2009 International Conference on Computer Engineering and Applications
IPCSIT vol.2 (2011) © (2011) IACSIT Press, Singapore

An investigation into Data Mining approaches for Anti Money Laundering

Nhien An Le Khac, Sammer Markos, M. O'Neill, A. Brabazon and M-Tahar Kechadi

University College Dublin, Ireland

{an.lekhac,sammer.markos, m.oneill, anthony.brabazon, 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 of the financial institutions internationally have been implementing anti-money laundering solutions (AML) to fight investment fraud activities. 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 on developing a new data mining solution for AML Units in an international investment bank in Ireland, we survey recent data mining approaches for AML. In this paper, we present not only these approaches but also give an overview on the important factors in building data mining solutions for AML activities.

Keywords: data mining, anti money laundering, clustering, classification, SVM

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 after Currency Exchange and Auto Industry. According to the United Nations Office on Drug and Crime, worldwide value of laundered money in a 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 according to the fraud that has already been exposed. Nowadays, it poses a serious threat not only to financial institutions but also to the nations. 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 the 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 a various 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. However, 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 the literature. Hence, in this paper, we present and discuss the current approaches. We then conclude important points for developing an AML solution based on DM.

The rest of this paper is organised as follows: Section 2 is a brief review of applying DM in banking and finance. Section 3 deals with the challenges of exploiting data mining to investigate money laundering. We present and analyse current approaches of DM within anti-money laundering contexts in Section 4. We review DM frameworks for detecting money-laundering activities in section 5. Finally, we conclude in section 6.

2. Data mining in banking and finance

Today, financial institutions manage a huge banking data and more datasets are being recorded daily. The growth of financial data collected by far exceeds human capacities to manage and analyse them efficiently in a traditional way. Global competitions, dynamic markets, and rapidly increase in the technological innovation become important challenges for these organizations. They need to apply new business intelligent solutions, as the traditional statistical methods do not have the capacity to analyse large datasets.

In banking and finance, we can use DM to solve business problems in finding patterns, causalities and correlations in financial information that are not obviously apparent to managers because of the volume of data. DM can firstly be used to analyse huge datasets and build customer profiles of different groups from the existing data. It can generate rules and models that can be used for understanding business performance, making new marketing initiatives, market segmentation, risk analysis and revising company customer policies. DM methods used for customer profiles can be listed as: clustering [3], classification [3], regression, association rule discovery and sequential pattern discovery [3]. Another important contribution of DM in banking and financial is the prediction, e.g. using DM's results to predict the future trends of a service and the risk factor that it belongs to, based on its previous behaviour. For example, Decision Tree [3], Rule Induction [3] and other classification methods can be used to build models that can predict default risk levels of loan services. Furthermore, with the capability of identifying trends and patterns in large datasets, DM is also suitable for detecting and/or predicting suspicious patterns in ML activities.

3. Challenges of using data mining to Investigate Money Laundering

3.1. Data Quality

In banking and finance, datasets has a different set of quality problems at the instance level. Some of them can be listed as: *missing values*, dummy values or null. This would happen in most of data fields in all databases except the CID, the customer type (corporate, individual and joint) and the fund name; *misspellings*, usually typos and phonetic errors. For instance: "MACAO" vs. "MACAU", "11 1101" vs. "11-1101", "Bloggs Corporation A/C 001" vs. "Bloggs Corporation 001", etc.; *abbreviations*, e.g. "A/C" vs. "AC" and "Account". Besides, banking datasets are normally managed in distributed way for the flexibility and security reasons. The independence and heterogeneity of each data source can also be data quality issues when an integrating task is required, as all conflicts must be solved. Fundamentally, data preprocessing step is applied to deal with data quality issues.

3.2. Data volume and heterogeneity data

Large and growing volume of datasets in financial institutions with regard to the relatively small number of suspicious ML cases in them become a challenge because the analysis of such large volumes is a time-consume tasks for AML experts. Moreover, these large data needed for analyzing is normally not available at

a single place. The distribution of datasets requires integration process in data preprocessing and that can lead as consequence to data quality issues as mentioned in the section above. Furthermore, financial datasets for investigating ML are usually high dimension and heterogeneous. For instance, [5] defines a massive dimensional support vector that consists of "customers x accounts x products x geography x time". Data type of account value is continuous; meanwhile the geography obtains discrete value.

