Towards a New Data Mining-Based Approach for Anti-Money Laundering in an International Investment Bank

Nhien-An Le-Khac, Sammer Markos, and Mohand-Tahar Kechadi

School of Computer Science & Informatics, University College Dublin
Belfield, Dublin 4, Ireland
{an.lekhac,sammer.markos,tahar.kechadi}@ucd.ie

Abstract. Today, money laundering (ML) poses a serious threat not only to financial institutions but also to the nation. This criminal activity is becoming more and more sophisticated and seems to have moved from the cliché of drug trafficking to financing terrorism and surely not forgetting personal gain. Most international financial institutions have been implementing anti-money laundering solutions (AML) to fight investment fraud. However, traditional investigative techniques consume numerous man-hours. Recently, data mining approaches have been developed and are considered as well-suited techniques for detecting ML activities. Within the scope of a collaboration project for the purpose of developing a new solution for the AML Units in an international investment bank based in Ireland, we propose a new data mining-based approach for AML. In this paper, we present this approach and some preliminary results associated with this method when applied to transaction datasets.

Keywords: data mining, anti money laundering, clustering, neural network.

1 Introduction

Money laundering (ML) is a process of disguising the illicit origin of "dirty" money and makes them appear legitimate. It has been defined by Genzman as an activity that "knowingly engage in a financial transaction with the proceeds of some unlawful activity with the intent of promoting or carrying on that unlawful activity or to conceal or disguise the nature location, source, ownership, or control of these proceeds" [17]. Through money laundering, criminals try to convert monetary proceeds derived from illicit activities into “clean” funds using a legal medium such as large investment or pension funds hosted in retail or investment banks. This type of criminal activity is getting more and more sophisticated and seems to have moved from the cliché of drug trafficking to financing terrorism and surely not forgetting personal gain. Today, ML is the third largest “Business” in the world following Currency Exchange and the Auto Industry. According to the United Nations Office on Drug and Crime, worldwide value of laundered money in one year ranges from $500 billion to $1 trillion [1] and from this approximately $400-450 Billion is associated with drug trafficking. These figures are at times modest and are partially fabricated using statistical models, as no one exactly knows the true value of money laundering, one can only forecast

S. Goel (Ed.): ICDF2C 2009, LNICST 31, pp. 77–84, 2010.

according to the fraud that has already been exposed. Nowadays, it poses a serious threat not only to financial institutions but also to the nation. Some risks faced by financial institutions can be listed as reputation risk, operational risk, concentration risk and legal risk. At the society level, ML could provide the fuel for drug dealers, terrorists, arms dealers and other criminals to operate and expand their criminal enterprises. Hence, the governments, financial regulators require financial institutions to implement processes and procedures to prevent/detect money laundering as well as the financing of terrorism and other illicit activities that money launderers are involved in. Therefore, anti-money laundering (AML) is of critical significance to national financial stability and international security. Traditional approaches to AML followed a labor-intensive manual approach. These approaches can be classified into identification of money laundering incidences, detection, avoidance and surveillance of money laundering activities [14]. Indeed, given that the volume of banking data and transactions have increased in number of ways, such approaches need to be supported by automated tools for detecting money laundering's pattern. Meanwhile, AML software tools in the market are normally rule-based that make the decisions using some sets of predefined rules and thresholds.

Besides, data mining techniques (DM) [3] have been proven to be well suited for identifying trends and patterns in large datasets. Therefore, DM techniques are expected to be applied successfully in AML. Nevertheless, there is still little research concerning this bias especially a DM framework/solution for supporting AML experts in their daily tasks. Recently, there are some AML approaches based on DM that have been proposed and discussed in literature. Most of these approaches try to recognize ML patterns by different techniques such as support vector machine [10], correlation analysis [16], histogram analysis [16]… They aim to provide techniques for detecting a variety of ML by exploring a massive dimensionality of datasets including customers x accounts x products x geography x time. However, these approaches are more or less appropriate for the cash world and not scaled well for investment activities due to the lack of good methods in choosing parameters and they still have performance issues. Hence, in this paper, we present a new approach basing on a combination of clustering and classification techniques for analysing ML patterns in an international investment bank. Customer behaviour in investment activities is complicated because it is influenced by many factors. We also show that by choosing suitable dimensions, simple DM techniques can be applied efficiently together to detect suspicious ML cases in investment activities.

The rest of this paper is organised as follows: Section 2 presents recent works on this subject. Section 3 deals with our framework for detecting money-laundering activities including two main steps: customer identification and transaction monitoring. We analyse the second step in Section 4. Preliminary results of our approach are presented and discussed in section 5. Finally, we conclude in section 6.

