Tax Evasion Detection with Graph Based Approach

Kartik A. Damdhar
Sinhgad Institute of Technology, Lonavala, India

Abstract— A tax is the source of government funding. The purpose of tax is to increase revenue to fund government. The money paid by taxpayers in taxes goes to many places. It is used to paying the salaries of government workers, tax money also help to support common resources, such as police and firefighters. Tax money helps to ensure the roads you travel on are safe and well-maintained. Taxes fund public libraries and parks. Tax evasion is increased so tax evasion detection is very important in current status to avoid loss of government funding. Taxpayers are required to store and update, on an annual basis, a set of documents and information relating to international transactions or specified domestic transactions. In recent work on tax evasion detection is done but it is not effective some drawbacks are there. This article gives an introduction to related work done in tax evasion detection and describes the methods of tax evasion. Auditing is very important to find out tax evasion, and data mining techniques are applied to select record for audit, also data mining techniques are applied in tax evasion detection.

Keywords: KDD, VAT, Tax Evasion Detection with Graph Based Approach

I. INTRODUCTION

The main reason that data mining has attracted a great deal of attention in information industry in recent years is due to the wide availability of huge amounts of data and the need for turning such data into useful information and knowledge. Data mining is nothing but the extracting or mining knowledge from large amounts of data. Many people uses data mining as an alternate for term, Knowledge Discovery in Databases", or KDD. Alternatively, data mining is essential and important step in the process of knowledge discovery in databases.

Tax evasion and tax fraud have been a constant issue for tax administrations, especially when pertaining to developing countries. While it is true that taxes are the source of government earning, the reality is that they send a very important signal about the commitment and effectiveness with which the State can carry out its functions and restrict access to other sources of income.

Tax evasion is illegal evasion of taxes by individuals and corporations. The number of annual tax related business records is up to 1 billion, the daily peak of these records is up to ten million. This volume of data challenges traditional data mining based methods of tax evasion. The results of the clustering based and neural network based methods are not explainable and their tax evasion identification efficiency is low. When talking about the properties of big data, traditional data mining- based methods have their limitations. The classification- based methods need a set of sample data for training, which means the data need to be manually labeled before training takes place. Moreover, the trained model is sensitive to the sample data and will be out- of-date if behaviors in tax evasion change. In addition, the results derived from clustering-based methods and neural network-based methods are difficult to explain and trace. The worse thing is that the above mentioned data-mining-based methods need to search and evaluate each transaction in the tax- oriented big data before\ reliable outcomes can be derived.

The proposed method is more effective and efficient than the existing approaches, as it aims to select the suspicious relations first via other related data sources and then identify those suspicious transactions.

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.

