Using Noise Addition Method Based on Pre-mining to Protect Healthcare Privacy

Likun Liu*, Kexin Yang*, Liang Hu*, Lina Li**

** College of Computer Science and Technology, Jilin University, Changchun, P. R. China (e-mail: likun_l@163.com, yangkx@jlu.edu.cn).*
*** Jilin Economic vocational and Technical College, Changchun, P. R. China*

Abstract: With medical device cyber-physical systems being more and more widely used, a lot of healthcare data are produced, making data sharing for health research a vital requirement. But, privacy concerns must be addressed before sharing and publishing any data set. Privacy-preserving data mining (PPDM) is an important technology to protect personal privacy. This paper begins with a proposal of two new noise addition algorithms for perturbing the original healthcare data, and then applies them to a two-step perturbation model. Experiments show that the algorithms given in this paper have much higher accuracy than existing ones under the similar privacy strength.

Keywords: privacy-preserving, data mining, healthcare, cyber-physical system.

1. INTRODUCTION

Cyber-physical system (CPS) is the seamless integration of networking technologies, embedded computer systems, sensor and actuator technologies. CPS research is revealing numerous opportunities and challenges in medicine and biomedical engineering. These include intelligent operating rooms and hospitals, image-guided surgery and therapy, fluid flow control for medicine and biological assays, and the development of physical and neural prostheses. Healthcare increasingly relies on medical devices and systems that are networked and needing to match the needs of patients with special circumstances (Radhakisan and Helen (2011)). The healthcare domain presents many promising applications for cyber-physical system, such as patient information management, real-time emergency reporting, elder living assistance (Shen et al. (2009)).

Gaining access to high-quality healthcare data is a vital requirement for healthcare institutes to extract or mine useful knowledge for research purposes. (YiYeh et al. (2010)) developed a decision support system to predict hospitalization of hemodialysis patients. However, healthcare data in its raw form often contains sensitive information about individuals, and mining or publishing such data will violate their privacy. As required by the Health Insurance Portability and Accountability Act (HIPAA), it is necessary to protect the privacy of patients and ensure the security of the medical data (Cios ans Moor (2002)).

To preserve the privacy on healthcare data, PPDM is one of the attractive techniques in data mining. PPDM perturbs the original dataset and then releases the result to the academic researchers. A trade-off between privacy and accuracy often needs 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 hand, accuracy needs that the "patterns" in the original data should be mined out in spite of the perturbation.

In this paper, there are two participants: government officials and academic researchers. The government officials collect the original healthcare data from different cyber-physical systems in regional healthcare centers, local hospitals and clinics, and then add noise to these original data. The academic researchers have only access to the obfuscated data and directly mine them (Fig. 1).

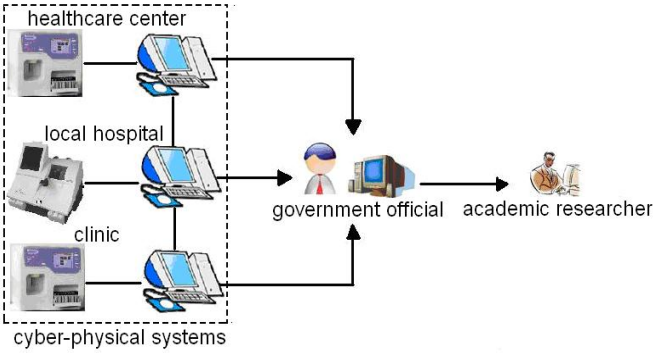

Fig. 1. *Cooperation among government, healthcare institutions and academia.*

We propose two new additive perturbation algorithms. Both of them can help the academic researchers mine out the "patterns" directly from the obfuscated data, and spare them from the usual 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. Further, our algorithms are applied to a two-step perturbation model, which composes additive perturbation with multiplicative perturbation, to enhance its privacy security.

2. RELATED WORK

CPS provides a closed interaction with ordinary people with embedded systems. Healthcare system is also a main application of CPS to enable patients to receive real-time medical care from the doctors and nurses. However, inappropriate sharing and usage of healthcare data could threaten personal privacy, so it necessitates the protection of the original healthcare data, especially for PPDM (Ge and Zhu (2011)).

