International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:12, 2009

A Heuristics Approach for Fast Detecting Suspicious Money Laundering Cases in an Investment Bank

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

***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, we proposed a data mining-based solution for AML. In this paper, we present a heuristics approach to improve the performance for this solution. We also show some preliminary results associated with this method on analysing transaction datasets.

***Keywords*—**data mining, anti money laundering, clustering, heuristics.

I. 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" [1]. 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 the Currency Exchange and the Automobile Industry. According to the United Nations Office on Drug and Crime, the worldwide value of laundered money in one year ranges from $500 billion to $1 trillion [2] 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 nobody exactly knows the true value of money laundering, one can only forecast 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 labour-intensive manual approach. These approaches can be classified into the identification of money laundering incidences, detection, avoidance and surveillance of money laundering activities [3]. 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) [4] 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 the area of 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 [5], correlation analysis [6], histogram analysis [6]… They aim to provide techniques for detecting a variety of ML by exploring a massive dimensionality of datasets including customers x

Dr. Nhien An Le Khac, is with the School of Computer Science & Informatics, University College Dublin, Ireland (corresponding author to provide e-mail: an.lekhac@ ucd.ie).

Ms. Sammer Markos, is with the University College Dublin, Ireland (e-mail: sammer.markos@ ucd.ie).

Prof. M-Tahar Kechadi, is with the University College Dublin, Ireland (e-mail: tahar.kechadi@ ucd.ie).

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. In our previous work [7][8], we proposed a new solution 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 together to detect suspicious ML cases in investment activities. In this solution, the same clustering algorithm is repetitively executed to analysis transactions depending on the characteristic of each transaction datasets. Hence, in this paper, we present a one-step clustering approach basing on some heuristics from AML experts to improve the performance of our previous solution in the term of running time.

The rest of this paper is organised as follows: Section 2 deals with recent works on this subject. Section 3 resumes and analyses our recent solution for detecting money-laundering activities. We present our new approach for improving the performance of our previous solution in Section 4. Preliminary results of this approach are presented and discussed in Section 5. Finally, we conclude in Section 6.

II. Related Works

[6] applied a discretisation process on their datasets to build clusters. 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 [9]. 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) [10]. In [11], authors propose an extension of SVM to detect unusual customer behaviour. They present a combination of an improved RBF kernel [12] with the definition of distinct distant [13] and supervised/unsupervised SVM algorithms. One-class SVM [10] 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.

III. Transaction Analysing

A framework for detecting ML activities is normally consisted of four layers [14][15] corresponding to four levels of analysis: transaction, account, institution and multi-institution. The most basic level is the transactions and the transaction analysis is an important task of all AML systems. As mentioned in [8], transactions and accounts cannot be separately investigated; they should be aggregated to give a general view of customers' behaviour. 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 are based on a decision tree using frequency and value of transactions as a marker, the thresholds for these markers are 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 banking such as political environment, market climate, fund prices, currency exchange rates, etc. Table I [8] is an example of the variety of transaction frequencies that exists among different funds of ***BEP***[1] bank's datasets.

Briefly, an efficient solution to investigate ML in investment banking is to determine relevant parameters to decrease the number of dimensions (attributes) and to improve performance. In our recent work [8], we proposed parameters that were used in detecting suspicious cases and we then applied DM techniques to determine the relevant thresholds. Concretely, we defined two parameters: Δ_1 (delta1), the proportion between the redemption value and the subscription value conditional on time (daily, weekly, monthly etc) and Δ_2 (delta2), the proportion between a specific redemption value and the total value of the investors' shares conditional on time as below:

TABLE I
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.

