International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 12 | Dec -2020 ISSN: 2582-3930

Big Data and Fraud

Chethana G

M.Tech 2nd SEM (CSE)

Department Of Computer Science and Engineering

University B.D.T. College Of Engineering, Davangere– 577004, Karnataka, India

(A Constitutional College Ofvisvesvaraya Technological University, Belagavi)

Mohammad Rafi

Department Of Computer Science and Engineering

University B.D.T. College Of Engineering, Davangere – 577004, Karnataka, India

(A Constitutional College Of visvesvaraya Technological University, Belagavi)

Abstract—Fraud is a criminal practice for illegitimate gain of wealth or tampering information. Fraudulent activities are of critical concern because of their severe impact on organizations,Communities as well as individuals. Over the last few years, various techniques from different areas such as data mining, machine learning, and statistics have been proposed to deal withfraudulent activities. Unfortunately, the conventional approaches display several limitations, which were addressed largely by advanced solutions proposed in the advent of Big Data. In thispaper, we present fraud analysis approaches in the context of Big Data. Then, we study the approaches rigorously and identify their limits by exploiting Big Data analytics.

Index Terms—Fraud Analysis, Big Data, Data Mining, Machine Learning, Statistical Modeling.

Introduction:

The meaning of fraud is broadly defined as the set of illegalactivities performed to gain personal benefits such as identitytheft of impersonating an authenticated user, shoplifting, illegaltransactions etc. Fraud detection and prevention has beena challenging task for many years. Retail sectors are widely affected by fraudulent activities.

A credit card is a payment card provided by every bank toeligible customers (cardholders) to make day-to-day transactions. Using the card, a cardholder can pay for goods and services without having money in their account at the particular moment and can be paid back to banks later point in time. The legitimate transactions made by the cardholder provide a pattern of his/her expenditures. If a card is stolen or accessed by some fraudsters, the transactions show an abnormal expenditure pattern and such transaction is called a fraudulent transaction. But, compared to large voluminous legitimate transactions, these types of transactions are relatively rare. Therefore,

identification of such fraudulent transactions is a quite complex task, and it is a part of fraud analytics. Due to the complexity involved in fraud analytics, identification of fraudulent transactions always has been an interesting research problem for banking and financial industries, research communities and academia. The fraud analytics can be achieved using different data mining tasks like classification, outlier detection, etc. Classification can be performed in various ways viz. binary, multi-, and One-Class Classification (OCC). Binary classification is the process to classify a set of samples into two classes. Similarly, the multi-class classification is used toclassify a sample into three or more classes. In the case of OCC, we have sufficient amount of samples available for one class, whereas samples for other classes will be rare. In this case, some rare samples do not belong to any of the known classes. Some examples of rare events area failure of a nuclear plant, credit card fraud, network intrusion, etc. Hence, whenever we came across the normal /regular samples in abundance and the targeted event in scarce, we can employ one-class classification approach to detect the rare event. In this study, we employed OCC for credit card fraud detection in Big Data framework. Big Data can be attributed using 4 V's viz. Volume, Velocity, Variety, and Veracity. Volume refers to a huge amount of data. Velocity implies how fast data are generated. Variety attributed to various formats of data. Finally, veracity represents the accuracy of data. In the case of credit card transactions, every bank deals with huge amount of credit card transactions every hour. Here, huge amount refers to volume and number of transactions per hour implies tovelocity. So, credit card transactions can be attributed to twoV's of Big Data i.e. volume and velocity. For our study, we relied upon one-class classification. Since there is a scarcity of fraudulent credit card transactions and a binary classificationapproach needs sufficient amount of historical data for both classes like legitimate and fraudulent, we extended one-class classification model in big data environment using Apache Spark framework for credit card fraud detection. Henceforth, we referred "Apache Spark" with Spark only. In this paper, we developed parallelization of a hybrid architecture involving Particle Swarm Optimization (PSO) and Auto-Associative Neural Network which is referred to as PSOAANN architecture. We implemented the AANN in a parallel manner over a Spark standalone cluster for one-class classification. The weight updation steps using PSO is implemented in a serial manner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACMmust be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org

Video verification or analytics at Point of Sale(POS) transaction is one of the loss prevention technique proposed by P. L. Venetianer *et al.* [18] have described a fraud detection tool toprevent merchandise loss at POS. They have proposed a video analytics approach to identify fraudulent activities near POS. Intelligent video software is provided by Object Video whichis called as RetailWatch which essentially combines POS data along with video data. Video surveillance has become one of the necessary tools to detect shrinkage of merchandise in ashop

near POS. The CCTV system is used to capture activities near POS to detect fraudulent transactions. The proposed solution detect most commonly used transactions-

• Refund transaction in absence of customer: Here retailemployee does the legitimate refund transaction at POS without the presence of a customer to earn money illegally.