2 Related Works

[16] applied a discretisation process on their datasets to build clusters. They map their feature space "customer x time x transaction" to $n+2$ dimensional Euclide space: n

customer dimensions, *1* time dimension and *1* transaction dimension. They firstly discretise the whole timeline into difference time instances. Hence, each transaction is viewed as a node in one-dimensional timeline space. They project all transactions of customers to the timeline axis by accumulating transactions and transaction frequency to form a histogram. They create clusters based on segments in the histogram. A local and a global correlation analysing are then applied to detect suspicious patters. This approach improves firstly the complexity by reducing the clustering problem to a segmentation problem [4]. Furthermore, it is more or less appropriate for analysing individual behaviours or group behaviours by their transactions to detect suspicious behaviours related to "abnormal" hills in their histogram. However, as we have to analyse many customers with many transactions with a variety of amounts for a long period, it is difficult to detect suspicious cases, as there are very few or no "peak hills" in the histogram. Firstly, another global analysis is needed and we can then apply this method for further analysis in this case.

Another approach for AML is using support vector machine (SVM) [7]. In [10], authors propose an extension of SVM to detect unusual customer behaviour. They present a combination of an improved RBF kernel [8] with the definition of distinct distant [15] and supervised/unsupervised SVM algorithms (C-SVM, one-class SVM). One-class SVM [7] is an unsupervised learning approach used to detect outliers based on unlabeled training datasets which is highly suitable for ML training sets. The advantage of this approach is that it can deal with heterogeneous datasets. However, there is a performance issue due to the lack of parameter selection.

3 AML Framework

A framework for detecting ML activities is normally consisted of four layers [11][13] corresponding to four levels of analysis: transaction, account, institution and multi-institution. The most basic level is transactions. At this level, transaction records are extracted for an investigation. However, they have a few analytical contexts because they do not provide links to accounts or other data. In the second level, multiple transactions are associated with specific accounts. Aggregation of transactions with individual accounts gives a general view of these accounts about their financial activity. This view shows the degree of association between various accounts based on frequencies of their transactions. At the institution level, the same customer (corporate or individual) may have multiple accounts. A consolidation of these accounts may show that an institution maybe in ML suspicious and may involve multiple accounts relating to different individuals. The last level investigates the ML involving multiple corporations, organizations and customers. The first three levels: transaction, account and institution are the most important where the last one depends more or less on the organisations and their policy.

3.1 Customer Identification

At the institution level, a customer-identification task is needed to determine whether a specific customer has multiple accounts and/or invests in different funds. Fundamentally, a customer is identified by querying customer databases using query tools

provided by DBMS. However, in the case where a specific customer is stored in separate databases that are managed independently, this will require a very large processing time due to the search operations initiated over all the databases. The users need firstly to login into the different databases, run the same query repeatedly, get the results separately, and displayed them independently. Furthermore, in large financial institutions, these databases are heterogeneous and have very complex architectures. This type of approach allows great flexibility, however it usually has poor performance. In addition, data quality is also another factor that makes this naïve approach unfeasible. In our previous work [18], we proposed an efficient approach to identify customers in ***BEP bank***[1]'s datasets. Therefore, in this paper, we present our approach to analyse customer transactions.

3.2 Transaction Analysis

Transaction analysis is an important task of all AML systems. As mentioned above, transactions and accounts cannot be separately investigated; they should be aggregated to give a general view of customers' behavior.

Normally, this analysis is based on two important characteristics: frequency of transactions and the value of each transaction. Current solutions apply these two characteristics in a set of rules to detect suspicious cases.

Generally, most of the vendor software approaches found in the market is based on a decision tree using frequency and value of transactions as a marker, the thresholds for these markers based on averages and the standard deviation. This approach only uses one way comparison i.e. customer X's behaviour against customer X's previous "normal" behaviour. This approach is reasonably adequate for the cash world (accounts). However they are not efficient for the investment market because there are many factors that influence the frequency of trades in investment activities such as political environment, market climate, fund prices, currency exchange rate, etc.

As shown in the Table 1, for instance, the frequency of transactions (subscription and redemption) from specific accounts weekly and monthly of some funds in the ***BEP bank***'s datasets. The frequency of transactions is firstly high (compared to cash account activity: less than 10 transactions weekly, for instance) and secondly it varies from fund to fund.

Briefly, an efficiently solution to investigate ML in investment banking is to determine relevant parameters to decrease the number of dimensions (attributes) and to improve performance.

Table 1. Transaction frequency of some investment funds in BEP bank (weekly and monthly)

Fund	Subscription/Redemption Weekly (Min-Max)	Subscription/Redemption Monthly (Min-Max)
A	0 - 98 / 0 - 70	0 - 287 / 0 - 287
B	0 - 43 / 0 - 40	0 - 95 / 0 - 74
C	0 - 29 / 0 - 38	0 - 105 / 0 - 38

[1] Real name of the bank can not be disclosed because of confidential agreement of the project.

