2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON)
Galgotias University, Greater Noida, UP, India. Oct 2-4, 2020

Anti Money Laundering detection using Naïve Bayes Classifier

Ashwini Kumar
Dept. of Computer Sc and Engg
Graphic Era Deemeed to be University
Dehradun, India
ashwinipaul@gmail.com

Sanjoy Das
Department of Computer Sc.
Indira Gandhi National Tribal
University-RCM
India
sdas.jnu@gmail.com

Vishu Tyagi
Dept. of Computer Sc and Engg
Graphic Era Deemeed to be University
Dehradun, India
tyagi.vishi@gmail.com

***Abstract*—Anti Money laundering is a very challenging issue in the banking system. Anti-money laundering is defined as a set of procedures, policies and ordinances designed to prevent the creating income during illegal actions, e.g. market operations, deal of illegal commodities, and corruption of public funds and tax evasion. Our objective of the paper is to classify a transaction as illegal or not. To achieve this, we have used big data analytics technique for a dataset to identify the money laundering activities. We have used a customized dataset with 10000 transactions used for analysis with data cleaning, statistical analysis, and data mining process. The logical operator and if-else conditions are used to analyze the relationships among various attributes in the dataset prior to analyze with big data analytics methods. Finally, we have used the Naïve Bayes classifier to find money laundering activities. The analysis has been done using R software and customized dataset. The result obtained through analysis is very significant and proposed model achieved accuracy is 0.8125.**

***Index Terms*—big data analytics, naïve bayes, decision making, Probabilities, anti money laundering**

I. Introduction

Now a days, the Money laundering has been spread like a spider net and affecting the global economy for few decades. During Money laundering huge amount of money are used to make illegally acquired funds and convert them legal and legitimate associated with criminal activities [1,2]. Banking institutions are compulsory to invest a lot in Anti-Money Laundering (AML) fulfillment today. The Anti-money laundering is one of the crucial task for many countries. Generally, money launderers divide the dirty money into multiple parts, and legalize through multiple small banking transfers or commercial transactions. So, this is very difficult to manually detect activities of money laundering is raise a challenging task [2].

There is no institution supporting money laundering or financing to terrorist, criminal organizations for little gain. Money earned illegally through various channels such as drug dealing, stealing tax requirements to be cleaned. This is very badly affects the capital, savings and investments of any nations, in opposite helps in the financial support of criminal activities. The banking frauds includes various activities like online transaction, credit/debit card fraud, money laundering etc,. [21]. Money laundering detection is also known as Anti-Money Laundering [18,19]. This is also referred as an action which stops, or goals to stop money laundering from incident. The definition of ‘illegal income’ that referred as judicature few actions depends on country to country. The major goals of AML includes detecting planned offence, to decrease drug dealing, to prevent terror campaign or to continue the status of the monetary services business. Day by day AML rules have evolved and turn into further difficult, expensive, and hard to adhere.

Banking organizations are encountering the load of Anti money laundering agreement requirements and reporting [23]. For instance:

- The AML processes and systems must support Know Your Customer (KYC) actions, as well as transaction monitoring at a minimum.
- Bribery, dishonesty and tax avoidance is also mandatory as part of anti money laundering behavior.
- Suspicious activity reports (SARs) have to now be file within 60 days rather than 90 days [14].

The big data analytics present banks an understandable path to quick, valuable, and cost-efficient agreement that can scale and adapt as requirements change. It is synthetic data set, we recommend an approach based on classification to detect the illegal actions. The major contribution in the paper includes the analysis on synthetic data using naïve bayes classification method for identifying money laundering transactions.

The paper is organized as: Section-2 literature based on AML is discussed. Section-3 includes experimental setup and methodologies. The system model is discussed in Section-4. The result analysis is discussed in section-5 and conclusion in section-6.

II. Literature Survey

In this section, we have included various works focused on detection of money laundering using machine learning and other algorithms. There are two algorithms mainly used for credit card fraud detection are Multilayer Perceptron Artificial Neural Networks and cluster Analysis (Iterative Naïve Bayesian Inference Agglomerative Clustering algorithm) [20]. The K-means clustering technique is used widely for its significant results. The activities related to AML day by day come with more advance techniques. To identify AML

978-1-7281-5070-3/20/$31.00 ©2020 IEEE

activities research is going based on Expectation Maximization (EM) [13]. The authors in [13] uses EM for AML and come up with promising results.

