The 27th annual workshop of the Swedish Artificial Intelligence Society (SAIS),
14–15 May 2012, Örebro, Sweden
Published by Linköping University Electronic Press
http://www.ep.liu.se/ecp_home/index.en.aspx?issue=071

Money Laundering Detection using Synthetic Data

Edgar Alonso Lopez-Rojas
School of Computing
Blekinge Institute of Technology
edgar.lopez@bth.se

Stefan Axelsson
School of Computing
Blekinge Institute of Technology
stefan.axelsson@bth.se

Abstract

Criminals use money laundering to make the proceeds from their illegal activities look legitimate in the eyes of the rest of society. Current countermeasures taken by financial organizations are based on legal requirements and very basic statistical analysis. Machine Learning offers a number of ways to detect anomalous transactions. These methods can be based on supervised and unsupervised learning algorithms that improve the performance of detection of such criminal activity.
In this study we present an analysis of the difficulties and considerations of applying machine learning techniques to this problem. We discuss the pros and cons of using synthetic data and problems and advantages inherent in the generation of such a data set. We do this using a case study and suggest an approach based on Multi-Agent Based Simulations (MABS).

Keywords: Machine Learning, Anti-Money Laundering, Money Laundering, Anomaly Detection, Synthetic Data, Multi-Agent Based Simulation

1 Introduction

Money laundering threatens the economic and social development of countries. The threat is due to the injection of illegal proceeds into the legitimate financial system. Due to the high amount of transactions and the variety of money laundering tricks and techniques, it is difficult for the authorities to detect money laundering and prosecute the wrongdoers. Thus, it is not only the amount of transactions, but the ever changing characteristics of the methods used to launder money that are constantly being modified by the *fraudsters*, which makes this problem interesting to study.

This paper aims to analyze the implications of using machine learning techniques for money laundering detection (also known as Anti-Money Laundering, AML) in a data set consisting of synthetic financial transactions.

Our case study is based on the company **AB** [1]. Company **AB** has developed a mobile money implementation that provides users with the ability to transfer money between mobile phone users, by using the phone as a sort of electronic wallet. The task at hand is to provide a tool that detects suspicious money laundering activities.

The mobile money service is currently running in a demo phase. Hence, real data from this system is not available at this stage and therefore the system does not produce representative data that can be used e.g. for the training of the machine learning algorithm. Thus, due to the lack of real data we turn to the generation of synthetic data as an alternative.

The use of synthetic data for machine learning has implications. In this paper we present our ideas about how to address some of the difficulties raised by the lack of real data.

Outline The rest of this paper is organized as follows: sections 2 and 3 introduce the topic of money laundering and present previous work. Sections 4 and 5 address the main topic, which is the use of synthetic data for Anti-Money Laundering. We finish with a discussion and conclusions including future work in sections 6 and 7.

2 Background

Money Laundering affects the finances of nations and it may contribute to an increase in the funding

[1] The identity of the Company AB unfortunately cannot be disclosed due to a Non-Disclosure Agreement

of criminal activities [3]. Due to issues such as the high amount of transactions taking place in any financial service, it is not a trivial task to find specific anomalous transactions that should be marked as suspicious. The reported suspicious activity needs to be supported by tangible evidence that allows specialized government agencies to investigate further.

In Sweden and other countries, most companies in the financial sector are required by law to address money laundering detection. The cost of implementing such controls for AML is quite high, mainly because of the amount of manual labor required. In Sweden alone it is estimated to be around 400 million SEK annually [13].

The most common method today used for preventing anomalous financial transactions consist in establishing thresholds for all transactions. Transactions that exceed these thresholds require extra scrutiny, whereby the client needs to declare the precedence of the funds. These thresholds are set by law without distinction made between different economic sectors or actors. However, this of course leads to fraudsters changing their behavior in order to avoid this kind of control, by e.g. making many smaller transactions that fall just below the threshold [9].

The specific domain covered here is the service *Mobile Money*[2], which is offered by Company **AB**. *Mobile Money* is a platform for transferring money between users, using their mobile phones. This is accomplished by the use of codes sent through text messages or the Internet.

