IEICE TRANS. INF. & SYST., VOL.E92–D, NO.4 APRIL 2009

LETTER

Privacy Protection by Matrix Transformation

Weijia YANG[†a)], ***Student Member***

SUMMARY Privacy preserving is indispensable in data mining. In this paper, we present a novel clustering method for distributed multi-party data sets using orthogonal transformation and data randomization techniques. Our method can not only protect privacy in face of collusion, but also achieve a higher level of accuracy compared to the existing methods.

key words: *data mining, privacy preserving, randomization*

1. Introduction

Privacy preserving data mining is becoming a popular research direction these years [1]. In this paper, we protect data privacy in the following scenario: A miner collects data distributed among multiple parties and performs clustering. To protect their privacy, the owners perturb their original data in such a manner that the miner cannot see the original data but can still get the same clustering results as those from the original data. A privacy preserving approach in the above scenario was recently proposed in [4]. In that method, we designed an "RD" matrix . Combining the RD matrix with orthogonal transformation, we proposed effective data processing ways for the participant and miner. However, the method is quite sensitive to the data dimension and does not have an ideal level of accuracy.

In this paper, we introduce a new method of data transformation. This method can not only protect the data privacy in face of collusion but also achieve a high accuracy level in the mining result.

2. Method

First of all, we briefly present the method in [4]. Then we analyze its disadvantages and introduce our method.

RD matrix In the work of [4], we developed the RD matrix to transform the source data matrix. Its diagonal elements $rd_{i,i} = 1$, and off diagonal elements $rd_{i,j}$, $i \neq j$ are realizations of mutually independent variables with the same normal distribution $N(\mu, \delta^2)$.

Data transmission Before the participants transmit their data, all of them are arranged in a circle. For the ith participant with data matrix X_i, it generates an orthogonal transformation matrix H_i and sends $Y_i = X_i \cdot H_i$ as its transformed data to the miner. Then, it generates two RD matrices RD_1 and RD_2, and sends $(RD_1 \cdot H_i)^T$ to its right neighbor $(i + 1)$. Next, it receives a matrix from its left neighbor $(i - 1)$, post-multiplies it with $RD_2 \cdot H_i$ and sends the product as the "perturbation matrix" $T_{i-1,i}$ to the miner.

Data integration The transformed data of each participant can be regarded as data in different coordinate systems. When the miner receives the perturbed data from all n participants, he arbitrarily chooses one party, and designates its data as the target coordinate system (TCS). For the ith party which is not TCS, the miner finds the shorter of the two paths to TCS, $(i, i+1, \ldots, TCS)$ or $(TCS, \ldots, i-1, i)$. Suppose former is the shorter path, then the miner transforms data matrix Y_i into the TCS coordinate system by multiplying $T_{i,i+1} \cdot T_{i+1,i+2} \cdot \ldots \cdot T_{TCS-1,TCS}$. After that, common mining algorithms are applied to discover knowledge.

Many privacy preserving methods including [4], use variance to quantify the privacy level of the data protection method. Suppose $X(i)$ is the ith column of the original data and $Y(i)$ is its perturbed form, both of which are normalized. Then, the greater the variance $Var(X(i) - Y(i))$, the more difficult it is for the miner to guess $X(i)$ from $Y(i)$. However, we find that if the perturbation is dependent on the original value, even a high value of variance will have problem. For example, if the perturbation: $Y(i) - X(i) = \alpha X(i)$, then a large value in $X(i)$ will also have a large amount of perturbation. As a result, the adversary can easily compare the original values in $X(i)$ by knowing $Y(i)$.

When we change the perturbation to $Y(i) - X(i) = \alpha X(j)$, where $X(j)$ and $X(i)$ are independent. Then the problem will be solved, because we can hardly scale the values in $X(i)$ any more. Even when α is known by the adversary, he can not infer $X(i)$ without first knowing $X(j)$. Therefore, the independence of the perturbation is more important to the privacy protection than the amount of the variance. Based on this understanding, we propose a new flexible random transformation matrix "FR" to replace "RD" in the above data transformation and integration processes.

FR matrix The generation of an $m \times m$ FR matrix is as follows: For the ith ($i \in [1, m]$) column in FR, we search in the original data D for the column which has the least dependence on the ith column of D. Let it be the kth ($k \neq i$) column. Then, we generate the kth value of the ith column in FR by normal distribution $N(0, \delta_i^2)$ and other off-diagonal elements to 0. Moreover, all diagonal elements of FR are set to 1 .

Since only one off-diagonal element in each column of

Manuscript received October 2, 2008.
Manuscript revised November 27, 2008.

[†]The author is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200030, China.
a) E-mail: weijia_yang@sjtu.edu.cn

DOI: 10.1587/transinf.E92.D.740

FR is randomly generated , both of the accuracy and privacy of data mining can be improved. We further analyze them in the following section.