• No-sale or void transaction: This kind of transaction also in absence of customer where the employee holds up until client leaves the shop after doing legitimate transaction then report that transaction as void or no sale to refund the amount to the client but actually goes to employee pocket.

• Unauthorized transaction: Many retail transactions required manager authorization for a transaction but in certain illegal transactions employee of the retail store use manager code to validate transaction for personal benefits.

• Refund transaction from wrong place: In this kind of fraudulent transaction involve buying a product from a retail store then taking it out and later come back to store, pick another item of the same type and ask for a refund.

•"Sweethearting": In this type of transactional fraud, a retail employee himself gives an expansive item to the friendly customer without ringing up.

Literature survey:

Fraud analysis approaches in the age of big data

Fraudulent activities (*e.g.*, suspicious credit card transaction,financial reporting fraud, and money laundering) are critical concerns to various entities including bank, insurance companies, and public service organizations. Typically, these activities lead to detrimental effects on the victims such as a financial loss. For instance, the Americans currently lose $50 billion a year to a variety of fraudulent practices 1. According to a recent study, the rapid growth of financial fraud will worsen the economic situation. Furthermore, research predicts that online fraud alone will climb from $10.7 billion in 2015 to $25.6 billion in 2020 [1]. Thus, the automated fraud detection systems have gained enormous popularity especially within financial institutions. These systems analyze complex events or actions over historical facts (*i.e.,* data) and discover fraud patterns. The analysis of fraud is a process consisting of a sequence of functions to predict or discover potential or explicit threats of fraudulent activities. The process relies on techniques from a wide variety of areas including *datamining*, *statistics*, *machine learning* etc. The efficacy of a fraud detection system largely depends on the efficiency of the used techniques and relevant data.

Examples of fraud:

• ***Computer Intrusion*:**This action can be considered a fraud itself like data or identity theft or a way to commit fraud by using the stolen information. Computer intrusion can be as simple as attempting to guess passwords, or sophisticated like introducing malicious software viaemail attachments.

• ***Credit Card Fraud***: This fraud is defined as an unauthorized use of a credit card account. It occurs when the card holder and the card issuer are not aware that the card is being used by a third party. Therefore, fraudsters can obtain goods without paying, or gain illegal access to a certain account.

• ***Telecommunication Fraud*:** In telecommunication, a fraud is characterized by the abusive use of any carrier serviceswithout the intention of paying. Sim cloning is a typicalexample of telecommunication fraud.

• ***Financial Statements Fraud***: Financial statements are official reports that reflects the financial activities and position of a business. Financial statement fraud also known as accounting fraud is defined as an *intentionalmisstatements of financial statements for the purpose ofmisleading investors and creditors, and to create a falseimpression of an organization's strength*.

• ***Securities Fraud*:** Securities market is a place where financial instruments (stocks, bonds, options, futures *etc.*) can be bought and sold between participants, and where prices are determined on the basis of demand and supply. Securities fraud, also known as financial markets fraud or investment fraud refers to *deceptive practices like thatinduces investors to make purchase or sale decisions onthe basis of false information*.

• ***Insurance Fraud*:** This behavior can be described as a misuse of an insurance policy. Insurance is made to cover losses and to protect against risks. Fraud occurs when the
Insured use the insurance contract as a tool to gain illegal profit, for example when receiving more money than the actual loss by overstating claims.

• ***Money Laundering***: This type of fraud is the scheme in which criminals try to disguise the source of money gained through illegal activities, with the intention of making it seem legitimate. Money laundering is a very complex phenomenon, because it is governed by multipleSocial and economic conditions, and its impact is a lot more dangerous because it is the main pillar to others crimes such as arms or drug trafficking, and terrorism financing.