The previous work for PPDM can be divided into two categories: data perturbation and data encryption (Elmisery and Fu (2010)). Data perturbation is more widely used than data encryption because of its low cost on computation and communications. It includes (but not limited to): additive noise (Haisheng (2010), Agrawal and Aggarwal (2001)), multiplicative noise (Kim and Winkler (2003)), matrix multiplication (Mohammad and Somayajulu (2010c)), data swapping (Fienberg and McIntyre (2003)), data shuffling (Muralidhar and Sarathy (2006)), k-anonymization (Poovammal and Ponnavaikko (2009), Mohammed et al. (2010)), blocking (Agrawal et al. (2004)). This paper focuses on two of them: additive noise and matrix multiplication, and their application to numeric continuous data will also be concerned.

(Khatri et al. (2010)) proposed architecture for privacy preserving in data mining by combining horizontal data distribution and vertical data distribution for breast cancer data set. Unfortunately, this method does not allow data owners to choose their desired privacy levels. An approach resolves this problem by reconstructing the original data distribution (Kargupta et al. (2003)), but the data needs to be separated from the random noise. Reconstruction of original distribution has been questioned for potential privacy breaches and the applicability (Liu et al. (2009)). In order to mine the data directly from the perturbed data, without reconstructing the original data distribution, (Liu et al. (2009)) proposed a threshold algorithm which uses a threshold to categorize a record by computing its probability. The choice of the threshold is crucial to the mining result, but not easy because the proper threshold value varies from case to case and can be set by no rules but experience. (So Mohammad and Somayajulu (2010a)) proposed a new noise addition scheme in which government officials firstly build a decision tree T by exploring the original data, and then for each record, add a noise 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 (Mohammad and Somayajulu's paper (2010a)), the result of mining the obfuscated data is close to mining the original data. But this method is limited by data sparsity. 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) (Kargupta et al. (2003)), singular value decomposition (SVD) filtering (Guo et al. (2006)), and principal component analysis (PCA) filtering (Huang et al. (2005)).

In the area of matrix multiplicative perturbation, distance-based preserving data perturbation (Yang (2009), Chen and Liu (2005), Liu et al. (2006)) has gain a lot of attention because it guarantees better accuracy. The transformed data is used as input for many important data mining algorithms, such as k-mean classification (Su et al. (2009)), k-nearest neighbor classification (Chong et al. (2010)) and distance-based clustering (Raaele et al. (2010)), and the corresponding output is exactly as same as the result of analyzing the original data. However the security issue of how much the privacy loss has caused researchers' concern. Kun Liu (Liu et al. (2006), Liu et al. (2008)) studied that how well an attacker can recover the original data from the transformed data and prior information. He proposed three different attack techniques based on prior information. (Giannella and Liu (2009)) made further study. They proposed a closed-form expression for the privacy breach probability and indicated that even with a small number of known inputs, the attack can achieve a high privacy breach probability.

Either additive perturbation or matrix multiplicative perturbation has the potential possibility of being attacked. (Chen et al. (2007)) considered a combination of matrix multiplicative and additive perturbation: $Y = MX + R$. This method makes it better to hide the original data. They also discussed a known I/O attack technique, and pointed out that $\hat{M}$, an estimate of M , can be produced using linear regression and then X is estimated.

Mohammad's (2010a) method is only applicable to building privacy-preserving decision tree. The two additive perturbation algorithms we proposed expand its application to security mine patients' information. The original data is pre-mined by the government officials to get the "patterns", and then after being added noise, the data is adjusted properly to keep the clusters similar to the ones in the original data. The academic researchers only need to mine the perturbed data directly without any extra work, so the step of reconstructing the original data distribution with its high computation cost and the step of modifying mining algorithm are both not needed any more. To protect privacy better, we address the application of our algorithms to a two-step model: $Y = M(X + R)$ which is not fit for building decision tree, but fit for statistical analysis. The first step of it gets the perturbed data by our algorithms, and the second step protects Euclidean distance of the perturbed data. In this way, computation cost is minimized and privacy is better preserved. 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 models.

