A Novel Approach for Detecting Suspicious Accounts in Money Laundering using Frequent Patterns and Graphs

Arati Gade[1], Prapti Nirmal[2], Pragati Bharambe[3], Rutuja Mhaske[4]

[1,2]AISSMS IOIT COLLEGE, PUNE
Dept. of IT Engineering, AISSMS IOIT College, Maharashtra, India

---

***Abstract* — In today's world banking system is equipped with highly intelligent and driven software solutions. This includes cash depositing, Printing passbooks, cheque book requisitions, cash withdrawing and many more which doesn't need any human interface. Even though on having so many technical advancements in the banking sector, many loopholes are still existed. One of the major from all this is Money laundering, This money laundering happens with the inclusion of multi hop transfers and many other patterns of debits and credits. Most of the time due to huge amount of the transaction data banking servers are unable to identify the actual cause or the root of the money laundering. So as a boon to this Machine learning emerges with some good solution to identify the fraud accounts that actually involve in this activity. So as a tiny step towards this proposed model put forwards an idea of suspicious account detection, which involves in the money laundering process using Frequent itemset mining and Hyper graph generation process. This technique is supported by information gain theory and Decision Making model which eventually enhances the process of suspicious account detection in money laundering.**

Keywords: Entropy Estimation, Linear clustering, Hyper graph, Frequent patterns.

INTRODUCTION

Money laundering is the method of making the appearance that huge quantities of cash received from crook interest, such as terrorist or drug trafficking pastime, derived from a legitimate supply. The finance from the unlawful interest is considered bad, and the system "launders" the cash to make it appear smooth. Money laundering is essential for crook corporations who desire to apply for unlawfully earned money successfully. Dealing in huge amounts of unlawful cash is ineffective and threatening. The criminals want a manner to deposit the cash in economic institutions, but they can most effectively accomplish that if the cash seems to come from legitimate sources.

There are many simple to complex methods ranging for money laundering technique. Most common practices for money laundering used by the criminal organization are a cash-based business. For example, if an institution buys a restaurant, it would inflate the everyday money receipts to channel its illegal cash via the restaurant to the bank. Then they are able to distribute the budget to the proprietors out of the restaurant bank account. Those varieties of organizations are regularly called "fronts."

Smurfing is another general practice for money laundering, in which the individual divides the big amount of cash into small multiple deposits and spread into multiple accounts to hide from detection. Another commonly used practice for money laundering is wire transfers, currency exchanges, cash smugglers or mules are the individuals who involved in cross borders money laundering practice. Other money laundering techniques involve making an investment in commodities inclusive of gemstone and gold that may be effortlessly shifted to other jurisdictions, cautiously spending in and selling precious belongings such as actual property, playing counterfeiting and creating shell agencies.

While conventional money-laundering methods are still used, the internet has positioned a brand new spin on an antique crime. Using the net allows money launderers to effortlessly keep away from detection. The upward thrust of the establishment of online banking, anonymous online price services, peer-to-peer transfers the usage of cellular telephones and using digital currencies like Bitcoin, have made finding the illegal transactions of money even greater difficulty. Moreover, the use of proxy servers and anonymizing software makes the 1/3 factor of cash laundering, integration, almost impossible to discover, as cash may be transferred or withdrawn leaving less or no hint of an IP address. Gambling websites, online auction and sales, and online virtual gaming websites are also involved in money laundering practice. They converted the ill-gotten money towards gaming currency, then black into real transactions taken place, untraceable and usable "clean and white" money.

Anti-money-laundering legal guidelines (AML) had been sluggish to trap up to these types of cybercrimes, seeing that maximum AML legal guidelines try to uncover dirty cash because it passes through conventional banking institutions. As money launderers try to stay undetected with the aid of converting their technique, retaining one step in advance of law enforcement, international businesses and governments are running together to discover a new perspective to locate them.

