***I.J. Wireless and Microwave Technologies*, 2012, 3, 28-33**
Published Online June 2012 in MECS (http://www.mecs-press.net)
DOI: 10.5815//ijwmt.2012.03.05

Available online at http://www.mecs-press.net/ijwmt

Two Noise Addition Methods For Privacy-Preserving Data Mining*

Likun Liu[a], liang Hu[b], Di Wang[c], Yanmei Huo[d], Lei Yang[e], Kexin Yang[f,+]

[a,b,d,e,f] *College of Computer Science and Technology, Jilin University, Changchun 130012, P.R.China*
[c] *Changchun Vocational Institute of Technology, Changchun 130033, P.R.China*

Abstract

In the last decade, more and more researches have focused on privacy-preserving data mining(PPDM). The previous work can be divided into two categories: data modification and data encryption. Data encryption is not used as widely as data modification because of its high cost on computing and communications. Data perturbation, including additive noise, multiplicative noise, matrix multiplication, data swapping, data shuffling, k-anonymization, Blocking, is an important technology in data modification method. PPDM has two targets: privacy and accuracy, and they are often at odds with each other. This paper begins with a proposal of two new noise addition methods for perturbing the original data, followed by a discussion of how they meet the two targets. Experiments show that the methods given in this paper have higher accuracy than existing ones under the same condition of privacy strength.

Index Terms: Privacy-preserving; Data mining; Data Perturbation; Additive Noise

1. Introduction

With increasing concern about privacy, various privacy-preserving data mining techniques have been developed. They are used in many areas such as medical, business and sociology. PPDM has a long history of preventing privacy disclosure by perturbing the original dataset and then releasing the result to the data analyst. A trade-off between privacy and accuracy often need to be made. On the one hand, privacy requires that the original data records must be fully obfuscated before data mining analysis. On the other, accuracy needs that the “patterns” in the original data should be mined out in spite of the perturbation.

In this paper, the participants of PPDM are divided into data provider and data analyst. The data provider owns the original data and adds noise to the original data, while the data analyst has access to the obfuscated data and mines them.

We propose two new additive perturbation methods which can be applied in the area of secure statistical databases. Both of the methods can help the data analysts mine out the “patterns” directly from the obfuscated data, and spare them from the work of reconstructing the original data distribution as an intermediate step or trying to modify data mining algorithm, which are very general in many perturbation techniques.

*This work is supported by the National Grand Fundamental Research 973 Program of China under Grant No. 2009CB320706, the National Natural Science Foundation of China under Grant No. 60873235 and 60473099, Program of New Century Excellent Talents in University of Ministry of Education of China under Grant No. NCET-06-0300, and the Fundamental Research Funds for the Central Universities of China under Grant NO.200903179.

+ Corresponding author. phone: 86-431-88858785; fax: 86-431-85166494;
E-mail address: [f]yangkx@jlu.edu.cn; [a]likun_l@163.com

The paper is organized as follows: Section 2 presents the related work. Section 3 describes our additive perturbation methods, followed by some experimental results in Section 4. Finally, we summarize the research in Section 5 with a discussion of the future work.

2. Related Work

Data perturbation is widely used for PPDM. It includes(but not limited to):additive noise[1,2], multiplicative noise[3], matrix multiplication[4], data swapping[5], data shuffling[6], k-anonymization[7,8], Blocking[9]. This paper focuses on additive noise and its application to numeric continuous data will also be concerned.

In order to mine the data directly from the perturbed data, without reconstructing the original data distribution, Li Liu proposed a threshold algorithm[10] which uses a threshold to categorize a record by computing its probability. A shortcoming of this method is that the choose of the threshold which will affect the mining result is not easy because the proper threshold value varies from case to case and can be set by no rules but experience. So Mohammad Ali Kadampur proposed a new noise addition scheme[11] in which data provider firstly builds a decision tree T by exploring the original data, and then for each record, adds a noise(i.i.d) to get the modified data which needs to be adjusted according to T. Decision tree T' will be drawn by mining the modified data, and it is similar to T. According to [11], the result of mining the obfuscated data is close to mining the original data. But this method is limited by data sparseness. If data is intensive, the deviation, caused by additive noise, may lead to more incorrect split. In this case, the similarity between T and T' will be reduced. Also this method is not safe enough, because it can be attacked by some attack techniques such as spectral filtering(SF) [12], singular value decomposition(SVD) filtering [13], and principal component analysis (PCA)filtering [14]. For similarity, Zhai Fangwen el [15] proposed a similarity measure framework using a specific formula to compute the similarity.

Mohammad Ali Kadampur's method is only applicable to building privacy-preserving decision tree .The two additive perturbation methods our proposed expand its application to the security control of statistical database. The original data is pre-mined by the data provider to get the "patterns", and then after being added noise in a proper way in order to maintain these "patterns", the data is wrapped and released to the data analyst. So the data analyst only needs to mine the perturbed data without doing anything else. With our methods, the step of reconstructing the original data distribution which has high computation cost and the step of modifying mining algorithm are not needed any more. Our experimental results have shown that this model not only has a higher degree of accuracy, but also guarantees that its privacy security is as good as, if not better than, the other methods.

