ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

Big Data Analytics Based Approach to Tax Evasion Detection

[1] Yashashwita Shukla, [2] Neena Sidhu, [3] Akshita Jain, [4] T.B. Patil, [5] S.T. Sawant-Patil
[1][2][3] U.G. Student, [4] Assistant Professor, Department of Information Technology, Bharati Vidyapeeth Deemed To Be University College of Engineering, Pune, Maharashtra, India
[5] Assistant Professor, Department of Electronics & Telecommunication, Smt. Kashibai Navale College of Engineering, Pune, Maharashtra, India

Abstract: **- Tax evasion is the illegal underpayment or non-payment of tax to the government. Detecting these illegal tax evasion activities is an important as well as challenging issue for the tax administration system of any country and the biggest challenge is the increasing volume of tax data. In this paper, we propose a system that can help to characterize and detect the probable tax evaders using the information in their tax payment through big data analysis techniques. After pre-processing the available tax data, K-mean clustering algorithm is applied to form clusters of taxpayers with similar behaviour. Then decision trees are used to classify each cluster into tax payers with and without fraud and patterns in their associated behaviour are detected. Using these characteristics and patterns, the artificial neural network is trained and potential fraudsters could be detected based on the information available. This system will help detect tax fraudsters and enhance knowledge on their patterns of fraud which will be helpful in the prevention of fraud.**

Keywords: **- Tax Evasion, Big Data, K-Mean Clustering, Decision Trees, Multi-Level Feed Forward Neural Network.**

I. INTRODUCTION

In extremely controlled sectors like financial, insurance, healthcare, retail, and social security, tackling fraud is essential as there is a multitude of compliance, regulations, risk management and fiscal consequences to be dealt with. In today's world, transactions and documents are mostly recorded in a digital form in one way or another and therefore evidences are always out there to aid investigators in identifying the damaging fraud activities. Big data analytics offers a feasible platform to fraud detection as it can analyse vast amounts of data in need of extracting and identifying various fraudulent patterns. It uses statistical, mathematical, artificial intelligence and machine learning techniques to extract and identify useful information which can help any administration take preventive measures. Techniques like Artificial Neural Networks, Self-organizing Maps, Decision Trees, Support Vector Machine, Logistic Regression and Bayesian Learning have been used and found beneficial in various sectors for fraud detection. One of the most exploited monetary sector is the tax administration sector. Tax evasion [10] is prevalent as it provides financial gains to the individual by lowering one's tax liability. As tax evasion hinders the economic growth of a country, it holds a very high priority and therefore there is an immediate need for enhancing the tax fraud detection system. The ever-increasing volume of tax-related data or records is the most challenging issue as it includes huge volumes of data integrated from various sources in both structured and unstructured form. Big Data Analytics [9] can offer great potential in tackling tax fraud as it allows extraction and generation of knowledge from large volumes of data. Generally, big data is processed using clustering and classification algorithms to partition the data into logical groups and form a logical structure before analysing it. Clustering is an unsupervised learning technique used to group similar data points on the basis of features and it is performed on unlabelled data. Commonly used clustering techniques are K-means clustering, Outlier detection method, Self-organising map and Hidden markov model. On the other hand, classification is a supervised learning technique used to assign predefined classes to data points on the basis of features and it is performed on labelled data. Commonly used classification methods are Naive Bayes, K-nearest neighbour, Decision tree, Support Vector Machine, etcetera. There are two approaches of combating the tax fraud: Tax Fraud Prevention and Tax Fraud Detection. Tax Fraud Prevention involves some mechanisms to avoid fraud occurrence, while Tax Fraud Detection comes into picture when Tax Criminals overcome Tax Fraud Prevention. Efficient techniques can be applied to detect tax fraud and identify suspicious activities being committed as soon as possible and effective decision can be made using the identified patterns. The objective of this paper is to classify the characteristics of the good and bad behaviour of the taxpayers, identify some distinguishing patterns between them and detect the tax fraudsters. To characterize the behaviour of the taxpayers efficient clustering and

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

classification methods can be used and some distinctive patterns can be identified between these clusters. Based on these characteristics and the patterns, the system can be trained and tax fraudsters can be detected using a machine learning algorithm. Machine learning makes it possible to analyse various ways in which fraud can be performed. Hence, a supervised classification can help determine whether a new transaction is legitimate or fraudulent.