Mobile money brings several benefits for users, including the simplicity of transferring money between themselves and others. One user only needs to know the mobile phone number of the receiving user in order to send money. If the receiving users are registered in the system then the money can be deposited right away in their account. Otherwise, the users receive a code via SMS that enables the recipient to collect the money in cash at one of the nearby local stores that are affiliated with the mobile money operator. It does not require a user to own a bank account, which is beneficial for many people in the world who do not have sufficient assets to warrant a bank account. However, if the user wants to refill their account or withdraw money, then an existing bank account can be connected to the mobile money account, and used in conjunction with it. There are also other alternatives such as top-up card or credit cards connected with the service, that can be used to deposit or withdraw money from the mobile money system.

[2] *Mobile Money* is a generic name that we use in this study and it is not the true name of the service provided by Company **AB**

3 Related work

A number of basic countermeasures against money laundering have been proposed, including basic statistical analysis which constrains the amount of the transactions as well as restricting their frequency [5]. Other methods that complement these basic security measures are based on checking every customer against a black list originating from previous investigated cases and a white list to e.g. avoid mistakes when faced with persons with the same name. Unfortunately, these and other methods have proved to be insufficient [13].

Several machine learning techniques have been used for detecting fraud, and more specifically money laundering, [17]. From the point of view of machine learning, the application is interesting, due to the successful classification rate (high *True Positives* and low *False Positives*) that the classification model can achieve compared to other methods such as simple rule based detection that compares transactions against fixed thresholds.

Data mining based methods have also been used to detect fraud [15, 20, 19]. This leads to the observation that machine learning algorithms can identify novel methods of fraud by detecting those transactions that are different (suspicious) in comparison with the benign transactions. This problem in machine learning is known as novelty detection. Supervised learning algorithms have been used on a synthetic data set to prove the performance of outliers detection [1].

There are tools such as IDSG (IDAS Data and Scenario Generator [11]) which was developed with the purpose of generating synthetic data based on the relationship between attributes and their statistical distributions. IDSG was created to support data mining systems during their test phase and it is been used to test fraud detection systems.

Gao [8] proposed one of the frameworks used for AML introduces the terms *legal transaction, usual transaction, unusual transaction, suspicious transaction and illegal transaction* for describing differ-

ent possible categories of transactions. This framework aims to rank the likelihood that a transaction would be illegal on a scale from 0 to 100, which enables prioritization.

We wish to stress that a detector cannot be completely certain that a transaction corresponds to money laundering. This task is delegated to the legal authorities. Instead of doing that, we intend to flag customers and transactions with a label of suspicion that focus the attention of the operator for further investigations.

Despite the possible bias injected in the data set during the simulation, synthetic data has been previously used with similar reasons as the ones presented by Barse [2]. As in that work, the *lack of real data* and the *low probability of real instances of fraud in the real world data obtained*, are some of the reasons discussed further in this paper.

In general the availability of financial information for research is very restricted by the corporate policies and even the law. Customers are usually protected by the financial organizations and the disclosure of their private information is limited by internal and government policies. In order to get access to such data, anonymization techniques should be used on the data set in order to allow for the preservation of privacy [18, 10].

4 AML for mobile money

The detection of money laundering in the mobile money service is not trivial due to the difficulty of classifying transactions that are intended to appear as normal and legal. In this paper we address this problem with the approach of learning from the experiences of past detected patterns of illegal behavior in order to gain more knowledge about the possible rules or new patterns of fraud that could emerge in a mobile money system.

The first step to address our problem is to start by clearly formulating the learning problem.

4.1 Problem definition

Regarding the mobile money AML domain, we formulate the learning problem as:

Task T Classification of transactions as normal or suspicious based on the known pattern of legal transactions. The aim is to find anomalies or outliers inside a data set of mobile money financial transactions.

Performance Measure P Percentage of transactions correctly classified as anomalous, also known as *True Positives* (TP) and the percentage of *False Positives* (FP) i.e. transactions that are not anomalies and are misclassified as anomalies.

Experience E Synthetic data generated with transactions labeled as legal (normal) and/or illegal (suspicious).

The main weakness is that the experience gained by using a synthetic data set can be biased and in some cases it may not match a realistic situation that would occur in a real-world data set. However in the following sections we present our analysis of how this can also be used to our advantage by allowing us to gain information about unseen but expected situations in a real-world data set.