3. Two Noise Addition Methods

The data provider replaces the original data X with

$$Y = X + R \tag{1}$$

Where R is the noise, generally satisfies some distribution(i.i.d).

We suppose that D is the original data set and $C(C_1, C_2 ... C_k)$ is the result of clustering(using k-mean clustering algorithm). Our effort will be to modify D to $D^{'}$, and if mining $D^{'}$, we will get a new cluster result $C^{'}(C_1^{'}, C_2^{'} ... C_k^{'})$, which is similar to C. After getting the cluster sets C, we propose two noise addition methods to perturb the original data: random distance in distance domain (RDD) and rotation around the center of clustering (RACC). Both of them are used to perturb numeric attributes.

3.1. RDD

In RDD, records will be traversed. After K-mean clustering, each record will be categorized, and then Gauss noise will be added to it. In order to keep the “pattern” unchanged before and after perturbing, we try to keep the record in the same category. Therefore, we adjust the distance between the center of category and the point of perturbed record, to make it in the domain of category, which is the range of values of distance from records to the center of category. C_i is the center of the category, R is a record. Noise N_x and N_y will be added to its attributes X and Y , then we get point $P(X+N_x,Y+N_y)$. Three cases will be considered, P in $D_{in}(i), D(i)$ *and* $D_{out}(i)$

$$dis(C_i,P^{'}) = \begin{cases} dis(C_i,P), P \in D(i), 1 \le i \le k \\ 2D(i).left - dis(C_i,P), P \in D_{in}(i), 1 \le i \le k \\ 2D(i).right - dis(C_i,P), P \in D_{out}(i), 1 \le i \le k \end{cases} \tag{2}$$

After computing the distance C_i and P', the coordinates of point P', which will be published to data analyst, will be got.

RDD Algorithm

```
divide the dataset into k categories using k-mean algorithm
for each Instance do
    find which category $x_j$ is in
    identify the domain of the category
    Add a small Gauss noise with certain mean and variance
    Compute $dis(C_i,P)$
    if( $dis(C_i,P) < D(i).left$ ) then
        $dis(C_i,P^{'}) = 2D(i).left - dis(C_i,P)$
    else if ( $dis(C_i,P) > D(i).right$ ) then
        $dis(C_i,P^{'}) = 2D(i).right - dis(C_i,P)$
    else
        $dis(C_i,P^{'}) = dis(C_i,P)$
    end if
    end if
    According $dis(C_i,P^{'})$ and the coordinate of $C_i$ ,compute coordinate of $P^{'}$
end for
```

3.2. RACC

RACC is a little different from RDD. The perturbed point lies on the circle, whose center is C , R is a record and radius is $CQ(CQ \le CR)$. In RACC, we do not directly generate the noise N_x and N_y to perturb X and Y , but the random noise $\theta, \theta \in (0, 2\pi)$ and random distance ratio d_j. Therefore, we can get the point Q . We compute Q using (3), (4) and (5).

$$r = d_j \times dis(R, C) \tag{3}$$

$$\alpha = \begin{cases} \arctan(\frac{R_y - C_{iy}}{R_x - C_{ix}}), R_x > C_{ix} \\ \arctan(\frac{R_y - C_{iy}}{R_x - C_{ix}}) + \pi, R_x < C_{ix} \end{cases} \tag{4}$$

$$\begin{cases} Q_x = C_{ix} + r\cos(\theta + \alpha) \\ Q_y = C_{iy} + r\sin(\theta + \alpha) \end{cases} \tag{5}$$

In this method, the radius r can be kept unchanged. The result of mining is better similar to true data than RDD. Because the perturbed effect of RDD depends on the choice of random function, different variance, different extent of data is modified. When the variance is small, RACC will get much more data changed than RDD, because it has more chance to reach the remote point. On the contrary, RDD has more change space.

```
RACC Algorithm
1: divide the dataset into k categories using k-mean algorithm
2: for each Instance $x_j$ do
3:     find which category $x_j$ (point $R$) is in, and get the center of category $C_i$ 4:     Generate random Noise $\theta, \theta \in (0, 2\pi)$ and random distance ratio $d_j$
5:     compute $r = d_j \times dis(R, C)$, $\alpha = \arctan\frac{R_y - C_{iy}}{R_x - C_{ix}}$
6:     if( $R_x < C_{ix}$ ) then
7:         $\alpha += \pi$
8:     end if
9:     using (4), compute the coordinate of $Q$
10: end for
```

4. Experiments

Our experiments are implemented on Weka tool and the perturbed matrix is tranformed by using Matlab 7.0. We will use three real-world datasets (Iris, Glass and Yeast), which were assembled from the UCI machine learning repository.