Fraud analysis techniques

Over the years, a number of fraud analysis techniques have been proposed in a large body of literature. In our study, we categorized the existing fraud analysis techniques based on the *area of study*.
The two main categories *statistical modeling* and *machine learning* are explained in the following.
• Statistical modelling: This is an area of mathematics that deals with collecting and analyzing data, according to distributions that follows certain assumptions.
• Machine learning: These techniques of programming models that learns from the data, and build systems that can solve complex problems. It consists of several methods that uses either supervised or unsupervised types of learning.

– ***Supervised learning*:** A sample of labelled data,which means attributes and associated labels indicating fraud or legitimate activities, is used to train the model that will classify a certain operation into one of the two classes conditionally to known values of the attributes.

– ***Unsupervised learning***: These techniques are performed solely using unlabeled data, the labels indicating the classes are not available. The aim is to find groups of similar customers according to the values of their attributes, or finding outliers

indicating unusual behavior that requires more investigation.

Here we describe themethods that are commonly used in fraud detection:

• Outlier detection methods: These techniques are basically unsupervised learning methods like *Peer GroupAnalysis* and *Break point analysis*. In the first method, individuals with similar behavior are grouped together. Then, individuals that start to behave differently than theothers in the same group are detected. The second method is a tool that catches changes in the behavioraccording to the information of a single individual.

• K-means clustering: K-means is the simplest clustering algorithm whose aim is to group data in k categories. It consists of initializing k centroids randomly, each point is then associated to the closest centroid forming groups, and at each iteration, centroid are recalculated this timeaccording to the groups formed, and so on until no more centroids change. Then, new observations are assigned to the group with the nearest centroids.

• K-nearest neighbor: This is a method that consists of assigning a class (classification) or predicting a value (regression) using the k nearest points to the one that we aim to predict. A specific norm is used to measure the distance between points, and a specific *k* is used topick the nearest points. Voting is used for classification or average is used to predict a numeric value or regression.

• Naıve Bayes: It is a classification tool that simply uses *Bayes conditional probability rule*. Each attribute and class label are considered random variable, and assuming That the attributes are independent, the naıve Bayes finds a class to the new observation that maximizes itsProbability given the values of the attributes.

• Bayesian belief network: These networks are graphical models for probabilistic relationships between a set of variables. Random variables are represented as nodes and conditional dependencies between variables are represented as arcs between nodes. Each node is linked toa probability function that generates probabilities of the node's variable conditionally to values of the parent's node.

• Hidden markov model: It differs from the normal statistical Markov model by having invisible states, but each state randomly generates one of the visible states. A hidden Markov model can be presented as the simplest dynamic Bayesian network.

• Decision tree: The basic idea of a decision tree is a recursive partition of the data. First all the observations are grouped together in the root node. Then accordingto the values of attributes, this node is decomposed in two child nodes, that classify the observations in better way than the root node, such that the uncertainty about the class at the child nodes are reduced. This procedureis repeated at every node, until all nodes of the tree are transformed into leaves with specific classes.

• Logistic regression: Logistic regression is a type of generalized linear model. Using simple linear regression is inappropriate when the variable to be predicted is binary; due to normality assumptions. Therefore, a *logit* link function is introduced, to force the linear combinationof the attributes to take values between 0 and 1, and the obtained logistic regression model is then used to predict if a certain operation is fraudulent or not.

• Support vector machine: As a classifier, the aim is to find a hyperplane that separates the data according to labels (fraud / legitimate) in the best possible way, which means with maximum distance between both categories. The hyperplane is then used as a discriminant for newobservations. One of the most important characteristic of support vector machine is the *Kernel trick* that consists of using a Kernel function such as the radial basis function as a map for a high-dimensional space allowing for nonlinear classification.

• Artificial neural network: Artificial neural networks were first created with the purpose to imitate the behavior of the human brain. A neural network is the connection of elementary objects called *the simple neuron*.The simple neuron consists of finding a connection between inputs (attributes) and outputs (labels). Each input is multiplied by a specific weight. These weighted inputs are then added together with a bias and then an activation function is applied to obtain the final result of the neuron. The network consists of consecutive layers containing these simple neurons. Each layer transfers the signals to the next layer, and the output layer provides the system's response. – Self-organizing map: This is a type of artificial neural network that uses unsupervised learning and aims to transform an incoming signal pattern of arbitrary dimension (attributes) into two dimensional discrete map (two clusters).

