Send Orders for Reprints to reprints@benthamscience.ae

The Open Cybernetics & Systemics Journal, 2015, *9*, 443-447

Open Access

A New Algorithm-independent Method for Privacy-Preserving Classification Based on Sample Generation

Guang Li* and Meng Xi

School of Electronic and Control Engineering, Chang'an University, Xi'an, Shaanxi, 710064, P.R. China

Abstract: With the development of data mining technologies, privacy protection is becoming a challenge for data mining applications in many fields. To solve this problem, many PPDM (privacy-preserving data mining) methods have been proposed. One important type of PPDM method is based on data perturbation. Only part of the data-perturbation-based methods is algorithm-irrelevant, which are favorable because common data mining algorithms can be used directly. This paper proposes a new algorithm-irrelevant PPDM method for classification based on sample generation. This method is a data-perturbation-based method and has three steps. First, it trains classifiers use the original data. Then, it generates new samples as the perturbed data randomly. Finally, it use the classifiers trained in the first step to predict these samples' category. The experiments show that this new method can produce usable data while protecting privacy well.

Keywords: Data mining, data perturbation, privacy preserving.

1. INTRODUCTION

Data mining is the process of extracting patterns from data. With the rapid development of data mining, preserving data privacy poses an increasing challenge to its application in many fields. To solve this problem, PPDM (privacy-preserving data mining) methods have been studied [1-3]. PPDM technology can perform data mining without accessing the details of the original data directly.

In the past decade, many PPDM methods have been developed. They can be divided into two main categories. The methods in the first category are based on data perturbation [4-6]. In these methods, the original data are not open, and users can only access perturbed data. The data mining is done on the perturbed data to extract patterns about the original data. The methods in the second category are based on SMC (secure multi-party computation) [7, 8]. They are often used for distributed databases. They assume that there are multiple nodes, each of which contains only a part of the global data set. These nodes want to carry out data mining on the global data set, but each node does not want others nodes to know its data. In these methods, all of the nodes exchange the information required by the mining algorithm through information exchange protocols based on SMC. These protocols allow the information to be exchanged privately, without allowing any node to obtain data directly from other nodes.

Many PPDM methods are algorithm-relevant. That means that, when extracting the patterns from the data, the common data mining methods cannot be used directly, and we have to modify them to fit the privacy protection methods. Obviously, it is inconvenient to apply algorithm-relevant PPDM method.

*Address correspondence to this author at the School of Electronic and Control Engineering, Chang'an University, Xi'an, Shaanxi, 710064, P.R. China; Tel: +86-18717365351; E-mail: hit6006@126.com

Only part of the PPDM methods are not algorithm relevant, which includes two main kinds: the k-anonymity model [4-6] and methods based on matrix decompositions and transformation [9-11], both are perturbation-based methods. They overcome the shortcoming of the algorithm-relevant methods.

This paper presents a new algorithm-irrelevant PPDM method based on data perturbation for classification. Our method generates and opens a new data set that is different from the original data set, and it retains the information that is important for training classifiers.

Our method has three steps. First, classifiers are training by using the original data. Then we generate new samples randomly. Finally, these samples' categories are predicted by using the classifiers trained in the first step. The experiments show that this method has a high performance. It can maintain good data utility and can protect privacy well.

The rest of this paper is organized in the following way: Section 2 introduces the new PPDM method proposed by this paper, Section 3 shows the experimental results and Section 4 contains the conclusion.

2. THE NEW METHOD

This paper believes that the basic idea of the perturbation-based PPDM method is that the data contains a lot of information and only part of it is important for data mining. If the perturbed data contains that part of information which is important for data mining, the data mining task can be finished by using the perturbed data. And if the relation between the original and the perturbed data is weak, it is very difficult to get the original data from the perturbed one and the privacy is protected.

This paper also believes that if the perturbed data contains that useful information in the same way as that in the

1874-110X/15

original data, the PPDM method should be algorithm-independent. The ordinary data mining method, which is designed for original data, can be used directly on the perturbed data. On the other hand, for the algorithm-relevant method, the perturbed data also contains that useful information but in a different way from that in the original data. So the ordinary data mining method cannot be used directly.

