Detection of money Laundering transactions using outlier detection

Nikee K Katre
Student
Dept. of computer science & Engineering
G.H Raisoni College of Engineering, Nagpur

Prof. Snehlata Dongre
Asst. Professor
Dept. of Computer Science & Engineering
G.H Raisoni College of Engineering, Nagpur

Mandar Deo
Director, Technical
Trust systems and software, Nagpur

***Abstract*—Money laundering is increasingly becoming a problem for banking systems throughout the world. The purpose of such a transaction is to evade taxes by masking the source and destination of funds. In view of this definition ,this study attempts to study and detect money laundering transactions. A simple but effective approach has been adopted to detect suspicious transactions, namely behavioral pattern analysis. The accounts exhibiting behavior different from others are identified as suspicious, these suspicious transactions are then investigated further to ascertain their involvement in money laundering operations.**

***Index Terms*—Behavioral pattern analysis.K-Means,Z-score,**

I.INTRODUCTION

Money laundering transactions are detected using a plethora of approaches however the particular solution to be applied depends upon the type of data .Many studies focus on machine learning,neural networks and artificial intelligence to effectively identify and flag suspicious transactions. Behavioral pattern analysis can be more discreet and practical to handle for a large database. Outliers have been effectively used to study trends regarding behavior pattern changes and money laundering transactions. However the need of more simple and less computationally extensive approaches is often felt, while the role of detection models is limited to reducing the workload of financial intelligence units.Some models though effective for one sphere of detection are completely out of context for another, meaning the approach to detection changes with data type and application in particular. Actionable leads need to be generated by the model by filtering the dataset. The time required for exact detection is particularly large in some models, this leads to severe lapses in the overall effect.The detection based on trends within the dataset are less dependent and suitable incorporation into machine learning applications may prove to be of greater utility. The study emphasizes on detection of money laundering transactions using behavioral patterns. An algorithm is devised to separate and detect accounts different from other accounts. These different accounts are termed as suspicious accounts.

The paper is divided into methodology, followed by a brief discussion about the algorithm, this is succeeded by implementation .Lastly results and efficiency along with the advantages of the model are elaborated.

II.RELATED WORK

The different solutions studied give an insight into the advancements in the field of data analytics and the allied domains which help in realising its true potential. The methods focussing on behavioural patterns are more feasible to apply. However those which are based on classifiers, neural networks require greater computational effort. The availability of sufficient training data may limit implementation of a variety of methods, without further avenues to create simulations and synthetic data, performance may be compromised. Therefore solution delivering optimum performance with least amount of computational costs and deployment challenges are most desirable.

The methods of detecting money laundering transactions discussed mostly comprise of machine learning and neural networks, the emerging arenas of research are the study and advancements of these two domains. While some methods detect behavioural patterns others utilise data mining from multiple sources to generate profiles of suspicious sources. The solutions studied on banking system use transaction records whereas other models rely on multiple sources for data collection. Some authors have studied the more practical side of data and its application along with integration. Whereas a fewtry to understandthe merits of centralisation. The methods involving complex models focus on focus on the sheer capacity of data handling. Certain methods try to improve accuracy of conventional models. The development of more complex and multi-dimensional approaches to the same problem is quite positive, this can ensure a more easeful integration and data-sharing among different research groups. Approach towards a particular money laundering pattern may differ in a large way if the data type varies.Several studies suggest advanced versions of classifiers. Improvements in data processing techniques have also played a part in reducing the complexity of detection algorithm in certain cases.The more advanced methods discuss higher data mining and tax record studies, for them a change in tax filing suggests a particular account is involved in money laundering. In part specific trends are extremely difficult to replicate in a model working on real-world data,some studies also face limitations in the form of non-availability of suitable data.

III.METHODOLOGY

The methodology followed in the study aims at reducing the time required and the human interaction with the process ,however the results are simpler to understand and provide an actionable knowledge base for further investigation of suspicious accounts .Not only this saves time but also improves the accuracy of financial detection models to detect trends large amounts of banking data. The ultimate goal of the study is to study behavioural patterns of the accounts in general and to draw inferences based on a deviation to this pattern by certain accounts. The accounts are termed as suspicious through this algorithm,they can differ from surrounding data in terms of activity,volume of transactions or the sheer number of transactions. These accounts facilitate an alternate route to discretise large transaction into smaller ones, thus confusing the traditional still rule based detection algorithms. The behavioral pattern analysis minimises these errors and provides specific accounts to investigate further.

