Applying Data Mining in Money Laundering Detection for the Vietnamese Banking Industry

Dang Khoa Cao and Phuc Do

University of Information Technology, Viet Nam National University-HCMC

Abstract. The applying of data mining techniques in banking is growing significantly. The volume of transaction data in banking is huge and contains a lot of useful information. Detecting money laundering is one of the most valuable information which we can discover from transaction data. This paper will propose the approaches on money laundering detection techniques by using clustering techniques (a technique of data mining) on money transferring data of banking system. Besides, we present an implemented system for detecting money laundering in Viet Nam's banking industry by using CLOPE algorithm.

Keywords: data mining, money laundering, money transferring data.

1 Introduction

In 2006, The World bank warned that Vietnam is becoming an easy target for money laundering activities because of the weakness in the inspection, supervision, auditing and customer relationship management systems. The level of using cash and several kinds of payment method make transactions become out of control [11]. Meanwhile banking plays an important role in cleaning dirty money in the money laundering process. Thus, the requirement for a mechanism to detect money laundering is growing with great importance for all over the world (especially in Vietnam).

This paper will research on anti - money laundering models, we combine the CPU process time of data mining techniques and the human's analyst ability to create an effective method to detect money laundering. Because of the large volume of data, banking creates a convenient environment to hide the original of money laundering. This fact makes money laundering techniques become more sophisticated and hard to trace the crime of money laundering. So the solution for detecting money laundering must be balanced between accuracy and the process time. Finding a suitable algorithm for data mining in banking is the most important step for the overall solution of the problem. This paper will propose a solution for money laundering detection system in Vietnam.

2 Preliminaries

2.1 Data Approaching

Following the research of Linard Moll about the approaches on money laundering models [3], depending entirely on each sort of bank's data, we have 4 approaches as follows:

J.-S. Pan, S.-M. Chen, N.T. Nguyen (Eds.): ACIIDS 2012, Part II, LNAI 7197, pp. 207–216, 2012.

1. **Supervised approaches on labeled data:** This method requires an existing training set (labeled data). The author used one of these techniques: data mining, expert systems, statistic models etc… on the training set [3]. This approach is suitable for experienced banks in detecting money laundering. Therefore data will be in well-form before being mined.
2. **Hybrid approaches with labeled data:** These approaches are similar to "Supervised approaches on labeled data" regarding the aspect of data. The most important difference is that this approach combines multi-techniques to increase the accuracy of overall process. It requires the experience of bank in money laundering detection and the budget invests to the anti money laundering system.
3. **Semi-supervised approached with legal (non-fraud) data:** This approach is different from two approaches before. This approach just requires the valid training set (this means that all members in the training set must be valid data). New transactions will be considered invalid if their behavior doesn't match with the training set. So this approach requires banks which already had a mechanism to distinguish normal and abnormal transactions. With this approach banks can distinguish between valid and invalid data so they have only set up detecting methods on invalid data and let valid data work in normal capacity.
4. **Unsupervised approaches:** This approach is suitable for banks without having any methods for reviewing data (it means that these banks don't have any training sets).

Following the survey at bank X, the approaches of the anti laundering models that require training sets can't be applied for the Vietnamese banking industry (because Vietnam have lacked experience in anti money laundering). Thus, "unsupervised approaches" is the most suitable approach for detecting money laundering of Vietnamese banking industry.

2.2 Money Laundering Process

2.2.1 General of Money Laundering

By the decree 74/2005/NĐ-CP about preventing and anti money laundering of Vietnam's authority, money laundering is defined as: "*Money laundering is the behavior of persons or organizations that tries to validate or cleanse dirty money (money was earned by criminal activities).*

Basically, the money laundering process consists of 3 steps: placement, layering and integration.

Placement: Distributing money from illegal activities in the banks that have weak management mechanisms. Usually, the money will be divided into several parts that fall below the bank alert level. [11]

Layering: In this step, money will be transferred to several banks or several accounts. The real purpose of this step is to hide the illicit money origin by creating a transaction sequence. By creating a sequence of transactions, the origins' money will be hard to detect. [11]

Integration: In this step, money will be invested in legal business, and its profit will be used for criminal activities again to begin new cycle [11].