3. TWO NOISE ADDITION ALGORITHMS

3.1 Approach overview

In additive noise algorithms, the original data X is replaced with

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

where R is the noise, generally satisfies independent and identically distribution (i. i. d).

If the original data set is D, government officials mine D with a k-mean clustering algorithm and get the result of clustering $C(C_1, C_2 ... C_k)$. This step is called pre-mining. We add noise to D and adjust the outcome according to $C(C_1, C_2 ... C_k)$, and then get $D^{'}$ which is different from D but has a similar cluster result $C^{'}(C_1^{'}, C_2^{'} ... C_k^{'})$. The government officials release the modified data set $D^{'}$ to academic researchers and are certain of its utility and privacy. We propose two noise addition algorithms 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.

3.2 RDD

After pre-mining, all records are categorized and form different clusters. Suppose there are K clusters and C_i is the center of cluster$_i$ ($1 \le i \le k$). Then we add noise to each record. Let $R = (R_1, R_2, ..., R_n)$ is an n-dimensional record and $N = (N_1, N_2, ..., N_n)$ is the n-dimensional noise. After adding N to R we get $P = (R_1 + N_1, R_2 + N_2, ..., R_n + N_n)$. In other words, P is the perturbed record of R. To simplify the demonstration, Fig. 2 shows a cluster, a ring with inner radius $r_1 = \min(dis(C_i, R))$ and outer radius $r_2 = \max(dis(C_i, R))$, formed by 2-dimensioal records. Obviously P falls into one of the three areas: the inner circle $D_{in}(i)$, the ring $D(i)$ and the area $D_{out}(i)$ further away from the center. To keep the "Pattern" unchanged before and after perturbing, we need to adjust P to keep it staying in the original cluster. Let $P^{'}$ is the final outcome after proper adjustments to P, and $P^{'}$ can be got through the distance between C_i and $P^{'}$ and the coordinate of C_i. The distance between C_i and $P^{'}$ can be computed by

$$dis(C_i, P^{'}) = \begin{cases} dis(C_i, P), P \in D(i), 1 \le i \le k \\ 2r_1 - dis(C_i, P), P \in D_{in}(i), 1 \le i \le k \\ 2r_2 - dis(C_i, P), P \in D_{out}(i), 1 \le i \le k \end{cases} \tag{2}$$

While $dis(C_i, P) = \sqrt{(C_{i1} - P_1)^2 + (C_{i2} - P_2)^2 + ... + (C_{in} - P_n)^2}$

Now, $P^{'}$ can be published to the academic researchers.

One parameter of Gauss noise is the mean. It is set to a fixed value (often zero) because it does not affect the Gaussian distribution. The other parameter is the variance which is related to the original data. The variance is regarded as the key factor which will affect the mining result of RDD. The larger the variance is, the lower the mining accuracy is.

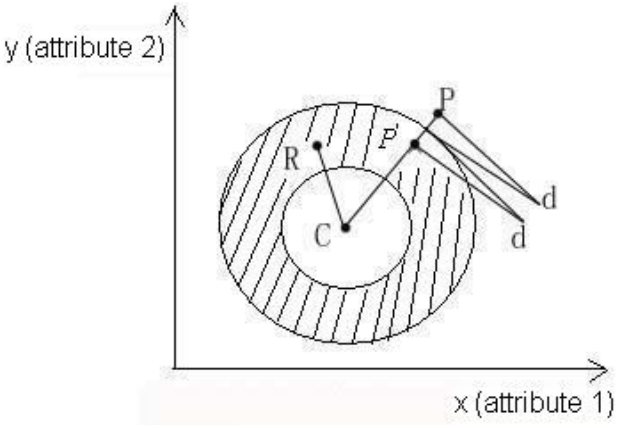

Fig. 2. RDD algorithm.