Dataset

The dataset is a typical banking transaction log ,comprising of time, amount, origin account, destination account ,along with opening and closing balances of accounts.The data reflects the banking habits of real time customers. The origin and destination of all transactions are routine banking operations. Being in a high volume they are not scanned before initiation for known laundering traits.

Clustering

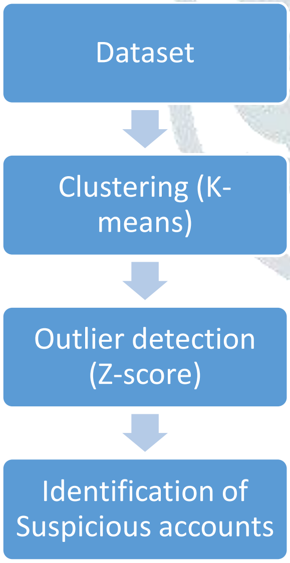
The transactions must be grouped together to understand the hidden underlying trends. Clustering will group the data into smaller but similar groups, these will denote the patterns more regular among a particular set of accounts.

Outlier detection

The data-points which differ from surrounding data areidentified using mathematical methods. These outliers signify the deviation of behaviour from regular or routine ,this may be an indication of money laundering activity.

Suspicious accounts

The accounts identified as outliers are regarded as suspicious, they differ significantly from other accounts in the database. They are investigated further to ascertain their role in money laundering operations .Sometimes entire groups involved in money laundering can be uncovered while studying these trends.


Dataset
Clustering (K-means)
Outlier detection (Z-score)
Identification of Suspicious accounts

III. ALGORITHM

The experiment is developed on a scale which is simpler to study and the data attributes are more realistic in unison with the needs of the actual problem. The data is taken from a real data set comprising of nearly 1,00,000 accounts having the transaction time frame ranging from 1 to 10 hrs.

The model consists of a K-means algorithm in the first step,K-means is an unsupervised machine learning clustering algorithm.It uses *Euclidian distance* as a means to calculate the distance between the centroids of clusters and the various datapoints. Clusters are used to group similar data in order to get a clear understanding about the range and orientation of data.

The *K-means* algorithm can be summarized as follows:

1)The K points (number of clusters) are initialized randomly, and the points known as centroid are defined.

2)The Euclidian distance of each point is calculated and the point closest to a particular centroid is assigned to that cluster.

3)The mean previously calculated ,is recalculated to find a new centroid.

4) 2&3 are repeated until no change is observed in the centroid and the clusters are stable.

The number of clusters needs to be defined prior to initialization of algorithm ,the number of clusters must be so that the intra-cluster variance must be minimum ,while keeping the model viable for near accurate outlier detection.Defining too many clusters can result in a hypothetical situation where each datapoint forms its own cluster.

A method known as the elbow curve method is used to find out the optimum number of clusters,the plot is a curve of variance of clustered data vs number of clusters. With every increase in number of cluster the variance of data changes, ultimately a point is reached at which no change occurs in the variance for an increase in number of clusters. This point gives us the optimum number of cluster *K*.

The relation used to calculate Euclidian distance is,

$$\mathbf{D=\sqrt{}(\sum (X_i - Y_i)^2 \text{ , where } i= 1 \text{ to } n}$$

D-Euclidian Distance

X- co-ordinate of centroid

Y- co-ordinate of datapoint

The data-point may have more than attribute, other co-ordinates can be similarly subtracted to find distances. Here in this particular example $\mathbf{X (X_1,X_2….X_n)}$ the centroid &$\mathbf{Y(Y_1,Y_2….Y_i)}$ is a typical datapoint.

The attributes considered for clustering of data are as follows:

Amount sent: The amount sent from source to the destination, sending account can send funds in multiple forms,the dataset has entries regarding each transaction ,the preliminary variability trends can be studied from the data itself.

Amount received: The change in balance or amount,final closing balance is considered to cross check.

Time difference/Frequency:

The time stamps attached to the transaction are considered to calculate frequency.

Transaction type:The type of transaction also conceptualizes the trends within the data, clusters are formed keeping in view the various routines that accounts follow to circulate funds.

The clusters are a preliminary step towards grouping of various transaction similar in nature,however the grouping is not exhaustiveas it only provides an approximate idea about various behavioral patterns within the dataset. Each centroid represents a group of data similar in attributes, this similarity underlines the behavior similarity of the account holders. The grouped data can be studied to form a general consensus about the prevailing trends in the data.