• Artificial immune systems: Artificial immune systems are concerned with extracting the role of the immune system to create computational systems, and to use them as tools to solve complicated problems; trying to imitate the immune system's ability to distinguish between *self*and *non-self*. The main developments have focused on three immunological theories: clonal selection, immunenetworks and negative selection.

• Model combination: It is a common approach to buildone model by combining several algorithms, with an attempt to obtain better performance and higher accuracy rates. Every model combination requires two phases. The first is a diversification like bagging, boosting [2], or hybrid diversification [3]. The second is an integration phase that aims to combine the results generated by the diversified models using uniform or weighted voting (for classification), mean or weighted average (for regression), or meta learning. Sometimes a hybrid architecture is builtby combining genetic algorithms, optimization algorithms or fuzzy logic with other models [4], [5].

Fraud detection in big data using supervised and semi-supervised learning techniques

Nowadays, most of the organizations, companies and government agencies are conducting a large portions of their activities using information systems, for example, accounting, human resources, customer relationship management, marketing, and sales [13]. Following this perspective, organizations have adopted practices associated to electronic commerce to increase their productivity or efficiency in trading products or services [14]. Simultaneously, the accessibility, omnipresence and high performance of the personal computing are motivating individuals to do their banking, shopping, and entertainment electronically. This increasing trend has

luredcriminals to move their efforts into this new territory making the electronic fraud a serious problem in several contexts. Based on the review presented in [14], Figure 1 shows a taxonomy of the areas where the electronic fraud can be found.

Database and experimental setup

The dataset used in this work was provided by a Colombianpayment gateway company, as part of a joint research project. The original database provided by the partner company is a relationaldatabase. This database has a structure made to support the business model of company, therefore a pre-processing of the information was required. After the characterization stage, the dataset was exported from the relational engine o Apache Hive2, a data warehouse based on map-reduce.

The dataset on Hive have the distribution of the three classes: fraud, legal and unlabeled, shown in Table I

TABLE I. DISTRIBUTION OF CLASSES OF THE DATABASE USED

Class	N° of samples	%	% (Fraud And Legal)
Legal	6488341	80.288	99.927
Fraud	4726	0.058	0.073
Unlabeled	1538104	19.033	NA

All the experiments were carried out using a cross validationstrategy with 5 folds. In order to get a baseline model to compare the BRF and the co-training approaches, the first model trained and validated is the Apache Spark RF implementation. Also, this first model is trained to evaluate the influence of the four parameters of Apache Spark RF implementation. A grid search of thefour parameters was performed with the following values: depth=[10; 20; 30]), bins=[32; 128], impurity= [entropy; gini], trees[2; 10; 50; 100]. The total number of experiments was 240. The values of the grid for BRF are the same excluding the number of trees, due to the fact that in this model the trees are pre-calculated is shown in equation (1). In the co-training model the parameters evaluated were = [0:75; 0:85; 0:95] and the condition of whether both classes are added to the new training set or only the fraudulent one.The different approaches tested were evaluated in terms of sensitivity, specificity and accuracy.

The sensitivity, specificity and accuracy are defined as

S =TP/TP + FN; E =TN/TN + FP

……(1)

Acc =TP + TN/TP + TN + FP + FN

……(2)

Sensitivity (eq. (1) left) and specificity (eq. (2) right) measure the accuracy on the positive (fraud) and negative (legal) cases. A trade-off between these true positives and truenegatives is typically sought. Therefore, the set of metrics able to quantify the performance in imbalance scenarios includes the G-mean, which can be defined as

G-mean = (S .E)1=2

…….(3)

and the weighted-Accuracy expressed as

wtdAcc =BS + (1 _B)E

…….. (4)

Both G-mean and wtdAcc provide summary performance indicators of the tradeoff between S and E, taking into account the relative weight of each class in the test sample. The _ used is 0:7 to indicate higher weights for accuracy on the fraud cases. Well known Area under the ROC Curve (AUC) was also included. AUC is often considered a better measure of overall

performance [30], and is independent of specific classification cutoff values.