Program 1: RDD Algorithm
1: divide the dataset into k clusters using k-mean algorithm
2: for each Instance x_j do
3: find which cluster x_j is in
4: identify the domain of the cluster
5: add a small Gauss noise with mean zero and different variances
6: compute $dis(C_i, P)$
7: if($dis(C_i, P) < r_1$) then
8: $dis(C_i, P^{'}) = 2r_1 - dis(C_i, P)$
9: else if ($dis(C_i, P) > r_2$) then
10: $dis(C_i, P^{'}) = 2r_2 - dis(C_i, P)$
11: else
12: $dis(C_i, P^{'}) = dis(C_i, P)$
13: end if
14: end if
15: compute the coordinate of $P^{'}$ using $dis(C_i, P^{'})$ and the coordinate of C_i and line CP
16: end for

RDD, aiming at improving data privacy while maintaining data utility during the data perturbation process, achieves its goal by adding noise and adjusting the perturbed data. Noise-adding hides the real data, and adjusting twists the noise distribution to stop the attackers from recovering the original data from the perturbed data. Meanwhile because $R \in D(i)$ and $P^{'} \in D(i)$, adjusting also keeps all data, before and after perturbing, staying in the same cluster, so the "pattern" of the whole cluster is not changed after the perturbation process.

3.3 RACC

RACC does not directly perturb the original record R by a noise N, but by a random noise $\theta(\theta_1, \theta_2 ... \theta_n), \theta_i \in (0, 2\pi)$ and a random distance ratio $d_j (0 < d_j \le 1)$. The perturbed

record Q is computed by using (3) and (4). To simplify the demonstration, Fig. 3 shows a cluster, a ring whose center is C, formed by 2-dimensioal records.

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

$$\begin{bmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_n \end{bmatrix} = \begin{bmatrix} C_{i1} \\ C_{i2} \\ \vdots \\ C_{in} \end{bmatrix} + r \begin{bmatrix} \cos(\theta_1) \\ \cos(\theta_2) \\ \vdots \\ \cos(\theta_n) \end{bmatrix}, 0 < \theta_i < 2\pi \tag{4}$$

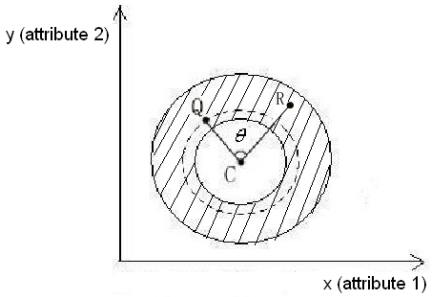

Fig. 3. RDD algorithm.

Program 2: RACC Algorithm
1: divide the dataset into k clusters using k-mean algorithm
2: for each Instance x_j *do*
3: find which cluster x_j *(point R) is in, and where the center* C_i *of that cluster is*
4: generate random Noise $\theta, \theta \in (0, 2\pi)$ *and random distance ratio* d_j
5: using (3), we compute $r = d_j \times dis(R, C)$
6: using (4), we compute the coordinate of Q
7: end for

RACC has the same goal as RDD has and it achieves it by rotating the record around the cluster center and increasing the density of the cluster. Rotation makes the perturbed records apart from the original ones and accordingly every attribute value changes, so privacy is preserved. But rotation does not throw the perturbed records away from their original clusters and so data utility does not be damaged. Increasing the density of the cluster also has many benefits. On the one hand, moving the records inwards varies every attribute value. On the other hand, each record still remains in their clusters. Furthermore, density-increasing makes clustering process more effective and efficient (Tan, Steinbach and Kumar, 2006).

4. TWO-SETP PERTURBATION

In finite field F^n, function T: $F^n \to F^n$, if for all $x, y \in F^n$, $\|x - y\| = \|T(x) - T(y)\|$, then it is said that Euclidean distance is preserved. In general matrix multiplicative perturbation model, the original data is replaced with

$$Y = MX \tag{5}$$

where M is a $p \times m$ matrix.

For (5), if $p = m$, M is a random orthogonal matrix($M^T M = I$), generated from a distribution function (i. i. d) with mean zero and variance σ^2 (Kargupta et al. (2003)). For any columns x_1, x_2 in original data X, by left-multiplication M, we get the columns y_1, y_2 in Y, satisfying $\|x_1 - y_1\| = \|y_1 - y_2\|$. In this method, Euclidean distance will be preserved with either small or no error, so it allows many important data mining algorithms to be applied to the perturbed data and produce results very similar to, or exactly the same as those produced by the original algorithm applied to the original data (Aggarwal and Yu (2008)).