3. Privacy and Accuracy

Privacy By using FR, both the columns and data values are protected. In a two-party (A, B) situation, the miner gets $H_A^T \cdot FR_A^T \cdot FR_B \cdot H_B$, $X_A \cdot H_A$ and $X_B \cdot H_B$. When A colludes with the miner by sharing its own matrices, the miner is only able to derive $X_B \cdot FR_B$ instead of X_B directly. Then, each value in the original data is protected by:

$$y_{i,j} = x_{i,j} + fr_{k,j} \cdot x_{i,k}$$

where $fr_{k,j}$ is the value generated by $N(0, \delta_j^2)$ in the jth column of FR. Since column $X(k)$ has the least dependence on $X(j)$ (according to the generation of FR), the individual privacy in the jth column is determined almost independently of the elements themselves.

Accuracy During clustering, the miner multiplies the transformed matrices with the perturbation matrices to estimate the dot product between original data matrices. In the above example, $Y_A \cdot T_{A,B} \cdot Y_B^T = X_A \cdot FR_A^T \cdot FR_B \cdot X_B^T$ is used to estimate the original $X_A \cdot X_B^T$. Thus, the error of the dot product is caused by a product of FR matrices. Suppose the off-diagonal elements in FR_A, FR_B are generated by the same $N(0, \delta^2)$. Then, for the diagonal elements of $FR_A^T \cdot FR_B$: $d_{i,i}, i \in [1, m]$ and off-diagonal elements: $d_{i,j}, i \neq j$, we can derive:

$$E(Var(d_{i,i})) = \frac{\delta^4}{m-1};\ E(Var(d_{i,j})) \approx \frac{2\delta^2 + \delta^4}{m-1}$$

Thus, for $X_A \cdot FR_A^T \cdot FR_B \cdot X_B^T$, we have its expected variance proportional to m (since it involves $m \times m$ elements), while in the RD matrix way, the corresponding variance increases by m^2.

The above analysis shows that the FR matrix provides a much better level of accuracy than RD, while the privacy is also protected even in face of collusion.

4. Experiment

We first compare the k-means clustering results between our FR method and RD method with different data dimensions. We randomly select 20, 30, 40 and 50 columns from the test data to form a new data set, and horizontally partition the data to simulate $n = 10$ different parties. For each party, we generate a random orthogonal matrix and FR matrices. All parties are assumed to use the same distribution $N(0, \delta^2)$ in FR matrices generation. As shown in Fig. 1 (a), with the increase of m, the error rate in FR increases much more slowly than that in RD. This further demonstrates our analysis in Sect. 3 that the FR method is more adaptive to larger dimensions than RD.

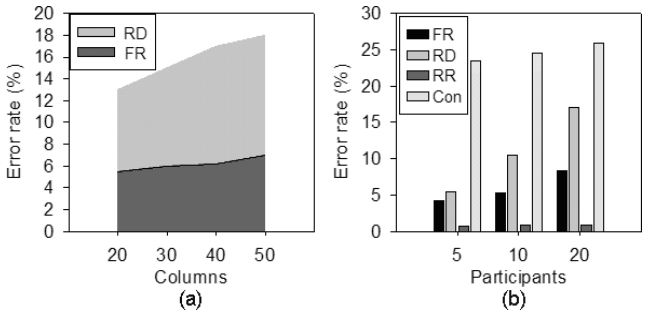

Fig. 1 Methods comparison. (a) Clustering with different number of columns. (b) Comparison of FR with other methods.

Then we further compare FR with RD, "Random Rotation" (RR) [3] and "Condensation" (Con) [2]. The RR method only uses orthogonal matrices to transform the data while the Con method combines the statistics from data partitions. Since our method actually reserves the distances among data records, we use k-NN classifier to conduct the comparison and set $k = 5$. By simulating $n = 5, 10, 20$ participants with the same distribution $N(0, 1/100\,n)$ to generate the FR and RD matrices, we find in Fig. 1 (b) that the FR way achieves a much lower error rate than RD and Con. This can be attributed to the lower variance achieved in the dot product in our method. While RR seems to well preserve the accuracy, it is vulnerable to adversaries [4].

5. Conclusions

In this paper, we introduce a novel method of privacy preserving clustering in homogeneous data sets. Our method of FR matrix leverages the orthogonal transformation in the normal conditions to avoid the compromise between privacy and accuracy, and also protects data from malicious adversaries with independent perturbation. Our method can further adapt well to the increase in data columns. Experiments demonstrates that our method is also able to achieve a much better level of accuracy compared to the existing methods.

References

[1] C.C. Aggarwal and P.S. Yu, "A general survey of privacy-preserving data mining models and algorithms," in Privacy-Preserving Data Mining Models and Algorithms, pp.11–52, Springer US, 2008.

[2] C.C. Aggarwal and P.S. Yu, "A condensation approach to privacy preserving data mining," Proc. 9th International Conference on Extending Database Technology, 2004.

[3] K, Chen and L, Liu, "Privacy preserving data classification with rotation perturbation," Proc. IEEE ICDM'05, 2005.

[4] W. Yang and S. Hang, "Privacy preserving clustering for multiparty," Proc. 12th International Conference on Database Systems for Advanced Applications, 2007.