3.3. The nature of ML

Most industries deal with funds in some ways whether it is cash, cheque, credit card or electronic transfers. In a banking and finance environment all mediums are used, this is why building an AML solution is not an easy task because ML instances are not self-revealing. Instances of ML reporting are likely to be rare [13]. Today, ML activities are more and more sophisticated because of this reason. ML crimes are well hidden within a normal distribution as it mimics normal behaviour. Hence, they exist in the large majority of legal transactions. Therefore, data volumes and the nature of ML are challenges to the first generation of AML solutions that are rule-based mechanisms based on predefined sets of fixed thresholds for example, the using mean and standard deviation rules for volume and quantity of transactions in a period of time.

4. Current data mining approaches in anti money laundering

4.1. Clustering

Clustering is the process of grouping the data into classes so that objects within the same cluster have high similarity and objects within different clusters are very dissimilar. There are different clustering methods in the literature and they have been successfully exploited for scientific datasets, spatial datasets, business datasets, etc. In AML, clustering is normally used for grouping transactions/accounts into clusters based on their similarities. This technique helps in building patterns of suspicious sequence of transactions and detecting risk patterns of customers/account. One of the most challenges in clustering financial datasets is their size, this technique, for instance, should deal with millions of transactions during hundreds/thousands of time instances. [16] applied a discretization 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 (Figure 1). They firstly discretize 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. This approach improves firstly the complexity by reducing the clustering problem to a segmentation problem [4]. Next, it avoids the iterative search existing in other clustering algorithms such as K-means [9]. Furthermore, it is more or less appropriate for analysing individual behaviors or group behaviors by their transactions to detect suspicious behaviors related to "abnormal" hills in their histogram. However, as we have to analyse many customers with many transactions of variety amounts for a long period, it is difficult to detect suspicious cases, as there are very few or no "peak hills" in the histogram. Another global analysis is firstly needed and we can then apply this method for further analysis in this case.

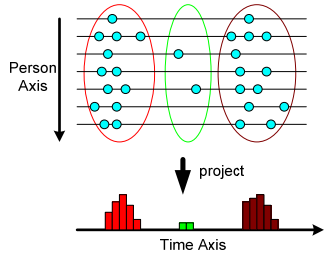

Fig. 1 Histogram segmentation based clustering [16]

4.2. Support Vector Machine

Support Vector Machine (SVM) [12] is a kind of statistical learning that is widely used for classification and regression. As mentioned in Section 3 above, AML task involves the detection of unusual behavior of all dimensions (transactions, accounts, product types, etc.). Hence, the AML problem becomes a pattern

classification and divides datasets in two parts: normal and abnormal sets. Besides, results of classification depend strongly on the training datasets. Therefore, the training set should be large enough in order to get stable results with high accuracy. Meanwhile, in money laundering, finding a popular training dataset is a challenge. In some financial institutions, for instance, there is only one or two suspicious transactions per month compared to thousands of "clean" transactions per day. This is why SVM, which is a classification method based on small training datasets is suitable for classifying normal and abnormal patterns in AML. Moreover, SVM is also not sensitive to the dimensionality disorder feature that is popular in financial datasets.