For preserving privacy better, in two-step perturbation model, the original data will be replaced with

$$Y = M(X + R) \text{ or} \tag{6}$$

$$\begin{aligned} Y &= MX + R \\ &= M(X + M^T R) \end{aligned} \tag{7}$$

where M is a random orthogonal matrix, and the value of R is decided by function RDD or RACC. It is easy to verify that $M^T R$ is the rotation of R. $M^T R$ can be replaced with a new perturbed matrix $R^{'}$. Thus, (6) and (7) are equivalent. We will choose (6) as the complete version of perturbation.

5. MEASURES

5.1 Privacy Measures

Privacy Loss of Addition Perturbation (PLAP): A key privacy measure (Agrawal and Aggarwal, 2001) is based on the differential entropy of a random variable. The differential entropy $h(A)$ of a random variable A is defined as follows

$$h(A) = -\int_{\Omega_A} f_A(a) \log_2 f_A(a) da \tag{8}$$

Where Ω_A is the domain of A. Actually $h(A)$ is a measure of uncertainty inherent in the value of A in the statistics. In [7], it was proposed that $2^{h(A)}$ is a measure of privacy inherent in the random variable A. This value is denoted by $\Pi(A)$.

Given a random variable B, the conditional differential entropy of A is defined as follows

$$h(A \mid B) = -\int_{\Omega_{A,B}} f_{A,B}(a,b) \log_2 f_{A|B=b}(a) da\, db \tag{9}$$

Thus, the average conditional privacy of A given B is $\Pi(A \mid B) = 2^{h(A|B)}$. This motivates the following metric $P(A \mid B)$ for the conditional privacy loss of A, given B

$$P(A \mid B) = 1 - \Pi(A \mid B) / \Pi(A) = 1 - 2^{h(A|B)} / 2^{h(A)} = 1 - 2^{-I(A;B)} \tag{10}$$

where $I(A;B) = h(A) - h(A \mid B) = h(B) - h(B \mid A)$. $I(A;B)$ is also known as the mutual information between the random variables A and B . Clearly $P(A \mid B)$ is the fraction of privacy of A which is lost by revealing B . More details can be found in (Agrawal and Aggarwal's paper (2001)).

This paper chooses this privacy measure to quantify the privacy for addition perturbation in the experiments.

5.2 Accuracy Measures

This section describes a set of metrics that reflect the utility achieved in the perturbed datasets (Mohammad and Somayajulu (2010b)).

1. Average Loss of Distance (ALD): It measures the average loss of distance between the perturbed and original records. If N is the total number of records then

$$ALD = \frac{\sum_{=1}^{i} \sum_{=1}^{j} (d_{i,j} - \overline{d_{i,j}})}{Total\ Number\ of\ records\ compared} \tag{11}$$

Where, Total Number of records compared is the value $\frac{N!}{2!\times(N-2)!}$. $d_{i,j}$ is the distance between the i^{th} and j^{th} record of original data set. $\overline{d_{i,j}}$ is the distance between the i^{th} and j^{th} record of perturbed data set.

2. Fmeasure: The quality of clustering is measured using this metric. It is defined as

$$F_{i,j} = 2\frac{P_{i,j}R_{i,j}}{P_{i,j} + R_{i,j}}$$

Where, Precision

$$P_{i,j} = \frac{C_i \bigcap \overline{C_j}}{\overline{C_j}}$$

and the Recall

$$R_{i,j} = \frac{C_i \bigcap \overline{C_j}}{C_j}$$

The F measures of a class C_i is $F_i = \max(F_{i,j})$ and the over all Fmeasures is

$$F = \sum_{i=1}^{n} \frac{|C_i|}{N} F_i \tag{12}$$

3. Classification Accuracy (CA): It is a measure of how well the classifier labels the class for the test inputs. Higher the accuracy is, better the classier is. It is defined as

$$Accuracy = \frac{Number\ of\ samples\ correctly\ classified}{Total\ Number\ of\ samples} \tag{13}$$

6. EXPERIMENTAL RESULTS

Our experiments are implemented with Weka and the perturbed matrix is transformed by using Matlab 7.0. We use three real-world datasets (Liver Disorders, Pima Indians Diabetes and Thyroid Disease, denoted by Datasets 1 , Datasets 2, and Datasets 3 respectively), which were assembled from University of California Irvine (UCI), machine learning repository. Liver Disorders has 345 instances, Pima Indians Diabetes has 768 instances and Thyroid Disease has 7200 instances.

