Journal of Theoretical and Applied Information Technology
31st January 2013. Vol. 47 No.3
© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ALGORITHM-IRRELEVANT PRIVACY PROTECTION METHOD BASED ON RANDOMIZATION

JING GUO
Lecturer, Chongqing Technology and Business University, Chongqing 400067, China
E-mail: dorothy-guo@163.com

ABSTRACT

Privacy preserving classification mining is one of the fast-growing subareas of data mining. The algorithm-related methods of privacy-preserving are designed for particular classification algorithm and couldn't be used in other classification algorithms. To solve this problem, it proposes a new algorithm-irrelevant privacy protection method based on randomization. This method generates and opens a new data set that is different from the original data set independently as the perturbed data. The perturbed data and the original data have the same distribution. Users get the models of the original data from the perturbed data. Experimental results demonstrate that the classification algorithms can be used on the perturbed data directly. And this method reduces the privacy data disclosure risk more effectively.

Keywords: *Privacy Protection, Data Mining, Classification, Randomization, Algorithm-Irrelevant*

1. INTRODUCTION

Privacy preserving classification mining is one of the fast-growing subareas of data mining. There are two kinds of privacy protection methods for classification mining. One is the algorithm-related method, the other is the algorithm-irrelevant one. Each algorithm-related method is designed for particular classification algorithm and other classification algorithms couldn't be used in it. It's not flexible, so algorithm-related methods are not conducive to the actual application.

Using algorithm-irrelevant methods is a powerful way to solve this problem, and it has a significant advantage. Existing algorithm-irrelevant method is basically based on data perturbation. In such methods, the ordinary classification mining algorithms can be directly applied to perturbed data, in order to get the pattern of the original data. Currently, such methods mainly include the methods based on matrix decomposition and transformation and k-anonymization method[1]. In the methods based on matrix decomposition and transformation, the data are arranged in a matrix form. Analysis of the data uses matrix decomposition, and retaining the important information in data mining, deleting those unimportant in terms of data mining to achieve perturbed data. The methods based on matrix decomposition and transformation mainly include Singular Value Decomposition methods, non-negative matrix factorization method[4] and Wavelet-based method[5]. Singular Value Decomposition methods mainly include BSVD (Basic Singular Value Decomposition) method[2] and SSVD (Sparsified Singular Value Decomposition) method[3].

Data mining is mainly concerned about the trend rather than the details of the data. So in terms of data mining, the statistics information is very important.

Target of data perturbation based on randomization is not only obtaining the statistics of the original data from perturbation data to complete the data mining, but also obtaining the exact value of the original information not from perturbation data. Existing data perturbation methods based on randomization are related to the algorithm[8-11]. In these methods, the statistical information of perturbed data and the statistical information of the original data are not the same, but using the statistical information of perturbed data can derivate and calculate the statistical information of the original data. During performing classification mining, the process of derivation and calculation, the statistical information must be embedded into the mining algorithm. Ordinary classification mining algorithm couldn't be directly applied to the perturbed data to obtain the original data model. It must be converted and embedded the process of derivation and calculation of the statistical information. For each classification algorithm, whether it can be applied and how it be applied to perturbed data should be researched into.

© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

This paper combine algorithm-irrelevant and randomization. It proposes a new algorithm-irrelevant privacy protection method based on randomization (AIBR). The rest of the paper is organized as follows. Section 2 describes AIBR method in detail and presents the workflow of AIBR method. Section 3 compares AIBR method to other existing methods through experiments. Finally, Section 4 concludes the paper.

2. DESCRIPTION OF THE METHOD

The basic idea of AIBR method is similar to the methods based on matrix decomposition and transformation. The distribution of the data for data mining is the important information, and the specific value of the data is not important for data mining. If perturbed data can not only retain the distribution of the original data, but also differ greatly from the original data, we can retain the availability of data at the same time in privacy protection. The method attempts to independently generate a set of new data with the same distribution of original data, and discloses it as the perturbation data. Perturbation data does not depend on the original data; therefore, the lack of information about the exact value of the original data from the perturbation data can protect privacy data. Perturbation data maintains the distribution of the original data, ordinary data mining method without transformation can be directly applied to perturbation data to find the original data.

The original data are often multi-dimensional, and between each dimension often are not independent. Therefore, there will be the "curse of dimensionality" phenomenon; the number of samples is insufficient and difficult to directly get the distribution of the original data. AIBR method uses a two-stage strategy to solve this problem. First of all, we gain respectively the statistical distribution of each attribute without considering the links between each attribute. At the same time, a set of independent and identically distributed data for each attribute should be generated. Subsequently, the link between the various attributes is restored by the relations of sequences. This is an approximate method which couldn't guarantee that the generated data and the original data must be strictly independent and have a same distribution. However, the experiments show that, running in multi-algorithm classifiers by using the generated data have similar classification accuracy with that by using the original data. So, distribution differences between the generated data and the original data are not enough to affect the application of data mining algorithms. On the other hand, it can provide better protection for the privacy of data as a result of the difference between the generate data and the original data makes not getting the exact value of the original data.

