Identifying Money Laundering Accounts

Chih-Hua Tai
Dept. of Computer Science and Information Engineering
National Taipei University
New Taipei City, Taiwan
hanatai@mail.ntpu.edu.tw

Tai-Jung Kan
Dept. of Computer Science and Information Engineering
National Taipei University
New Taipei City, Taiwan
j7400660@gmail.com

***Abstract*—Money laundering is often associated with criminal activities. Anti-money laundering is thus regarded as an important task in many countries. However, as it is common that money launderers divide the dirty money into multiple parts and make sequences of banking transfers or commercial transactions, manually detecting activities of money laundering is challenging. To ease the task, this work establishes a two-phase intelligent method based on machine learning and data analysis techniques for identifying suspicious money laundering accounts from the transaction data. The first phase emphasizes on identifying every suspicious money laundering account while the second phase further retrieves highly suspicious ones so that both the recall and precision for the identification of money laundering accounts can be somewhat taken care of. Evaluated on the data given by Bank SinoPac, the established intelligent method achieves a recall rate of 26.3%, which is three times the recall rate (8.6%) of the Money Laundering Control Act in Taiwan, in the first phase, and later the precision rate can be increased up to 87.04% in the second phase.**

Keywords—anti-money laundering, data analysis, machine learning

I. INTRODUCTION

Dirty money is the money earned or obtained through illegal ways and often associated with criminal activities. Money laundering is the process of making the dirty money appear "clean" (i.e. legal) [1]. To combat crime, anti-money laundering is thus regarded as an important task in many countries.

To detect and prevent money laundering, creating particular laws against money laundering, such as the Money Laundering Control Act [4][6][8], the Money Laundering Prevention Act [2], and the Proceeds of Crime Act [3], to name a few, is a traditional method for the governments of countries in the past centuries. However, although such a method can increase the difficulty of money laundering and provide somewhat prevention in advance, it cannot detect the money laundering act effectively. Therefore, the task of anti-money laundering involves a large amount of manual effort from anti-money laundering specialists, who require long-time training.

Nowadays, due to the rise of machine learning and big data analysis, there are new chances to ease the task of anti-money laundering. However, to the best of our knowledge, no existing work addresses this issue from such aspects. Motivated by the emergent need, this work aims at developing an intelligent method to discover suspicious money laundering accounts, so that the required manual effort for anti-money laundering can be alleviated.

Specifically, this work proposes the Machine learning cooperating Layering Simulation (McLayS) method to discover suspicious money laundering accounts based on the transactions. Note that similar to other detection problems, somewhat there is a trade-off between avoiding missing the identification of suspicious money laundering accounts (i.e., recall) and focusing on identifying real money laundering accounts (i.e., precision). The McLayS method is designed as the two-phase identification. The first phase is developed by using machine learning techniques to target on avoiding missing the identification of every suspicious money laundering account (i.e., recall). Those identified accounts in this phase are then recommended to be long-term monitored systematically. After that, the second phase aims at identifying real money laundering accounts (i.e., precision) and thus further retrieves highly suspicious accounts (from those identified in the first phase) using data analysis techniques. Consequently, anti-money laundering specialists can put the effort mainly on the accounts that are identified as highly suspicious. Evaluated on the data given by Bank SinoPac, the proposed intelligent method achieves a recall rate of 26.3%, which is three times the recall rate (8.6%) of the Money Laundering Control Act in Taiwan [4][6], and the precision rate can be increased up to 87.04%.

In summary, the contributions of this work are as follows:

- To the best of our knowledge, we are the first to address the problem of identifying money laundering accounts by machine learning and data analysis techniques.
- For the addressed problem, we propose the Machine learning cooperating Layering Simulation (McLayS) method, which provides the two-phase identification of money laundering accounts and is capable of simultaneously recommending suspicious accounts to be long-term monitored systematically and retrieving real money laundering accounts.
- We demonstrate the effectiveness of the proposed McLayS method on the real data set provided by Bank SinoPac. The evaluations show that McLayS achieves a recall rate of 26.3%, which is three times the recall rate (8.6%) of the Money Laundering Control Act in

978-1-7281-0525-3/19/$31.00©2019 IEEE

Taiwan, and the precision rate can be increased up to 87.04%.