II. LITERATURE SURVEY

[1] This research analysed various classification and clustering methods to distinguish between tax payers who have good and bad financial behaviour associated with the usage of false invoices. It used the neural gas technique to identify some relevant attributes and self-organizing maps to detect patterns in the form of clusters. Decision tree technique was used to detect variables and classify between fraud activities and no fraud activities. In case of small enterprises, the significant attributes were mostly related to the percentage of tax credits and previous inspections of the negative behaviours. Other important parameters in the research were the number of invoices delivered during the fiscal year, the net amount of Value Added Tax declared, the percentage of average tax credit balance and positive behaviour audits. Whereas, in medium or large enterprises, the significant parameters were the amount of excess credits, percentage of credit linked with invoices delivered and the relationship between costs and properties owned. It recommended a combination of the results achieved with decision trees and artificial neural networks in order to inspect the individuals who were identified as fraudsters. [2] The objective of this study was to apply data mining techniques to detect suspicious Value Added Tax (VAT) evasion reports for inspection. They designed a screening model which offers a more scientific and resource saving approach to detect potential tax evaders compared to the manual screening methods. Therefore, the model could help reduce unnecessary wasting of staff resources and also improve the accuracy rate of the fraud detection. Due to the budget restrictions, the current study had a few limitations related to the data mining tool used in the implementation process. The tool wasn't very efficient and advanced. Hence, it recommended other efficient data mining software's to enhance the tax evasion detection performance and accuracy. [3] This study proposed a solution to tax evasion and implements a heterogeneous information network to distinguish between the behaviours of the different tax payers. It also implemented trail-based pattern recognition to detect the doubtful groups of taxpayers. Through fusion and reduction of multiple social relationships, it simplified the vast information network into a coloured model with two colours for the nodes and two colours for the edges. After the experimental analysis, it concluded that the system detects the suspicious fraudsters by scaling down the number of suspected individuals or enterprises and then scaling down the number of doubtful transactions associated to them. This method has been implemented on multiple levels in several provinces of mainland of China in its tax monitoring and management system. [4] This study presented a fraud detection system for transactions carried out with credit cards. The system was designed to tackle the three most challenging problems related with fraud detection: processing a large number of transactions, inclusion of labelled and unlabelled samples of data and a strong class imbalance in data sets. A Balanced Random Forest (RF) technique was implemented in order to deal with the problem of class imbalance and a co-trained Balanced Random Forest was adopted to deal with the problem of unlabelled sample data. The strategy based on the combinations of both Balanced RF and co-trained Balanced RF attained a better performance. [5] This study designed a hybrid framework to solve the problem of online fraud detection. The system worked in two stages, detection and training. The framework aimed at combining different detection algorithms to improve accuracy and implemented a four-layered design structure to handle data storage, system training, sharing of data and online fraud detection at each layer. The framework was implemented using the latest big data techniques to build a scalable and high performance system with fault tolerance. As future work, proposed to use the combination of better detection algorithms and tools, testing the system with real transactional data and a systematic optimization of all components of the given framework.

III. PROPOSED SYSTEM

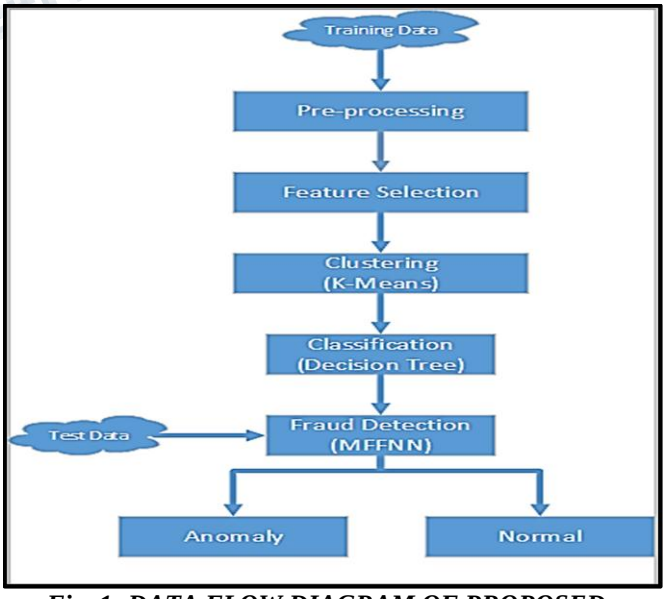

Fig. 1: DATA FLOW DIAGRAM OF PROPOSED SYSTEM

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

i) PRE-PROCESSING:
Pre-processing of training data is an important phase to make the dataset an appropriate input for the classification phase. The main objective of pre-processing is to minimize ambiguities and provide accurate information that can be analysed. The data is transformed into an appropriate format by grouping and labelling during the data pre-processing phase. This phase also handles the missing or incomplete dataset by replacing missing values with some mean values.