AIBR method can be divided into two steps. In the first step, generated independent and identically distributed data for each attribute, not considering the links between the attributes. In the second step, the link between the various attributes is be restored by the relations of sequences. AIBR method is described as follows:

Input: The original data is *A*. *A* includes n samples, each of which contains *m* attributes. There is a parameter *k* given in advance.

Output: *MA* is the perturbation data of *A*.

1. $MA = \phi$;
2. for i=1:m //the first step of AIBR
3. R_i is the ith attribute of *A*;
4. D_i is the projection of the original data *A* in R_i;
5. MD_i is the perturbation data of D_i. $MD_i = \phi$;
6. x^i_{max} and x^i_{min} are the maximum and minimum value of D_i;
7. Interval [x^i_{min}, x^i_{max}] is divided into *k* subintervals equally, denoted as $I^i_1, I^i_2, ..., I^i_k$.
8. for j=1:k
9. N^i_j is the number of data in D_i falled into the interval I^i_j;
10. Data sets T^i_j is generated, data of which obey uniform distribution in I^i_j. $|T^i_j| = N^i_j$;
11. $MD_i = MD_i \cup T^i_j$;
12. end for;
13. end for;
14. for i=1:n //the second step of AIBR
15. The ith samlple of *A* is denoted as $X_i = (x^1_i, x^2_i, ..., x^m_i)$;
16. The ith samlple of *MA* is denoted as $Y_i = (y^1_i, y^2_i, ..., y^m_i)$;
17. for j=1:m
18. x^j_i is the jth component of X_i;
19. if $x^j_i \in I^j_p$ then
20. Select y^j_i randomly, $y^j_i \in MD_j$ and $y^j_i \in I^j_p$;
21. $MD_j = MD_j \setminus \{y^j_i\}$;

© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

22. end if;
23. end for;
24. $MA = MA \cup \{Y_i\}, Y_i = (y_i^1, y_i^2, ..., y_i^m)$;
25. end for;

There is a parameter *k* in AIBR method, represents the number of attributes subinterval in the first step. The larger the value of k is, the more precise the statistical distribution of each attribute is. Therefore, the new distribution of the generated data for each attribute has closer than it of the original data. The availability of final perturbation data is higher. On the other hand, the more subintervals are divided, the smaller the width of each subinterval is, owing to the range of attributes is fixed. In this case, the numerical gap between the perturbation data and the original data is smaller, and privacy protection is weaker. In summary, there is conducive to maintaining the availability of data, but not conducive to protecting privacy information when the value of the parameter *k* is too large.

3. EXPERIMENTAL RESULTS AND ANALYSIS

We used the experiments to compare AIBR method, BSVD[2], SSVD[3] and kCG method[6]. The basic idea of the experiment is selecting the parameter values for previous four methods firstly to strengthen privacy protection under the premise of keeping data availability. Subsequently, it compares previous four methods through comparing privacy metrics under selected parameter values in each method. There is also a parameter k in BSVD method. The basic idea of BSVD is executing the singular value decomposition of the original data matrix firstly, and then reserving k components corresponding to the absolute value of the largest singular value. In BSVD method, the larger the parameter k is, the better the data availability is, the worse the protection of privacy data is. There are two parameters k and d in SSVD method. The basic idea of SSVD is perturbing data by using BSVD algorithm in parameter k firstly, then all the absolute value of element in matrix after singular value decomposition is less than d being setting zero. In SSVD method, the larger the parameter k is and the less the parameter d is, the better the data availability is, the worse the protection of privacy data is. In experiments, we use another parameter e to determine the value of parameter d in the SSVD method. E represents the matrix element ratio of the absolute value of element in matrix after singular value decomposition is less than d to all elements after the BSVD perturbation. Obviously, the bigger the value of the parameter e is, the larger the value of the parameter d is correspondingly. There is also a parameter k in kCG method. The basic idea of kCG is clustering the origin data and ensuring that each class contains at least k sample firstly, then generating a set of perturbed sample for each class separately. In kCG method, the smaller the parameter k is, the better the data availability is, the worse the protection of privacy data is.

Experiments use two actual data sets from UCI (University of California atIrvine) machine learning database. One is WBC (the Original Breast Cancer Wisconsin Data Set), and the other is PID (the Pima Indians Diabetes Data Set). Among them, WBC contains 9 attributes and 699 samples. Experiments use only one complete not duplicate samples, a total of 449. PID contains 8 attributes and 768 samples.