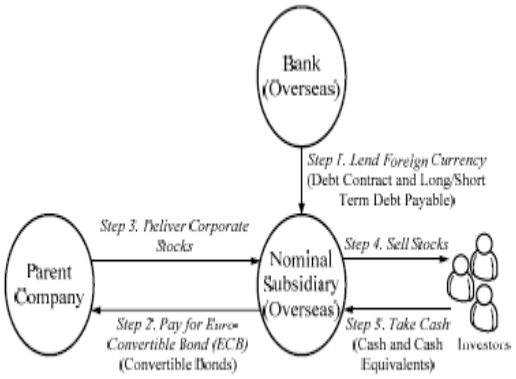
On big data fraud detection method for financial statements of business groups

Recently, various approaches have been developed todetect fraud in corporate financial statements [15] [16] [17] [18] [19] [20] However, these recent studies focused primarily on using DMtechniques to detect financial statement fraud within a single enterprise; studies have seldom focused on detecting such fraud within an entire business group by using the diversity andvalue of big data. This phenomenon often results in increasingfinancial statement fraud when business groups search for buyers in corporate mergers, acquisitions, or stock undertaking,and increases the investment risk for investors in capital markets. Therefore, this study develops a big data-based fraud detection approach for financial statements of business groups to efficiently improve fraud detection efficiency accuracy, and thereby reducing investment losses and risks and enhancing investment decision-making benefits for investors and creditors. This objective is achieved using the following tasks:

(i)Designing a big data-based fraud detection process for financial statements of business groups.

(ii) Developing big data-based fraud detection techniques for use with such statements.

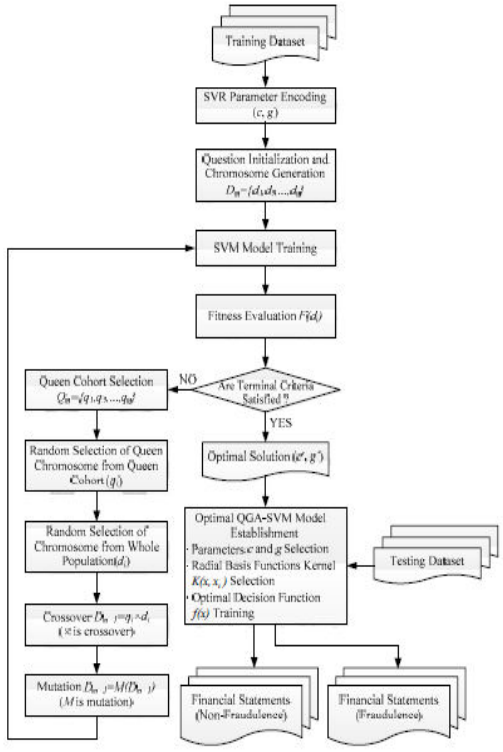
(iii) Demonstrating and evaluating use of this proposed big data-based fraud detection approach. From a comprehensive survey of infamous financial statement fraud cases relating to business groups, such as the Procomp and Enron cases, four financial statement fraud models for business groups are deduced, including exaggerated profit, undue deposit and debt, false financial statement information, and irrational balance sheet through ECB (Euro- Convertible Bond). In this study, the financial statement fraudfor business groups "Irrational Balance Sheet through ECB" isused as an illustrative example, as shown in Figure 1 presents the big data-based fraud detection process forirrational balance sheet through ECB.



Development of big data based fraud detection

ALGORITHM FOR BUSINESS GROUPS' FINANCIAL STATEMENTS

Based on the big data-based fraud detection process for financial statements of business groups, this section presents the core algorithm involved in the process, that's the algorithm for clustering financial statements.



Algorithm for clustering financial statements.

In this section, the financial statement fraud for businessgroups "Irrational Balance Sheet through ECB" is used toexplain the feasibility of our method. Based on reports to
Shareholders, stocks trading volume, debt structure indicators, corporate governance indicators, and the financial ratio of business groups in Taiwan retrieved from the Taiwan Economic Journal (TEJ) database [21] and the Taiwan Stock Exchange Corporation [22].

This study considers the characteristics of variety and value of big data used in finance and economics to develop a big data-based fraud detection approach for the financial Statements of business groups to more precisely detect the financial statement fraud of business groups, and thus reducing investment losses and risks and enhancing investment decision makingbenefits for investors and creditors.