III. PROPOSED SYSTEM

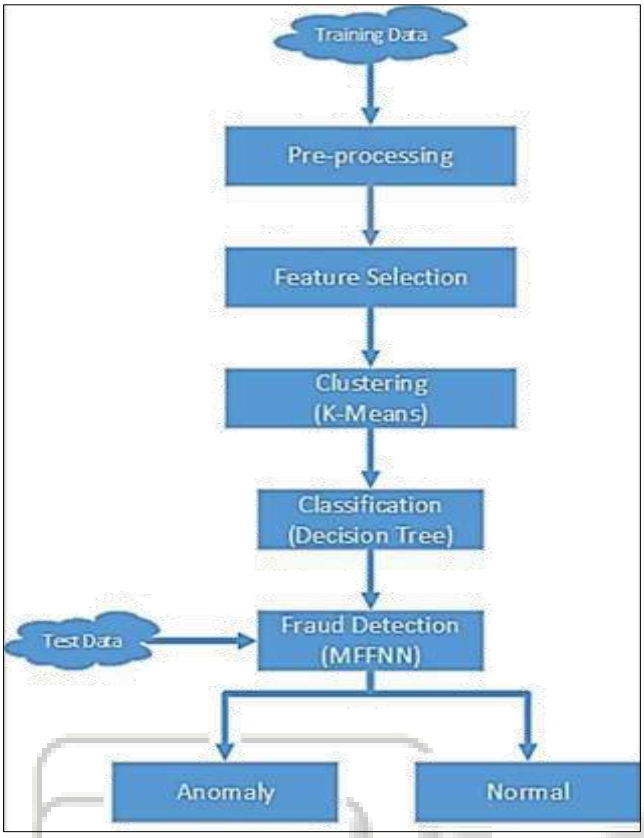

Fig. 1: Data Flow Diagram of Proposed System

A. 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.

B. 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.

C. 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 associaed 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.

D. 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.

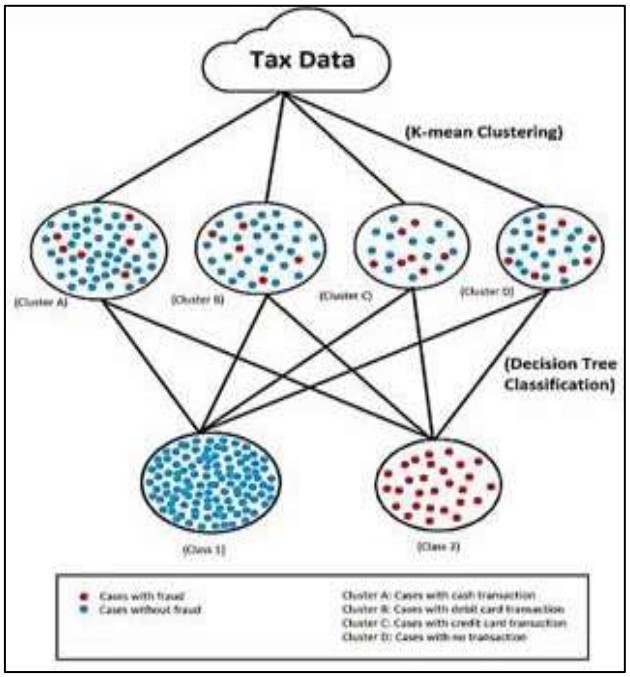

Fig. 2: An Illustration of Clustering and Classification

E. 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 define these characteristics, several number of cycles and nodes in the hidden layer should be taken int o 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. PROPOSED GRAPH BASED APPROACH

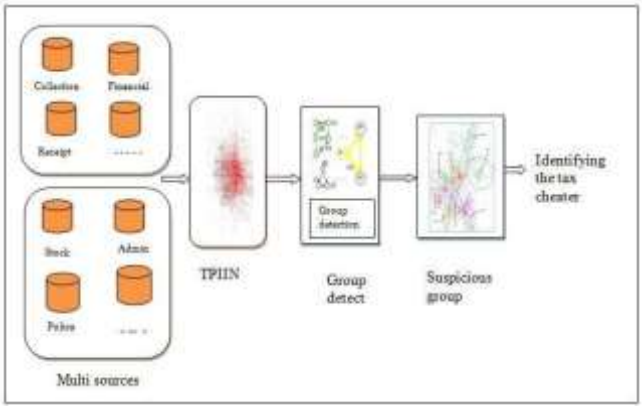

Fig. 3: System architecture for proposed system

Proposed work of finding suspicios group based on graph theory.

It contains two main phases:

A. Building TPIIN

Using database that contains company, director and transaction database graph will be generated. Graph shows nodes with different colors. Companies and directors will show with nodes and edges shows relation between nodes.

B. Finding suspicious groups

In this phase graph will be taken as input. That graph is called taxpayers interest interacted network. Using that heterogeneous network patterns will be generated and matched to find out suspicious tax evasion groups.

V. ANALYSIS OF ALGORITHM

INPUT: Database (company, directors, transaction data).