The rest of paper is organized as follows. Section II introduces the data provided by Bank SinoPac for this work. Section III proposes the McLayS method. Section IV provides some insights into the characteristics of money laundering act and evaluates the performance of McLayS. Finally, Section V concludes the work.

II. Preliminary About Data

This section provides a brief introduction for the transaction data, in which the cue to the act of money laundering is implicitly left.

This work studies the identification of money laundering accounts using the transaction data. Without loss of generality, each transaction record consists of anonymized account id, the act (i.e., either deposit or withdrawal), the date, and the amount. Provided by Bank SinoPac, the data used in this study contains 739,142 transaction records, where 339,869 transaction records are for deposit and 399,273 transaction records are for withdrawal. Correspondingly, there are 8,440 different accounts in total, among which 48% of the accounts are manually tagged as suspicious money laundering accounts by anti-money laundering specialists. The tags are then regarded as the ground truth in the study.

III. The McLayS Method

A. Overview

The purpose of the McLayS method is to facilitate the identification of money laundering accounts from the transactions recorded in the bank sites. That is, rather than replacing anti-money laundering specialists, narrowing down the search space to save the required manual effort is the object. Therefore, discovering every possibly suspicious money laundering account (i.e., the recall) and identifying real money laundering accounts (i.e., the precision) are both important. Simultaneously, it is also expected to provide somewhat insights into the act of money laundering.

Motivated by the needs, the McLayS method is designed into two-phase identification based on machine learning and data analysis techniques. The first phase recommends every possibly suspicious money laundering account for the systematical long-term monitoring, while the second phase provides highly suspicious accounts for the manual detection.

The details of the two phases are introduced in the following.

B. Recall Phase

This phase aims at having more insights on the act of money laundering and building a predictor to figure out every possibly suspicious account. For the purposes, this phase focuses on the characteristics of accounts and performs the tasks of feature creation and feature selection in sequence.

Specifically, the feature creation is to extract more characteristics of accounts from the transactions in kinds of aspects, so that the actor of the account can be implicitly drafted as much as possible. In this work, we create 18 features of three categories as follows.

- For each account, the basic statistics on the transaction amounts include:
  1. The average amount of transaction deposits
  2. The average amount of transaction withdrawals
  3. The average amount of all transactions (including deposits and withdrawals)
  4. The median amount of transaction deposits
  5. The median amount of transaction withdrawals
  6. The median amount of all transactions (including deposits and withdrawals)
  7. The total number of transactions
  8. The standard deviation of the transaction amounts
- For each account, the appearance frequencies of particular digits [7] include:
  9. The average of the digit sums of transaction amounts
  10. The average number of digits of transaction amounts
  11. The average number of the digit ‘0’ in the transaction amounts
  12. The average ratio of the number of digit ‘0’ to the number of digits in the transaction amounts
- For each account, the statistics on the transactions with high amounts include:
  13. The appearance frequencies of the digit ‘0’ in the top 10% high transaction amounts
  14. The average ratio of the number of digit ‘0’ to the number of digits in the top 10% high transaction amounts
  15. The average amount of top 10% high transaction amounts
  16. The appearance frequencies of the digit ‘0’ in the top 5% high transaction amounts
  17. The average ratio of the number of digit ‘0’ to the number of digits in the top 5% high transaction amounts
  18. The ratio of ‘the average amount of top 10% high transaction amounts’ to ‘the median amount of transaction amounts’

Given the above 18 features, the feature selection then targets on finding critical features with respect to the recall by training corresponding predictors using machine learning methods (such as Support Vector Machine (SVM)).