The government has become increasingly vigilant in its efforts to confront money laundering through the years through passing anti-money-laundering rules. These regulations need financial institutions to have structures in location to discover and document suspected money-laundering activities. In 1989, the organization of 7 (G-7) formed a global committee referred to as the financial movement assignment force (FATF) in a try to fight cash laundering on an international scale. Inside the early 2000s, its purview turned into accelerated to tackle the financing of terrorism. The US exceeded the Banking Security Act in 1970, requiring monetary establishments to file some transactions to the branch of the Treasury, like cash transactions above $10,000 or any transactions they consider suspicious, on a SAR (Suspicious Activity Report). FCEN (Financial Crimes Enforcement Network), used data provide by banks to the treasury department and forward this information to foreign financial intelligence, criminal investigators. Until 1986, the US didn't declare money laundering practice illegal, though after passing the Money Laundering Control Act, these laws are helpful in tracking financial transactions.

Today there's a lot of statistics on global tendencies in money laundering, economic crime, and terrorism financing and lots of work is completed in an attempt to produce precise approximate of terrorism financing flows and money laundering, but, even though a numerous of in large part various estimates have been provided, but no one can be indisputably tested. Also, the quantitative problems which have been arising by means of anti-cash laundering and the combat towards terrorism financing have yet to be definitively replied (Biagioli, 2008) and no extensively accepted dimension method has yet been evolved (Fleming, 2009).

Quantifying terrorism financing and money laundering is a very important and profitable exercising, however, finding and growing uniform strategies and strategies for speedy and without problems detailing, categorizing and sharing new terrorism financing and money laundering techniques and behaviors with the broader worldwide AML/CTF network is similarly, if no longer more critical, especially whilst the structures, techniques, and methods employed by way of adversaries exchange rapidly and are getting more complex (Nardo, 2006).

In India also the government bodies like Securities and Exchange Board and RBI Reserve Securities have listed out diverse tips to the economic institutions. All the banks gather the list of transactions which isn't according with the RBI and then put up it to financial investigation Unit (FIU) for further investigation. The FIU recognize the money laundering technique from the statistical data received from numerous banks. This system is turning into more complicated because the matter of doubtful transactions is increased appreciably and the policies imposed with the aid of RBI solo aren't sufficient to monitor this criminal activity.

This research paper dedicates section 2 for analysis of past work as literature survey, section 3 deeply elaborates the proposed technique and whereas section 4 evaluates the performance of the system and finally section 5 concludes the paper with traces of future enhancement.

LITERATURE SURVEY

This section of the literature survey eventually reveals some facts based on thoughtful analysis of many authors work as follows.

G. Krishnapriya and Dr. M. Prabakaran [1], analyze various methodologies to recognize money laundering crime. They identify that all methods have scalable in accuracy and efficiency. They proposed a Time deviant behavioral model based money laundering recognition framework which generates transactional behavior patterns according to the different time window. Based on generated patterns the recognition of money laundering is performed. Their technique produced higher efficient results and with accurate findings and with less time complexity.

Xingrong Luo [2], proposed an algorithm depend on classification to effectively determine the suspicious transaction in the account. They consider the all financial transaction as a data stream and form classifier which depends on a mined frequent rules set. The result of their experiment on simulated dataset proved the effectiveness of their method.

Reza Soltani, Uyen Trang Nguyen, Yang Yang, Mohammad Faghani, Alaa Yagoub and Aijun An [3], proposed an MLD framework to search out the money laundering groups between a huge number of the financial transaction. To enhance the efficiency of the framework, case reduction methods such as matching transaction detection and balance score filter are used to narrow down the list of potential ML accounts. Next, by taking benefit of structural similarity, they can identify possible group money laundering accounts. Their preliminary experimental outcome shows a high degree of correctness in the detection of ML accounts.

Anu and Dr. Rajan Vohra [4], proposed a technique to search out the illegal transaction in the banking sector. The outcomes received for the given problem helps in identification of a chain of activities that contribute to the occurrence of any kind of financial crime. Numbers of the transaction which are prone to major bank fraud. This problem primarily identifies the different networks which contain the suspicious and non-suspicious transactions into cluster 0 and cluster 1 respectively. The final result of this problem includes the total no transactions in cluster 0 are 447 i.e. these are the transactions which are suspicious.

A total of 71% contributes to the suspicious activity occurring for a particular instance of database In cluster 1 there are the total of 181 transactions which are identified as non-suspicious as these don't fulfill the criteria of the transaction being suspicious.29% of the whole amount contribute to the non-suspicion category of the transactions.