Based on above analysis, this paper presents a new algorithm-irrelevant PPDM method for classification. Our method has three steps. First, classifiers are trained by using the original data. We think that the useful information for classification is recorded in these classifiers. Then we generate new samples, which will be the perturbed data, randomly. Because the perturbed data is generated randomly, so the relation between it and the original data is very weak and privacy is protected well. In the third step, the perturbed samples' category is predicted by the classifiers trained in the first step. In this step, the important information extracted in the first step is added to the data generated in the second step in the same way as that in the original data. Fig. (**1**) shows the workflow of this algorithm in detail.

Input: The original data set A, which has n records and d attributes. A's category vector c.

Output: The perturbed data set MA, and its category vector Mc

classifiers $C_1, C_2, \ldots, C_k$ is trained by using A and c

for each attribute R_j

D_j is A's projection on R_j

x^j_{max} and x^j_{min} are, respectively the biggest and the smallest value in D_j

interval $I^j = [x^j_{min}, x^j_{max}]$

end for

$SampleNum = 0$

$MA = \emptyset$

while $SampleNum < n$

Generate $Y = (y^1, y^2, \ldots, y^d)$, y^j is uniformly distributed on I^j

classifier C_i predicted Y belong to category pc_i

if $pc_1 = pc_2 = \ldots = pc_k$

$SampleNum = SampleNum + 1$

$MA = MA + Y$

$Mc[SampleNum] = pc_1$

end if

end while

Fig. (1). The new PPDM algorithm.

Let A be the original data containing the records $\{X_1, X_2, \ldots, X_n\}$. Let c be the category vector for A, that means $c[i]$, the i-th element of c, is the category of X_i. Let us also assume that each original record X_i contains the d dimensions, which are denoted by $(x^1_i, x^2_i, \ldots, x^d_i)$. Then, A's projection on the j-th attribute is $D_j = \{x^j_1, x^j_2, \ldots, x^j_n\}$. Let x^j_{max} and x^j_{min} be the biggest and the smallest values in D_j. Let $C_1, C_2, \ldots, C_k$ be classifiers trained by using A and c.

New samples are generated one by one. Let $Y = (y^1, y^2, \ldots, y^d)$ be a new sample. The y^j is generated randomly from the uniform distribution of the interval $[x^j_{min}, x^j_{max}]$. We will use classifiers $C_1, C_2, \ldots, C_k$ to predicted category of Y. Each classifier will give its own prediction. Only if all the classifiers give the consensus forecast, the sample will add to the perturbed data set and its category is the predicted one. If there are two classifiers in $C_1, C_2, \ldots, C_k$ give different prediction, this sample will be discarded. This process will be stopped when generating new samples as many as the original samples.

3. EXPERIMENTS

3.1. Utility Measures

Data utility measured whether a data set yields similar performance for data mining techniques after data distortion, *e.g.*, whether the patterns of the original data can be extracted from the perturbed data. This paper chose three kinds of classifiers, namely, the j48 decision tree in WEKA (Waikato Environment for Knowledge Analysis) [12], the nearest neighbour (nn) classifier and the SVM (Support Vector Machine) for measuring data utility. We assumed that the classification accuracy of classifiers trained on the perturbed data and the original data are R_p and R_o, respectively. If $\Delta R = R_p - (1 - \epsilon)R_o \geq 0$ for all the three classifiers, the perturbation method is believed to maintain data utility. This paper sets $\epsilon = 0.02$.

3.2. Privacy Measures

We used the privacy measures of the PPDM methods based on matrix factorization [9-11]. We assumed that the original data are denoted by A and the modified data are denoted by MA. A and MA are both $n \times m$ matrices. There are five privacy measures, namely, VD, RP, RK, CP and CK.

The first measure, VD, is the ratio of the Frobenius norm of the difference of MA from A to the Frobenius norm of A. It is calculated as Equation (1).

$$VD = \frac{||A - MA||_F}{||A||_F} \tag{1}$$

The Frobenius norm of an $n \times m$ matrix A, where the i-th row and j-th column entry is a_{ij}, is calculated as Equation (2).

$$||A||_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2} \tag{2}$$

If $Rank^i_j$ and $MRank^i_j$ denote the rank in the ascending order of the j-th element in the i-th attribute in A and MA, respectively, the second measure, RP, is defined as Equation (3).

$$RP = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left|Rank^i_j - MRank^i_j\right|}{nm} \tag{3}$$

The third measure, RK, represents the percentage of elements that maintain their ranks in each column after data distortion. RK is computed as Equation (4).

$$RK = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} Rk_j^i}{nm} \tag{4}$$

where,

$$Rk_j^i = \begin{cases} 1 & Rank_j^i = MRank_j^i \\ 0 & Rank_j^i \neq MRank_j^i \end{cases}$$