4.2 Data Preprocessing

One of the tasks that need to be addressed is data preprocessing. This task includes the selection of the attributes, discretization, noise removal and, in certain domains, data fusion.

Company **AB** has a design with a database that aims to store all the log information about the users interactions with the service. For this study we need to select the attributes that best contribute to the correct classification of suspicious transactions.

The customers are originally differentiated in the system by a profile. The profile for each customer is specified at the opening of the account by internal criteria. Customers with certain profiles are limited e.g. in the amount and the frequency of the transactions that they can perform. Additionally, there are specific methods that detect anomalies based on profiling customers which can be applied when faced with prior profiles such as these [4].

In addition to the generation of profiles, we added the following attributes to our simulation: Customer ID, Profile, Date of the Transaction, type of transaction (e.g. deposit, withdraw, transfer), Amount of the Transaction, Location (city), Account Age (months since the creation of the account) and Customer Age (years).

For each transaction of type *transfer* there is also a *deposit* transaction with the same value for a different customer. This transfer transaction describes a social network between customers. The rest of the fields are generated according to the given parameters of the simulation and random op-

erations with range validation to guarantee consistent data that follows a realistic model.

Data labeled as anomalous should be added in the transaction database in order to run supervised algorithms. These anomalous records are created with the intention to replicate some of the common patterns used by fraudsters. One example of these patterns is an unverified user profile which makes either large deposits or large withdraws in comparison to a predefined threshold. Some other known problematic patterns of usage include: several withdraws from the same profile above the average normal value for transactions by young customers, a verified customer that performs a single large withdraw, and finally a chain of transactions that deposits money in a single account followed by consecutive withdrawals from that same account.

In a realistic situation we would be handling millions of transactions in a data set and in most of the cases only a small testing samples of the whole data set can be processed.

Although computational scaling performance is a topic that is not addressed here, the learning algorithms selected are profoundly affected by the amount of data provided for the training and the cross validation phase.

4.3 Learning with Synthetic Data

For a real world data set the selected algorithms should produce the best accuracy, i.e. TP rate, in comparison with the other algorithms. This tells us directly that our classifier can detect a significant number of suspicious transactions. However, the FP rate which is represented by the misclassified number of instances from the normal data, is also an important indicator of performance because we do not want a situation where the high number of FP will consume the time of an investigator and leads to possible missed cases and lack of trust by the investigators in the detector.

Thus, we are interested in providing an accurate method to improve the detection rate (TP) and reduce the misclassification rate of the benign data (FP) counted on the data collected from the simulation. A synthetic data set can be used to train the classifier and test different scenarios. One example of such a scenario is when the customer population have low income and only a few customers have large assets.

The results of this classification algorithm with a synthetic data set should be interpreted in a broader perspective than results with a real-world data set. In a real scenario the results are used for prosecuting and reporting individuals. But when doing research using a synthetic data set the purpose is different. One of the goals is to identify the measures of detection and control that could be added to the system, given a set of conditions, with the clients.

We studied possible algorithms for our detection research using the same data set. The algorithms analyzed here are based on *Decision Tree* learning and clustering techniques. Other methods such as Support Vector Machine (SVM), Neural Networks, Link Analysis and Bayesian networks are not addressed here, but we expect to improve our approach in a future work by including these methods.

Decision Tree algorithms construct a tree that contains branches with rules that correctly classifies the most of the data set [16, 6]. The main advantage of using these algorithms (in comparison with other machine learning algorithms) for the domain of mobile money AML is the possibility for an investigator to determine common rules that classify suspicious behavior.

There could be situations where *fraudsters* start to behave according to a new pattern based on e.g. a specific location or city, combined with other attributes such as age, profile and others that inspected singly seem to be normal, but in combination could lead to the detection of a new fraud trend. However some of the requirements of these algorithms are that the data set used for the training should be representative enough of the whole data set in order to get rules that sufficiently generalize to the real scenario. These rules should be refreshed frequently in order to detect new possible fraud trends.

Clustering techniques such as distance based clustering and density based clusters can be useful to classify natural clusters that appear in the data set. The disadvantage with these techniques is the hard task of finding the parameters that expose abnormal behavior clusters due to the class unbalance problem of the distribution of the classes [7]. In a normal situation most of the records in a data set are instances of the normal behavior of a customer, with only a few representing anomalous (i.e. interesting) behavior.

