Research On Suspicious Financial Transactions Recognition Based On Privacy-Preserving Of Classification Algorithm

Chunhua JU
Zhejiang Gongshang University; College of Computer Science & Information Engineering
Hangzhou , China
Jch@mail.hzic.edu.cn

Lili ZHENG
Zhejiang Gongshang University; College of Computer Science & Information Engineering
Hangzhou , China
Zhenglili378@163.com

***Abstract*—Classification basing on Privacy-preserving is one of the hottest spots in the field of data mining in recent years. This paper studies how to identify suspicious financial transactions under the privacy-preserving of classification algorithm. The research is studied on multi-data that from several parts by using Scalar Product Protocol under the privacy-preserving. The experimental results show that this method is a effective way for financial institutions to improve the efficiency of identify Money-laundering transactions.**

Keywords-Money- laundering, Scalar Product Protocol, Privacy-Preserving, Decision tree

I. INTRODUCTION

With the development of network technology and economic globalization, money-laundering has become a major problem that financial institutions authorities must face. Using data mining technology, such as classification, clustering analysis, can enhance the efficiency and accuracy of the identification of money laundering transaction. But because of the privacy-preserving, all parties of the collaboration don't want the others or any third party to learn much about their private data. It hindered the mining work to carry out. So how to protect private information or sensitive information during the mining process becomes a very significant study in data mining [1, 2].

Recently, using data mining techniques to improve the efficiency of the Anti-Money Laundering is still a hotspot [3].Bolton R J and Hand D J introduce some tools to solve the financial fraud detection problem, specially describe about decision tree [4]. Although the study in the area of anti-money-laundering application in our country is not early, there are some results [5, 6]. But most of the studies were a theoretically studies, not taking the privacy-preserving into account. In the classification based on privacy-preserving, for the vertical partitioned data, Jaideep Vaidya propose privacy preserving association rule mining in vertically partitioned data [1].Du Wenliang ,Zhan Zhijun propose a solution to the privacy preserving classification problem using the Scalar Product Protocol, a powerful tool developed by the secure multi-party computation[7]. As for the horizontal partitioned data, Lindell Y, Pinkas B propose a solution to the problem of privacy preserving classification by using encryption [8]. Another approach of solving the privacy preserving classification problem was proposed by Zhangqiang Yang [9].

In this paper, on the basis of the above study, we propose a new application to identify suspicious financial transactions by using classification algorithm over privacy-preserving. The experimental results show that this method pays more attention to privacy and accuracy than other methods. It provides a effective way for financial institutions to improve the efficiency of identifying Money-laundering transactions.

II. RELATED WORK

A. Classification problem Over Private Data

Classification is an important problem in the field of data mining. The decision tree models are found to be the most useful in the domain of data mining since they obtain reasonable accuracy and they are relatively inexpensive to compute. It performs classification in two phases: Tree Building and Tree Pruning. In tree building, the decision tree model is built by recursively splitting the training data set based on optimal information entropy until all or most of the records belonging to the same class label. To improve generalization of a decision tree, tree pruning is used to avoid over-fitting by pruning the leaves and branches responsible for classification of single or very few data vectors.

The classification problem based on privacy described in the following:

Suppose A and B each have a private data set (denoted by S1 and S2 respectively), want to collaboratively conduct decision tree classification on the union of their data sets. But they are not willing to disclose its raw data set to the others. So, the following assumptions are made on the data sets S_1 and S_2:

1) S_1 and S_2 contain the same number of data records. Let N represent the total number of data records.

2) S_1 contains some attributes for all records, S_2 contains the other attributes. Let n represent the total number of attributes.

3) Both parties share the class labels of all the records, also the names of all the attributes, so the collection of S_1, S_2 is a complete dataset.

978-0-7695-3557-9/09 $25.00 © 2009 IEEE
DOI 10.1109/ETCS.2009.378

B. Scalar Product Protocol

There are many methods of the Secure Multi-party Computation (SMC) [10], this paper uses Scalar Product. Its security is based on both side are able to calculate their private scalar product of vectors.

