Predicting the Fraudulent Claims of Tax Payers A Case of Boditi Town Revenue Authority, SNNPR, Ethiopia

Biruk Wodajo, Dr.Sreedhar Appalabatla, Dr. Telkapalli Murali Krishna

Abstract: Tax is the main source of revenue for any government. Hitches in tax collection directly affects administrative activities. Fraudulent practices such as lowering down the income declaration are commonly observed. Fraudulent claims account for a significant portion of all claims received by auditors, and cost billions of birr annually.This study is initiated with the aim of exploring the potential applicability of data mining technologies to predict fraudulent claims of its kind. Boditi Town Revenue Authority (BTRA) is used as a case for the study. K-Means clustering algorithm is employed to find the natural grouping of the different tax claims as fraud or non-fraud. The resulting cluster is then used for developing the classification model. The classification task of this study is carried out using the J48 decision tree and Naïve Bayes algorithm in order to create model that best predicts fraud / suspicious tax claims. Primary data was collected using interview and observation. For the experiment, the collected tax payers' dataset is preprocessed to remove outliers, select relevant attributes, integrate data and derive attributes. In this study, different characteristics of the Boditi Town Revenue Authority (BTRA) customers' data were collected from their database system called SIGTAS. A total of 5374 tax payers' records are used for training the models. The model developed using the J48 decision tree algorithm has showed highest classification accuracy of 98.79% further tested with the 1612 testing dataset and scored a prediction accuracy of 98.39%. The results of this study have showed that the data mining techniques can be valuable for tax fraud detection. Hence future research directions are pointed out to come up with an applicable system in the area.

Keywords: Data Mining Techniques, BTRA, Fraudulent claims, SIGTAS, Tax. Classification, Prediction

1 INTRODUCTION

Boditi town revenue authority office has been established in 1996 E.C as alone .But before 1996 E.C it was in Boditi town finance office as one department by collecting tax from taxpayers for the development of town administration and to pay salaries for government employers. Since taxes are important sources of public revenue, the existence of collective consumption of goods and services necessitates putting some of our income into government hands.

Fraud is one of the main problem in tax collection activity. Outdated ways of data analysis have been in practice since long time as a means of detecting fraud. They require multifaceted and time-consuming researchto deal with different spheres of knowledge like financial, economics, business practices and law.

Frauds are perpetrated by parties and organizations to secure personal or business advantage through unlawful act to obtain money, property, services or to avoid payment or loss of services.

- *Biruk Wodajo, is currently pursuing Master's degree program in Information Technology in Woliata Sodo University, Ethiopia. E-mail: birukwodajo@gmail.com*
- *Dr. Sreedhar Appalabatla is currently working as Assistant Professor in School of Informatics in Woliata Sodo University, Ethiopia. E-mail: appalabatla.s@gmail.com*
- *Dr. Telkapalli Murali Krishna is currently working as Assistant Professor in School of Informatics in Woliata Sodo University, Ethiopia. E-mail:* murali2007tel@gmail.com

Fraud management is a knowledge-intensive activity. AI techniques used for fraud management includes data mining to classify, cluster, and segment the data and automatically find associations and rules among the data that may signify interesting patterns, including those related to fraud. Organizations use this information to detect existing fraud and non-compliance so as to prevent future occurrences.

Two primary functions of Data mining are prediction and description. Prediction involves finding unknown values/relationships/patterns from known values, and description provides interpretation of a large database. Classification is useful for prediction, whereas clustering, pattern discovery and deviation detection are for description of patterns in the data. Its primary goal is to extract knowledge from data to support the decision-making, planning and problem solving process

Data mining enables data exploration and analysis without any specific hypothesis in mind, as opposed to traditional statistical analysis, in which experiments are designed around a particular hypothesis. While this openness adds a strong exploratory aspect to data mining projects, it also requires that organizations use a systematic approach in order to achieve usable results. The application of Data Mining techniques for commercial classification is a productive research area.

2 STATEMENT OF THE PROBLEM

According to Wolaita Zone Revenue Authority (WZRA) annual plan (2004-2008 E.C) different documents described,

currently Boditi Town Revenue Authority (BTRA) is one of the three reform towns in the zone which don't collect enough amount tax from the taxpayers properly. The main responsibility of Boditi town revenue authority (BTRA) is to collect tax from all business sectors to change the government plan into practice. Now a day Boditi town revenue authority (BTRA) has given to collect 43,232,023.78 birr but the total budget required is 87,022,222 birr to the town for the government employers and for the development of the town so which is around 50% of the total budget. In 3 - 4 years BTRA is expected to collect 100% budget for the town without government treasury in order to raise the revenue to finance the town government expenditure, redistribution of wealth income to promote the welfare and equality of the citizens and further regulating the economy there by creating enabling environment for business thrives.