The fourth measure, CP, is used to define the change in rank of the average value of the attributes. If $RankV_i$ and $MRankV_i$ are ranks in ascending order of the average value of the *i*-th attribute in A and MA, respectively, CP is defined as Equation (5).

$$CP = \frac{\sum_{i=1}^{m} |RankV_i - MRankV_i|}{m} \tag{5}$$

As in the case of RK, we define the fifth measure CK to quantify the percentage of attributes that maintain their ranks of average values after data distortion. It is calculated as Equation (6).

$$CK = \frac{\sum_{i=1}^{m} Ck_i}{m} \tag{6}$$

where,

$$Ck_i = \begin{cases} 1 & RankV_i = MRankV_i \\ 0 & RankV_i \neq MRankV_i \end{cases}$$

To summarize, VD is the relative value difference in the Frobenius norm. RP, RK, CP and CK all measure the rank difference of data elements. Simply put, if privacy is protected better, VD, RP and CP will have larger values, and RK and CK will have smaller values.

3.3. Databases

We used two real-life databases for our experiments. They were the Pima Indians Diabetes Data Set (PID) and the Iris Data Set (Iris). They are both from the University of California at Irvine's Machine Learning Repository. The PID database has 8 attributes and 768 samples. And the Iris database has 4 attributes and 150 samples. In our experiments, for both databases, 20% of the samples were selected randomly as testing samples, and the other 80% of the samples were used as training samples.

3.4. Experiments Result

All the experiments were repeated 50 times and all the experimental data was the average of these 50 times.

The comparing methods are the SVD-based methods [9, 10]. The SVD-based methods take two main forms. One is the basic SVD-based method, and the other is the sparsified SVD-based method, or SSVD-based method. The SSVD-based method is an improvement over the basic SVD-based method. It did additional perturbation on the result of the basic SVD-based method.

Fig. (**2**) shows the utility measures for databases perturbed by the new method. When using the new method, the j48 decision tree and the nearest neighbor classifier are trained in its first step. Fig. (**2**) shows that our new method maintains data utility. For all cases, $\Delta R > 0$. Especially, even if the classification algorithm is not used in the first step of the proposed method, like the SVM, it still can be used di-

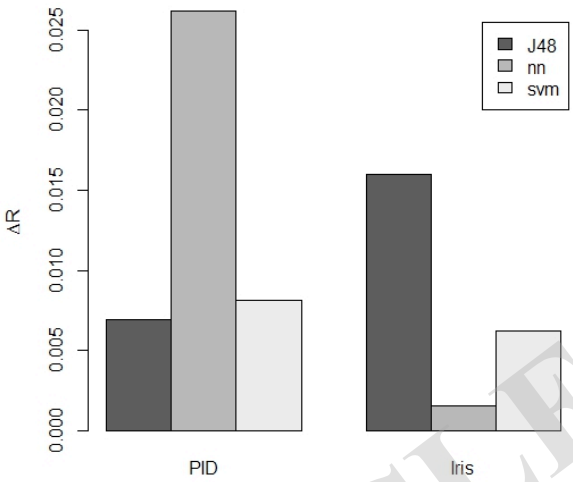

Fig. (2). The data utility measures for databases perturbed by our new method.