4 A New Approach for Transaction Analysis

In this section, we present our new solution for analysing investment transactions. We firstly define parameters used in our model and then present DM techniques that will be applied to determine the relevant threshold and detect suspicious cases.

X	T	Y	α	β	θ	Δ_1	Δ_2
Customer ID	Time	Fund	Subscription	Redemption	Value of the investors shares	Proportion red V sub	Proportion Reds v bal
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
001	**Week 30**	**A**	**100**	**80**	**200**	**0.8**	**0.9**
002	Week 30	A	52	20	300	0.3846154	0.06666667
001	Week 33	A	100	90	500	0.9	0.18
003	Week 37	A	100	80	600	0.8	0.13333333
004	Week 30	A	500	400	900	0.8	0.44444444
005	Week 500	A	700	300	1500	0.4285714	0.2
…	…	…	…	…	…	…	…
∞	Week 884	A	∞	∞	∞	∞	∞

Fig. 1. An example of Δ_1 and Δ_2 of the investors in the fund A by weekly

4.1 Parameter Definition

The value of transactions (subscription or redeemption) of each investor in an investment fund is aggregated by time: daily, weekly, monthly, quarterly, 6-monthly and yearly. For instance (Figure 1), in week 30, the investor 001 had a total subscription of \$100 and redeemed a total of \$80. Let X = Investor, Y = Fund, α_i = Subscription Value where $\alpha_i \in [0\ldots\infty]$, β_j = Redemption Value where $\beta_j \in (0\ldots\text{-}\infty]$, Θ_h = Value of the investors shares where $\Theta_h \in [0\ldots\infty]$, T_k = Time where k = (Days, Weeks, 1 Month, quarterly, 6 months, 12 months). We then define two parameters Δ_1, the proportion between the redemption value and the subscription value conditional on time (daily, weekly, monthly etc) and Δ_2, the proportion between a specific redemption value and the total value of the investors' shares conditional on time as below:

$$\Delta_1 = \begin{cases} \left|\frac{\alpha_i}{\beta_j}\right|_{\tau_k} & \text{Where } \alpha_i \leq \beta_j \\ \left|\frac{\beta_j}{\alpha_i}\right|_{\tau_k} & \text{Otherwise} \end{cases} \qquad \Delta_2 = \left|\frac{\beta_i}{\theta_h}\right|_{\tau_k}$$

We use these two parameters at two levels: investor and investment fund (set of investors).

4.2 Analysing Process

We firstly apply a clustering technique (K-Means family, for instance) for each Δ_1 and Δ_2 at both two levels: fund and investor. These outputs will be then used to feed in to a

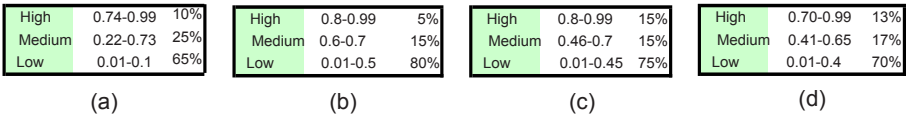

Fig. 2. An example of clustering results on Δ_1 and Δ_2 at fund and investor level

neural network (back propagation based) for training on suspicious and non-suspicious cases. These results are then stored in a knowledge-base that assists the AML experts to make a decision.

For instance, as shown in Figure 2a, there are three groups of Δ_1 (weekly) at fund ***A*** level: 10%, 25% and 65%. This means 10% of the population has $0.74 \leq \Delta_1 \leq 0.99$, 25% of population has $0.22 \leq \Delta_1 \leq 0.73$ and 65% of population has $0.01 \leq \Delta_1 \leq 0.1$. The interpretation of these results is that most investors (65%) in fund ***A*** have a small proportion of redemption versus subscription values (weekly) i.e. they are not suspicious. 10% of them have a high proportion of redemptions vs. subscriptions hence suspicious. Then, the analyser can refine this group to determine the extremely high proportion group e.g. Δ_1 is from 0.9 to 0.99 and suppose it is 1%, this group is a high suspicious. By analysing Δ_1 at the fund level, we know the behaviours of all investors on subscription and redemption value in this fund. Figure 2b shows the same results as Figure 2a but at the investor level i.e. we analyse each investor's behaviour. Figure 2c and 2d relate to Δ_2, the proportion between the redemption values against total values, at fund and investor level. This parameter helps us to investigate the cases where the redemption values is around or greater than the total values (negative balance). These four parameters (Δ_1, Δ_2 at fund and investor level) of each record (i.e. each line in Figure 1) from the highest suspicious level and lowest one are used to train the neural network as suspicious and non-suspicious class respectively.