In addition, the complexity of the patterns used

by *fraudsters* represent a challenge, due to the fact that the fraudsters' patterns intend to mimic normal behavior in order to pass undetected by law enforcement.

Besides, from the perspective of our detection tool, we cannot be 100% certain of the illegal precedence of the funds in a transaction, that is why our detector should include a *suspicious* rank that allows an investigator to prioritize more relevant and important cases.

5 Mobile Money simulation

During this research, we found a number of difficulties that affect working with this domain. We found the lack of access to real data, the poor quality of samples in the real-world data we could access, and the many possible scenarios that we would like to explore, are some of the problems found in our case study. Such a problems make the process of research more difficult. This is the reason behind the discussion of using synthetic data as an alternative to further research in this area.

Important issues arise when analyzing the use of machine learning for money laundering such as: volume and complexity of data, class imbalance, concept drift, class overlap and class mislabeling [17].

However, many other considerations also play a role in the simulation. One of the most important challenges that we need to address is: *How can we make our model realistic and as close as possible to a desired scenario?*

In order to answer this question, we present the following benefits and disadvantages of using synthetic data for our research.

5.1 Benefits of using synthetic data

When using synthetic data one of the benefits we identified is the possibility of selecting attributes that reduce considerably the complexity of the data structures involved. Furthermore, this simplifies the tasks of data preparation and extraction from real sources. The volume of the data can be tuned to comply with different experimental setups.

The class imbalance problem can be reduced by setting up a simulation that produces enough records of each of the interesting classification classes. Class overlap is still a remaining issue with simulations. However, simulations that properly represent fraud behavior can avoid class mislabeling.

We summarize our findings as:

- The data that represent realistic scenarios are readily available.
- The privacy of the customer is not impacted.
- The disclosure of results is not affected by policies or legal issues.
- The data set is available for other researchers to reproduce experiments.
- Different scenarios can be modeled with parameters controlled by the researcher.
- Injection of enough abnormal data to address the class unbalance problem.
- Simulation of abnormal behavior prevent the problem of mislabeled classes.

5.2 Disadvantages of using synthetic data

Unfortunately other issues arise that are important to consider when using synthetic data. Some of these issues are:

- The data generated might be nor representative or realistic.
- Data can have biased information.
- It is difficult to build a realistic model due to the complexity of variables and parameters.
- The simulated suspicious data cannot be investigated further by the government agency. In a real scenario these results could be used for improving the accuracy of the existing classification algorithms.
- It is unknown if we can transfer the learning from a simulated data set to a real-world data set.

Some of the disadvantages presented can be minimized if we can build a simulation with records that can represent a real-world situation. It is important to understand that the purpose of the simulation is not to reproduce a view of the real world, but to provide an alternative simplified scenario that is designed according to a model as it is presented in the following section.

5.3 Multi-Agent Based Simulation

Multi-Agent Based Simulation is an approach that involves the use of autonomous and interactive agents and it has been used to model complex sys-

tems. These agents are described by their state, behavior and their interaction with other agents, which generates complex global behavior usually found in different domains [12].

Previous work has shown the use of Multi-Agent based simulation in the task of simulating social networks and analyzing social behavior [14]. *Mobile Money* resembles a social network of connected clients where the connections are represented by the transactions (money sent or received) and the nodes are represented by the clients.

The synthetic data from a simulation aims to represent the interactions of the customers of a mobile money system. The graph shown in Figure 1 is used to represent a desired scenario that can be used to study a certain phenomenon nor existing in a real world data set. This graph is an hypothetical situation where 2000 clients from 7 different cities perform legal transactions with customers inside or outside their cities. The simulation allows the researcher to follow one agent and keep track of its behavior and also store all the transactions for further analysis.