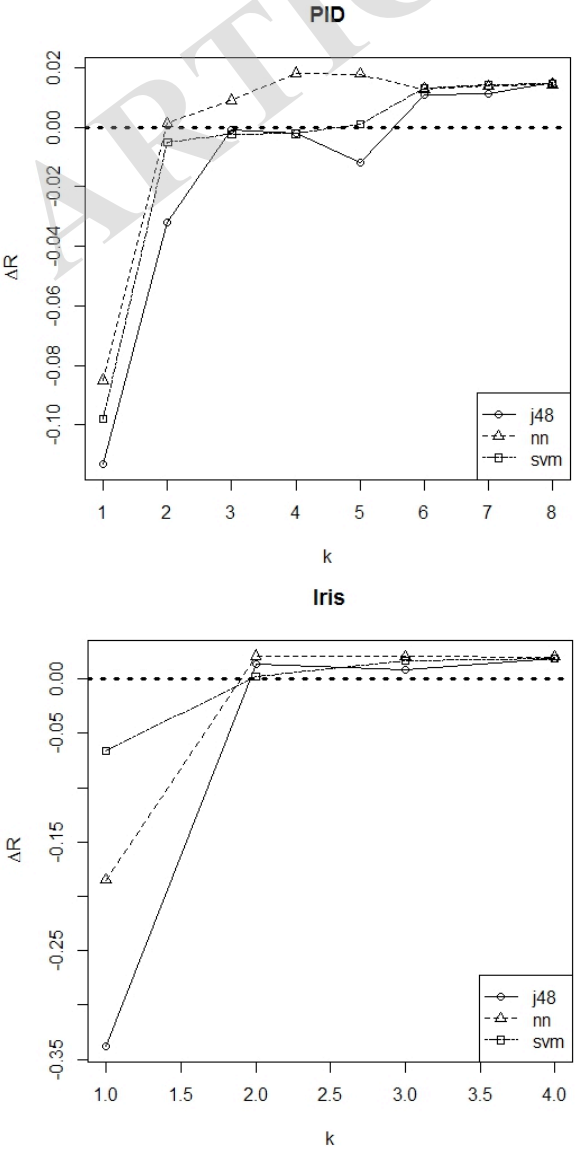

Fig. (3). The data utility measures for databases perturbed by the basic SVD-based method with different values of *k*.

rectly on the perturbed data to get good classifier for the original data.

The basic SVD-based method has only one parameter k. The larger values of k result in better data utility and worse privacy protection. Fig. (**3**) shows the utility measures for databases perturbed by the basic SVD-based method with different values of k. We selected the appropriate parameters' values, which would guarantee data utility and optimize privacy measures, with these experiment results. The appropriate value of k is 6 for PID and 2 for Iris.

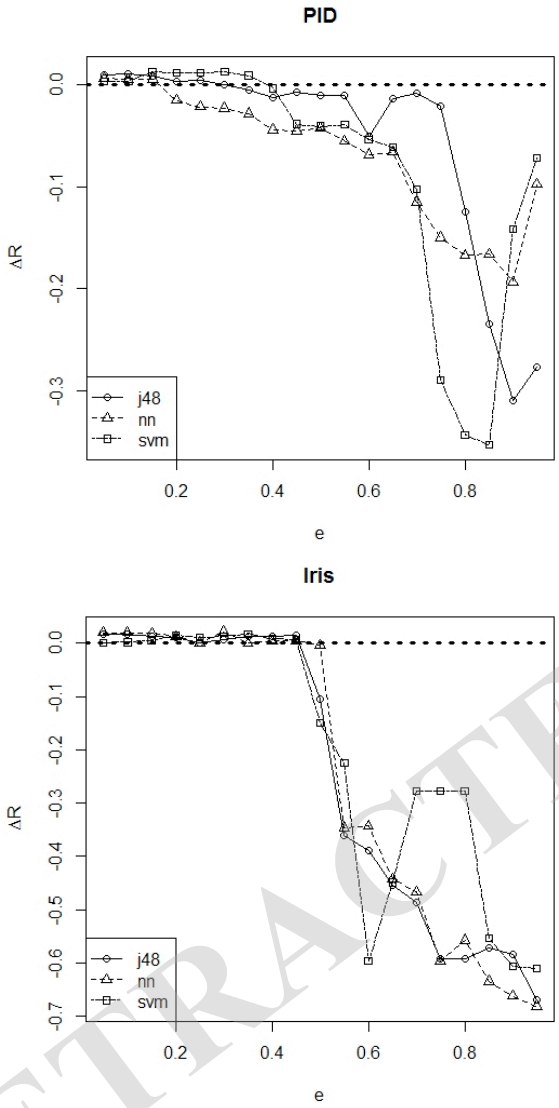

Fig. (4). The data utility measures for databases perturbed by the SSVD-based method with different values of e and the appropriate value of k.

The SSVD-based method has two parameters, which are k and e. The SSVD-based method firstly use the basic SVD-based method with parameter k to perturb data, then, it turn some minor elements of the reserved components to zero in order to implement additional perturbation of the modified data generated by the basic SVD-based method from the first step. The parameter e is the proportion of the elements turned to be zero. In the SSVD-based method, the parameter k should be as small as possible and the parameter e should be as large as possible, while guaranteeing data utility. We use a greedy strategy to select the parameters' appropriate values. Let the value of k in the SSVD-based method be equal to the appropriate value of k in the basic SVD-based method to maximize the perturbation in its first step while maintaining data utility. That implies that $k = 6$ for the PID and $k = 2$ for the Iris. Then select as large a value of e as possible to guarantee data utility. Fig. (**4**) shows the utility measures for databases perturbed by the SSVD-based method with the appropriate value of k and different values of e. The appropriate value of e should be 0.15 for PID and 0.45 for Iris.