In order to investigate one case, its transaction is firstly put in a suitable time frame (weekly, for instance). Its relevant Δ_1, Δ_2 are determined and used to investigate by comparing it with knowledge-base content. At the first level, if it is always in a highly suspicious group (both fund and investor level), we can conclude then this is a suspicious case. If not, a neural network related to this case is used to determine its suspicious degree.

5 Performance Evaluation

We evaluate our approach with transactions from 16 funds of ***BEP bank*** with two millions of transaction records. The testing platform is Windows XP with 2Gb RAM, 2.4GhZ Intel Dual Core. In each fund, we take about 5-10% of population as a training set and the reminder as testing set.

Our preliminary results in fund A, for instance, the suspicious cases detected was approximate 0.5%. These cases were then investigated further and most of them have inadequate starting values (i.e. the first transaction is a redemption) causing by mapping and loading error. After the refinement process, the real suspicious cases were approximate 5. This is suitable with reports from BEP's bank by manual process that

takes more than a week to detect. It takes only less than 5 minutes using our approach (semi-automatic).

6 Conclusion and Future Work

In this paper, we have presented an approach for analysing transactions in an investment bank to detect ML. In our approach, we determined first of all, the important factors for investigating ML in the investment activities. Next, we proposed an investigating process based on clustering and neural network to detect suspicious cases in the context of ML. From our experimental results obtained on parts of the ***BEP***'s transaction datasets, we can conclude that our approach is promising and it satisfies the needs of the AML unit.

Experimental results on real-platforms of BEP for all transaction datasets are also being produced and these will allow us to test and evaluate the robustness of our approach. We are currently working on improving the learning process to tackle the problem of very large datasets.

References

1. Baker, R.: The biggest loophole in the free-market system. Washington Quarterly 22, 29–46 (1999)
2. Brabazon, A., O'Neill, M.: Biologically inspired algorithms for financial modelling. Springer, Heidelberg (2006)
3. Han, J., Kamber, M.: Data Mining: Concept and Techniques, 2nd edn. Morgan Kaufmann Publishers, San Francisco (2005)
4. Jain, R., Kasturi, R., Schunck, B.G.: Machine Vision. Prentice Hall, Englewood Cliffs (1995)
5. Kingdon, J.: AI Fights Money Laundering. IEEE Transactions on Intelligent Systems, 87–89 (2004)
6. `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
7. Scholkopf, B.: A short tutorial on kernels, Microsoft Research, Rech. Rep.: MSR-TR-200-6t (2000)
8. Scholkopf, B., Plattz, J.: Estimating the support of a high dimensional distribution. Neural Computing 13(7), 1443–1472 (2001)
9. Steinhaus, H.: Sur la division des corp materiels en parties. Bull. Acad. Polon. Sci., C1. III IV, 801–804 (1956)
10. Tang, J., Yin, J.: Developing an intelligent data discriminating system of anti-money laundering based on SVM. In: Proceedings of the Four International Conference on Machine Learning and Cybernetics, Guangzhou, August 2005, pp. 3453–3457 (2005)
11. Tang, J.: A Framework on Developing an Intelligent Discriminating System of Anti Money Laundering. In: International Conference on Financial and Banking, Czech Rep. (2005)
12. Vapnik, V.: The Nature of Satistical Learning Theory. Springer, NewYork (1995)
13. Vidyashankar, G.S., Natarajan, R., Sanyal, S.: Mining your way to combat money laundering. DM Review Special Report (October 2007)

14. Watkins, R.C., et al.: Exploring Data Mining technologies as Tool to Investigate Money Laundering. Journal of Policing Practice and Research: An International Journal 4(2), 163–178 (2003)
15. Wilson, D.R., Martinez, T.R.: Improved Heterogeneous distance functions. Journal of Artificial Intelligence Research 6(1), 1–34 (1997)
16. Zang, Z., Salermo, J.J., Yu, P.S.: Applying Data mining in Investigating Money Laundering Crimes. In: SIGKDD 2003, Washington, DC, USA, August 2003, pp. 747–752 (2003)
17. Genzman, L.: Responding to organized crime: Laws and law enforcement. In: Abadinsky, H. (ed.) Organized crime, p. 342. Wadsworth, Belmont
18. Le-Khac, N.-A., Markos, S., O'Neill, M., Brabazon, A., Kechadi, M.-T.: An Efficient Search Tool for an Anti-Money Laundering Application of an Multi-National Bank's Dataset. In: The 2009 International Conference on Information and Knowledge Engineering (IKE 2009), LA, USA, July 13-16 (to appear, 2009)