Specifically, to figure out critical features leading to higher recall, this phase performs top-down feature elimination. That

is, this phase regards the 18 features as the initial feature set. Accordingly, an initial predictor is first trained. Next, each feature is temporally concealed in turn from the training of another predictor. Then, by comparing the recall of the 17-feature-trained predictor with that of the 18-feature-trained predictor, the most uncritical feature can be figured out. If the concealment of the most uncritical feature can help train a predictor of higher recall, the feature is eliminated from the feature set. Such a feature elimination process is iteratively performed until the removal of any further feature can decrease the recall. Consequently, when the process terminates, the feature set contains the features highly relevant to the identification of every possibly suspicious money laundering account, and the corresponding predictor can recall the most so far.

The discovered critical features in this work will be presented in the section of evaluation.

C. Precision Phase

After every possibly suspicious money laundering account is figured out, this phase aims at identifying highly suspicious ones so that anti-money laundering specialists can focus more on specific accounts and save much effort as possible.

For the purpose, note that the process of money laundering usually consists of three steps: placement, layering, and integration [5]. Placement means introducing the dirty money into financial systems. Layering refers to performing complex financial transactions to conceal the source of the dirty money. Finally, integration is the acquirement of the money after the layering. This phase then tries to identify the highly suspicious money laundering accounts by simulating the layering step. The more frequently an account is involved in the layering step, the more suspicious the account is.

Specifically, the McLayS performs as follows to simulate the complex financial transactions among suspicious accounts. First, based on the transaction records of suspicious accounts, the link between the transferor and transferee of a transaction is re-constructed if there exist two transaction records with the 'deposit' and 'withdrawal' acts, respectively, having identical transaction amount on the same date. After the links are re-constructed, the transferring paths are further inferred by iteratively extending links with other links. Note that by intuition, an amount of money can be withdrawn only after it has been deposited. Therefore, a link $\boldsymbol{l_1}$ (a path $\boldsymbol{p}$) is extended with the other link $\boldsymbol{l_2}$ if and only if (1) the transferee of $\boldsymbol{l_1}$ (the latest transferee of $\boldsymbol{p}$) is the transferor of $\boldsymbol{l_2}$ and (2) the date of $\boldsymbol{l_2}$ is not earlier than the date of $\boldsymbol{l_1}$ (the date of the latest transferring of $\boldsymbol{p}$). As the financial transactions for the layering are usually quite complex, the McLayS extends the links to the paths as many and long as possible. Finally, the frequency of each suspicious account involving in the paths is calculated. The accounts with the frequencies larger than a threshold $\boldsymbol{T}$ are regarded as highly suspicious money laundering accounts.

IV. Evaluations

A. Experimental Settings

The data used for the evaluations are introduced with the details in Section II. All the evaluations are implemented in R and conducted using the 5-fold cross validation.

For comparisons, we implement a Rule-Based method as the baseline according to the concepts of rules lists in the Money Laundering Control Act, Taiwan [4][6]. In addition, to show the advantage of two-phase identification, a version of the Rule-Based method cooperating the Precision Phase of McLayS is also carried out.

B. Selected Critical Features

To provide some insights into the act of money laundering, the following lists the discovered most critical features regarding the recall.

- The average amount of transaction withdrawals
- The total number of transactions
- The standard deviation of the transaction amounts
- The average of the digit sums of transaction amounts
- The average number of the digit '0' in the transaction amounts
- The average ratio of the number of digit '0' to the number of digits in the transaction amounts
- The appearance frequencies of the digit '0' in the top 10% high transaction amounts
- The average amount of top 10% high transaction amounts

C. Performances

In the following, the proposed McLayS method is then compared with the Rule-Based method (with and without the cooperation with the Precision Phase of McLayS) in terms of recall, precision, accuracy, and F1 score. Table I shows the performance comparisons, where the first two columns list the performance regarding the Recall Phase (RP) and Precision Phase (PP) of McLayS, respectively, with the threshold $\boldsymbol{T}$ set as 25.