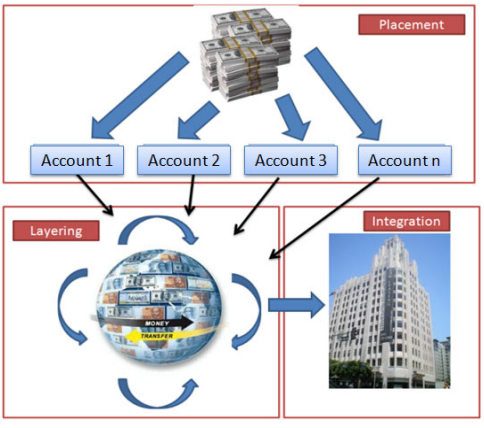

Fig. 1. General money laundering process in real world

2.2.2 Money Laundering Process

The sophistication of money laundering activities depends on the transaction's sequence that is used to hide the relationship between dirty money and its origin. In addition, because of the sophistication of money laundering to be mentioned above, layering becomes the hardest step for applying money laundering mechanism. But by the perspective of the data, this step has the biggest chance for applying automatic money laundering mechanisms.

Following the survey of a bank in Vietnam, the real money laundering processes can be divided into smaller processes that can be considered as sub-processes of the money laundering process. These sub-processes basically have characteristics that make involving transactions become different from normal transactions. So it makes the accounts performing these transactions become suspicious account.

The advantage of this solution is that we can limit the suspicious transactions by checking the suspected account only. This convenience will decrease significantly the operation time and make anti-laundering become realistic. Basically, the processes of money laundering are shown in figure 2:

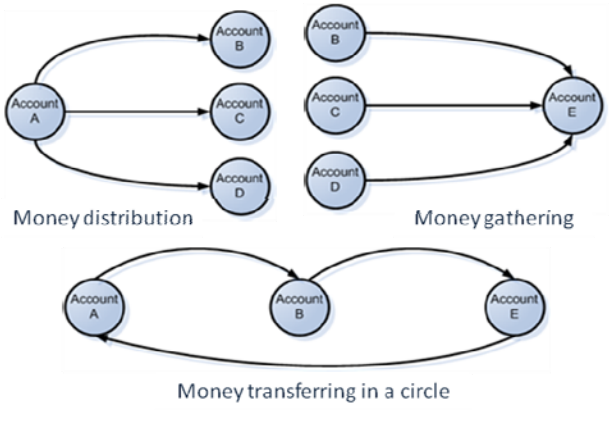

Fig. 2. The money laundering processes

Our paper focuses on the main problem of money laundering. We will identify the accounts that have potential for money laundering instead of finding all money laundering transactions directly.

2.3 The Learning Data Model

Bank transferring transaction consists of five main attributes:

"Sender account ID, receiver account ID, amount of money, money type, transaction date"

All attributes above can express the detail of transferring money data's information. When we take an instance of data, we can draw a graph to present relationship between accounts.

To determine accounts having the potential characteristics of money laundering or not, we must find out the behaviors of these accounts in one period of time. Thus, we will propose a new data set to be created by grouping accounts from transferring information. The new data set can express the special behavior of an account in a determined period of time. The new data has following attributes:

Table 1. Attributes of new data set

Attribute	Explaination
Account	Transaction account
Sum_sending	Sum of sending
Sum_receiving	Sum of receiving
Number_sending	Number of sending
Number_receiving	Number of receiving
Receiving_relationship	Number of account that send money to this account
Sending_relationship	Number of account that receive money from this account
| R_S |	| Sum_receiving – Sum_sending |

Money transferring data is converted to transaction data specifying a particular behavior of accounts. Each transaction data expresses specific behavior for a particular account. Multi-dimensional and large databases are two characteristics of transaction. In the next section we will introduce CLOPE algorithm and prove that CLOPE is the most comprehensive algorithm for processing the transaction data.

2.4 Introducing CLOPE Algorithm

CLOPE algorithm was invented by Yiling Yang, Xudong Guan and Jinyuan You [8]. This algorithm is used for clustering technique. It can works with the nominal variables (string variable) only. The main idea of the algorithm is based on the realistic situation from real life data in string data type. Moreover, the applying of data mining in real life always faces with a multi-dimensional (containing diversified information) data and a large database.

The authors of CLOPE algorithm proved that the distance approach for nominal variables isn't suitable, especially for finance data. Alternatively, CLOPE also defines a global criterion function that is used to optimize clustering process. Section below will introduce briefly about the criterion function and how it affects to the clustering processing. Typically, each clustering algorithm will define a criterion function and use it as a function to optimize the clustering process based on its calculation.