The method known as *Z-score* is used to detect local outliers within the dataset. Outliers are data-points which differ so much from the surroundings data-points so as to arouse suspicion. These apparent suspicious transactions are identified based on their Z-scores calculated as follows,

$$Z=(x-\mu)/\sigma$$

Z=Z-score for a typical datapoint
x =value of attribute for which score is being calculated
σ = Standard deviation of dataset
μ = Mean of dataset

Z-score method assumes that the data follows a normal distribution curve and the data-points lying outside the curve are considered as outliers. Outliers can differ in various ways from the surrounding data ,either they be different in magnitude or simply out of normal range for a particular account. Much study has gone into defining outliers, in this particular study they denote behavior pattern foran account which is different .Suspicion may arise based on this variance and prompt the detection mechanism to flag an alert.

Theorem of three standard deviation is used wherein the point is termed as an outlier if,-3>Z>3.The outliers are mathematically local outliers as Z-score is being calculated for each data-point within the clusters.It should be noted that the data is having more than one attribute and the score is calculated taking average of all columns in the row. The outliers are those accounts which exhibit behavior different than ordinary or normal, this identification lists them under suspicious accounts. Suspicious accounts signify the trend useful to detect money laundering transactions.

Suspicious transactions are then investigated further to ascertain their involvement,the detection is purely of indicative in nature.It acts to minimize the role of human interference and automate the process further.

Training of model

The model is trained through training data,this is an extremely important step for definig the threshols of a particular datatype on which the system is desired to work.In this particular study, dataset is partitioned into training as well as experimental data. While the presence of an entirely different different dataset is appreciated to train the model,it is difficult to come across one due to secrecy in banking laws.

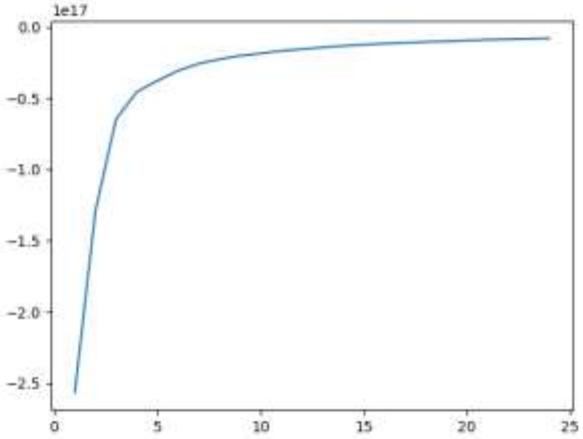
Hence,the system is trained on 90 percent of the dataset ,the remaining dataset is test data..

Training solves the most common problem with the model that is of invalid or incomprehendableresults,the need to preprocess large chunks of data also becomes less necessary.The model can be accustomed to the average normal behavior patterns for account holdersIn that way it can more accurately determine which accounts exhibit abnormal behavior patterns and term them as *suspicious.*

IV.IMPLEMENTATION

The project is implemented in *Python* on the *pyQT* platform, it was preferred due to the ease of availability of Machine learning libraries.These libraries make it less intensive to use advanced algorithms wthin relatively simpler codes.

Python allows for better adaptability and simpler way to incorporate changes into the code.The versatility of the platform makes is easy to reduce the size of code. The process to train the dataset is done by initiating the K-means algorithm for a random number of clusters. The efficiency of experiment depends upon the number of clusters obtaining just the right amount of variance to maintain the viability of outlier detection mechanism. A well-known method known as the*Elbow Curve* method is used to decide the number ofclusters.



The variation within the clustered data is unchanged after the number of clusters are increased beyond 20.Hence,number of clusters are taken as 20.

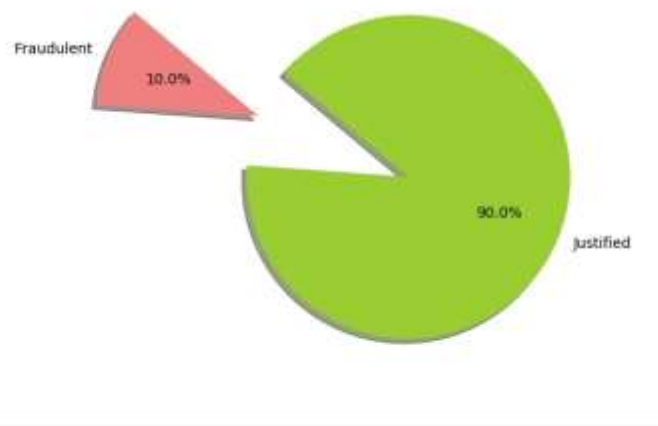
Besides elbow curve ,dendogram may also be used to predict the suitable number of clusters in a suitable way, instead of calculating variance directly it looks for the relative size of obtained clusters,this method is of little use for methods where the number of clusters are specified by user. Dendogram may be studied for a dataset to examine the validity of current distribution.