A has a vector a and B has another vector b, both of the vectors have n elements. A and B want to compute the scalar product between a and b, such that A gets V_1 and B gets V_2, where $V_1+V_2 = a \cdot b$ and V_2 is randomly generated by B. That is, the scalar product of a and b is divided into two secret pieces, with one piece going to A and the other going to B. The following computation is assumed in the real domain.

1) The semi-credible third-party generates two random vectors R_1 and R_2 of size n, and let $r_1 + r_2 = R_1 \cdot R_2$, where r_1 (or r_2) is a randomly generated number, then sends (R_1, r_1) to A, and (R_2, r_2) to B.

2) A sends $A = A + R_1$ to B, and B sends $B = B + R_2$ to A.

3) B generates a random number V_2, and computes $A \cdot B + (r_2 - V_2)$, then sends the result to A.

4) A computes $(A \cdot B + (r_1 - V_2)) - (R1 \cdot B) + r1 = A \cdot B - V_2 + (r_2 - R_1 \cdot R_2 + r_1) = A \cdot B - V_2 = V_1$.

Evidence: It ensure that A, B can not know other's data. Because B obtain $A' = A + R_1$, as R_1 is a result of randomness and confidentiality, B can not obtain related information with A. Omitted to prove.

III. Decision Tree Classification Over Privacy-Preserving

A. Basic Algorithm

The following is the procedure for building a decision tree on (S_1, S_2).

(1)If all of the samples belong to the same class or meet to other rules of termination, and then no longer divided, just form a leaf node;

(2)Otherwise, A and B respectively calculated the information gain of each attributes in subset of S_1, S_2;choose the largest information gain to make the division until build a tree;

Description of the code is as follows:

1. A computes information gain for each attribute of S_1. B computes information gain for each attribute of S_2. Initialize the root to be the attribute with the largest information gain
2. Initialize queue Q to contain the root
3. While Q is not empty do {
4. Let the first node A from Q
5. For each attribute B[i] (for i = 1… k), evaluate splits on attribute B[i]
6. Find the best split among these B[i]
7. Use the best split to split node A into A1, A2… Am
8. For i = 1… m, add Ai to Q if Ai is not well classified
9.}

B. Basic Formula

If a data set S contains examples from n classes, the Entropy(S) and the Gain(S) are defined as followings:

$$E(s) = -\sum_{j=1}^{n} p_j * \log p_j \qquad (1)$$

$$Gain(S, A) = E(S) - \sum_{v \in A} (\frac{|S_v|}{|S|} * E(S_v)) \qquad (2)$$

Where Pj is the relative frequency of class j in S. where v express any possible values of attribute A; Sv is the subset of S for which attribute A has value v; |Sv| is the number of elements in Sv; |S| is the number of elements in S.

C. Best Split Attribute Choose

Choose the best split attributes is the key to building a decision tree over privacy data. Because of private data, the process of choose properties become more complicated.

Let S represent the set of the data belonging to the current node a. Let R represent the set of requirements that each record in the current node has to satisfy. To compute the information gain for the attribute B[i], if all the attributes involved in R and B[i] belong to the same party, then this party can compute the information gain for B[i] by herself. However, it is unlikely that R and B[i] belong to the same party except the root node; so, there are two problems we need to solve: one is to compute Entropy(S), and the other is to compute Gain(S, B[i]) for each candidate attribute B[i](i = 1…k).

R is divided into two parts, R1 and R2, where R1 represents the subset of the conditions that only included in A attributes, and R2 represents the subset of the conditions that only included in B attributes. Let V1 represents a vector of size N. if the ith record satisfies R1, $V_1(i)=1$,else $V_1(i)=0$.Because R1 belong to the A, it can compute V_1 by herself. Similarly, the definition of V_2 is made, also about V_j. A, B will be able to compute Vj, let $V = V_1 \square V_2$ ($V(i) = V_1(i) \square V_2(i)$, i = 1…N),it means the corresponding record satisfies both R1 and R2. To build decision trees, it needs to compute how many entries in V is non-zero. This is equivalent to computing the scalar product of V_1 and V_2:

$$V_1 \cdot V_2 = \sum_{i=1}^{N} V_1(i) * V_2(i)$$

Neither part will disclose her private data to the other. The Scalar Product Protocol introduced in section 2.2 can enable A and B to compute $V_1 \cdot V_2$ without sharing information between each other.