For each dataset, we choose 3 as the number of clusters. Both of additive noise and multiplicative noise are Gaussian distribution noise data. Our results are compared with Chen's Geometric Data Perturbation (Chen et al. (2007)). Ten groups of noise data are generated to perturb the original data. Privacy loss is measured by PLAP and PLMMP, while mining accuracy is measured by computing ALD, Fmeasures and CA.

As shown in Table 1, both of our algorithms have similar privacy loss with Chen's. And in Table 2, 3 and 4, the results show that in most cases our two algorithms can get higher accuracy.

Table 1. Privacy measure of different data sets (PLAP)

	RDD	RACC	Chen's
Datasets 1	0.21	0.198	0.201
Datasets 2	0.163	0.109	0.17
Datasets 3	0.185	0.132	0.174

Table 2. ALD (addition perturbation/two-step perturbation)

	RDD(%)	RACC(%)	Chen's(%)
Datasets 1	0.006/0.174	0.005/0.175	0.008/0.174
Datasets 2	0.008/0.082	0.004/0.08	0.01/0.082
Datasets 3	0.025/0.118	0.026/0.118	0.033/0.118

Table 3. Fmeasure (addition perturbation/two-step perturbation)

	RDD	RACC	Chen's
Datasets 1	0.66/0.66	0.643/0.643	0.639/0.62
Datasets 2	0.762/0.721	0.762/0.724	0.759/0.694
Datasets 3	0.54/0.466	0.545/0.479	0.536/0.462

Table 4. CA (addition perturbation/two-step perturbation)

	RDD(%)	RACC(%)	Chen's(%)
Datasets 1	98.67/94.06	96.67/96.02	86.00/79.69
Datasets 2	89.25/83.41	89.25/79.66	88.32/77.18
Datasets 3	74.39/60.69	82.01/73.37	73.92/59.38

When Gauss noise is used as random noise data, the variance can dramatically affect the result. The results (see Figs. 4, 5 and 6) show that with the increase of noise level, the accuracy has a decreasing trend in both Chen's algorithm and RDD, while the accuracy of RACC keeps relatively steady. In addition, mining accuracy is affected by the number of

instances. The more instances a dataset has, the less fluctuation the curves have.

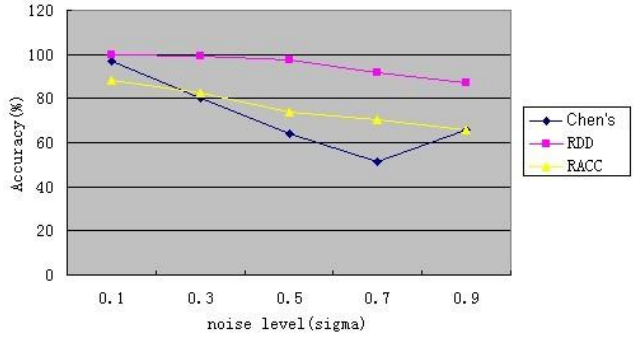

Fig. 4. Data mining accuracy of three perturbation algorithms on Liver Disorders.

In Fig. 4, when variance is 0.1, the mining accuracy of RACC is lower than that of Chen's because of small number of instances and small variance, so the original data is modified less in Chen's and the perturbed data is closer to cluster center.

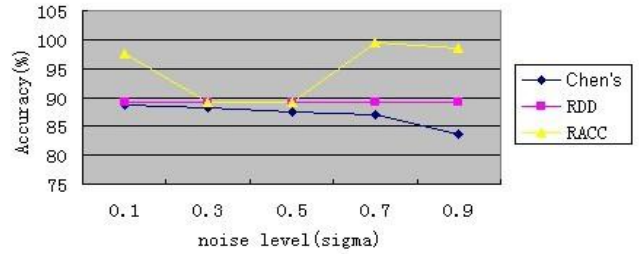

Fig. 5. Data mining accuracy of three perturbation algorithms on Pima Indians Diabetes.