In [1]various machine learning algorithms are categorized, and summarized by authors, which are used in detecting suspicious transactions. In this survey the following things are comprehensively discussed and analyzed are solutions of AML typologies, link analysis, user behavioral modeling, risk scoring, detection of various anomalies and geographic capability. The major aspects of data preparation, transformation, and data analytics techniques have been discussed. Through their analysis and various future research directions is highlighted. In [2] authors identifying suspicious money laundering in two-phase intelligent Method using machine learning and data analysis techniques. Initially, the model emphasizes on identifying every suspicious money laundering activities. The second phase segregate highly suspicious. The recall and precision value is considered for the identification of money laundering . The recall rate is 26.3 percent and 87.04 percent is precision rate. In [3]authors proposed a novel and open multi-agent architecture based on the following properties like autonomy, reactivity and pro-activity which are very essential to control money laundering prevention controls. In [4] authors discussed a case study on ML by using data mining and various natural computing techniques. In [5]authors proposed a cross validation method. This method is useful in finding the optimal parameters for SVM classifier to solve suspicious financial activities from transactions. The grid search achieves the highest classification accuracy rate.This method effectively ignore the over-learning and less learning. This technique significantly improve the performance of the classifier. In [6] determination rules associated with money laundering is derived by using a decision tree method in commercial bank of china. The dataset used for the analysis consists of total 28 customer's information with four attributes. The result shows effectiveness of decision tree. In [7] shows data mining approaches are better for detecting money laundering activities. An efficient data mining-based technique has been proposed. Experiment is done on real transaction datasets. In [8], authors proposed algorithm based on decision tree to identify money laundering activities.. Authors uses clustering algorithm using BIRCH and K-means combined together. Effective laundering patterns and money laundering rules are derived; this is Identifying abnormal transaction very effectively.

In [9] a capital flow hierarchical model is proposed for AML. An entropy-weight method is used to evaluate each account entity to draw their importance and experiment is done on simulated dataset. In [10] authors have introduced an anti-money laundering system for union-bank centre. The AML system uses dynamic data for proper analysis and provides support to the bank. This technique utilized various modern tools and methodologies. A few techniques are multi-agent neural network, text mining. Along with this genetic algorithms, velocity analysis and case-based reasoning, methodologies are used. This AML is very useful in finding more accurate money laundering activities. In [11], authors combined distributed mining techniques for distributed environment with the AML.

III. Experimental Setup and Methodology

In this part, we have discussed the methods that are used in the anti money laundering analysis by the Big Data Analytics method named text analytics algorithm. The dataset used is the synthesized bank transactions, R language is used to classify the activities of anti money laundering, and its probabilities. The step wise activities performed in the AML are shown in Figure 1.

R programming is the most powerful and popular platforms used for statistical programming, machine learning and visualization [24]. R is the open source and free [22] and popular because of its vast number of powerful algorithms and libraries called packages [15].

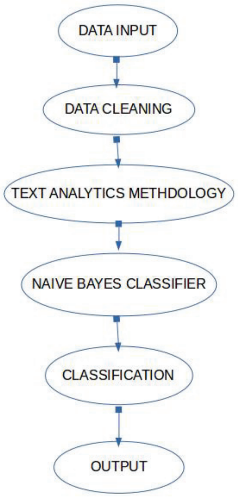

Fig. 1. Anti Money Laundering Activities

Initial phase, identify the input dataset required for our analysis. The second step is data cleaning to remove the noisy data from the dataset. In third step, big data analytics method, i.e. text analytics algorithm considers the naïve bayes classifier is applied on anti money laundering dataset to find legal and illegal transactions. Finally, data are classified and produce the desired output or result [20][22].

A. Text analytics methods

Mostly all information is accessible in text format in the databases. Manual Analytics is not possible for information extraction. Text analytics mention the process of obtaining significant knowledge from text data. It will apply to extract considerable data from the content. From all textual data it will get better knowledge. After recovering facts it will be classified. And from those classified facts we can take the business decision which helps to improve the business [12].

The following steps are involved in Text Analytics algorithm.

a) Text:The first stage, the data is in formless format.

b) Text processing: In the second stage, data is processed into information.

c) Text transformation: this stage refers that the essential text will mine for upcoming references.

d) Feature selection: this stage data is selected and show in numeric form.

e) Data mining: the entire data is confidential.