The criterion function will be separate into global criterion function and local criterion function. Global criterion function will determine the optimization for overall clustering instead of each cluster. Otherwise, the local criterion function focuses on optimizing each cluster. Because of the purpose of local criterion function, so it is harder to calculate than the global criterion especially with multi-dimensional data. The use of the global criterion function of CLOPE algorithm proves that it's suitable for multi-dimension data and large database. CLOPE algorithm will model each cluster to histogram that will be displayed in figure 4:

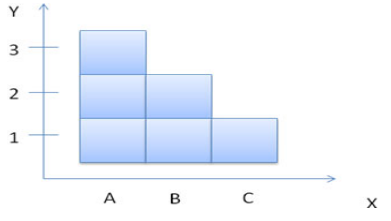

Fig. 3. Modeling clusters to histogram by CLOPE algorithm

- Axis (X): cluster's members ∈ D(C)
- Y-axis (Y): the frequent appearance of cluster members ∈ D(C) C is a cluster.
- Assumption:

$$S(C) = \sum_{i \in D(C)} Occ(i, C) = \sum_{t_i \in C} |t_i| \tag{1}$$

$$W(C) = |D(C)| \tag{2}$$

- S(C): Number of member in cluster C
- W(C): Number of member on x-axis.
- Occ(i,C) :The appearance frequency of member i in cluster C
- H(C) (the high): = S(C)/W(C)

CLOPE's criterion function is as follows:

$$Profit(C) = \frac{\sum_{i=1}^{k} \frac{S(C_i)}{W(C_i)^r} \times |C_i|}{\sum_{i=1}^{k} |C_i|} \tag{3}$$

r: **Repulsion** is a real number (r > 0). In case arguments $S(C_i)$, $|C_i|$, $W(C_i)$ are given. If *r* increases the more similar data will be grouped or the clustering will bring more profit. Otherwise, if *r* decreases, more similar members will be separated into different groups. The highest purpose of the algorithm is to optimize the clustering to make the highest profit with a given *r*. Because the purpose doesn't focus on managing number of clusters (a cluster will be created if its existence and it makes the overall profit of clustering increasingly). So more cluster was created, it doesn't mean the profit is increased. [8]

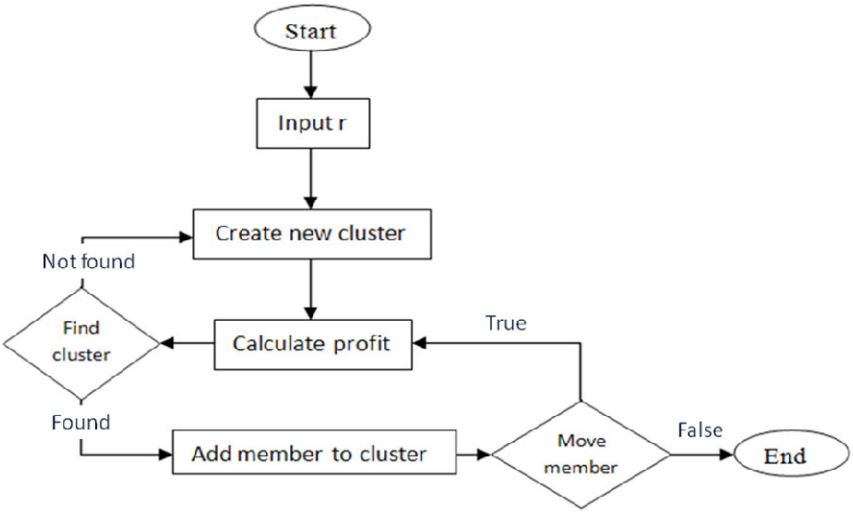

Fig. 4. The process flow of CLOPE algorithm

3 System Implementation

This section will explain the architecture and work flow of the money laundering detection system that utilizes clustering technique. The work flow consists of four stages: data converting, data fragmentation, clustering & analyst, checking the relationship of suspicious accounts:

1. **Data converting:** In this stage, transferring data will be converted to transaction data by separating each pair of accounts (sending and receiving) then make statistics for each account. The final purpose is to create a new data set (transaction data) expressing the behaviors of each account. The new data set will contain *m* records that $n \leq m \leq 2n$ (n is the record of the old data set).
2. **Data fragmentation:** In this stage, the system will convert all number data types into nominal data type by separating number data type to several parts (fragments) that contain specific meaning. Eg: Converted data of attributes "Sum_sending" ∈ [1.000.000.000, 10.000.000.000] into a text like "[1.000.000.000 =>

10.000.000.000]" and assign a meaning for this fragment such as "a big money transaction" (this conversion will make data to become meaningful for data mining before applying CLOPE algorithm).[6]

3. **Clustering & analysis:** In this stage, data will be grouped into clusters by using the CLOPE algorithm. It depends on the optimization of clustering (r argument in criterion function).