Table **1** shows the privacy measures for our new method and the SVD-based methods with their parameters' appropriate values determined in the above experiments. In Table **1**, the "Our Method" means our new method, the "BSVD" means the basic SVD-based method, and the "SSVD" means the SSVD-based method. Based on Table **1**, note that when the data utility is maintained, our method can protect privacy better than the comparing ones.

CONCLUSION

This paper proposes a new algorithm-irrelevant PPDM method based on data perturbation. The basic idea of our method is to generate a new data set and open it instead of the original data. The new data has weak relation with the original data and the important information for classification is retained in it in the same way as that in the original data. Because the open data has weak relation with the original data, the privacy is protected. And because the important information for classification is retained in the same way in both the open and the original data, the classification can

Table 1. The privacy measures for our method and the comparing methods.

Data	PPDM Method	VD	RP	RK	CP	CK
PID	BSVD	0.01	48.3	0.126	0	1
PID	SSVD	0.03	56.2	0.064	0	1
PID	Our Method	1.93	204.6	0.002	0.60	0.48
Iris	BSVD	0.04	7.82	0.11	0	1
Iris	SSVD	0.35	14.62	0.05	0	1
Iris	Our Method	0.38	39.96	0.01	0	1

finished only using the open data and the common classification algorithms can be used directly.

The new algorithm has three steps. First, a serious of classifiers is trained by using the original data. Then, new samples are generated randomly. Finally, we use the classifiers trained in the first step predict the category of the samples generated in the second step. Using experiments, we demonstrated that our new method can maintain good data utility while preserving privacy.

CONFLICT OF INTEREST

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

ACKNOWLEDGEMENTS

This work was financially supported by the National Nature Science Foundation of China (51407012) and the Chang'an University's Fundamental Research Funds (0009-2014G6114024).

REFERENCES

[1] B. Fung, K. Wang, R. Chen, and P. Yu, "Privacy-preserving data publishing: a survey of recent developments", *ACM Computing Surveys*, vol. 42, pp. 14:1-14:53, 2010.

[2] E. Bertino, I. Fovino, and L. Provenza, "A framework for evaluating privacy preserving data mining algorithms", *Data Mining and Knowledge Discovery*, vol. 11, pp. 121-154, 2005.

[3] V. Verykios, E. Bertino, I. Fovino, L. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining", *SIGMOD Record*, vol. 33, pp. 50-57, 2004.

[4] B. Fung, K. Wang, and P. Yu, "Anonymizing classification data for privacy preservation", *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, pp. 711-725, 2007.

[5] K. Babu, N. Reddy, N. Kumar, M. Elliot, and S. Jena, "Achieving k-anonymity using improved greedy heuristics for very large relational databases", *Transactions on Data Privacy*, vol. 6, pp. 1-17, 2013.

[6] S. Kisilevich, L. Rokach, Y. Elovici, and B. Shapira, "Efficient multidimensional suppression for K-anonymity", *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 334-347, 2010.

[7] B. Pinkas, "Cryptographic techniques for privacy-preserving data mining", *ACM SIGKDD Explorations Newsletter*, vol. 4, pp. 12-19, 2002.

[8] F. Emekci, O. Sahin, D. Agrawal, and A. Abbadi, "Privacy preserving decision tree learning over multiple parties", *Data & Knowledge Engineering*, vol. 63, pp. 348-361, 2007.

[9] S. Xu, J. Zhang, D. Han, and J. Wang, "Singular value decomposition based data distortion strategy for privacy protection", *Knowledge and Information Systems*, vol. 10, pp. 383-397, 2006.

[10] J. Wang, J. Zhang, S. Xu, and W. Zhong, "A novel data distortion approach *via* selective SSVD for privacy protection", *International Journal of Information and Computer Security*, vol. 2, pp. 48-70, 2008.

[11] J. Wang, W. Zhong, and J. Zhang, "NNMF-based factorization techniques for high-accuracy privacy protection on non-negative-valued datasets", In: *Proceedings of the 6th IEEE International Conference on Data Mining - Workshops*, Hong Kong, pp. 513-517, 2006.

[12] I. Witten, and E. Frank, Data Mining: *Practical Machine Learning Tools and Techniques* (3rd ed), Massachusetts: USA, 2011.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014