OUTPUT: Suspicious groups

- Step 1: Create graph from database.
- Step 2: abstract trading relationship and saving in trading list.
- Step 3.abstract antecedent relationship save as a antecedent list data.
- Step 4: Generate patterns from graph
- Step 5: pattern matching algorithm applied on patterns to find out suspicious groups.
- Step 6: Return suspicious groups

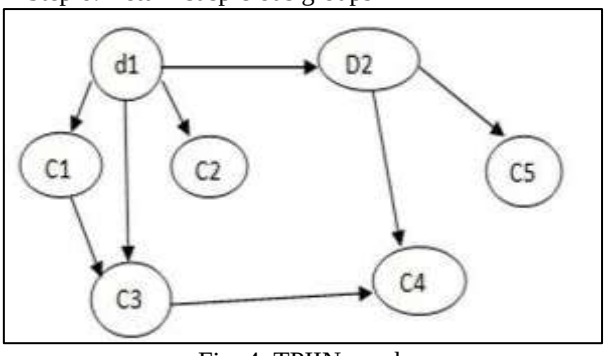

Fig. 4: TPIIN graph

- Step 1.Using data1
- Step 2. By using above generated graph trading data and antecedent data is generated.

Trading data

1	C1	C3
2	C3	C4

Antecedent data

1	D1	C1
2	D1	C2
3	D1	C3
4	D1	D2
5	D2	C4
6	D2	C5

- Step 3.Pattern sgenerated during trading and antecedent data.
  D1->c1->c3
  D1->c3->c4 D1->c2
  D1->d2->c4
  D1->c1->c3->c4
- Step 4. Patterns D1->d2->c4 and D1->c1->c3->c4 are matched. It has suspicious group.

VI. CONCLUSION

Tax evasion is illegal evasion of payable taxes. Due to large database finding tax evasion is difficult for tax administrator. Auditing and Tax inspection is important and effective but checking all records is time consuming. Large volume of data is challenge for traditional data mining methods Graph based approach helps to find out tax evasion. Graph based approach is based on graph theory. It finds out patterns from data and suspicious groups of tax evasion using pattern matching. So to find out tax evasion it is very useful and reduces time with increased accuracy.

To successfully prevent tax fraud, we first need to detect the fraud criminals and their patterns of committing the fraud. This proposed system follows 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 detected the 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.

VII. FUTURE SCOPE

For future work, we aim towards considering new attributes and a idea 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] W. F. Fox, L. Lunab, and G. Schau,"Destination Taxation and Evasion: Evidence from U.S. Inter-State Commodity Flows,"Journal of Accounting and Economics,vol.57, no. 1, pp. 43–57, Feb. 2014.

[2] P. C. Gonz´l´cleza and J. D. Vel´l´csquez, "Characterization and Detection of Taxpayers with False Invoices Using Data Mining Techniques, "Expert Systems with Applications,vol.40, no. 5, pp. 1427–1436, Apr.2013.

[3] Yizhou Sun and Jiawei Han, "Meta-Path-Based Search And Mining in Heterogeneous Information Networks," Tsinghua Science and Technology, vol. 18, no. 4, pp. 329-338, Aug. 2013.

[4] Yun Xiong, Yangyong Zhu, and Philip S. Yu, "Top-k Similarity Join in Heterogeneous Information Networks," IEEE Trans. Knowledge and Data Engineering, vol. 27, no. 6, pp. 1710-1723, Jun. 2015.

[5] Andrea Trevino, "Introduction to K-means Clustering", 2016, [Online] Available: https://www.datascience.com/ blog /k-means- clustering, [Accessed: February 15, 2018].

[6] Komal Rokade and Rashmi Mane ,"More Focus on Tax Evasion Detection with Graph Based Approach".

[7] Mehdi Samee Rad and Asadollah Shahbahram"HIGH PERFORMANCE IMPLEMENTATION OF TAX FRAUD DETECTION ALGORITHM".

[8] Yashashwita Shukla, Neena Sidhu, Akshita Jain, T.B. Patil, S.T. Sawant-Patil," Big Data Analytics Based Approach to Tax Evasion Detection" Vol 5, Issue 3, March 2018