ii)FEATURE SELECTION:
Feature selection is one of the critical stages and is equally important as the efficiencies of the algorithms to be used. Generally, the input data is in a high dimension feature space but not all the features are relevant to the classes that are to be classified. The data sometimes includes irrelevant and redundant features which can introduce noise/errors during the learning phase. Therefore, the feature selection phase eliminates irrelevant, noisy or redundant features and sometimes also decreases the number of attributes. Selected set of attributes or features are used as an input vector for further analysis. Feature selection improves the system by speeding up the process, improving learning accuracy and comprehensibility.

iii)CLUSTERING
Clustering phase can be implemented by using the K-means clustering algorithm [6]. K-means clustering technique is a kind of unsupervised learning, which is used to find groups in the data, where K represents the number of groups. These groups are known as clusters. The algorithm iterates to assign each data point to one of these K clusters based on the parameters that are provided. Data points are clustered based on the similarity of parameters. K-means clustering algorithm results in the centroids of the K clusters, these centroids are used to label the new data. Each centroid of a cluster is a collection of parameter values which defines the resulting group. Centroids can be examined to interpret what kind of data a cluster consists. Once the system becomes stable, the training is complete and it can assign a cluster to a data point with nearest centroid.
Clustering phase is be applied to the world of tax payers to categorize groups of individuals with similar behaviour K-means clustering will initialize a few centroids randomly, each data point will be then associated to the closest centroid forming group. As the algorithm will iterate through the training data, centroids will be updated with each iteration according to the clusters formed. The algorithm will keep iterating through the data until no more values of centroids change.

iv)CLASSIFICATION
Classification of these clusters can be implemented using the Decision Tree Method. Decision tree method is a non-parametric supervised learning technique used to classify data into predefined classes. It uses a predictive modelling approach to go from observations about the given data item to conclusions about the target value of the data item. In the tree structure, leaves represent the target class labels and the branches represent conditions of features that lead to those targets. A decision tree results in grouping the homogenous data items together and maintaining different classes for heterogeneous data items. This technique is the most widely used classification technique given its efficiency and simplicity. It requires a very little data normalization and is able to handle both numerical and categorical data.
Therefore, once the clusters are formed with the individuals having similar behaviour, we can classify these clusters into two classes. The first class will be formed by the individuals with no fraud and the second by the individuals with fraud. For classification phase, decision tree algorithm can be applied by specifying the known classes. This algorithm should be applied to each cluster formed in the clustering phase to distinguish the individuals with good and bad behaviour. The basic idea of decision tree is the recursive partitioning of data. The algorithm will start by grouping the data together in the root node and then decompose it in two child nodes according to the values of the attributes. This procedure would be repeated at every node until all nodes of the trees are transformed into leaves with specified classes i.e. fraud and no fraud.