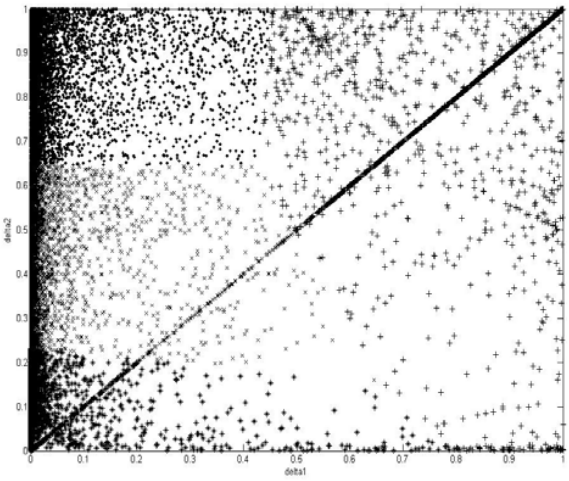
$$\Delta_1 = \begin{cases} \left|\frac{\alpha_i}{\beta_j}\right|_{\tau_k} & \text{Where } \alpha_i \leq \beta_j \\ \left|\frac{\beta_j}{\alpha_i}\right|_{\tau_k} & \text{Otherwise} \end{cases} \qquad \Delta_2 = \left|\frac{\beta_i}{\theta_h}\right|_{\tau_k}$$

where α_i is the subscription value and $\alpha_i \in [0\ldots\infty]$, β_j is the redemption value and $\beta_j \in (0\ldots\text{-}\infty]$, θ_h is the value of the investors shares and $\theta_h \in [0\ldots\infty]$, τ_k is time where k = (Days, Weeks, 1 Month, 3 months, 6 months or 12 months). Note that the value of the transactions (subscription or redemption) of each investor in an investment fund is aggregated by time: daily, weekly, monthly, 3 monthly, 6-monthly and yearly.

In the next step, we apply a clustering technique (centre-based family) for each Δ_1 and Δ_2 at two levels: fund and investor. These outputs will be then fed into a 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 making a decision. The clustering stage includes a repetitiveness of a clustering algorithm on transaction datasets to determine the suspicious group. This is one of the most time-consuming steps in our solution. We need, moreover, interaction with AML experts at each loop of this stage. Consequently, this step affects the overall performance of our solution in the term of running time.

Figure 1, for instance, shows four clusters of fund S[2] datasets (~50000 elements) based on two variables Δ_1 and Δ_2 (by week) after the first running of a centre-based clustering algorithm. Generally, the most suspicious cases should obtain high values in Δ_1 and Δ_2. In this example, Cluster 1 contains not only elements with high values in Δ_1 and Δ_2 but also the elements with low values in Δ_1 and Δ_2. Hence, this clustering algorithm is required to perform several times on this cluster 1 and its subsets to determine the suspicious group. In addition, in this solution, two kinds of investors: individual and corporate are investigated together. However, they are relatively different in their investment behaviour.

For instance, the Figure 2 and Figure 3 shows respectively the distribution of the corporate and individual investors by use of their Δ_1 and Δ_2 (by week) values. Therefore, corporate and individual investors should be separately analysed in order to improve the performance.



+ Cluster 1 * Cluster 2 x Cluster 3 ♦ Cluster 4

Fig. 1 Clustering of fund S datasets based on two variables Δ_1 and Δ_2

IV. Heuristics Approach for Transaction Analysis

As mentioned in the previous section, our recent solution is required to improve the performance in terms of reducing running time and interaction with AML experts. In this section, we present a heuristics approach where a centre-based clustering algorithm performs only one time for analysing investment transactions.

Firstly, transaction datasets are divided into two groups according to two kinds of investors: individual and corporate. In the rest of this paper, we only use the corporate datasets as examples because it is more popular than individual one. Besides, we can still apply the same approach to both kinds of datasets. Secondly, we refine the parameter Δ_1 so that it is the proportion between the redemption value in the time τ_k and the maximum of the subscription values from time τ_{k-l} to τ_k instead of the proportion between the redemption value and the subscription value in the time τ_k as in our previous solutions. Besides, the parameter l in τ_{k-l} is adjustable and is defined by AML experts. It normally varies from 3 to 5. For instance, in the Table II, Δ_1 of the customer A01 at the week 33 is not the proportion between the redemption value and the subscription value in week 33 but now is the maximum of subscription values from week 30 to week 33 (l=3 in this case). We apply this first heuristics because of AML experts' experience: the relevant subscriptions of a redemption activity in suspicious cases are normally not only in the current investigation term (week, month...) but also in its short previous term (two, three weeks or two three month ago).