B. Naïve bayes classification

It is based on Bayes theorem. It is a supervised learning method and also statistical method for classification. Thomas Bayes(1702-1761), who suggested the Bayes Theorem. Naïve bayes provides pragmatic learning algorithms and initial knowledge and declared data can be joined. It offers a useful discernment to understanding and assessing many learning algorithms. Naive Bayes text classification is used as probabilistic classification [16]. This classification algorithm used for binary (two-class) and multi-class classification problems. [17].

M and N are two events for Bayes theorem relates two conditional probabilities as follows: P(M N) = P(M) P(M — N) is equivalent to P (N — M) = P (N) P (M — N) P (M).

In a classification model , we have some independent variables or functions and dependent or target variable. Each iteration has some values for the input(independent variable) and an output(class). In Naive Bayes, first we evaluate a probability for each output(class) based on independent value .We can use target variable having highest probability.

Suppose there are n independent variables, denoted by Z1 Z2, Z3. . . , Zn.. and the outcome variable y, coming from any one of k outputs(classes) denoted by C1,C2,C3,. . . ,Ck. Now suppose the independent values are given as follows:

Z1= z1, Z2 = z2, Z3 = z3, . . . , Zn = zn

Evaluate the probability of given observation from any of the K ouputs(classes), say Ck. We can write conditional probability notation in equation-1 as follows.

P(y = Ck, Z1= z1, Z2 = z2, Z3 = z3, . . . , Zn = zn) (1) Above Bayes formula, if we put in N = Ck and M= z1z2,. . . ,zn = z1, z2,. . . ,zn. The equation-(1) can rewritten as equation(2) as follows:

P(Ck — z1, z2, z3, . . . , zn) = P(Ck) P(z1, z2, ... , zn— Ck) P(z1, z2, ... , zn) (2)

Now the numerator in Equation(2) is the joint probability of Ck and z1, z2, . . . , zn that is

P(Ckz1, z2, . . . ,zn) = P(Ck, z1, z2,. . . .. , zn) (3)

Now, Bayes formula applied in the equation (3) respectively, and obtain the following results.

P(Ck, z1, z2, z3, ..., zn)=P(z1, z2, ..., zn, Ck)

=P(z2, z3,, ..., zn, Ck) P(y1 — z2, z3,, ..., zn, Ck)

=P(z3, z4, ..., zn, Ck) P(z2—z3, z4, ..., zn, Ck) P(z1—z2, z3,, ..., zn, Ck)

=P(z4, z5, ..., zn, Ck) P(z3, — z4, z5, ..., zn, Ck) P(z2 — z3, z4, ..., zn, Ck) P(z1 — z2, z3,, ..., zn, Ck)

=P(Ck) P(zn —Ck) P(zn1 — zn, Ck). . . P(z1 — z2, z3,, ..., zn, Ck) (4)

The Conditional Independence Assumption

The conditional independence assumption given a output or class, say Ck and the feature/independent values are independent of each other. Class and features having no correlation. event M and N are independent conditional on event O, then the following equation :

P(M — N, O) = P(M — O) P(M — N, O) = P(M — O) (5)

Now, we will apply equation-5 in our model. We assumed that all the predictors z1, z2, . . . , zn are independent conditioned on class Ck. Therefore,

P(z1 — z2, z3,, ... , zn, Ck) =P(z1 — Ck)

P(z2 — z3, z4, ... , zn, Ck)=P(z2 — Ck) (6)

and so on.

Now, we can write

P(z1, z2, ... , zn, Ck) = P(Ck) P(zn — Ck) P(zn1 — Ck) P(zn-2 — Ck). . . . . P(z1 — Ck)

P(z1, z2, ... , zn, Ck) = P(Ck) j = log P(zj — Ck) (7)

So, we have

P(Ck — z1, z2, ... , zn)=P(Ck) j =log P(zj — Ck)P(z1, z2, ... , zn) (8)

We have evaluate the numerator part for all values of k1,2,. . . ,K and selected for the one with the highest value.

The prior and the likelihood

The prior and the likelihood can be defined as P(Ck) j = log P(zj — Ck). The prior could be estimated as P (Ck) =n output(class) kn total. For the likelihood, we required the conditional probability distributions, f(Zj — Ck) for j=1,2,. . . ,n. If Zj is continuous,common assumption. If Zk is discrete, as a multinomial distribution.

IV. System Model

To devise a model to classify the transaction based on the given data set of attributes using naïve bayes classifier. We have money laundering dataset with 45 objects and 7 variables. The output variable is the transaction. We used naïve bayes classifier to classify the transaction as illegal or legal and also finding out the probabilities for given attribute combination. The following steps are involved in our AML.