Fraud detection and prevention by using big data analytics

The meaning of fraud is broadly defined as the set of illegalactivities performed to gain personal benefits such as identity theft of impersonating an authenticated user, shoplifting, illegaltransactions etc. Fraud detection and prevention has been a challenging task for many years. Retail sectors are widely affected by fraudulent activities. Consumers and retail sectors are equally affected by the recent threats encounter in shopping malls and virtual product stores. Shoplifting is another problem found in the retail sector where staff or customer pick an item without knowledge of retail shop and do not pay for it. A tremendous amount ofdata need to be analyzed to find pattern involved in fraudulent transactions. In the retail sector, a significant amount of loss happens worldwide due to check out related fraud. The itemswhich are being sold are being registered at Point of Sale (POS) it is possible that merchandise loss can happen for unregistered items at POS which may not be checked at security checkout. A lot of human activities are involved at POS such as (i) picking up items (ii) scanning bar code onitem (iii) putting items into bag (iv) payment process. Video Analytics is used to evaluate the checkout process. They have proposed a video analytics approach to identify fraudulent activities near POS. Intelligent video software is provided by ObjectVideo which is called as RetailWatch which essentially combines POS data along with video data. -

• Refund transaction in absence of customer: Here retail employee does the legitimate refund transaction at POS without the presence of a customer to earn money illegally.
• Unauthorized transaction: Many retail transactions required manager authorization for a transaction but in certain illegal transactions employee of the retail store usemanager code to validate transaction for personal benefits.
• Refund transaction from wrong place: In this kind of fraudulent transaction involve buying a product from retail store then taking it out and later come back to store, pick another item of the same type and ask for a refund.
•"Sweet hearting": In this type of transactional fraud, retail employee himself gives an expansive item to them friendly customer without ringing up. mvandro et *al.* [24]. Hasanalyzed fraudulent electronic transactions.

Big data and its contribution

Big Data is initially characterized regarding 3V by DougLaney in 2001: Volume, Velocity and Variety. Volume refers to the measure of data which can be huge in amount, Velocityrefers to the speed at which different data arrives and Variety refers to the complexity of organized and unstructured data arrive from various data sources. Later a few definitions of BigData were presented, one of them proposed by Zikopoulous [25] that adds Veracity and Value to the existing 3Vs of Volume, Velocity, and Variety. Veracity refers to the correctness of data got and can address the quality issues, for example, noise or missing values. Value is next quality which is basic to locate the relevant data for investigation. Data mining and MachineLearning techniques such as Classification, Prediction, Clustering, Dimensionality Reduction, Regression, Artificial Neural Network and Outlier Analysis are used for identification of apattern in the fraudulent transaction using skimmed data. Big Data analysis tools provide an effective approach to analyze unusual patterns for detecting retail fraud. Fig 1 describesthe fraud detection mechanism using bigdata architecture. The model has three major components data gathering, fraud detection and communicating to the user through APIs.

Analyzing retail fraud detection tools

Fraudulent activities are critical concerns to many entitiessuch as the retail division, banking, and public sector establishments. These dishonest activities led to a financial loss to anorganization. Over the years many techniques were developed to analyze and detect fraud.Saad Mohamed Ali *et al.* [26] have proposed anomaly detection system based on k-mean clustering algorithm and Sequential Minimal optimization (SMo) to detect online networkanomaly. The detector system does its tasks with accuracy