[2] Again, real name of fund can not be disclosed because of confidential agreement of the project

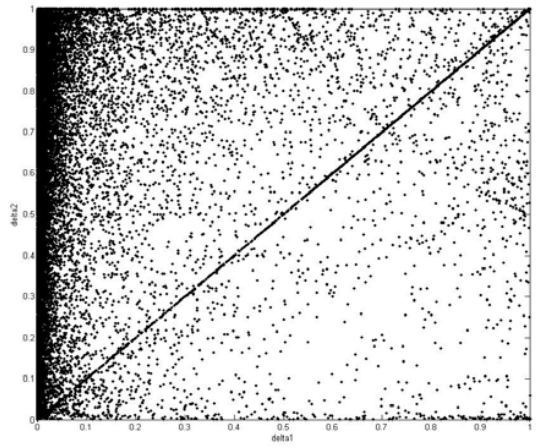


Fig. 2 Distribution of corporate investors (Δ_1 and Δ_2) of the fund S

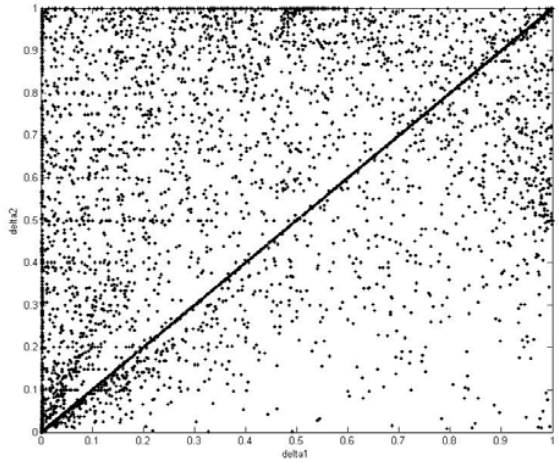


Fig. 3 Distribution of individual investors (Δ_1 and Δ_2) of the fund S

A. Suspicious screening

Let V be the set of transactions datasets aggregated by Δ_1 and Δ_2:

$$V = \bigcup_{i=1}^{n} v_i$$

where each point v_i is represented by a pair of values (Δ_{1i} and Δ_{2i}).

By analysing the distribution of V (Figure 2) as well as referring to the AML expert's experience, suspicious cases often gets high value in both Δ_1 and Δ_2. Concretely, their points v_i gather around the point v_H (Δ_{1H}=1, Δ_{2H}=1). If a customer X has Δ_1=1 and Δ_2=1 in the time τ_k then it can be deducted as "in the time τ_k, X redeemed all of his/her total value of shares and this total value is equal to his/her subscriptions in the short period from τ_{k-l} to τ_k". This case is clearly very highly suspicious. Besides, the number of suspicious cases is very small compared to the total transactions. Consequently, we have a high density of points v_i spreads from position (0,0) to position (0,1) and low density around (1,1). Based on this observation, we can focus our analysis on a subset with low density $V' \subset V$ and the process of determining this subset is called suspicious screening. V' is formally defined as:

$$V' = \{v_i\,(\Delta_{1i},\Delta_{2i}) \in V \mid s \le \Delta_{1i} \le 1 \wedge S \le \Delta_{2i} \le 1,\ s, S \in \mathrm{R} \wedge s \in [0..1] \wedge S \in [0..1]\}$$

For instance, we can only analyse a subset V' with $s = S =$ 0.5 on the datasets of the corporate customer from fund S above. Two parameters s and S are adjustable and defined by AML experts with the support of a simple analysis tool.