Angela Samantha Maitland Irwin and Kim-Kwang Raymond Choo [5] describe the use of modeling to give an optical representation of important features and easy to follow the side of money laundering pattern extracted from actual money laundering and terrorist financing typologies.

Rafa Dreewski, Jan Sepielak and Wojciech Filipkowski [6], present a structure which favors money laundering identification. In this structure, bank statements importer was executed. Clustering algorithm used imported data that is used further for examination of money transactions. The generated clusters are in graph form. The standards account by the algorithm are money transaction condition, whole amounts of money into individual account and how much commission acquired by the entities present in the money laundering procedure. They implemented a total of six algorithms that mine frequent sets and that provide all frequent patterns.

Quratulain Rajput, Nida Sadaf Khan, Asma Lari1 & Sajjad Haider [7], proposed ontology depended on the structure to find the suspicious transactions. They used the ontology to develop the structure more systematic and it needs less calculation. The ontology-based system also reuses the information base in distinct applications in the same domain. The proposed ontology contains the transactions information and rules written in SWRL. They utilize pellet reasoner to conclude new knowledge for financial transaction nature. They test their system on a real-time dataset based on a commercial bank. The results show that the system is capable of suggesting transactions that can be further analyzed by the head of the compliance department to label transactions as suspicious or not.

Mahesh Kharote [8], proposed a system that generates an alarm or notifies for money laundering within the time before the actual money laundering has taken place. The proposed system contains seven modules: Preprocessing, Data Importer, Clustering and Suspected sequence/sets, Data visualization, learning from the user decision, company/organization profile generation, extracting behavior. Preprocessing makes sure the data provided to the system is always the correct one and reduced as possible. The less the data the faster are the operations performed. At the output of the clustering module along with graph structure which will be having all the records involving huge amounts of money transfer, the rest further process is to learn exactly which account has been indulging in money laundering practices. To cluster the records K-means algorithm is sufficient. And to extract a pattern from records and learns user decision pattern, Frequent Pattern Mining and the Association Rules would be used. A minimum support threshold for the transfers from an account or by individual over several accounts is predefined and then measures the deviation from the normal profile. The software intervention for the analysis purpose will provide more accuracy and save more time. Finally, customer behavior is extracted by combining association rule mining and clustering.

Arafat Al-Dhaqm, Shukor Abd Razak, Siti Hajar Othman, Asri Nagdi and Abdulalem Ali [9], developed a generic model specific for forensic inspection process of the database, known as the DBFIPM. To construct the DBFIPM, numerous existing investigation process models related to the database is reviewed. From a thorough investigation against these models, the DBFIPM reveals that the database forensic inspection process has 5 common process phases which include the: identification, collection, preservation, analysis phase, and presentation phase. To validate the wholeness of the DBFIPM model, the FBS technique is applied against the model. The future works of this research are to detail out all idea and relationships in each of the identified phases (in DBFIPM) by adopting a software engineering approach term as a metamodel.

Nhien An Le Khac, Sammer Markos, M. O'Neill, A. Brabazon and M-Tahar Kechadi [10], present a survey of utilizing DM techniques for AML. Firstly they find out the problems of DM in finance and banking and noticed the challenges faced on using the DM to investigate ML. They also analyzed some current DM methods that have been present and executed for AML.

Nhien-An-Le-Khac and Tahar Kechadi [11], present a study of knowledge depended on approach to analyze and find a suspicious case of ML in investment bank transactional data. To give an effective solution for AML, multi-methods like a neural network, clustering, genetics algorithm, heuristic are utilized together. Important information recovers from their knowledge are those by choosing appropriate parameters, general knowledge-based techniques can be utilized to detect ML cases within the investment activities. Besides, the combination of main and subordinate parameters helps to improve the learning process. Form their outcome they conclude that their method is more promising and satisfies the AML unit needs.

Nhien An Le Khac, Sammer Markos and M-Tahar Kechadi [12], present a survey report on methodologies used to recognize financial fraud. They also categorize the behavior pattern of fraudulent, find out the major sources and attributes of the data depending on which fraud detection has been organized.