If R_o and R_p are the classifier classification accuracy in the original data and the disturbance data, $r = (R_o - R_p)/R_o$ is used as a perturbation data usability metrics. The smaller r represents better data availability. In experiments, three classification algorithms are used to calculate data usability metrics, which are the nearest neighbor classifier, support vector machine and J48 decision tree in WEKA(Waikato Environment for Knowledge Analysis)[7]. Supported when r<0.02, data availability can be accepted. At this time, availability metric can also be denoted by $\Delta R = R_p - (1-r)R_o$, in which r = 0.02. If $\Delta R > 0$, then the data availability can be accepted.

In experiments, each data set 80% samples are selected randomly as training sample, in which the remaining 20% samples as test samples. All experiments were repeated 50 times, experimental results are the average of the 50 experiments. Figure 2 and Figure 3 represent the data availability of WBC and PID perturbed by AIBR with different values of k. In experiments, the minimum value of parameter k is 1, the maximum value to the nearest integer n/20, in which n is the number of training samples and k increases by 1. As can be seen from the figures, the bigger k is, the better the data availability is, and the worse the protection of privacy is, just as the previous analysis. So in order to keep the data availability and maximize privacy protection strength, the value of k should be as large as possible. From figure 1 and 2, k should take 2 by using WBC, and take 4 by using PID in AIBR method.

Journal of Theoretical and Applied Information Technology
31st January 2013. Vol. 47 No.3
© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

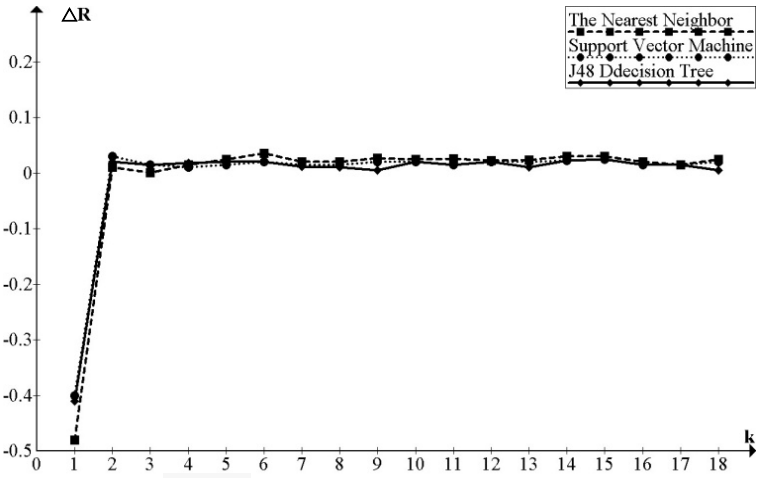

Figure1 The Data Usability Of WBC Perturbed By AIBR With Different Values Of K

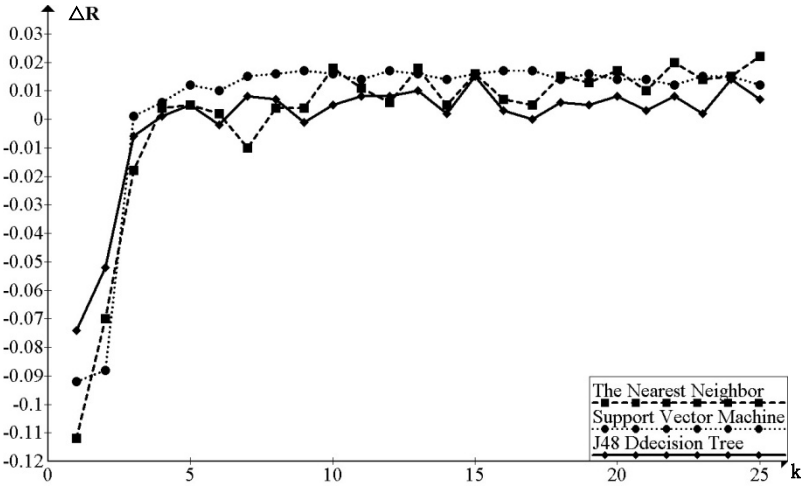

Figure2 The Data Usability Of PID Perturbed By AIBR With Different Values Of K

In order to keep the data availability and maximize privacy protection strength, k should take 7 by using WBC, and take 6 by using PID in BSVD method. e should take 0.45 by using WBC, and take 0.15 by using PID in SSVD method with k taking the same value as in BSVD method. k should take 9 by using WBC, and take 5 by using PID in kCG method.