B. Clustering process

After determining the subset V', one more analysis step is required to determine the suspicious group. Since there is no prior knowledge of customer behaviour in each investment fund, a clustering technique is applied. We choose the centre-based technique because of its simplicity and efficiency [16]. Besides, the shape of cluster (convex) does not really affect on the final decision. As mention in [16], the performance of centre-based clustering algorithm depends strongly on the initialisation of centres. Therefore, we apply a heuristics in choosing initial centres instead of using random ones: the point v_H (Δ_{1H}=1, Δ_{2H}=1) is chosen as the first initial centre. As analysed above, the suspicious cases always gather around this point. The second initial centre is the furthest point to the first point v_H because value points gathering around this second centre are often clear cases in the subset V'. Briefly, in the case of 3-centres, the set of centre points C can be formally defined as:

$$C = \{c_1, c_2, c_3 \mid c_1 \equiv v_H;\ c_2 \in V' : \forall\, c_i \in V',\ i \neq 2,\ ||c_1 - c_i|| < ||c_1 - c_2||;\ c_3 \in V' : c_3 \neq c_1,\ c_3 \neq c_2\}$$

By applying heuristics in the clustering process as well as the suspicious screening, we only need to perform the clustering algorithm one time to determine the suspicious group. Briefly, these heuristics help to improve the running time of clustering process. We evaluate and analyse this approach in Section 5. Suspicious and non-suspicious groups are then fed into a neural network for training and their results are stored in a knowledge-base. The rest of the ML detecting process was presented in [8] and can be resumed as following: in order to investigate one case, its transactions are firstly placed in a suitable period (weekly, for instance). Its relevant Δ_1, Δ_2 are then calculated and used in the investigation by comparing it with stored knowledge-base content. At the first level, if it is always in a highly suspicious group (at both fund and investor level), we can then conclude that this is a suspicious case. If not, a neural network related to this case is used to determine its suspicious degree.

V. Evaluation and Analysis

We evaluate our approach using transactions from funds administered by ***BEP bank*** with two millions transaction records. The testing platform is Windows XP with 2Gb RAM, 3.4GhZ Intel Dual Core. In each fund, we use approximately 30-40% of the population as a training set and the reminder as a testing set.

We analyse the fund S, the largest fund (~ 1 million transactions), by week (transactions are aggregated by week to determine Δ_1, Δ_2). In this example, we choose $s = S = 0.5$ (cf. IV.A) and the number of elements of V' after screening are approximate 5% of total elements. Figure 4 shows the clustering results of V'. The clustering time on this subset is only 0.718s compared to 3.5s of clustering on all elements.

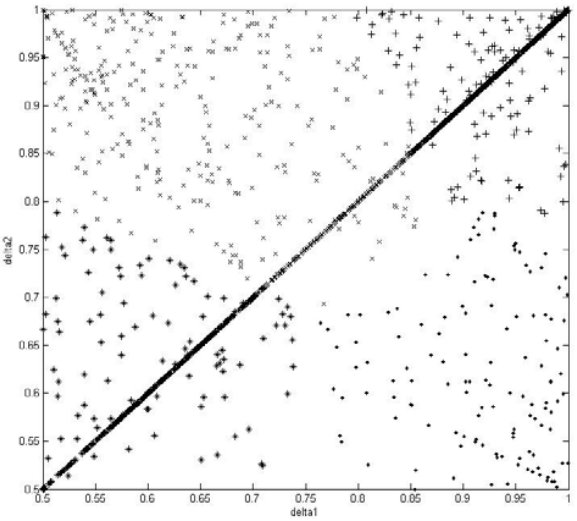


Fig. 4 Clustering of the fund S where $s = S = 0.5$

The suspicious cases are gathered in the Cluster 1 (Figure 4) which was approximately 0.02% of the population and this significantly reduces the number of elements compared to 0.5% in the previous solution [8]. Besides, we need only perform the clustering algorithm one time while the previous one takes from 3 to 4 times. Furthermore, the overall process takes less than one minute (semi-automatic) to detect suspicious cases comparing to 5 minutes in our previous solution. These cases were then investigated further and most of them have exchange transactions i.e. one can redeem his/her entire share from one sub-fund and invest into another sub fund. Both two sub-funds are in the same investment fund. After the refinement process, the real suspicious cases were approximate 5. This is consistent with reports from ***BEP***'s bank by using a manual approach that takes more than a week to detect.

