Send Orders for Reprints to reprints@benthamscience.ae

Research on a New Privacy-Preserving Algorithm of Association Rules Based on Parameter Perturbation

Jingjing Yang, Beibei Dong, Xiao Zhang*, Zhonghua Li and Shangfu Hao

School of Information Science and Engineering, Hebei North University, Zhang Jiakou, Hebei, 075000, China

Abstract: Due to the reason that the randomness of the parameters in the MASK algorithm always leads to the volatility and uncertainty of the mining results, this paper proposed an optimization algorithm for the maximum likelihood estimation of the parameters to choose a parameter that is most approximate to the common parameters from the parameter group that has been generated randomly. Such a parameter generated as above represented all of the parameters in the parameter group. The simulation experiment proves that the application of such a parameter has reduced the great volatility hidden in the mining results to some extent.

Keywords: Association rules, Privacy preserving, Parameter optimization.

1. INTRODUCTION

Ever since the first KDD (Knowledge discovery in database) conference, where privacy-preserving data mining was proposed in 1995, privacy-preserving data mining has been developed to be a specialized research topic in the field of knowledge discovery. In 1999, Rakesh Agrawal gave a brilliant keynote speech [1] on KDD, considering privacy-preserving data mining as one of the key researches in the future. Ever since then, privacy-preserving data mining has attracted increasing attention from researchers [2-4], becoming promptly one of the hot researches in the field of data mining in recent years. However, the relationship between privacy preserving and data mining is quite contradictory. Such problems have been found constantly in data mining that what must be dig out fails to be mined and the data that must be preserved fails to be protected well [5]. Then how to tackle this contradictory problem has been the focus of research on the privacy-preserving mining algorithm of association rules. Moreover in many cases, the effect of privacy-preserving data mining has been applied in the following industries, including bio medicine, DNA detection, finance, retail sales and telecom etc. [6-7]. Then it is inevitable that the effects will be verified in the practical application as well.

2. PRELIMINARIES

Currently, the following have been the research hotspots in the field of privacy-preserving data mining: the research on the technical details of a specific data mining algorithm, the research on the different data mining algorithms, the utilization of data mining tools to detect the potential possibility of inference attack in database and the utilization of the existing password security technology to develop such a secure multi-party computation protocol that can be used in special circumstances and to carry out researches on the privacy preserving data mining algorithm [8].

MASK (Mining associations with secrecy constraints) algorithm is a random transformation-based method with the advantages including: the preservation of high privacy and the obtaining of accurate mining results. Through a randomized method, the Mask algorithm is able to distort the transaction data: Assume that a database tuple consists of 0 and 1 with 1 representing attribute occurrence and 0 referring to the nonoccurrence of attributes, then the probability for each data item remains to be the original value that is P, and the inversion probability is 1-P. In this way, all of the database tuples will be distorted to form a new database, based on which data mining is conducted after the distortion. Use different P values to perform distortion for different attributes. However ,for simplification purpose, assume that among all of the probabilities, P are with the same value [9].

2.1. Association Rules

Estimation on the support degree of frequent itemset: Assume that the matrix of the real dataset is represented by T, then distort T to get Matrix O and the distortion probability is P. Indicate the number of 1 in the i th Column of Matrix T by C_1^T and the number of 0 is indicated by C_0^T. Also indicate the number of 1 in the i th Column of D by C_1^D and the number of 0 is represented by C_0^D. Through the data distortion process, it reveals that:

$$C_1^T \times p + C_0^T \times (1-p) = C_1^D \; C_0^T \times p + C_1^T \times (1-p) = C_0^D \tag{1}$$

To deduce (2)

$$C^D = \begin{bmatrix} C_1^D \\ C_0^D \end{bmatrix}, \; C^T = \begin{bmatrix} C_1^T \\ C_0^T \end{bmatrix} \tag{2}$$

*Address correspondence to this author at the School of Information Science and Engineering, Hebei North University, Zhang Jiakou, Hebei, 075000, China; Tel: +8618032322093; E-mail: r78z-yang@126.com

1874-4443/15

According to Equation (3), we are able to estimate the support degree of 1-item set in the real matrix based on the distortion matrix D. Similarly, the support degree of N-item set can be calculated just like this 1-item set:

$$M=\begin{bmatrix} m_{0,0} & m_{0.1} & \cdots & m_{0,2^{k-1}} \\ m_{1,0} & m_{1,1} & \cdots & m_{1,2^{k-1}} \\ \vdots & \vdots & \ddots & \vdots \\ m_{2^{k-1},0} & m_{2^{k-1},1} & \cdots & m_{2^{k-1},2^{k-1}} \end{bmatrix}$$

$$C^D=\begin{bmatrix} C^D_{2^{K-1}} \\ \vdots \\ C^D_1 \\ C^D_0 \end{bmatrix}, \quad C^T=\begin{bmatrix} C^T_{2^{K-1}} \\ \vdots \\ C^T_1 \\ C^T_0 \end{bmatrix} \tag{3}$$