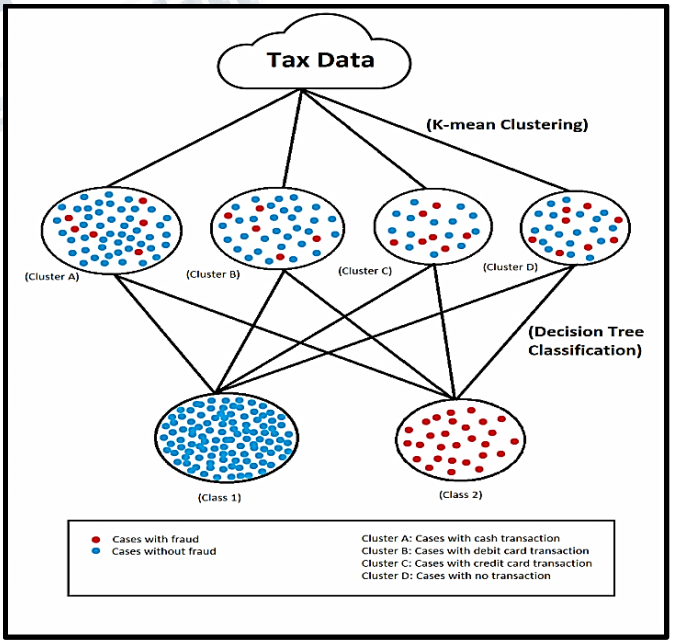

Fig. 2: AN ILLUSTRATION OF CLUSTERING AND CLASSIFICATION

v) FRAUD DETECTION

For fraud detection, multilayer feed forward neural networks [7] can be applied. The multi-layer feed-forward network is a type of neural network model commonly used for classification and grouping. The network consists of neurons ordered into layers: input, output and hidden layers. For each input case, it associates the input attribute with the desired output. This association is done by adjusting the weights of the network in order to minimize the prediction error; this learning method is called back propagation. This method works in two stages: training stage and prediction stage. During the first stage, output is calculated based on the inputs and the initial weights provided to the network and the prediction error is also calculated. During the second stage, the error is calculated backwards through the network from the output nodes to the input nodes, getting an error at each node. The weights are updated with each iteration through a gradient descent method until the network converges to a state with a minimum error, called a stable state. This state will allow the classification and grouping of all training patterns with least chances of errors. The network in the proposed system can be trained using the refined data by the classification phase. One of the major complexity with the neural networks is to define the number of layers and hidden nodes and also the number of iterations or epochs. To definethese characteristics, several number of cycles and nodes in the hidden layer should be taken into account through trial and error in order to establish an appropriate value. Once the system reaches a stable state, it can detect the cases with fraud and no fraud when provided with new test data.

IV. CONCLUSION

To successfully prevent tax fraud, we first need to detect the fraud criminals and their patterns of committing the fraud. This proposed systemfollows a conventional flow of data analysis in order to characterize and detect the probable tax evaders. Using the available data on tax fraud, features are identified that can be useful in the characterization of taxpayers. The two algorithms, K-mean Clustering grouped together the tax payers with similar behaviour and decision tree classification classified the fraudulent and innocent tax payers and detectedthe patterns in their transactions. These characteristics and patterns are used to train the Multilevel Feed Forward Neural Network and once the network reaches its stability, it can identify the tax fraud.

V. FUTURE SCOPE

For future work, we aim towards considering new attributes and a ide coverage for the tax domain in order to cope with the upcoming trends in tax fraud. We can also extend the study by exploring and implementing other methods or techniques of data analysis with fewer complexities and more accuracy in order to improve the tax fraud detection system.

REFERENCES

1. Pamela Castellón González, Juan D. Velásquez, “Characterization and detection of taxpayers with false invoices using data mining techniques”, Expert Systems with Applications: An International Journal, Volume 40, Issue 5, Pages 1427-1436, April, 2013.
2. Roung-Shiunn Wu, C.S. Ou, Hui-ying Lin, She-I Chang, David C. Yen, “Using data mining technique to enhance tax evasion detection performance”, Expert Systems with Applications: An International Journal, Volume 39, Issue 10, Pages 8769-8777 , August, 2012.
3. Fend Tian, Tian Lan, Kuo-Ming Chao, Nick Godwin, “Mining suspicious Tax Evasion groups in Big Data”, IEEE Transactions on Knowledge and Data Engineering, Volume 28, Issue 10, Pages 2651 - 2664, October 1st, 2016.
4. Freddy Duitama-Munoz, Julian D. Arias-Lundono, “Fraud detection in big data using supervised and semi-supervised learning techniques”, IEEE Columbian Conference on Communications and Computing, August, 2017.
5. You Dai, Jin Yan, Xiaoxin Tang, Han Zhao, Minyi Gua, “Online credit card fraud detection: A hybrid framework with Big Data technologies”, TrustCom/BigDataSE/ISPA 2016 IEEE, August, 2016.
6. Andrea Trevino, “Introduction to K-means Clustering”, 2016, [Online] Available: https://www.datascience.com /blog /k-means-clustering, [Accessed: February 15, 2018].
7. Bc. Miroslav Hlav´a˘cek, “Multilayer feedforward neural networks based on multi-valued neurons”, 2014, [Online] Available: https://is.muni.cz /th/359995/ fi_m/dp_text.pdf, [Accessed: February 16, 2018].
8. Barry de Ville, “Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner”, SAS Institute Inc., Cary, NC, USA, 2006.
9. Amir Gandomi, Murtaza Haider, “Beyond the hype: Big data concepts, methods, and analytics”, International Journal of Information Management, Volume 35, Issue 2, Pages 137-144, April 2015.
10. Investopedia, “Tax Evasion”, [Online] Available: https://www.investopedia.com/terms/t/taxevasion.asp, [Accessed: February 15, 2018]