Table 1 The Contrast Of AIBR And The Other Methods Through Privacy Metrics

Data	Method	VD	RP	RK	CP	CK
WBC	AIBR	0.42	72.5	0.006	0.8	0.4
WBC	kCG	0.41	57.9	0.007	0.3	0.7
WBC	BSVD	0.11	31.9	0.019	0.3	0.8
WBC	SSVD	0.25	37.3	0.015	0.3	0.8
PID	AIBR	0.50	111.3	0.007	0.3	0.8
PID	kCG	0.19	107.1	0.008	0	1
PID	BSVD	0.01	48.3	0.126	0	1
PID	SSVD	0.03	56.2	0.064	0	1

Five privacy metrics are used in experiments, which are VD, RP, RK, CP and CK[3-6]. Assuming that the original sample matrix is A, disturbed sample matrix is MA, then VD is the relative error between A and MA under F norm. RP, RK, CP and CK are used to measure the rank difference between A and MA elements. To protect the privacy better, VD, RP and CP should be larger, RK and CK should be smaller. We contrast AIBR with other methods through 5 privacy metrics in table 1. It founds that privacy protection of AIBR is best, using not only WBC but also PID.

4. CONCLUSIONS

To solve privacy protection problems in classification mining, this paper uses the data perturbation method based on randomization, and proposes an algorithm-irrelevant privacy protection method. The method attempts to independently

© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

generate a set of new data with the same distribution of original data, and discloses it as the perturbation data. Because of the perturbation data and the original data are independently generated, therefore, the exact value of the original data couldn't be obtained from the perturbation data directly. In addition, due to the perturbation data and the original data having the same distribution, ordinary data mining method can be applied directly to the perturbation data, in order to find the original data model.

This method is divided into two steps. First of all, a set of independent and identically distributed data for each attribute are generated without considering the links between each attribute. Subsequently, the link between the various attributes is restored by the relations of sequences, before generating the perturbation data. Experimental results show that the perturbation data of this method have good availability. The ordinary classification methods can be applied directly to the perturbation data of the method. The classifier has similar classification accuracy with the classifier original data trained. Moreover, the existing algorithm independent privacy protection method is compared, this method can provide stronger privacy protection comparing with existing algorithm-irrelevant privacy protection method, and it keeps the availability of data at the same time.

REFRENCES:

[1] HAN Jian-min,CENTing-ting,YUHui-qun. Re-search in Microaggregation Algorithms for k-Anonymization[J]. *Acta Electronica Sinica* , 2008, 6(10):2021–2029.

[2] S.Xu, J.Zhang, D.Han, et al. Singular Value Decomposition Based Data Distortion Strategy for Privacy Protection[J]. *Knowledge and Information Systems*, 2006, 10(3):383–397.

[3] J.Wang, J.Zhang, S. Xu, et al. A Novel Data Distortion Approach via Selective SSVD for Privacy Protection[J]. *International Journal of Information and Computer Security*,2008, 2(1):48–70.

[4] J.Wang, W.Zhong, J.Zhang. NNMF-based Fac-torization Techniques for High-accuracy Priva-cy Protection on Non-negative-valued Datasets [C]//*Proceedings of the sixth IEEE Internation-al Conference on Data Mining Workshops*. Washington, DC, USA: IEEE Computer Society, 2006:513–517.

[5] L. Liu, J. Wang, J. Zhang. Wavelet-based Data Perturbation for Simultaneous Privacy preserv-ing and Statistics preserving[C]//*Proceedings of the eighth IEEE International Conference on Data Mining-Workshops*. Washington, DC, USA: IEEE Computer Society, 2008:27–35.

[6] C. Aggarwal, P. Yu. A Condensation Approach to Privacy Preserving Data Mining[J].*Lecture Notes in Computer Science*, 2004, 2992:183–199.

[7] I.Witten, E.Frank. Data Mining: Practical Mach-ine Learning Tools and Techniques (second Edition)[M]. Burlington, Massachusetts, USA: Morgan Kaufmann, 2006.

[8] R. Agrawal, R. Srikant. Privacy-preserving Data Mining[J]. *ACM SIGMOD Record*, 2000, 29(2):439–450.

[9] D. Agrawal, C. Aggarwal. On the Design and Quantification of Privacy Preserving Data Mining Algorithms[C]// *Proceedings of the 20th ACM SIMOD Symposium on Principles of Database Systems*. New York, NY, USA: ACM, 2001:247–255.

[10] W. Du, Z. Zhan. Using Randomized Response Techniques for Privacy-preserving Data Mining [C]//*Proceedings of the ninth ACM SIGKDD international conference on Knowledge discov-ery and data mining*. New York, NY, USA: ACM, 2003:505–510.

[11] Ge Weiping, Wang Wei, Zhou Haofeng, et al. Privacy Preserving Classification Mining[J]. *Journal of Computer Research and Develop-ment*,2006, 43(1):39-45.