C^D_K is defined to be the number of the N-item set in the distortion matrix, where the item sets are represented by N numbers in binary form, which corresponds to the decimal number of K. M is an Nth order matrix, and represents the probability that the N-item set j (indicated by N numbers in binary form) is distorted to be the N-item set i (indicated by N numbers in binary form). For example, 2-item set indicates the probability that the 11-item set is distorted to be 00. It can also be expressed as $(1-p)^2$.Then through Equation (3), we are able to work out $C^T_{2^{n-1}}$, the support degree of the N-item set in the original dataset.

2.2. Shortcomings of the Traditional MASK Algorithm in Parameter Selection

(1) Regarding the privacy-preserving problem, different users might have different requirements on the preservation of data privacy. However, in the MASK algorithm, data "1" and "0" has been distorted by using the same parameter. But in essence, the data protection degree for "1" must be higher than that for data "0". In this case, the users might adjust the perturbation parameter for "1" and "0" according to the actual situation to balance the privacy preservation and the accuracy of mining results [10].

(2) Low efficiency in the running of algorithm: The change in the density of the data sets is the main reason to cause the increase in the running time of the program. For the perturbation transformation, a k-item set might bring about 2k of combinations, which will undoubtedly increase the computational overhead when the support degree is reconstructed and recomputed [11].

3. TRADITIONAL PARAMETER SELECTION METHOD

Before data mining, make sure to transform and conceal the original data. The specific method for the selection of randomized parameters is provided as follows:

Provide that the randomized parameters $0\leq p_1$, p_2, $p_3 \leq 1$ and $p_1+p_2+p_3=1$. Also for Item x□{0,1}, assume that $r_1=x$, and $r_3=0$. Then value of the randomized function will be set to be r_j according to the probability p_j, where j=1,2,3. When the total number of items is indicated by k, the transactions expressed with 0-1 sequence can be indicated by $X=(x_1,x_2,\cdots,x_k)$. After the perturbation, the transaction can be obtained when calculation is based on Y=R(X), where $y_i=r(x_i)$. That is to say that the value of y_i is set to be based on the probability and the value is set to be 1 when it's based on probability p_2, while the value is 0, when it's based on probability p_3.

3.1. New Parameter Optimization Method

Through the traditional parameter selection method, the results are always uncertain. The selection of any parameter will lead to different results, which will affect significantly the analysis of the data mining result. This paper optimizes the parameter set to work out a new parameter, which is approximate to the result of the whole parameter set so that no significant influence will be exerted on the data mining result.

3.2. Principle of the Maximum Likelihood Function

The maximum likelihood estimation starts from the distribution of a given phenomenon that is observed. However, the parameter involved is unknown. In this case, the maximum likelihood method will utilize the parameter value of the observations (samples) that are with the highest probability to estimate the parameter for such a distribution. In this way, this paper provides an approach to obtain a group of parameters that are used to estimate and describe a distribution.

Based on the idea of the maximum likelihood method, assume $f(x,\theta)$ is the density function of the random variable X, where θ is an unknown parameter for this distribution. Then in the case that the following random samples $x_1,x_2,\cdots,x_n$ are provided, the maximum likelihood estimate of θ is actually the value of the highest probability for the generation of this sample. Or in other words, the maximum likelihood estimate of θ is such a value that maximizes the density function $f(x,\theta)$.

First, obtain randomly the parameter set θ, based on which use the formula for the maximum likelihood function to obtain $\hat{\theta}$,the maximum likelihood value of Set θ:

$$\frac{d}{dp}\ln L(p)=\frac{\sum_{i=1}^{n}x_i}{p}-\frac{n-\sum_{i=1}^{n}x_i}{1-p}=0 \tag{4}$$

Then obtain the maximum likelihood estimate of p as below:

$$\hat{p}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x} \tag{5}$$

For example, choose randomly multiple groups of parameters and optimize them.