Computing P_j after getting V_1, V_2, V_j, where Pj is the number of appearance of class j in partition S.

$$P_j = V_1 \cdot (V_2 \Lambda V_j) = (V_1 \Lambda V_j) \cdot V_2$$

After knowing P_j, P_j and $|S| = \sum_{i=1}^{N} P_j$ could be computed, Pj is the relative frequency of class j in S.

Therefore, Entropy(S) could be computed by using Equation (1). Repeat the above computation, obtaining $|S_v|$ and E (S_v) for all values v of attribute B[i]. Then compute Gain(S, B[i]) using Equation (2).

The above process will be repeated until all the information gain for all attributes B[i] for i = 1…k are got. Finally, choosing the split attribute for the current node of the tree, where Gain(S, B[m]) = Max {Gain(S, B [1]), … , Gain(S, B[k])}.

IV. RESEARCH ON MONEY LAUNDERING RECOGNITION BASED ON THE ABOVE ALGORITHM

A. Decision Tree Classification Process

According to the process of money laundering in suspicious financial transactions, from the tax point of view, the above data mining algorithms is applied to detect suspicious movement of funds on financial transactions data which belong to bank and tax data which belong to tax department. The experimental data is from the Zhejiang Province. Part of the data show in Table1 and Table 2.

TABLE I. A FINANCIAL TRANSACTION DATA

USER ID	TOTAL VOLUME OF TRANSACTIONS (SUM)	THE TOTAL NUMBER OF TRANSACTIONS (NUM)	SUSPICIOUS FINANCIAL TRANSACTIONS LABEL (C)
1	>35000	Normal	NO
2	>35000	Less	YES
3	5000-35000	Frequent	NO
4	>35000	Frequent	NO
5	<5000	Normal	NO
6	5000-35000	Less	YES
7	5000-35000	Normal	YES
8	>35000	Frequent	YES

TABLE II. B TAX DATA

USER ID	AGE	TAX	SUSPICIOUS FINANCIAL TRANSACTIONS LABEL (C)
1	Mid	>3500	NO
2	Old	<3500	YES
3	Mid	>3500	NO
4	Young	>3500	NO
5	Old	>3500	NO
6	Old	<3500	YES
7	Young	<3500	YES
8	Young	>3500	YES

Application steps:

(1)According to the formula (1), (2), A and B respectively compute information gain of its attributes, such as SUM,NUM,AGE,TAX. Then AB exchange their data by using Scalar Product Protocol and choose the largest information gain of the attribute as a root node. In this case, sum is the best split attribute.

(2) According to the value of SUM, dataset A is divided into three subsets S_{sum}>35000, S_{sum}5000-35000, S_{sum}<5000, then calculate Gain（S_{sum}>35000, NUM）, Gain（S_{sum}>35000, AGE）, Gain（S_{sum}>35000, TAX）, the first A can compute separate, but the latter; neither party can compute the information gain by herself. Since A knows all the necessary information to compute Gain（S_{sum}>35000, NUM）, A can compute it by herself. In order to compute Gain（S_{sum}>35000, AGE）, Gain（S_{sum}>35000, TAX）, A and B has to collaborate;

For instance：

$$Gain(S_{sum>3500,Tax}) = E(S_{sum>35000}) - |S_v = Tax > 3500| / |S_{sum>35000}| \cdot E(S_{sum>35000,Tax>3500}) - |S_v = Tax < 3500| / |S_{sum>35000}| \cdot E(S_{sum>35000,Tax<3500})$$

$$E(S_{sum>35000,Tax>3500}) = -\sum_{i=1}^{2} q_i \log_2 q_i$$

$$q_1 = (V_a(S_{sum>35000\_yes}) \cdot (V_b(Tax > 3500)) / (V_a(S_{sum>35000}) \cdot (V_b(Tax > 3500))$$