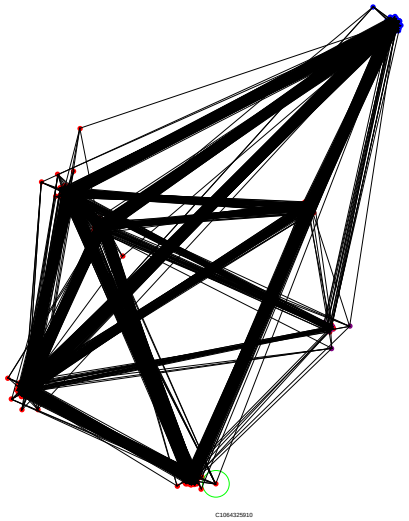

Figure 1: Scenario 1 - 2000 clients distributed across 7 cities and multiple edges connecting clients that produces legal transactions

Figure 2 represent a small simulation of 20 malicious agents distributed across 3 cities. The behavior of these agents can be modeled as a cooperative network of agents which aim to move a certain amount of money from the red nodes to the blue nodes. By doing this, the malicious agents avoid the threshold controls present in the system.

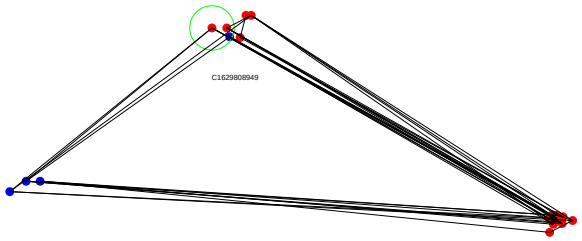

Figure 2: Scenario 2 - 20 accounts distributed across 3 cities generating suspicious transactions

There are several agent-based frameworks that incorporate toolkits to aid the development of these kind of systems. Some of them are freely available and are widely used in academic simulations (e.g. MASON[3], Repast [4], Swarm[5]). We used MASON for the network graphs and simulations presented in Figure 1 and 2.

6 Discussion

One of the difficulties with this domain is the lack of available data sets to use for training the machine learning algorithms and compare results with other researchers. This is the reason why a synthetic data set is proposed as an alternative. We are unable to obtain any data set for the mobile money system at the moment.

Before, we have done some preliminary work based on a simplified statistical simulation. Because of the elementary of this simulation we are not going to discuss any result here. This simulation was a biased representation of the domain that was used to run different machine learning algorithms in order to experiment with different outputs. But from this experience, we gained interest in producing a better simulation that shows real behavior and characteristics of clients from this domain.

The concept of MABS presented in section 5.3 make us consider that we can obtain a model based

[3]MASON `http://cs.gmu.edu/~eclab/projects/mason/`
[4]Repast `http://repast.sourceforge.net/`
[5]Swarm `http://www.swarm.org/`

on multi-agents that would be realistic enough for our purpose of analyzing the possible scenarios and build a detector for money laundering. This led us to work on a simulation based on the concept of Multi-Agents Systems.

The problem of finding anomalies within the domain of money laundering is really a challenge. Every time a new pattern of money laundering is detected by the authorities and a new control mechanism is implemented, the fraudsters change their modus operandi and create a new method that is undetectable by the current rules of a rule based AML system.

When doing research in the domain of *mobile money* a number of difficulties arise, including the lack of access to real data that can be used to evaluate the learning algorithms. Even with real data, the lack of anomalous transactions can be a problem. This is why the injection of synthetic anomalies in a real-world data set is an alternative to overcome the problem of e.g. an unbalanced number of classes.

We do not expect to be able to identify all anomalies but we intend to identify abnormal behavior from the customers that can lead to the detection of these new mutated methods.

7 Conclusions

We have presented an analysis of the use of a synthetic data set from the domain of mobile money for experimentation with machine learning algorithms. Through the use of simulation of different scenarios we can discover flaws in the current system. This can also lead to the finding of new policies and legislation that could detect the appearance of previous detected patterns of money laundering in the future.

We pretend to illustrate the methods that can be used to evaluate the accuracy of different algorithms, without going into specific details. Our analysis covers *Decision Trees*, *Clustering* techniques and *Decision Rules* that are more understandable by human operators than other machine learning algorithms.

When working with synthetic data there is always a risk of generating a data set that does not represent the real world data set. This can lead to results that are biased by the way the data was generated. On the other hand a synthetic data set can also simulate different scenarios that are not available for experimentation and analysis as they are unusual, catastrophic etc.

As shown before, the benefits presented in section 5.1 make us conclude that using synthetic data for machine learning experimentation is a good alternative in domains where the lack of real data is a problem.