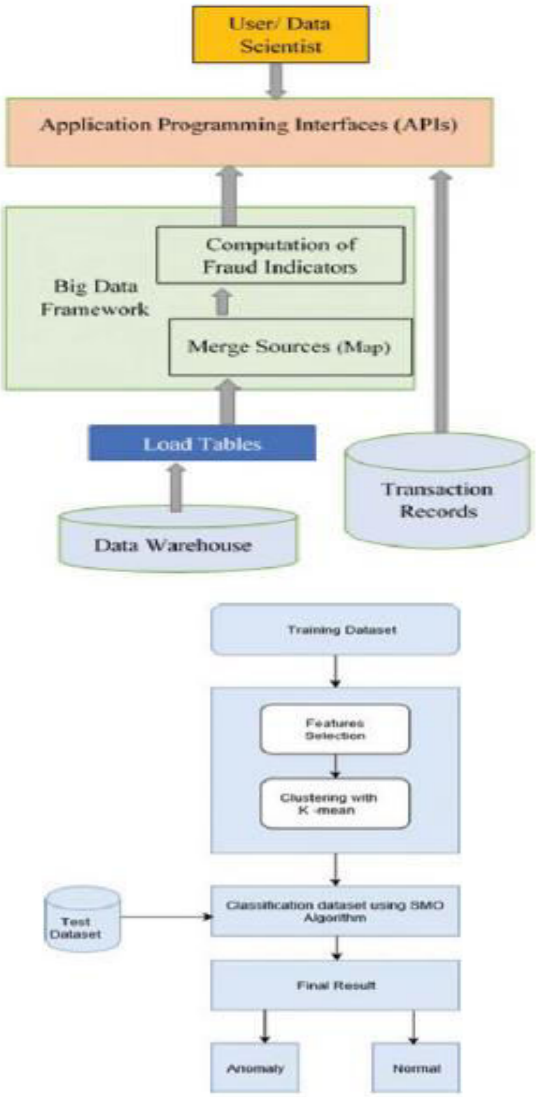


Figure: anomaly detector system

Sequential Minimal optimization (SMo) [27] technique isused to identify network outlier online with high accuracy. NSL KDD dataset is used in place of KDD CuP. The advantage of utilizing this data set is that duplicate record is expelled and an adequate number of data set is utilized forthe training purpose. Outlier detection accuracy rate of SMo is 73.82%. Hybrid approach (K mean + SMO) even give an accuracy of 97.36%. Richard Zuech et *al.* [28] reviewed the main technological need for intrusion detection. Hadoop framework reviewed for its need in detecting intrusion. The framework comes with several technologies bundled with it such as HDFS.

In this survey, we have addressed different types of fraudand their consequences. We have analyzed the recent trends in retail fraud and various technologies used in detectingthem. In this paper, we have also discussed the various advantages and disadvantages of these technologies. The main goal of this work is to make a survey on the detection andprevention of fraud using data mining, machine learning, and big data analytics approach. Here we have also addressed the underlying tools and techniques of pattern analysis of variousproblem domains. We have also analyzed various data source for big data analytics (BDA).

Conclusion:

Fraud analysis is of critical importance to the banking sector– as well as many others – since fraudulent activities are becoming a more frequent occurrence that lead to disastrousimpact to the organizations, society, and individuals. They presented a review of state of the art approaches used within Big Data analytics which detects a widevariety of fraudulent activities. They provided a comprehensive overview of frauds where we explained what is fraud and its classification. We also presented data mining, statistical, an machine learning methods commonly used in analytics. We explained different approaches that are proposed in literature to address issues such as reducing data, implementing data miningmethods within Big Data frameworks, and under or oversampling. We provided a detailed discussion on challenges and issues that remain unsolved. To the best of our understanding,efficiency remains an important

issue which should be focused on to reap especially the benefits of data driven fraud analytics in real-time. By surveying this paper I can conclude current trending frauds like network fraud currently happening with google malicious can be prevented.

References:

[1] N. Ivanov, "Preventing financial fraud with in-memory computing,"2017.

[2] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques - chapters 8 and 9*. Morgan Kaufman Publishers, 2003.

[3] R. Patidar and L. Sharma, "Credit Card Fraud Detection Using Neural Network," *International Journal of Soft Computing and Engineering*, vol. 919414552184, no. May, pp. 13–14, 2011.

[4] Y. Dhanalakshmi and R. Babu, "Intrusion Detection Using Data Mining Along Fuzzy Logic and Genetic Algorithms," *International Journal ofComputer Science and Network Security*, vol. 8, no. 2, pp. 27–32, 2008

[5] H. He and E. Garcia, "Learning from imbalanced data," *IEEE Transactions on Data and Knowledge Discovery*, vol. 12, no. 9, pp. 1263–1284,2009.

[6] N. V. Chawla, N. Japkowicz, and A. Kolcz, "Editorial: Special issue on learning from imbalanced data sets," *Sigkdd Explorations*, vol. 6, no. 1, pp. 1–6, 2004.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal ofArtificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[8] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-smote: A new oversampling method in imbalanced data sets learning," in *Proceedings ofthe International Conference on Intelligent Computing*, 2005, pp. 878–887.