0.3000 0.7000 0.8000 0.8000 0.6000

0.4000 0.3000 0.7000 0.3000 0.2000

0.5000 0.7000 0 0.8000 0.7000

0.5000 0.7000 0.6000 0.5000 0.2000

0.9000 0.1000 0.4000 0 0.9000

0.5000 0.3000 0.9000 0.2000 0.3000

0.9000 0.2000 0 0.7000 0.8000

0.6000 0.7000 0.5000 0.5000 0.2000

1.0000 0.8000 0.4000 0.2000 0.3000

0.2000 0.3000 0.5000 0.3000 0.1000

For each of the above samples, obtain separately the parameters as follows: 0.8,0.3,0.7,0.5,0.9,0.3,0.8,0.5,0.4,0.3. Then choose 0.3, which is the sample median determined according to the original sample as the estimate of the overall median. When its standard error is given, the Bootstrap estimation can be provided as below:

$$\hat{\sigma}_{\hat{\theta}} = \sqrt{\frac{1}{9}\sum_{i=1}^{10}(\hat{\theta}_i^* - \overline{\theta^*})^2} \approx 0.731 \tag{6}$$

All of the above shows that the standard error is rather small. Therefore, we choose 0.55 as the parameter value.

In Table **1**, take the shopping page for each commodity as an item, which is also assigned with a definite SN. The customer's each purchase will be regarded as a transaction, which will be expressed by a Boolean sequence with the length representing the total varieties of commodities: When the commodity with the corresponding SN is purchased, set the value to be 1, otherwise the value will be set as 0.

Table 1. The customer shopping session set.

IPID	SID	PAGES
100000001	10001	P1,P2,P3
100000001	10002	P1
100000001	10003	P3,P4,P5
100000002	10001	P1,P2,P3
100000002	10002	P1,P2,P3,P5

In Table **2**, provide that the randomized parameters are 0<x, y, z<1 and x + y+ z=1. Assume that F(i, j) is the value of the pseudo instance Fi when the transaction number is set to be TID=j. P(i,j) refers to the value of the original column Pi when the transaction number is set to be TID=j. For Item $b \in \{0,1\}$, choose 0 as the value of the randomized function $R(b)$ when the probability is x. When it's based on the probability y, the value will remain unchanged as b. But when it's based on the probability z, the value will be set to be 1 with the formula provided as below:

Table 2. The shopper session set by a Boolean matrix.

TID	IPID	SID	P1P2P3P4P5
1	10000001	10001	1 1 1 0 0
2	10000001	10002	1 0 0 0 0
3	10000001	10003	0 0 1 1 1
4	10000002	10001	1 1 1 0 0
5	10000002	10002	1 1 1 0 1

$F(i, j) = R(P(i, j))$ generates the padding data for perturbation purpose in the pseudo instance. Where the value of Y is set to be 0.55, which is an optimal approximate value. Then set the threshold for the minimum support to be 0.4.

The result of mining association rules is as below: Frequent 1-item set {N1}, {N3}, {N5}, {N6}, {N7}, {N8}, {N9}; Frequent 2-item set: {N1，N6}, {N1,N7}, {N1,N8}, {N1,N9}, {N3,N7}, {N3,N8}, {N6,N7}, {N6,N8}, {N6,N9}, {N5,N6}, {N7,N8}, {N7,N9}, {N8,N9}.

Through the permutation of the frequent 1-item set: {P2}, {P3}, {P1}, and frequent 2-item set: {P2, P3}, {P2, P1}, {P3, P1} in the mapping table, are shown as Tables **3-5**.

Table 3. The transaction set after distortion.

TID	P1F1P2F2P3F3P4F4P5F5
1	1011110000
2	1100000000
3	0000111111
4	1011110000
5	1110100011

Table 4. Table for the permutation mapping relationship.

The Original Column Name	After the Replacement of the Column Name
P1	N8
F1	N1
P2	N3
F2	N9
P3	N7
F3	N6
P4	N2
F4	N0
P5	N4
F5	N5

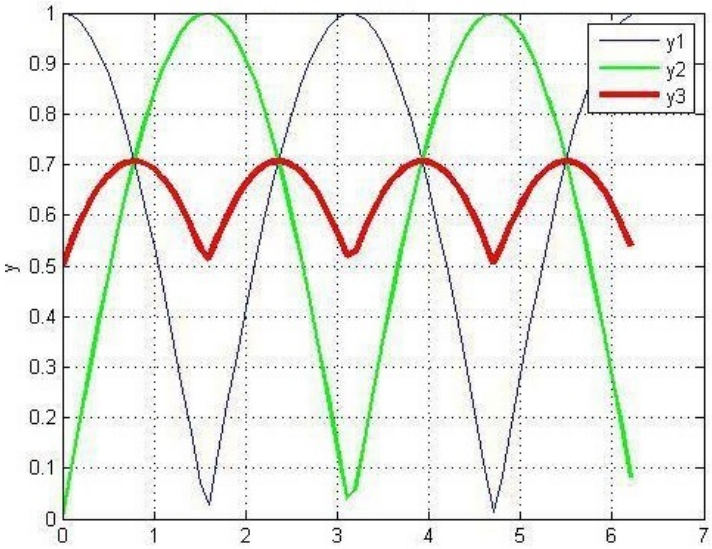

Fig. (1). Comparison of parameters before and after optimization.

