ratio of the Frobenius norm of the difference of M and $\bar{M}$ to the Frobenius norm of M.

$$VD = \frac{\left\| M - \bar{M} \right\|}{\left\| M \right\|} \qquad (1)$$

Where || denotes the Frobenius norm of the enclosed argument

B. Position Difference

The order of the value of the data element changes after data distortion. We use several metrics to measure the position difference of the data element.

1. RP- RP parameter is used to represent the average change of rank for all attributes after data distortion. For a dataset M with q data object and p attributes. Let O_j^i is the rank (in ascending order) of the j[th] element in attribute i. Similarly $\overline{O_j^i}$ is the rank of the corresponding distorted element. Then the RP parameter is given by:

$$RP = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q}\left| O_j^i - \overline{O_j^i} \right|}{p * q} \qquad (2)$$

2. RK- RK parameter represents the percentage of elements that keeps their rank in each column after distortion. The RK parameter is given by:

$$RK = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q} Rk_j^i}{p * q} \qquad (3)$$

Where $Rk_j^i = 1$ if $O_j^i = \overline{O_j^i}$ otherwise $Rk_j^i = 0$

3. CP- CP parameter is used to measure how the rank of the average value of each attributes varies after data distortion. CP represents the change of rank of the average value of the attributes. CP parameter is given by:

$$CP = \frac{1}{p}\sum_{i=1}^{p}\left| OM_i - \overline{OM_i} \right| \qquad (4)$$

Where OM_i and $\overline{OM_i}$ represent the rank of the average value of i[th] attribute before and after data distortion respectively.

4. CK- Similar to RK, CK is used to measure the percentage of the attributes that keep their rank of average value after data distortion. CK parameter is given by:

$$CK = \frac{1}{p}\sum_{i=1}^{p} Ck^i \qquad (5)$$

Where

$Ck^i = 1$ if $OM_i = \overline{OM_i}$

Otherwise $Ck^i = 0$

C. Utility Measure

After the conduction of certain perturbation the data utility measures indicate the accuracy of data mining algorithms on distorted data. In this paper we choose the accuracy of a NBTree (Naive Bayes Classifier) [5] as our data utility measure.

V. NORMALIZATION METHODS

Data transformation such as Normalization [4] is a data preprocessing tool used in data mining system. An attribute of a dataset is normalized by scaling its values so that they fall within a small-specified range, such as 0.0 to 1.0. Normalization is particularly useful for classification algorithms involving neural networks, or distance measurements such as nearest neighbor classification and clustering. There are many methods for data normalization includes min-max normalization, z-score normalization and normalization by decimal scaling.

(I) Min Max Normalization:

Min-max normalization performs a linear transformation on the original data. Min-max normalization maps a value d of P to d' in the range [new_min(p), new_max(p)]. The min-max normalization is calculated by the following formaula:

$$d' = \frac{[d - \min(p)] * [new\_\max(p) - new\_\min(p)]}{[\max(p) - \min(p)]} + new\_\min(p) \qquad (6)$$

Where

min(p) = minimum value of attribute

max(p) = maximum value of attribute

In our case min-max normalization maps a value d of P to d' in the range [0,1], so put new_min(p) =0 and new_max(p) =1 in the above equation (6). Now we get the simplified formula of min-max normalization.

$$d' = \frac{d - \min(p)}{\max(p) - \min(p)} \quad (7)$$

Min max normalization preserves the relationship among the original data values.

(II) Z-Score Normalization:

Z-score normalization is also called zero-mean normalization. In Z-score normalization the values for an attribute P are normalized based on the mean and standard deviation of P. A value d of P is normalized to d' by the following formula:

$$d' = \frac{d - mean(P)}{std(p)} \quad (8)$$

Where mean (p) = mean of attribute P and std (p) = standard deviation of attribute P

(III) Normalization By Decimal Scaling:

Normalization by decimal scaling normalizes by moving the decimal point of value of attribute P. The number of decimal points moved depends on the maximum absolute value of P. A value d of P is normalized to d' by the following formula:

$$d' = \frac{d}{10^{m}} \quad (9)$$

Where m is the smallest integer such that Max ($|d'|$)<1.

VI. THE PROPOSED PRIVACY RESERVING METHOD

Let M be a matrix of dimension q x p, representing the original dataset. The rows of the matrix represent objects and the column of the matrix represent attributes.

Now the original data matrix M whose size is q x p, must be first transformed by min max normalization transformation to get transformed matrix $\bar{M}$ whose size has the same size q x p as the original data matrix. The min max normalization transformed each element of the original data matrix M into the specific range between 0.0 and 1.0.

Now after applying the min max normalization on the original data matrix, each element of the original data matrix M has been now perturbed into the small specific range between 0.0 and 1.0, i.e. min max normalization scaled each element of the original data matrix M into a specific range such as [0.0,1.0].

Now we have obtained a new matrix $\bar{M}$, which is very similar to the original data matrix M, but not identical. More importantly, $\bar{M}$ preserve the properties of M. Thus $\bar{M}$ can work as a distorted version of the original data matrix M.