Further work will focus on building a model for the simulation of mobile money transactions. Multi-Agent Based Simulation (MABS) is an interesting technique that can be used to improve the results of the generation of realistic synthetic data sets for this domain. We aim to test in the future the performance of several machine learning algorithms such as Support Vector Machine (SVM), Neural Networks, Link Analysis and Bayesian Networks. These algorithms have been used successfully in previous studies and it is of our interest to evaluate them in future research.

Acknowledgments This research is supported by the research group DISL (Distributed and Intelligent Systems Laboratory) from BTH (Blekinge Techniska Högskola) in Sweden. We also thank PhD Niklas Lavesson (BTH) for his contribution and reviews of this article.

References

[1] Naoki Abe, Bianca Zadrozny, and John Langford. Outlier detection by active learning. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, page 504, 2006.

[2] E.L. Barse, H. Kvarnstrom, and E. Johnson. Synthesizing test data for fraud detection systems. *19th Annual Computer Security Applications Conference, 2003. Proceedings.*, pages 384–394, 2003.

[3] B. Bartlett. The negative effects of money laundering on economic development. *Asian Development Bank Regional Technical Assistance Project No*, (5967), 2002.

[4] RJ Bolton and DJ Hand. Unsupervised profiling methods for fraud detection. *Conference on credit scoring and credit control*, 2001.

[5] R.J. Bolton and D.J. Hand. Statistical fraud detection: A review. *Statistical Science*, 17(3):235–249, 2002.

[6] L Breiman. Random forests. *Machine learning*, pages 5–32, 2001.

[7] NV Chawla, KW Bowyer, L.O. Hall, and W.P. Kegelmeyer. SMOTE: synthetic minority oversampling technique. *Journal of Artificial Intelligence Research2*, 16:321–357, 2.

[8] Zengan Gao and Mao Ye. A framework for data mining-based anti-money laundering research. *Journal of Money Laundering Control*, 10(2):170–179, 2007.

[9] John Hunt. The new frontier of money laundering: how terrorist organizations use cyberlaundering to fund their activities, and how governments are trying to stop them. *Information & Communications Technology Law*, 20(2):133–152, June 2011.

[10] L I U Junqiang. Optimal Anonymization for Transaction. *Chinese Journal of Electronics*, 20(2), 2011.

[11] P.J. Lin, B. Samadi, Alan Cipolone, D.R. Jeske, Sean Cox, C. Rendon, Douglas Holt, and Rui Xiao. Development of a synthetic data set generator for building and testing information discovery systems. In *Information Technology: New Generations, 2006. ITNG 2006. Third International Conference on*, pages 707–712. IEEE, 2006.

[12] C M Macal and M J North. Tutorial on agent-based modelling and simulation. *Journal of Simulation*, 4(3):151–162, September 2010.

[13] Dan Magnusson. The costs of implementing the anti-money laundering regulations in Sweden. *Journal of Money Laundering Control*, 12(2):101–112, 2009.

[14] J Pavon, M Arroyo, S Hassan, and C Sansores. Agent-based modelling and simulation for the analysis of social patterns. *Pattern Recognition Letters*, 29(8):1039–1048, June 2008.

[15] Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. A comprehensive survey of data mining-based fraud detection research. *Arxiv preprint arXiv:1009.6119*, 2010.

[16] J.R. Quinlan. Simplifying decision trees. *International journal of man-machine studies*, 27(3):221–234, 1987.

[17] Agus Sudjianto, Sheela Nair, Ming Yuan, Aijun Zhang, Daniel Kern, and Fernando Cela-Díaz. Statistical Methods for Fighting Financial Crimes. *Technometrics*, 52(1):5–19, February 2010.

[18] Yabo Xu, Ke Wang, Ada Wai-chee Fu, Hong Kong, and Philip S Yu. Anonymizing Transaction Databases for Publication. *International Journal*, pages 767–775, 2008.

[19] Dianmin Yue, Xiaodan Wu, Yunfeng Wang, Yue Li, and Chao-Hsien Chu. A Review of Data Mining-Based Financial Fraud Detection Research. In *2007 International Conference on Wireless Communications, Networking and Mobile Computing*, pages 5514–5517. Ieee, September 2007.

[20] ZM Zhang and JJ Salerno. Applying data mining in investigating money laundering crimes. *discovery and data mining*, (Mlc):747, 2003.