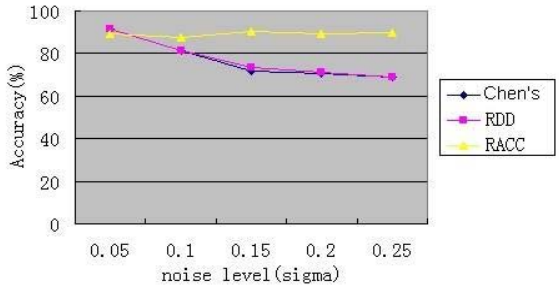

Fig. 6. Data mining accuracy of three perturbation algorithms on Thyroid Disease.

In Fig. 5 and Fig. 6, with the increase of noise level, the disturbance becomes stronger, and the original data is modified more, consequently for Chen's and RDD, the mining accuracy is decreasing, but for RACC, the mining accuracy is steady. RACC is not affected by noise level, because the perturbed data is generated by rotation which keeps data always in a small circle. The accuracy curves in Fig. 6 are smoother than Fig. 5 because the number of Thyroid Disease's instances is nearly ten times as many as that of Pima Indians Diabetes's.

7. CONCLUSIONS

To protect healthcare privacy in medical cyber-physical systems, we propose two additive perturbation algorithms RDD and RACC. It is proved that our two additive perturbation algorithms not only make the reconstruction with high computation cost unnecessary and keep the mining algorithm unmodified, but also have higher accuracy. In order to enhance the anti-attack capability, a two-step perturbation model which combines additive perturbation with matrix multiplicative perturbation is deployed. Besides improving its resilience to attack, matrix multiplicative strategy preserves Euclidean distance of the perturbed data with either small or no error.

In the future, we plan to expand them to the distributed data mining and apply our algorithms to the open, interoperable systems.

ACKNOWLEDGEMENT

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: Kexin Yang: phone: 86-431-88858785; fax: 86-431-85166494; email: yangkx@jlu.edu.cn

REFERENCES

Aggarwal, C. C. and Yu, P. S. (2008). Privacy-Preserving Data Mining. Models and Algorithms, *Vol. 34 of Advances in Database Systems*.

Agrawal, S., Krishnan, V. and Haritsa, J.R. (2004). On addressing efficiency concerns in privacy-preserving mining. In *Proceedings of the 9th International Conference on Database Systems for Advanced Applications*, 113–124.

Agrawal, D. and Aggarwal, C.C. (2001). On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems*. Santa Barbara, CA, 247-255.

Chen, K. and Liu, L. (2005). Privacy preserving data classification with rotation perturbation. In *Proceedings of the Fifth IEEE International Conference on Data Mining*. Houston, TX, 589-592.

Chen, K., Sun, G. and Liu, L. (2007). Towards attack-resilient geometric data perturbation. In *Proceedings of the 2007 SIAM International Conference on Data Mining*. Minneapolis, MN.

Chong, Z. H, Ni, W. W., Liu, T. T. and Zhang, Y. (2010). A privacy-preserving data publishing algorithm for clustering application. *Computer Research and Development*, 47(12), 2083-2089.

Cios, K. J. and Moor, G. W. (2002). Uniqueness of medical data mining. Artificial Intelligence in Medicine, 26(2), 1-24.

Elmisery, A. M. and Fu, H. G. (2010). Privacy Preserving Distributed Learning Clustering Of HealthCare Data Using Cryptography Protocols. In *Proceedings of the 34th International Computer Software and Applications Conference*. Seoul, Korea, 140-145.

E. Poovammal, M. Ponnavaikko. (2009). Task Independent Privacy Preserving Data Mining on Medical Dataset. In *Proceedings of the 2009 International Conference on Advances in Computing*, Control, & Telecommunication Technologies, 814-818.

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

Ge, X. J. and Zhu, J. M. (2011). Privacy Preserving Data Mining. New Fundamental Technologies in Data Mining, 535-560.

Guo, S., Wu, X. and Li, Y. (2006). On the lower bound of reconstruction error for spectral filtering based privacy preserving data mining. In *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*. Berlin, Germany, 520-527.