Step-1. Data acquisition

To collect the raw dataset from secured sited or to create the customized or synthesized dataset. It should be cleaned. It has various fields and their descriptions is given in Table I :

Step-2 Feature selection

Here we have filter out unnecessary columns which are not affect the transaction. We removed fields AMOUNT and OLD BALANCE, because these fields are factors which do not affect a transaction. The remaining fields will be used to build our model in Table II.

Column Name (money laundering)

Step-3 Divide Dataset We divided our entire dataset into two subsets as: Training dataset- to train the model, Test

TABLE I
DATA ACQUISITION TABLE

Field	Description
Type	Type of transactions like payment in/out,cash$_in/cash_outetc.$
Amount	How much have an amount customer's account
Account no.	Account no. in which transactions were held.
Old balance	Old balance
New balance	Amount after transaction
Date and time	Date and time of the transactions

TABLE II
FEATURE SELECTION TABLE

money$_laundering$	"Type"	"Amount"	"Account No."	"Old Balance"	"New Balance"	"Type Of Account"	"Date and Time"

TABLE III
MONEY LAUNDERING NB MODEL

Accuracy	0.8125
95 percent CI	(0.5435,0.9595)
No Information Rate	0.75
P - value	0.405
Kappa	0.5385
Mcnemar's Test P-value and Sensitivity	1.000 and 0.8333
Specificity and Pos Pred value	0.7500 and 0.9091
Neg Pred value	0.6000
Prevalance	0.7500
Detection Rate	0.6250
Detection Prevalence	0.6875
Balanced Accuracy	0.7917
'Positive' Class	ILLEGAL

dataset- to validate and make predictions. In a given dataset , consider 70 percent as training data and 30 percent as test dataset to measure the performance of our model.

id¡-sample(2,nrow(money laundering),prob=c(0.3,0.7),replace= T)

money laundering train¡-money laundering [id==1,]

money laundering test¡-money laundering [id==2,]

Step-4 Implement model

We designed the model with the help of the naïve bayes classified using the library 'e1071' on the training dataset.

The model is called money laundering nb.

Library (e1071)

Library (caret)

Step-5 Optimization model

Optimization model refers to modifying our model so as to achieve highest accuracy.

The Naïve bayes classifier can be further improved by including laplace correlation and normalization.

Step-6 Model validation

We can go ahead and check the validation. We will populate the confusion matrix which shows all the matrices to measure the accuracy, sensitivity, specificity, prevalence, etc.

V. RESULT ANALYSIS

The result is shown in Table III.

Our model has accuracy 0.8125 i.e. average for any model and got p-value 0.405. Its p-value is less than 0.05, so this model is best for this type of classification.

VI. CONCLUSION

In this paper, we used the Big Data Analytics method to classify money laundering actions into two categories i.e. legal or illegal. After using the Big Data Analytics, we explained that banking transactions can be detected in terms of, anti money laundering used by this classification in real time. It gives the best decision. It also helps to achieve the highest accuracy also. This technique may help in answering serious query related to the anti money laundering. Applied naïve bayes classification algorithm on some attributes of dataset and we found the classification and probabilities. This analysis revealed that sometimes this classification help to detect the fraud management such as anti money laundering or money laundering detection.

REFERENCES

[1] Z. Chen, L. D. Van Khoa, E. N. Teoh, A. Nazir, E. K. Karuppiah, and K. S. Lam, "Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review," Knowl. Inf. Syst., vol. 57, no. 2, pp. 245–285, 2018, doi: 10.1007/s10115-017-1144-z.

[2] C. H. Tai and T. J. Kan, "Identifying Money Laundering Accounts," Proc. 2019 Int. Conf. Syst. Sci. Eng. ICSSE 2019, pp. 379–382, 2019, doi: 10.1109/ICSSE.2019.8823264.

[3] S. Gao, D. Xu, H. Wang, and Y. Wang, "Intelligent anti-money laundering system," 2006 IEEE Int. Conf. Serv. Oper. Logist. Informatics, SOLI 2006, no. 7001805, pp. 851–856, 2006, doi: 10.1109/SOLI.2006.235721.

[4] N. A. Le Khac and M. T. Kechadi, "Application of data mining for anti-money laundering detection: A case study," Proc. - IEEE Int. Conf. Data Mining, ICDM, pp. 577–584, 2010, doi: 10.1109/ICDMW.2010.66.