[9] W. Fan, S. J. Stolfo, J. Zhang, and P. Chan, Misclassification cost-sensitive boosting," in *International Conference on MachineLearning*, 1999, pp. 97–105.

[10] G. Ditzler and R. Polikar, "Incremental learning of concept drift from streaming imbalanced data," *IEEE Transactions on Knowledge and DataEngineering*, vol. 25, no. 10, pp. 2283–2301, 2013.

[11] S. Chen, H. He, and E. Garcia, "Ranked minority oversampling in boosting," *IEEE Transaction on Neural Networks*, vol. 21, no. 10, pp. 1624–1642, 2010.
To Detect Medical Fraud and Abuse," *33rd Annual I* [40] H. Hormozi, M. K. Akbari, E. Hormozi, and M. S. Javan, "Credit cards fraud detection by negative selection algorithm on hadoop (To reduce the training time)," *The 5th Conference on Information and Knowledge Technology*, pp. 40–43, 2013.*processing systems II*, pp. 889–895, 1999.
[12] M. Behdad, L. Barone, M. Bennamoun, and T. French, "Nature-inspired techniques in the context of fraud detection," IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, vol. 42, no. 6, pp. 1273–1290, 2012.

[13] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," Journal of Network and Computer Applications, vol. 68, pp. 90–113, 2016.

[14] PwC, "Global Economic Crime Survey," Tech. Rep., 2016. [Online]. Available: https://www.pwc.com/gx/en/economic-

crimesurvey/pdf/GlobalEconomicCrimeSurvey2016.pdf

[15] PricewaterhouseCoopers, "PwC Global Economic Crime 2014 Survey: Cybercrime and Electronic Fraud," 2014. [Online]. Available: http://www.pwc.com/gx/en/economic-crime-survey/cybercrime.jhtml

[16] http://www.tej.com.tw/twsite/, Taiwan Economic Journal.

[17] http://www.twse.com.tw/ch/index.php, Taiwan Stock Exchange.

[18] F. H. Glancy and S. B. Yadav, "A computational model for financial reporting fraud detection," Decision Support Systems, vol. 50, No. 3, pp. 595-601, 2011.

[19] G. L. Gray and R. S. Debreceny, "A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits," International Journal of Accounting Information Systems, vol. 15, No. 4, pp. 357-380, 2014.

[20] S. Y. Huang, R. H. Tsaih and F. Yu, "Topological pattern discovery and feature extraction for fraudulent financial reporting," Expert Systemswith Applications, vol. 41, No. 9, pp. 4360-4372, 2014.

[21] E. Kirkos, C. Spathis and Y. Manolopoulos, "Data mining techniques for the detection of fraudulent financial statements," Expert Systems with Applications, vol. 32, No. 4, pp. 995-1003, 2007.

[22] P. Ravisankar, V. Ravi, G. R. Rao and I. Bose, "Detection of financial statement fraud and feature selection using data mining techniques,"Decision Support Systems, vol. 50, No. 2, pp. 491-500, 2011.

[23] W. Zhou and G. Kapoor, "Detecting evolutionary financial statementfraud," Decision Support Systems, vol. 50, pp. 570-575, 2011.

[24] Saad Mohamed Ali Mohamed Gadal, and Rania A. Mokhtar, "Anomaly detection approach using hybrid algorithm of data mining technique,"ICCCCEE, pp. 1-6, Jan. 2017.

[25] Richard Zuech, Taghi M Khoshgoftaar, and Randall Wald, "Intrusion detection and Big Heterogeneous Data: a Survey", Journal of Big Data, Springer Open Journal, 2015

[26] Abd. Charis Fauzan, Riyanarto Sarno, Nurul fajrin Aiyana Structurebased Ontology Matching of Business Process Model for Fraud Detection Technology and System (ICTS), pp. 221-225, 2017.

[27] Evandro Caldeira, Gabriel Brandao, Hudson Campos, Adriano Pereira,"Characterizing and Evaluating Fraud in Electronic Transactions", pp. 115-122, Eighth Latin American Web Congress, 2012, USA.

[28] P. L. Venetianer, Z. Zhang, A. Scanlon, Y. Hu, A. J. Lipton ObjectVideo. Inc.,and Reston, "Video Verification of Point of Sale Transactions", IEEE, pp. 411-416, VA, USA, 2007.