Table 5. Faked transaction set S'.

TID	N0N1N2N3N4N5N6N7N8N9
1	0001011111
2	0100011011
3	1110101101
4	0101010110
5	0101101111

EXPERIMENT

In order to prove the superiority of this optimization algorithm, an experiment has been conducted to compare it with the MASK algorithm. The platform for both of the algorithms is Dual-Core CPU2.20GHz, 2G memory and the running software is Microsoft Visual Studio 2008 with the parameters having been chosen randomly.

In the experiment (Fig. **1**), we choose the absolute value of COSX to be the parameter of Y1 and select the absolute value of SINX to be the parameter of Y2. Then Y3 is the optimized parameter for both the parameters of Y1 and Y2. This experiment has been carried out to figure out the influence of the parameters on the trueness, revealing the influence of different parameters on the data trueness and proving that the result is fixed when the mining of association rules is conducted with the same trueness. Therefore, trueness can be considered as equivalent to the excellent or bad mining result. According to the experiment, we can find that the trueness of the optimized data tends to be stable, while there are great changes in the data trueness of the parameters that haven't been optimized.

The experiment proves that during the mining of the association rules, this algorithm performs very well in the preservation of data privacy. The mining result also proves its excellent mining effect. Therefore, we might come to a conclusion that the parameters chosen after the optimization will play a positive role in the privacy preserving and the mining result.

CONCLUSION

Aimed at the research on the privacy-preserving data mining algorithm, this paper proposes a new MASK algorithm, which in itself is able to optimize the parameters based on parameter perturbation. Through the previous MASK algorithm, all of the parameters will be selected randomly and applied directly to the algorithm, leading to great fluctuation in the mining results when different parameters are used in the mining process. Therefore, it would be very difficult to make an effective analysis on the results. Then this paper suggests to optimize the parameters. First, generate the parameter groups randomly. Second, through the process of evaluation, obtain the maximum likelihood value for the parameter group and choose this value as the final parameter. Through the optimization of the parameters, only the most suitable parameter will be chosen every time from the parameter group. Therefore, the influence on the mining result is the least, which makes it possible to reflect the mining result objectively and exerts the least influence on the correctness of the analysis on the mining result.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

This work was supported by Hebei North University (No. Q2014002, No. ZD201301, No. ZD201302, No. ZD201303, No. Q2014005, No.Q2014008) and the Education Department of Hebei Province (No. QN2014182).

REFERENCES

[1] R. Agrawal, "Data mining," *Crossing the Chasm*, 1999.

[2] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," In: *Proceedings Of The Twenty-Second Acm Sigmod-Sigact-Sigart Symposium On Principles Of Database Systems*, ACM, pp. 211-222, 2003.

[3] Y. Lindell, and X.B. Lindell, "Secure multiparty computation for privacy-preserving data mining," *Journal of Privacy and Confidentiality,* vol. 1, no. 1, pp. 5, 2009.

[4] O. Jia, Y. Jian, and L. Shaopeng, "Limiting Privacy Breaches in Differential Privacy," In: *3rd International Conference on Computer Science and Service System*, Atlantis Press, 2014.

[5] J. Vaidya, C. W. Clifton, and Y. M. Zhu, "Privacy preserving data mining," vol. 19, pp. 1-5, 2006.

[6] C. C. Aggarwal, and S. Y. Philip, "A general survey of privacy-preserving data mining models and algorithms," Springer US, 2008, pp. 11-52.

[7] J. Wang, Y. Luo, Y. Zhao, and J. Le, "A survey on privacy preserving data mining," *Database Technology and Applications, First International Workshop on IEEE*, pp. 111-114, 2009.

[8] X. Zhang, and H. Bi, "Research on privacy preserving classification data mining based on random perturbation," *Inform. Networking Autom. (ICINA)*, 2010 International Conference on. *IEEE*, vol. 1, pp. 173-178, 2010.

[9] S. Agrawal, V. Krishnan, and J. R. Haritsa, "On addressing efficiency concerns in privacy-preserving mining," *Database Syst. Adv. Appl.*, Springer Berlin Heidelberg, vol. 2973, pp. 113-124, 2004.

[10] C. L. Huang, and C. J. Wang, "A GA-based feature selection and parameters optimization for support vector machines," *Exp. Syst. Appl.*, vol. 31, no. 2, pp. 231-240, 2006.

[11] H. Aytug, M. Khouja, and F.E. Vergara, "Use of genetic algorithms to solve production and operations management problems: a review," *Int. J. Prod. Res.*, vol. 41, no. 17, pp. 3955-4009, 2003.

Received: June 02, 2015 Revised: August 02, 2015 Accepted: September 05, 2015