Liu Keyan and Yu Tingting [13], proposed cross-validation method to find the optimal framework when the overall execution of the model is the best, they can successfully avoid the state of over-learning and low learning, and then finally get better classification results of test sets. Experimental outcomes prove that the model obtained by training SVM using the selection parameters by cross-validation technique is more successful than the model gain by SVM using the randomly selected attributes on the classification effect. It not only strengthens the classifier performance, but also greatly enhances the detection rate of suspicious transactions.

III PROPOSED METHODOLOGY

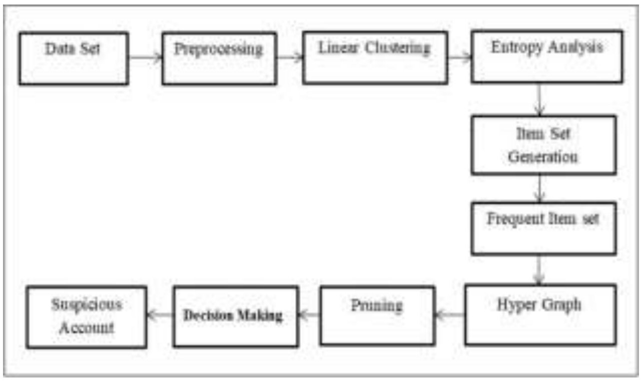
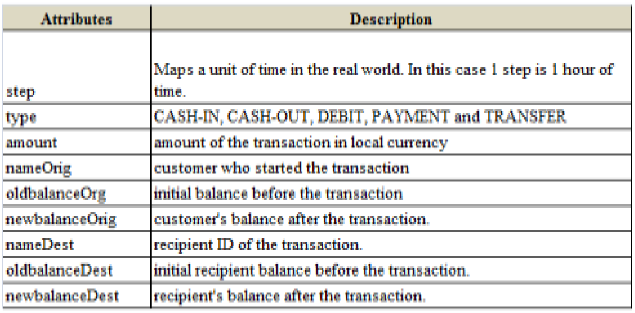


Figure 1: Overview of the proposed methodology

The proposed methodology of suspicious account detection in money laundering is depicted in figure 1 and it is explained in the below mentioned steps.

Step 1- Dataset selection and Preprocessing - This is the initial step of the proposed model where a synthetic Financial Dataset For Fraud Detection is downloaded from the following URL https://www.kaggle.com/ntnu-testimon/paysim1. The dataset is stored in a workbook which contain some attributes as mentioned below in Table 1.

Table 1: Attribute Description



Attributes	Description
step	Maps a unit of time in the real world. In this case 1 step is 1 hour of time.
type	CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER
amount	amount of the transaction in local currency
nameOrig	customer who started the transaction
oldbalanceOrg	initial balance before the transaction
newbalanceOrig	customer's balance after the transaction.
nameDest	recipient ID of the transaction.
oldbalanceDest	initial recipient balance before the transaction.
newbalanceDest	recipient's balance after the transaction.

In the process of preprocessing only required attributes are selected to form a list. Here in this process " step " attribute is skipped to select all other attributes for the further use and this list is called as preprocessed list.

Step 2: Linear Clustering and Entropy Analysis - Here in this step cluster are formed based on the attribute called " type ". All the data in the column of "type" attributes is collected in a list to estimate the hash set of the list to remove all the duplicate data to call this as the unique list.

Then a unique set list is searched for the specific column data to match with the attribute values. This forms a linear cluster, which is detailed in the algorithm 1.

Algorithm 1: Linear Cluster Formation

```
// Input : Unique Type list $U_{L}$, $P_L$ [ Preprocessed List ]
// Output : Linear Clusters $L_C$
Function : linearClusterFormation($U_L$)
Step 0: Start
Step 1: $L_C = \emptyset$
Step 2: for i=0 to size of $U_L$
    Step 3: $S_G = \emptyset$ [ Single Cluster]
    Step 4: for j=0 to size of $P_L$
        Step 5: ROW=$P_{LJ}$
        Step 6: IF $U_{Li}$ = $ROW_K$ ,THEN
            Step 7: $S_{G\,=}\, S_{G\,+}$ ROW
    Step 8: End for
    Step 9: $L_C = L_{C+}S_G$
Step 10: End for
Step 11: return $L_C$
Step 12: Stop
```

The formed clusters are subject to estimate the entropy or distribution factors based on the four attributes namely, amount, Oldbalanceorg, oldbalanceDest and newbalanceDest. A count is being calculated for a single cluster based on some rules like if the amount is equal to Oldbalanceorg. And oldbalanceDest is equal to zero, newbalanceDest is equal to zero. Based on this a count is estimated to calculate Entropy or distribution to determine the most important cluster.