Futher step involves detection of outliers,this process can be further automated if the need for processing large amounts of data arises.The process of reading the obtained results can be further automated to generate reports and even rank them according to the severity.

V.RESULTS

The model performs fairly well in detecting suspicious transactions with a fairly high accuracy as compared to other similar models, the transactions identified have a fairly higher variation in behavior relative to other accounts.

For the given dataset the number of outliers are found to be about 10 percent ,this value is in close agreement with other models .The transactions comprise of accounts which have out of ordinary or vastly different attributes from other accounts. This different behavior may be due to co-ordination of payment cycles for other destination accounts,or the lag effect while acquiring new assets. The main objective of such transactions is to mask actual source or destination account and evade taxation.



The above figure indicates the percentage of suspicious accountsagainst normal accounts. The accounts identified by the model can be termed as suspicious a detailed list mentioning each such particular account highlights them as well as their particular score in the index. Experts can very feasibly segregate and investigate these accounts further.

VI.DISCUSSION

The experiment very well establishes the fact that behavior pattern analysis is an efficient method to detect suspicious accounts within a banking system. The study highlights the following key advantages of behavior pattern analysis:

I) Being discrete the violating corporations or account holders will not be able to change their patterns often. The trends evident can also be applied to other accounts.The model uses routine transaction data minimizing the need to prepare data for pre-processing.

II) The results displayed are more direct to understand and can be easily analyzed with limited knowledge of system .This is in stark contrast to many existing models.

III) It is simpler to apply on a wide variety of datasets, due to no requirements of pre-processing or preparation of data.This enhances the viability of proposed solution to wide spectrum problems.

IV) Financial intelligence units can act decisively and large money laundering groups can be detected through studying of trends on related data through multiple banks.

V) The advanced or complicated trends that need to be replicated on relatively unrelated datasets for some methods may be ignored as the model is based on more concrete account activity parameters.

VI) The need for human supervision is negligible if the model is implemented with an increased automation.

VII.CONCLUSION

The main aim of the study has been duly achieved through identification of suspicious transactions within traditional banking system. The results given by the model are simpler to understand and offer an insight in the behavior pattern trends within the dataset. Such trends can be studied to infer more specific strategies for detection. Through advancements in data mining techniques more data can be collected towards building a precise profile of every account holder. These developments coupled with automation can further ease the demand of human interference to detection models.

VIII.REFERENCES

[1] Rui Liu, Xiao- Long Xian, Shu Mao &Shuai-zhengzhu, Research on Anti-money laundering based on core decision tree algorithm, Chinese control and decision conference (2011), IEEE.

[2]Shu Mao, Rui Liu , Dancheng Li, Shuaizen Zhu , Antimoney -Laundering software based on mainframe and SOA , International conference on computational Intelligence and computer network(2013), IEEE.

[3]Umadevi P, DivyaE,Money laundering detection using TFA system,2012, IEEE.

[4]ZenganGao , Application of Cluster-Based Local Outlier Factor Algorithm in Anti-Money Laundering,IEEE transactions ,2009

[5]Nhien An Le Khac, M-TaharKechadi ,Application of Data Mining for Anti-Money Laundering Detection: A Case Study, IEEE International Conference on Data Mining Workshops,2010

[6]LIN-TAO LV, NA JI, JIU-Long Zhang,A RBF NEURAL NETWORK MODEL FOR ANTI-MONEY LAUNDERING, International Conference on Wavelet Analysis and Pattern Recognition,IEEE,2008.

[7]Liu Keyan, Yu Tingting,An Improved Support-Vector Network Model for Anti-Money Laundering, International Conference on Management of e-Commerce and e-Government,IEEE,2011

[8]Xurui Li, Xiang Cao, XuetaoQiu, Jintao Zhao, JianbinZheng,Intelligent Anti-Money Laundering Solution Based upon Novel Community Detection in Massive Transaction Networks on Spark, IEEE, 2017.

[9]Ebberth L Paula, Marcelo Ladeira, Rommel N. Carvalho and ThiagoMarzag˜ao,Deep Learning Anomaly Detection as Support Fraud Investigation in Brazilian Exports and Anti-Money Laundering,IEEE 2016.

[10]Reza Soltani, UyenTrang Nguyen, Yang Yang, Mohammad Faghani, AlaaYagoub, Aijun An,A New Algorithm for Money Laundering Detection Based on Structural Similarity,IEEE 2016.