Where

$$q_2 = (V_a(S_{sum>35000\_no}) \cdot (V_b(Tax > 3500)) / (V_a(S_{sum>35000}) \cdot (V_b(Tax > 3500))$$

S_{sum}>35000_yes represent the value of attribute sum is sum>35000and the class label is yes. q_1 and q_2 are computed by using Scalar Product Protocol. This process complies with Scalar Product Protocol strictly and neither part discloses her private data to others. Repeat the process until it can no longer be divided, then the decision tree generated.

(3)Tree pruning: this is process unlike an ordinary decision tree pruning because of the confidentiality. Similarly, the process of the pruning use secure computing. Generation decision tree as shown in Figure 1:

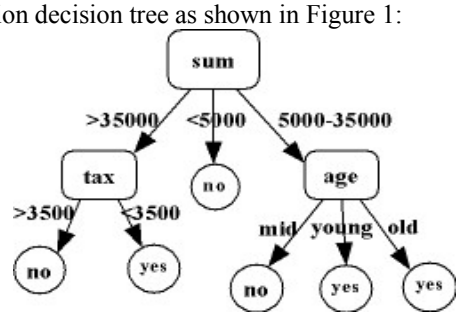

Figure 1. Decision tree

B. Analysis

Through the above analysis, the studies show that when the main features of the transaction meet the following conditions, it can be regarded as suspicious transactions.

(1)The total transaction value (sum) is greater than 35,000 but the tax is less than 3500;

(2)The total transaction value (sum) is between 5000 and 35000 but the age is young or old.

The results in line with the characteristics of suspicious transactions experts summed up. If there is a big gap between the total transaction value and tax, then it will be deemed to be suspicious transactions. If traders are too young or too old, accompanied by large and frequent trading, then it will be deemed to be suspicious transactions.

V. CONCLUSION

A new method of identifying suspicious financial transactions is put forward in this paper. This method combines decision tree algorithm with privacy-preserving strategy. On the premise of protecting the privacy of the merchandisers and the supervision departments, it improves the efficiency of recognizing money laundering transaction. The trades which identified by experiments is basically consistent with the records from supervision departments, thus this method is a effective way to identify money laundering behavior. Other decision tree building algorithms will be study, and see whether this method can be applied to the other areas.

REFERENCES

[1] Vaidya J, Clifton C, Privacy preserving association rule mining in vertically partitioned data, In :t he 8t h ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2002,639～644.

[2] Han Jiawei, Kamber M, Data mining and Technology,2001,8.

[3] Zdanowicz John S, Detecting Money Laundering and Terrorist Financing Via Data Mining [J], Communications of the ACM, 2004,(5):53-55.

[4] Bolton R J, Hand D J. Statistical Fraud Detection[J]. Statistical Science, 2002, (3):235-254.

[5] zheng yan, Application of data mining in the financial field [J], Computer engineering and Application, 2004, (18):208-221.

[6] Yang sheng-gang, Wang peng, He jiahui, Exploring decision trees as a tool to investigate money laundering [J], Journal of Hunan University (Social Sciences), 2006,20(1):65-71.

[7] Du Wenliang, Zhan Zhijun, Building decision tree classifier on private data, In : Procedings of t he IEEE ICDM Workshop on Privacy ,Security and Data Mining ,2002.

[8] Lindell Y, Pinkas B, Privacy preserving data mining, In : Advances in Cryptology-Crypto, 2000,36-54.

[9] Yang Zhangqiang, Zhong Zhong, Wright R N, Privacy-preserving Classification of Customer Data Without Loss of Accuracy In: Proceedings of the 5th SIAM International Conference on Data Mining, 2005, 21-23.

[10] DU W, ATALLAH M, Privacy-preserving cooperative scientific computations[M], 14th IEEE Computer Security Foundations Workshop, 2001,273 - 282.