This entropy is measured using the Shannon information gain theory, Which yields the values in between the 0 to 1. So values nearer to 1 indicates the row from the respective cluster is important. So this step assigns a threshold of more than the or equal to 0.5 to select the row for the further step, to store them in the information gain list. This entropy can be analyzed using the following equation of 1.

$$IG = -\frac{A}{C}\log\frac{A}{C} \; -\frac{B}{C}\log\frac{B}{C} \;\_\_\_\_(1)$$

Where

A= Count

C= Cluster Elements Size.

B= A-B

IG = Information Gain of the cluster

Step 3 : Frequent itemset mining- Here in this step the information gain list is used to extract the "type" attribute values like Payment, Cash in, cash out, Transfer and Debit. These values are used to form the frequent itemsets using the power set. Where a support is estimated to identify the important pattern of transactions. This pattern is identified by counting the pattern types and sort them in descending order to identify the pattern of transaction more efficiently.

Step 4 : Hyper graph , Pruning and Decision making - Once the pattern types are identified, then for money laundering activity these pattern types are mapped to form a graph in between the account numbers. While doing this the account number is considered as the nodes of the graph and protocols on which they are suspected is used as the edges in between them.

This list of nodes and edges are used to form the graph , which is stored in the neo4j advance database system for the graphs. The graph is pruned to estimate the weight of the nodes to filter the most suspicious account, which are decided through decision tree protocols.

IV RESULT AND DISCUSSIONS

The proposed model of suspicious account detection in money laundering activity is developed in Windows based machine. Which is powered by Pentium Core i5 Processor and 6GB of primary memory. To develop and deploy the system Java Language is being used and Netbeans is used as the IDE along with the Neo4j Graph database.

To measure the effectiveness of the proposed model some experiment is conducted to measure the error in percentage of precision. This error in percentage of precision is measured using a parameter called RMSE (Root mean Square Error).

RMSE is difference in error rate between the two continues correlated entities. In this experiment those entities are Actual suspicious accounts and Detected Suspicious accounts. The RMSE can be measured using the following equation 2.

$$\text{RMSE}_{fo} = [\sum_{i=1}^{N}(z_{f_i} - z_{o_i})^2/N]^{1/2} \quad \_(2)$$

Where

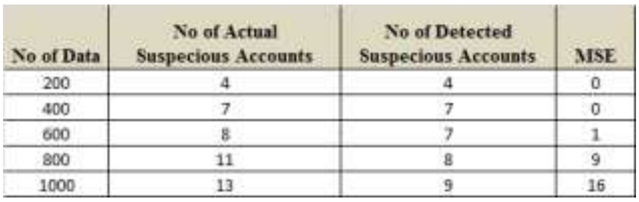
$\sum$ - Summation

$(Zfi - Z_{oi})^2$ - Differences Squared for the summation in between the Actual Suspicious account and Detected Suspicious Account

N - Number of samples or Trails

Some experiment is conducted to measure the RMSE, the values are recorded in the given table 2.

Table 2: MSE Measurement



No of Data	No of Actual Suspecious Accounts	No of Detected Suspecious Accounts	MSE
200	4	4	0
400	7	7	0
600	8	7	1
800	11	8	9
1000	13	9	16

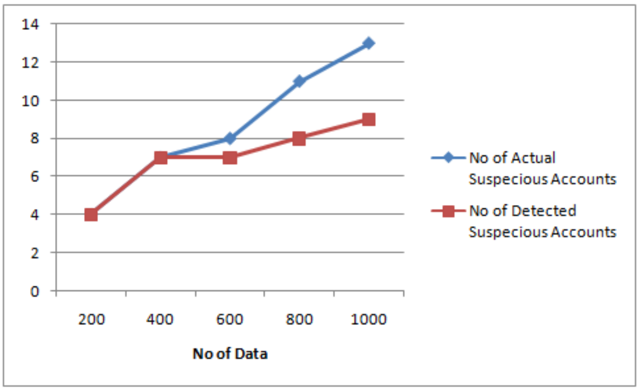


Figure 2: No. of Actual Suspicious Accounts V/s No. of Detected Suspicious accounts

On observing the table 2, it clearly yields the mean MSE (Mean Square Error) of about 5.2. And the RMSE of about 2.28. The obtained RMSE is too low and it is a good sign about the performance of the proposed model.