To find out which clusters have potential behaviors for money laundering, it must have a set of criteria for each case of money laundering. Based on the survey at bank X in Vietnam, we will examine an example set of criterias to validate cluster .The set of criterias is listed below:

Case 1: Suspicious transferring in a circular pattern: The attribute R_S (|Sum_receiving – Sum_sending|) will be focused on. Because each account in this case will send and receive the same amount of money so R_S attribute of these accounts will near to 0. When focusing on the behavior of account, we can find that all involve accounts belonging to this case will do the same actions: sending and receiving same amount of money. So when the system performs clustering, these accounts will belong to the same cluster as a result. The question is why R_S isn't equal to 0. R_S = 0 is not absolutely right in the real world since the intelligent crime will perform exchange money during the process and the exchange rate of money will affect to the value of original money (just a minute amount). Hence R_S will hardly equal to 0.

Case 2: Suspicious for money distribution: To determine the clusters for this case, we use the following attributes:

- Num_Sending: number of sending times of these account are higher than almost others accounts.
- Sending_Relationship: number of accounts that receive money from each account is higher than almost others account.
- Sum_Sending: Sum of sending is large or very large (higher than alert level of bank).But the sending of money each time usually small or normal when the system query each transferring transaction record of this account.

The complication of this case depends on how large the amount of money in money laundering activities and the state of current finance. Using data mining, the system doesn't care how large the amount of the money laundering activities is, instead of the similar behavior of account that performed these transactions.

Case 3: Suspicion for gathering money from several sources

- Num_Receiving: the frequency of receiving money is high.
- Receiving_Relationship: Number of accounts sent money to this account is higher than almost others account.
- Sum_Receiving: Sum of sending money is very large (larger than alert level of bank). But the receiving of each time is small or normal.

Similarity with the distribution of money case but in this case the current account plays a role to collect money from several accounts (maybe this account is original account in the chain of sequence).

We distinguished 3 basic cases of money laundering activities. Accounts that belong to transferring data in a circle case will have the same behavior so they will be easier to group. But for two remain cases the accounts will be separated into two groups with opposite behavior.

Thus, the result of clustering for money distribution or gathering money will create 2 clusters that have opposite characteristics. The first one contains accounts that send money to several accounts. The second one contains accounts receive money from several accounts. Figure 6.

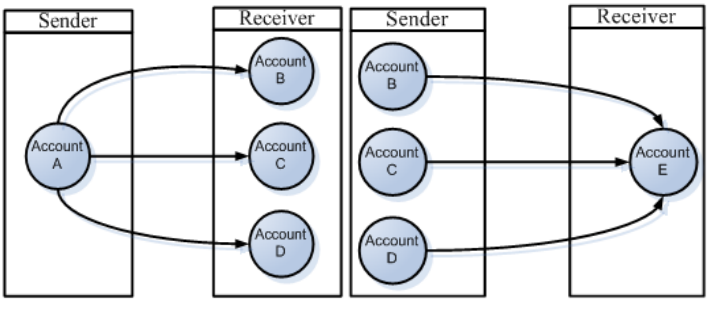

Fig. 5. Distributed & Gathered money difference

4. **Verify the relationships of the suspicious accounts:** After determining the suspicious cluster, the system must check the relationship of each account in each suspicious cluster to indicate which account participating in money laundering activities and in which case the activities belong to. We propose a solution that combines a data management system with n-tree data structure to increase performance for finding the relationship.

4 Experimental Results

Our tested data set contains 8020 records of transferring transaction of bank X. After converting to transaction data, the data set has 12.350 records. This test was performed on software that was implemented by the author based on WEKA open source (learning machine open source of Waikato University). We were simulated 25 records to present each case of money laundering; Structure of money laundering activities was shown in Figure 7:

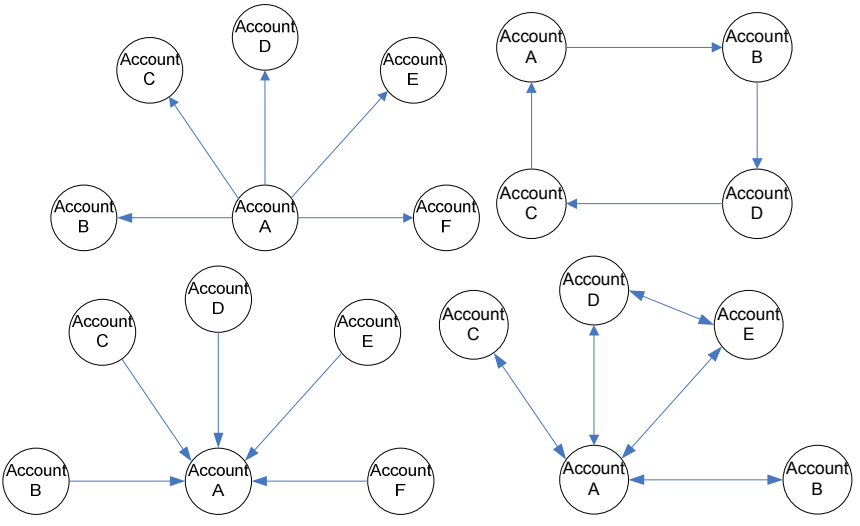

Fig. 6. Simulated money laundering structure cases

Table 2. Experimental result

Case	CLOPE		K-means	
Transferring on a circle	Cluster 1	12 member/12	Cluster 0	4 member /246
	Cluster 19	1 member /1	Cluster 9	9 member /143
	Sum : 13 member / 13		**Sum: 13 member / 389**	
Distributed money	Cluster 6	5 member /1804	Cluster 0	1 member /246
	Cluster 21	1 member /1	Cluster 1	1 member /3178
			Cluster 2	1 member /3039
			Cluster 3	2 member 1091
			Cluster 4	1 member /966
	Sum: 6 member / 1805		**Sum: 6 member / 8520**	
Gathered money	Cluster 2	5 member /1296	Cluster 0	1 member /246
	Cluster 18	1 member /1	Cluster 1	3 member /3178
			Cluster 2	1 member /3039
			Cluster 16	1 member /159
	Sum: 6 member / 1297		**Sum: 6 member / 6622**	

5 Conclusions

According the survey and the requirement specification of bank X (the provider of our data source), we recognized that finding solution for money laundering detection is growing more significantly nowadays for all over the world and especially in Vietnam.

We proposed a new training data set that is converted from banking data to suitable for applying CLOPE algorithms in money laundering detection. The experimented

result proved that CLOPE is a suitable algorithm for money laundering detection. But the system can't run stand alone absolutely, it must base on the ability of analysts in analyzing data and providing a set of rules (criteria set) to validate clusters after clustering. We hope our system can support the money laundering detection and in the near future it can discover automatically the money laundering problems when system receives new transactions.

References

1. Vimal, A., Valluri, S.R., Karlapalem, K.: An Experiment with Distance Measures for Clustering (2008)
2. Rosen, K.H.: Curriculum: Applying discrete mathematics into computer. Translator: Phạm Văn Thiều, Đặng Hữu Thịnh (2002)
3. Linard Moll from Switzerland, Master Thesis: Anti Money Laundering under real world conditions - Findingrelevantpatterns,Universitys of Zurich, 4-15 (2009)
4. Vu Lan, P.: Research and implement some algorithm of data mining. Ha Noi University of Science and Technnology (2006)
5. Le-Khac, N.-A., Markos, S., Kechadi, M.-T.: A Heuristics Approach for Fast Detecting Suspicious Money Laundering Cases in an Investment Bank (2009)
6. Do, P.: Data mining curriculum. National University of HCM City (2008)
7. Wiwattanacharoenchai, S., Srivihok, A.: Data Mining of Electronic Banking in Thailand: Usage Behavior Analysis by Using K-Means Algorithm
8. Yang, Y., Guan, X., You, J.: CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data. Shanghai Jiao Tong University (2002)
9. Webpage : Wikipedia – searching about transaction database, http://en.wikipedia.org/wiki/Database_transaction
10. Webpage :Researching for money laundering forms, http://www.vnecon.vn/showthread.php/3764-R%E1%BB%ADa-ti%E1%BB%81n-l%C3%A0-g%C3%AC-C%C3%A1c-h%C3%ACnh-th%E1%BB%A9c-r%E1%BB%ADa-ti%E1%BB%81n-hi%E1%BB%87n-nay
11. Webpage :anti money laundering in Vietnam (2009), http://www.hids.hochiminhcity.gov.vn/Noisan/32009/mach3.htm