Giannella C and Liu K. (2009). On the Privacy of Euclidean Distance Preserving Data Perturbation. *Compute Science-Cryptography and Security*.

Huang, Z., Du, W. and Chen, B. (2005). Deriving private information from randomized data. In *Proceedings of the 2005 ACM SIGMOD Conference. Baltimore*, MD, 37-48.

Kargupta, H., Datta, S., Wang, Q. and Sivakumar, K. (2003). On the privacy preserving properties of random data perturbation techniques. In *Proceeding of the IEEE International Conference on Data Mining*. Melbourne, FL, 99-106.

Khatri, A., Kabra, S. and Singh, S. (2010). Architecture for Preserving Privacy During Data Mining by Hybridization of Partitioning on Medical Data, 93-97.

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

Li, H. S. (2010). Study of privacy preserving data mining. In *Proceedings of the International Symposium on Intelligent Information Technology and Security*. Jinggangshan, China, 700-703.

Liu, K. Kargupta, H. and Ryan, J. (2006). Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on knowledge and Data Engineering*, 18(1), 92-106.

Liu, K., Giannella, C. and Kargupta, H. (2006). An Attackers View of Distance Preserving Maps for Privacy Preserving Data Mining. In *the Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*. Berlin, Germany, 297-308.

Liu, K., Giannella, C. and Kargupta, H. (2008). A survey of Attack Techniques on Privacy-Preserving Data Perturbation Methods. In: *Privacy-Preserving Data Mining: Models and Algorithms*.

Liu, L., Kantarcioglu, M. and Thuraisingham, B. (2009). Privacy Preserving Decision Tree mining from Perturbed Data. *In proceedings of the 42th Hawaii International Conference on System Sciences*.

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

Mohammad, A. K. and Somayajulu, D.V.L.N. (2010a). A Noise Addition Scheme in Decision Tree for Privacy Preserving Data Mining. *Journal of Computing*, 2(1), 2151-9617.

Mohammad, A. K. and Somayajulu, D.V.L.N. (2010b). Privacy Preserving Technique for Euclidean Distance Based Mining Algorithms Using a Wavelet Related Transform. LNCS, 6283(1), 202-209.

Mohammad, A. K. and Somayajulu, D.V.L.N. (2010c). Privacy preserving technique for Euclidean distance based mining algorithms using a wavelet related transform. In *Proceedings of the 11th International Conference on Intelligent Data Engineering and Automated Learning. Paisley*, United Kingdom, 202-209.

Mohammed N., Benjamin and FUNG, C.M. (2010). Centralized and Distributed Anonymization for High-Dimensional Healthcare Data. ACM Transactions on Knowledge Discovery from Data, 4(4), Article 18.

Raaele Giancarlo, Giosue Lo Bosco, Luca Pinello. (2010). Distance functions, clustering algorithms and microarray data analysis. In *Proceedings of the 4th International Conference on Learning and Intelligent Optimization. Venice*, Italy, 125-138.

Radhakisan Baheti, Helen Gill. (2011). Cyber-physical Systems. (2011). [Online]. Available: http://ieeecss.org/main/images/documents/IoCT-Part3-02CyberphysicalSystems.pdf

Su, C. H., Zhan, J. and Sakurai. K. (2009). Importance of Data Standardization in Privacy-Preserving K-Means Clustering. In the *Proceedings of International Workshops on Database Systems for Advanced Applications*. Brisbane, QLD, Australia, 276-286.

Shen, H. Y., Li, Z. and Yang, L. L. (2009). A Distributed Cyber-based Information Distillation and Control Architecture for Wireless Healthcare Systems. In *Proceedings of the 1st ACM international workshop on Medical-grade wireless networks*. New Orleans, Louisiana, USA, 33-37.

Tan, P., Steinbach, M. and Kumar, V. (2006). Introduction to Data Mining. Addison-Wesley, Reading, MA.

Yeh, Y. J., Wu, T. H. and Tsao, C. W. (2010). Using data mining techniques to predict hospitalization of hemodialysis patients. *Decision Support Systems*, Vol. 50, 439-448.

Yang, W. J. (2009). Privacy protection by matrix transformation. *IEICE Transactions on Information and Systems*, E92-D(4), 740-741.