V CONCLUSION AND FUTURE SCOPE

As we know the money laundering is the illegal activity that is carried out very carefully and smartly through the complex banking activities. Due to the huge size of the bank transaction data it is quite complex and difficult task to identify the suspicious account that are involved in this crime. Proposed model uses the Information gain theory on the linear clustered data to estimate the probability of finding the suspicious account number. Utilization of frequent pattern helps to evaluate the pattern of fraud which leads to form a hyper graph.

This graph eventually helps to manage the links between the suspicious accounts. The pruning and decision making protocols finally reveal the suspicious account numbers from the dataset. The Experimental analysis indicates the proposed model delivers around 2.28 of RMSE, that eventually shows the effectiveness of the proposed model.

In the future this model can be implemented in real time bank servers in distributed paradigm.

REFERENCES

[1] G. Krishna Priya, Dr. M. Prabakaran, “Money laundering analysis based on Time variant Behavioral transaction patterns using Data mining”, Journal of Theoretical and Applied Information Technology 2014.

[2] Xingrong Luo, “Suspicious transaction detection for Anti Money Laundering”, International Journal of Security and Its Applications 2014

[3] J Reza Soltani, Uyen Trang Nguyen, Yang Yang, Mohammad Faghani, Alaa Yagoub, and Aijun An, “A New Algorithm for Money Laundering Detection Based on Structural Similarity”, DOI: 978-1-5090-1496-5/16, IEEE, 2016.

[4] Anu and Dr. Rajan Vohra, “Identifying Suspicious Transactions in Financial Intelligence Services,” IJCSMS, Vol. 14, Issue 07, July 2014.

[5] Angela Samantha Maitland Irwin and Kim-Kwang Raymond Choo, “Modelling of money laundering and terrorism financing typologies”, Journal of Money Laundering Control, DOI 10.1108/13685201211238061, Vol. 15 No. 3, 2012, pp. 316-335.

[6] Rafa Dreewski, Jan Sepielak, and Wojciech Filipkowski, "System Supporting Money Laundering Detection", Research Gate, DOI: 10.1016/j.diin.2012.04.003, May 2012.

[7] Quratulain Rajput, Nida Sadaf Khan, Asma Lari1 & Sajjad Haider, "Ontology Based Expert-System for Suspicious Transactions Detection", Canadian Center of Science and Education, ISSN 1913-8989, Computer and Information Science; Vol. 7, No. 1; 2014.

[8] Mahesh Kharote, V. P. Kshirsagar, “Data Mining Model for Money Laundering Detection in Financial Domain", International Journal of Computer Applications (0975 – 8887), Volume 85 – No 16, 2014.

.

[9] Arafat Al-Dhaqm, Shukor Abd Razak, Siti Hajar Othman, Asri Nagdi and Abdulalem Ali, “ A Generic Database Forensic Investigation Process Model”, Research Gate, DOI: 10.11113/jt.v78.9190, eISSN 2180–3722, 1 Aug 2015.

[10] Nhien An Le Khac, Sammer Markos, M. O'Neill, A. Brabazon, and M-Tahar Kechadi, "An investigation into Data Mining approaches for Anti Money Laundering”, IPCSIT vol.2,2011.

[11] Nhien-An-Le-Khac and Tahar Kechadi, “Apply data mining and natural computing in detecting suspicious cases of money laundering in an investment bank: A case study”, Research Gate, DOI: 10.1109/ICDMW.2010.66, December 2010.

[12] Pankaj Richhariya and Prashant K Singh, “A Survey on Financial Fraud Detection Methodologies”, IJCA, Volume 45– No.22, May 2012.

[13] Liu Keyan and Yu Tingting, “An Improved Support-Vector Network Model for Anti-Money Laundering”, DOI 10.1109/ICMeCG.2011.50, IEEE,2011.