VI. Conclusion and Future Works

In this paper, we have presented a heuristics approach for improving the process of transaction analysis in an investment bank to detect ML. We continue our recent work where an investigating process based on the clustering and a neural network was proposed by using experience from AML experts. In our approach, we refined firstly the important factors for investigating ML in the investment activities. We also divide the datasets into two kinds of investors: corporate and individual as their investment behaviours are different. Next, we proposed two heuristics: suspicious screening and suspicious initial centres to detect suspicious cases in the context of ML. From our experimental results obtained on the greatest fund of ***BEP***'s transaction datasets, we can conclude that our approach is promising and it satisfies the needs of the AML unit. It can also improve significantly the performance from our previous solution in terms of running time. Experimental results for other fund datasets are also being produced.

References

[1] L. Genzman, *Responding to organized crime: Laws and law enforcement*. Organized crime, In H.Abadinsky (Ed.) Belmont, CA: Wadsworth, pp. 342.

[2] R. Baker, *The biggest loophole in the free-market system*. Washington Quarterly, 22, 1999, pp. 29-46.

[3] R. C. Watkins et al, *Exploring Data Mining technologies as Tool to Investigate Money Laundering*. Journal of Policing Practice and Research: An International Journal. Vol. 4, No. 2, January 2003, pp. 163-178.

[4] J. Han and M. Kamber, *Data Mining: Concept and Techniques*. Morgan Kaufmann publishers, 2nd Eds., Nov. 2005.

[5] J. Tang, J. Yin, *Developing an intelligent data discriminating system of anti-money laundering based on SVM*, Proceedings of the Four International Conference on Machine Learning and Cybernetics, Guangzhou, Aug. 2005: pp.3453-3457.

[6] Z. Zang, J.J. Salermo and P. S. Yu, *Applying Data mining in Investigating Money Laundering Crimes*, SIGKDD'03, August 2003, Washington DC, USA. pp: 747-752.

[7] N-A. Le-Khac, S. Markos, M. O'Neill, A. Brabazon and M-T. Kechadi, *An Efficient Search Tool for an Anti-Money Laundering Application of an Multi-National Bank's Dataset*, The 2009 International Conference on Information and Knowledge Engineering, July 13-16, 2009 (IKE 2009), LA, USA.

[8] N-A. Le-Khac, S. Markos and M-T. Kechadi, *Towards a new Data Mining-based approach for Anti Money laundering in an international investment bank*. a NY, USA (to appear).

[9] R. Jain, R. Kasturi and B.G. Schunck, *Machine Vision*, Prentice Hall, 1995.

[10] B. Scholkopf, *A short tutorial on kernels*. Microsoft Research, Tech Rep: MSR-TR-200-6t, 2000.

[11] J. Kingdon, *AI Fights Money Laundering*, IEEE Transactions on Intelligent Systems, 2004, pp. 87-89.

[12] B. Scholkopf and J. Plattz, *Estimating the support of a high dimensional distribution*, Neural Computing, Vol. 13, No. 7, 2001: pp1443-1472.

[13] D.R Wilson and T. R. Martinez, *Improved Heterogeneous distance functions*. Journal of Artificial Intelligence Research, Vol. 6, No. 1, 1997: pp 1-34.

[14] J. Tang, *A Framework on Developing an Intelligent Discriminating System of Anti Money Laundering*, International Conference on Financial and Banking, Czech Rep., 2005

[15] G.S. Vidyashankar, R. Natarajan and S. Sanyal, *Mining your way to combat money laundering*. DM Review Special Report, Oct 2007.

[16] G. Gan, C. Ma and J. Wu, *Data Clustering: Theory, Algorithms and Applications*. Siam publishers 2007, pp 161-182