As shown in Table Ⅰ, for each dataset, we choose number of cluster is 2 and 3, addition noise is from Gaussian distribution. Our results are compared with Keke Chen's Geometric Data Perturbation[16], ten groups of noise data will be generated to perturb original data. The effect of experiments is measured by computing the mean accuracy. The result shows that in most cases our two algorithms can get higher accuracy. When using Gauss noise as random noise data, the variance can dramatically affect the result. From Fig. 1, we can see that with the increase of noise level, both the accuracy of Keke Chen's algorithm and RDD algorithm have a decreasing trend, but RACC keeps relatively steady.

Table 1 Accuracy After Perturbtion

Datasets	Number of categories	RDD	RACC	Keke Chen
Iris	2	100%	96%	95.33%
Iris	3	98.67%	96.67%	86%
Glass	2	100%	100%	99.07%
Glass	3	89.25%	89.25%	88.32%
Yeast	2	88.32%	91.85%	83.02%
Yeast	3	74.39%	82.01%	73.92%

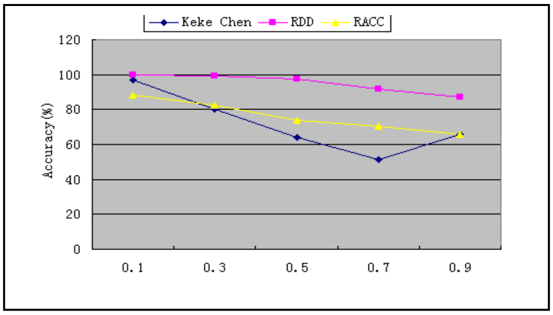

Fig 1. noise level (sigma)

5. Conclusions

We propose two additive perturbation methods RDD and RACC which are suitable in the field of security control of statistical database. Our methods modify the original data according to the result of the pre-mining the original data. It is proved that our two additive perturbation algorithms not only make the reconstruction with high computation cost unnecessary and the mining algorithm unmodified, but also have higher accuracy.

In the future, we plan to apply our algorithms to more statistical analysis by making more comparisons and expand them to the distributed data mining.

References

[1] Agrawal, R., Srikant,R. Privacy-perserving data mining. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, Texas, USA, May 16-18, 2000, pages 439-450. ACM, New York, 2000.

[2] D.Agrawal and C.C.Aggarwal," On the design and quantication of privacy preserving data mining algorithms," in Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems. pages. 247-255, Santa Barbara, CA, 2001.

[3] J.J.Kim and W.E.Winkler.Multiplicative noise for masking continuous data. Technical Report Statistics #2003-01, Statistical Research Division, U.S. Bureau of the Census, Washington D.C, April, 2003.

[4] M.Artin, Algebra. PrenticeHall, 1991

[5] S.E.Fienberg and J.McIntyre.Data swapping: Variations on a theme by dalenius and reiss. Technical report, National Institute of Statistical Sciences, Research Triangle Park, NC, 2003.

[6] K.Muralidhar and R.Sarathy.Data shuffing-a new masking approach for numerical data. Management Science,52(5):658-670,May,2006

[7] P.Samarati.Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering,13(6):1010-1027, November/December 2001.

[8] L.Sweeney.k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557-570,2002

[9] S.Agrawal, V.Krishnan, and J.R.Haritsa,"On addressing efficiency concerns in privacy-preserving mining," Proc.of 9th Intl.Conf. on Database Systems for Advanced Applications(DASFAA), pages. 113–124, 2004.

[10] Li Liu, Murat Kantarcioglu, Bhavani Thuraisingham. Privacy Preserving Decision Tree mining from Perturbed Data. In proceedings of 42th Hawaii International Conference on System Sciences. 2009.

[11] Mohammad Ali Kadampur, Somayajulu D.V.L.N. A Noise Addition Scheme in Decision Tree for rivacy Preserving Data Mining. Journal of Computing, vol 2, no1, pages.2151-9617.January 2010.

[12] H.Kargupta, S.Datta, Q.Wang, and K.Sivakumar. On the privacy preserving properties of random data perturbation techniques. In Proceeding of the IEEE International Conferenceon Data Mining(ICDM'03), pages 99-106, Melbourne, FL, November 2003

[13] S.Guo, X.Wu, and Y.Li. On the lower bound of reconstruction error for spectral filtering based privacy preserving data mining. In Proceedings of the10th European Conference on Principles and Practice of Knowledge Discovery in Databases(PKDD'06), pages 520-5277, Berlin, Germany, September, 2006.

[14] Z.Huang, W.Du, and B.Chen. Deriving private information from randomized data. In Proceedings of the 2005 ACM SIGMOD Conference, pages 37-48, Baltimroe, MD, June 2005

[15] Zhai Fangwen, Yang Zehong, Song Yixu, Liu Yi. A Novel Similarity Measure Framework on Financial Data Mining. 2010 Second International Conference on Networks Security, Wireless Communications and Trusted Computing,Wuhan,China,April 24-25,2010

[16] K.Chen, G.Sun,and L.Liu. Towards attack-resilient geometric data perturbation. In Proceedings of the 2007 SIAM International Conference on Data Mining(SDM'07), Minneapolis, MN, April 2007