[5] L. Keyan and Y. Tingting, "An improved support-vector network model for anti-money laundering," Proc. - 2011 Int. Conf. Manag. e-Commerce e-Government, ICMeCG 2011, pp. 193–196, 2011, doi: 10.1109/ICMeCG.2011.50.

[6] S. N. Wang and J. G. Yang, "A money laundering risk evaluation method based on decision tree," Proc. Sixth Int. Conf. Mach. Learn. Cybern. ICMLC 2007, vol. 1, no. August, pp. 283–286, 2007, doi: 10.1109/ICMLC.2007.4370155.

[7] N. A. Le Khac, S. Markos, and M. T. Kechadi, "A data mining-based solution for detecting suspicious money laundering cases in an investment bank," 2nd Int. Conf. Adv. Databases, Knowledge, Data Appl. DBKDA 2010, pp. 235–240, 2010, doi: 10.1109/DBKDA.2010.27.

[8] R. Liu, X. L. Qian, S. Mao, and S. Z. Zhu, "Research on anti-money laundering based on core decision tree algorithm," Proc. 2011 Chinese Control Decis. Conf. CCDC 2011, pp. 4322–4325, 2011, doi: 10.1109/CCDC.2011.5968986.

[9] Y. Jin and Z. Qu, "Research on Anti-Money Laundering Hierarchical Model," Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS, vol. 2018-Novem, pp. 406–411, 2019, doi: 10.1109/ICSESS.2018.8663895.

[10] Q. Yang, B. Feng, and P. Song, "Study on anti-money laundering service system of online payment based on union-bank mode," 2007 Int. Conf. Wirel. Commun. Netw. Mob. Comput. WiCOM 2007, pp. 4986–4989, 2007, doi: 10.1109/WICOM.2007.1223.

[11] C. W. Zhang and Y. B. Wang, "Research on application of distributed data mining in anti-money laundering monitoring system," Proc. - 2nd IEEE Int. Conf. Adv. Comput. Control. ICACC 2010, vol. 5, pp. 133–135, 2010, doi: 10.1109/ICACC.2010.5487272.

[12] Lopez-Rojas, Edgar Alonso, and Stefan Axelsson. "Multi agent based simulation (mabs) of financial transactions for anti money laundering (aml)." In Nordic Conference on Secure IT Systems. Blekinge Institute of Technology, 2012.

[13] Muriithi, R. ”The effect of Anti-Money Laundering regulation implementation on the financial performance of commercial banks in Kenya.” A Master’s Thesis, University of Nairobi(2013).

[14] omal, H. K. ”Big Data Analytics: Tackling Business Challenges in Banking Industry.” Business and Economics Journal 8, no. 2 (2017): 60-75.

[15] Saha AK, Kumar A, Tyagi V, Das S. Big Data and Internet of Things: A Survey. In2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) 2018 Oct 12 (pp. 150-156). IEEE

[16] naive-bayes-for-machine-learning. 2017. Retrieved from https://machinelearningmastery.com/

[17] Naive Bayes Algorithm.2015. Retrieved from https://software.ucv.ro/ cmihaescu/ro/teaching/AIR/

[18] Naive Bayes Classifier. 2017.Retrieved from https://rpubs.com/riazakhan94/

[19] Chen M, Mao S, Liu Y. Big data: A survey. Mobile networks and applications. 2014 Apr 1;19(2):171-209.

[20] Sameer, Ameer. “Big Data and Data Mining A Study of (Characteristics , Factory Work, Security Threats and Solution for Big Data ,Data Mining Architecture, Challenges Amp; Solutions with Big Data).” Unpublished, 2016. https://doi.org/10.13140/RG.2.1.3238.9525.

[21] Omran, Behzad Abounia, and Qian Chen. ”Trend on the implementation of analytical techniques for big data in construction research (2000–2014).” In Construction Research Congress 2016, pp. 990-999. 2016.

[22] Abourezq M, Idrissi A. Database-as-a-service for big data: An overview. International Journal of Advanced Computer Science and Applications (IJACSA). 2016;7(1).

[23] Mukherjee S, Shaw R. Big data–concepts, applications, challenges and future scope. International Journal of Advanced Research in Computer and Communication Engineering. 2016 Feb;5(2):66-74.

[24] Elgendy N, Elragal A. Big data analytics: a literature review paper. InIndustrial Conference on Data Mining 2014 Jul 16 (pp. 214-227). Springer, Cham.