Traditional SVM is a supervised learning method that requires labeled training datasets to create classification rules. 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. One-class SVM can be summarized as follow: Given a set of unlabeled training set $x_1, x_2 \ldots, x_n \in X$, and X is chosen in such a way that its most data have a common feature while a small number of elements are outlier. This approach attempts to find a kernel function *f* where *f(x)* takes the value +1 with most of the data $x \in X$ and it takes the value -1 on outlier. This problem is modeled as a quadratic optimization problem [7]. Moreover, an implementation of this approach can be found in [6]. The advantage of one-class SVM as mentioned above is that it requires a small set of unlabeled training data. However, finding efficient parameters for the optimization cost function to avoid the overfitting with a given datasets is still an open question. Besides, modern ML activities as analysed in Section 3.3, try to hide themselves by behavior as "clean" as possible so that outlier detection approach is more and more difficult to discover. Finally yet importantly, the financial datasets is normally heterogeneous with continuous and discrete data. Therefore, additional techniques are needed to extend this SVM-based approach for analyzing heterogeneous datasets. In [10], for instance, the authors present a combination of an improved RBF kernel [8] with the definition of distinct distant [15] to deal with both continuous and discrete data. [5] proposes an unsupervised learning approach by a support vector based on probability thresholds to detect suspicious case. Author exaggerates support vectors to generate an enormous adaptive probabilistic matrix (customer x account x products x geography x time) to compute the likelihood of each customer's behaviors based on simple weighted aggregations.

4.3. Other Data mining techniques

Other data mining techniques for AML are also presented in the literature. For instance, the association rules can be used to detect hidden relationships between financial transactions based on their co-occurrence. Frequent sequence mining finds patterns of transactions that occur frequently. Regression analysis is used to predict the possibility of an account being used as a conduit for ML based on demographic and behavioral variables. Furthermore, to the best of our knowledge, there are still no efficient implementations of other DM techniques listed above for AML.

5. Data Ming frameworks for detecting money laundering

In order to exploit DM techniques efficiently, they need to be integrated in a framework for detecting ML. A DM framework is normally consisted of four layers [11][13] corresponding to four levels of mining: transaction, account, institution and multi-institution. The most basic level is transactions. In this level, transaction records are extracted for an investigation. However, they provide a few analytical contexts because they do not constitute links to accounts or other data. In the second level, multiple transactions are associated with specific accounts. Aggregation of transaction with individual accounts gives a general view of these accounts on 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 (business 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 related to different individuals. The last level investigates the ML involving multiple corporations, organizations and customers. In [11], authors moreover propose three modules for DM-based AML systems: filtering, integrating and analyzing link analytical. Statistical methods have been used in the first module to select feature required. The second module deals

with the description the relationships between different accounts through a direct map and a suspicious score ranking. The last module is responsible for integrating this framework to others.

6. Conclusion

In this paper, we gave an overview of exploiting DM techniques for AML. We start with problems of DM in banking and finance and we mention then challenges of using DM to investigate ML. We also present and analyse of some current DM techniques that have been proposed and implemented for AML. According to these analyses, we can conclude that:

- Rule-based AML systems have been replaced by artificial intelligent approach for AML.
- Unsupervised learning with a small set of training data is suitable for building DM-based solutions for AML.
- Classification and clustering are two important mining methods that can efficiently applied for AML. Most of the current contributions are in these two methods.
- A cooperation of multi DM-techniques is needed to build an efficient solution for detecting ML patterns that are more and more sophisticated. For instance, biology-based techniques such as genetic programming, grammatical evolution [2] could be applied for improving the training sets and in classifying suspicious patterns. Finally yet importantly, few DM solutions take into account challenges analyzed in Section 3. For instance, there is only [10] that deals with the heterogeneous datasets problem.

Furthermore, we also need an efficient framework for integrating DM techniques that can deal with different levels of AML from transactions to multi-organizations. This problem is also a part of our current research activities.

7. References

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

[2] A. Brabazon and M. O'Neill. *Biologically inspired algorithms for financial modelling*, Springer Verlag, 2006

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

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

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

[6] *http://www.csie.ntu.edu.tw/~cjlin/libsvm*

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

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

[9] H. Steinhaus. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci.*, vol IV, 1956, pp.801– 804.

[10] 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

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

[12] V. Vapnik. *The Nature of Satistical Learning Theory*, Springer Verlag, NewYork, 1995

[13] G.S. Vidyashankar, et al. Mining your way to combat money laundering, *DM Review Special Report*, Oct 2007

[14] 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

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

[16] 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.

[17] L. Genzman, Responding to organized crime, *Organized crime*, In H.Abadinsky (Ed.) Belmont, CA: Wadsworth