First, the McLayS(RP) outperforms the Rule-Base in all aspects, showing the potentiality of machine learning techniques for the identification of money laundering accounts. Particularly, after selecting critical features (listed in Section IV.B) regarding the recall, the McLayS(RP) can achieve a recall rate of 26.3%, which is three times the recall rate (8.6%) of the Rule-Base. Next, given the suspicious accounts identified in the McLayS(RP), the McLayS(PP) improves the precision from 64.9% to 87.04%. Given the suspicious accounts identified by the Rule-Base, the mechanism of the Precision Phase of McLayS can improve the precision from 61.12% to 85.71%. These results show the capability of the layering simulation in helping identify money laundering accounts. The McLayS(PP) outperforms the Rule-Base(PP) in

the precision because there are more money laundering accounts successfully identified by the McLayS(RP). Nevertheless, the increase of the precision rate somewhat goes with the decrease of the recall rate due to the trade-off between the precision and the recall.

TABLE I. PERFORMANCE COMPARISONS ($T = 25$).

METHOD	MCLAYS(RP)	MCLAYS(PP)	RULE-BASE	RULE-BASE(PP)
Recall	**0.2630**	0.0117	0.0866	0.0060
Precision	0.6490	**0.8704**	0.6112	0.8571
Accuracy	0.5800	0.5275	0.5378	0.5251
F1 Score	0.3740	0.0230	0.1517	0.0118

V. CONCLUSION

Noted that the task of anti-money laundering involves a large amount of manual effort from anti-money laundering specialists, who require long-time training. This work addressed the possibility of applying artificial intelligence to facilitate the task and proposed the McLayS method based on machine learning and data analysis techniques. The McLayS provided the two-phase identification of suspicious money laundering accounts, where the phases particularly emphasized on the recall and precision, respectively. Therefore, every possibly suspicious money laundering account can be suggested to be systematically monitored further, and highly suspicious accounts can be identified for manual investigation. Evaluated on the data given by Bank SinoPac, the McLayS method can achieve a recall rate of 26.3%, which is three times the recall rate (8.6%) of the Money Laundering Control Act in Taiwan, in the Recall Phase, and the precision rate can be increased up to 87.04% in the Precision Phase.

ACKNOWLEDGMENT

This work is supported in part by Bank SinoPac and Atelier Future, National Cheng Kung University under the project " Enhancing the Mechanism of KYC and Money Laundering Prevention by AI - Customers' Properties and Data Collection and Analysis, KYC and Money Laundering Prevention, Deposit and Credit Card Fraud Protection", and by the Ministry of Science and Technology (MOST), Taiwan, under the grant number MOST 107-2221-E-305-016.

REFERENCES

[1] "History of Anti-Money Laundering Laws". United States Department of the Treasury. 30 June 2015. Retrieved 30 June 2015. https://www.fincen.gov/history-anti-money-laundering-laws

[2] "Money Laundering Act 2012 amended". Resource Portal of OGR Legal. OGR Legal. Retrieved 2 November 2015. https://resource.ogrlegal.com/money-laundering-act-2012-amended/

[3] "OPSI: Proceeds of Crime Act". Retrieved 14 February 2009. http://www.opsi.gov.uk/acts/acts2002/ukpga_20020029_en_1

[4] "Suspicious Money Laundering Transaction: Q&A". Anti-Money Laundering Division, Investigation Bureau, Ministry of Justice, Taiwan. Retrieved 20 November 2017. https://goo.gl/jbu6Ln

[5] "Money Laundering". Wikipedia. Retrieved 18 January 2018. https://en.wikipedia.org/wiki/Money_laundering

[6] "Suspicious Money Laundering or Terrorist Trading" Anti-Money Laundering Division, Investigation Bureau, Ministry of Justice, Taiwan. Retrieved 30 January 2018. https://goo.gl/SfDrvC

[7] "Don't Trust Your Instincts-Benford's law" PanSci. Retrieved 25 December 2017. https://pansci.asia/archives/54264

[8] "Money Laundering Control Act". Wikipedia. Retrieved 25 May 2019. https://en.wikipedia.org/wiki/Money_Laundering_Control_Act