Now the distorted matrix $\bar{M}$ is further shifted by multiplying it with a shifting factor, i.e. a negative number, to increase the security of data, because after applying the shifting factor (a negative number) on the matrix $\bar{M}$. The order as well as the value of each element of the distorted matrix $\bar{M}$ was changed, i.e. the greater number become lesser and vice-versa.

VII. EXPERIMENTAL RESULTS

We have conducted experiments to evaluate the performance of data distortion method. We choose four real-life Databases obtained from the University of California Irvine (UCI), Machine Learning Repository [6]. Datasets are the Glass Identification, Haberman's survival data, Bupa Liver Disorders and Iris Dataset. The summaries of the original database are given in Table [I] and Table [II] show the performance of distorted method. We use WEKA (Waikato Environment for Knowledge Analysis) [7] software to test the accuracy of distorted method. The privacy parameters are measured by a separate Java programme. We have constructed the classifier for NBTree classification, and a 10-fold cross validation to obtain the classification results.

We apply our data perturbation method with shifting factor S_f = -15 on the real life dataset mentioned in table I and get the results. All these result shown in table II.

Table [III],[IV],[V],[VI] show the performance of Galss identification, Haberman's Survival data, Bupa liver disorder, Iris respectively. In Haberman's Survival data and IRIS datasets the value difference VD is vary with respect to shifting factor S_f. But in case of Glass Identification and Bupa liver disorder dataset, the value difference VD as well as accuracy have changed with respect to Shifting factor S_f , because after applying shifting factor the value of each element of distorted matrix $\bar{M}$ was changed. Therefore we carefully choose the correct shifting factor S_f for better result.

Table I: The summary of the database

Database	Number of Instances	Number of Features	Number of Classes
Glass Identification	214	10	7
Haberman's Survival data	306	3	2
Bupa Liver Disorders	345	6	2
IRIS	150	4	3

.

Table II: How the privacy parameters and accuracy vary in four datasets

Data	VD	RP	RK	CP	CK	Acc in %
Glass dentification (Original)	-	-	-	-	-	94.3925
Glass dentification (Distorted)	1.091597	101.2514	0.00654	4.8	0	93.4579
Haberman's survival data (Original)	-	--	-	-	-	72.549
Haberman's survival data (Distorted)	1.119772	151.98257	0	1.3333	0.33	72.549
Bupa Liver Disorders (Original)	-	-	-	-	-	66.087
Bupa Liver Disorders (Distorted)	1.096625	172.40966	0	2.6666	0	65.2174
Iris (Original)	-	-	-	-	-	94
Iris (Distorted)	1.195294	74.7433	0	1.0	0.25	92.6667

Table III: Performance of Privacy Parameter and NBTree Accuracy of Glass Identification Dataset with S_f Original Accuracy of NBTree is: 94.3925 %

S_f	VD	RP	RK	CP	CK	ACC %
-5	1.029421	101.2514	0.0065	4.8	0	93.4579
-10	1.059986	101.2514	0.0065	4.8	0	93.4579
-15	1.091597	101.2514	0.0065	4.8	0	93.4579
-20	1.124167	101.2514	0.0065	4.8	0	93.9252
-25	1.157615	101.2514	0.0065	4.8	0	92.9907
-30	1.191867	101.2514	0.0065	4.8	0	93.4579

Table IV: Performance of Privacy Parameter and NBTree Accuracy of Haberman's Survival Data with S_f Original Accuracy of NBTree is: 72.549%

S_f	VD	RP	RK	CP	CK	ACC %
-5	1.039617	151.9826	0	1.3333	0.3333	72.549
-10	1.079553	151.9826	0	1.3333	0.3333	72.549
-15	1.119772	151.9826	0	1.3333	0.3333	72.549
-20	1.160245	151.9826	0	1.3333	0.3333	72.549
-25	1.200947	151.9826	0	1.3333	0.3333	72.549
-30	1.241855	151.9826	0	1.3333	0.3333	72.549

Table V: Performance of Privacy Parameter and NBTree Accuracy of Bupa liver disorder dataset with S_f Original Accuracy of NBTree is: 66.087%

S_f	VD	RP	RK	CP	CK	ACC %
-5	1.032079	172.4096	0	2.6666	0	65.2174
-10	1.064291	172.4096	0	2.6666	0	65.2174
-15	1.096625	172.4096	0	2.6666	0	65.2174
-20	1.129068	172.4096	0	2.6666	0	65.2174
-25	1.161614	172.4096	0	2.6666	0	65.2174
-30	1.194252	172.4096	0	2.6666	0	65.5072

Table VI: Performance of Privacy Parameter and NBTree Accuracy of IRIS dataset with S_f Original Accuracy of NBTree is: 94%

S_f	VD	RP	RK	CP	CK	ACC %
-5	0.54242	74.7433	0	1.0	0.25	92.667
-10	0.66116	74.7433	0	1.0	0.25	92.667
-15	1.19529	74.7433	0	1.0	0.25	92.667
-20	1.80803	74.7433	0	1.0	0.25	92.667
-25	2.44089	74.7433	0	1.0	0.25	92.667
-30	3.08149	74.7433	0	1.0	0.25	92.667

IX. CONCLUSION

In this paper we proposed a privacy preserving data distortion method based on min max normalization transformation. We conducted the experiment on four real life datasets and the experimental result show that the min max normalization transformation based data distortion method is effective for privacy preserving data mining. The privacy parameters used in this work show the degree of privacy protection by the proposed method. In addition, the proposed method also maintain the performance of data mining technique after data distortion, it is interesting to use the other normalization methods like Z-score normalization and Normalization by decimal scaling and also compare its result with Min-max normalization.

REFERENCES

[1] M. Chen, J. Han, and P. Yu, "Data mining: An Overview from a database Prospective", IEEE Trans. on Knowledge and Data Engineering, vol. 8, no. 6, pp. 866-883, Dec. 1996.

[2] S. Xu, J. Zhang, D. Han, J. Wang, "Data distortion for privacy protection in a terrorist analysis system", Proceeding of the IEEE International Conference on Intelligence and Security Informatics, pp. 459-464, 2005.

[3] W. Frankes and R. Baeza-Yates, "Information Retrieval: Data Structures and Algorithms", Prentice–Hall, Englewood cliffs, NJ, 1992.

[4] J. Han, and M. Kamber, "Data Mining: Concepts and Techniques", Second edition, 2006, Morgan Kaufmann, USA.

[5] Ian H. Witten, Eibe Frank,"Data Mining Practical Machine Learning Tools and Techniques", Second Edition, 2005.

[6] UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets.html

[7] The Weka Machine Learning Workbench. http://www.cs.waikato.ac.nz/ml/weka

[8] Jie Wang, Weijun Zhong, Jun Zhang, "NNMF- Based Factorization Techniques for High-Accuracy Privacy Protection on Non-negative-valued Dataset", Proceeding of IEEE Conference on Data Mining, International Workshop on Privacy Aspects of Date Mining (PADM2006), pp.513-517, 2006.

[9] Xiao-Bai Li, and Sumit Sarkar, "A Tree-Based Data Perturbation Approach for Privacy-Preserving Data Mining", IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 9, pp. 1278 – 1283, 2006.

[10] Saif M. A. Kabir, Amr M. Youssef, Ahmed K. Elhakeem, "On data distortion for privacy preserving data mining", Proceedings of IEEE Conference on Electrical and Computer Engineering (CCECE 2007), PP. 308-311, 2007.

[11] Shuting Xu, Shuhua Lai,"Fast Fourier transform based data perturbation method for privacy protection", Proceeding of IEEE International Conference on Intelligence and Security Informatics, pp. 221-224, 2007.

[12] Lian Liu, Jie Wang, Jun Zhang,"Wavelet-Based Data Perturbation for Simultaneous Privacy-Preserving and Statistics-Preserving", Proceeding of IEEE International Conference on Data Mining Workshop, PP. 27-35, 2008.

[13] Zhenmin Lin, Jie Wang, Lian Liu, Jun Zhang, "Generalized random rotation perturbation for vertically partitioned data sets", Proceeding of the IEEE Symposium on Computational Intelligence and Data Mining, pp:159- 162, 2009.

[14] Bo Peng, Xingyu Geng, Jun Zhang, "Combined data distortion strategies for privacy-preserving data mining", Proceeding of the IEEE International Conference on Advanced Computer Theory and Engineering (1CACTE), PP. V1-572 - V1-576, 2010.

[15] S. Xu, J. Zhang, D. Han and J. Wang, "Singular value decomposition based data distortion strategy for privacy protection", ACM Journal of Knowledge and Information Systems, vol. 10, no. 3, pp. 383-397, 2006.

[16] J. Wang, W. J. Zhong, J. Zhang and S.T. Xu, "Selective Data Distortion via Structural Partition and SSVD for Privacy Preservation", Proceedings of International conference on Information & Knowledge Engineering, pp. 114-120, June 2006.

[17] C. K. Liew, U. J. Choi, and C. J. Liew, "A Data Distortion by Probability Distribution", ACM Transaction on Database Systems (TODS), vol. 10, no. 3, pp. 395-411, Sep. 1985.

[18] R. Agrawal and R. Srikant, "Privacy-preserving data mining", Proceeding of the ACM SIGMOD Conference on Management of Data, pp. 439–450, May 2000.

[19] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," International Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.

[20] K. Chen and L. Liu, "Privacy preserving data classification with rotation perturbation", Proceeding of the 5th IEEE International Conference on Data Mining (ICDM 2005), pp. 589-592, 2005.

[21] Jie Wang, Jun Zhang," Addressing Accuracy Issues in Privacy Preserving Data Mining through Matrix Factorization", pp. 217-220, 2007.