During Tax collection the main problem of the authority is fraudulent protection on taxpayer's declaration. To prevent this problem, the first task should be investigating the taxpayer's declaration data.

2.1 General Objective

This research aims to create a predictive model that determine the fraudulent claims of taxpayers for the purposes of developing an effective tax collection policy by Boditi town Revenues Authority.

2.2 Research Questions

Therefore, the current research is intended to find answers for the following main research questions.

- What is the pattern that characterizes whether a given claim is fraudulent or not?
- Which data mining technique is more appropriate to identify /determining factors for fraud detection?

2.3 Scope and Limitation of Study

Current research is exploring the applicability of DM for fraud claim prediction and detection in BTRA focusing on the audit process and taxpayers financial statements task process owner.

2.4 Research Methodology

Primary data was collected by interviewing concerned auditors and tax collecting department as well as through observation. Relevant literatures on data mining techniques and fraud were reviewed. The potential of data mining in general and particularly successful data mining applications in fraud prevention were assessed. In this research WEKA (open source software) is employed to implement most of the technical aspects of the CRISP-DM standard data mining methodology that has been adopted. Business understanding, data understanding, data preprocessing, selection of modeling technique, model building and model evaluation has been undertaken.

3 BUSINESS AND DATA UNDERSTANDING

Boditi town revenue authority office has been established in 1996 E.C as alone .But before 1996 E.C it was in Boditi town finance office as one department by collecting tax from taxpayers for the development of town administration and to pay different government employers salary. BTRA has divided the type of taxes as, indirect and direct tax. Direct Taxes include Withholding Tax, Income Tax, rental Tax, Business Profit Tax, Cost Sharing, Sched D-Games of Chance. Indirect Taxes are Value Added Tax (VAT), Turnover Tax, and Excise Tax.

While the companies register in BTRA, they have their own Tax Identification Number (TIN). This number is unique. Every tax payer should have this number. This number is used to uniquely identify a tax payer for tax collection purpose.

The taxpayers categorized based on their yearly income. Category "A" taxpayers income is more than 500,000 Birr. Category "B" taxpayer's income is between 100,000 Birr and 500,000 Birr. Category "A" department has 2496 customers or taxpayers and Category "B" also 2878 customers or taxpayers.

Originally there were around 5374 records. From this, 30% of the total around 1612 of the records are randomly selected for testing purpose. The rest of the dataset is used for training purpose.Subsequently data has been preprocessed, handled missing values, outliers, data inconsistency, noise removal. Out of 15 attributes 4 of them have registered with missing values.Accordingly, the researcher performed an appropriate action to clean the data.

Table 1. Handling missing values

No	Attribute name and their data type	% of missing values	Data types	Reason /techniques applied
1.	Net income	7	Numeric	The mean of this attribute
2.	Profit income tax	9	Numeric	The mean of this attribute
3.	Total gross income	10	Numeric	The mean of this attribute
4.	Depreciation expenses	8	Numeric	The mean of this attribute

Relevant attributes have been selected using appropriate attribute selection and ranking algorithms viz. GainRatioAttributeEval attribute evaluator and Ranker

search method. In addition, required derived attributes have also been created to carry out the experimentation

The Oracle and SIGTAS databases were used to carry out the data integration process. Data is then formatted into WEKA supportable format in the form of .CSV and .ARFF formats.

4 EXPERIMENTATION AND RESULTS

The main objective of this research is, discovering regularities for predicting and detecting fraudulent claims within the taxpayers" dataset. The model building phase in the DM process is carried out by a two-step DM approach viz. unsupervised clustering technique - K-Means clustering algorithm and the supervised classification techniques - J48 Decision Tree and Naïve Bayesian classification which are widely applicable in solving the current problem.

First, the given dataset is segmented into different clusters based on their similarity. Then the output of this clustering process is used for the classification task as an input. These techniques are implemented using WEKA 3.7.4 DM tool.

4.1 Experiment Design

In order to perform the model building process of this research, 5374 training dataset is used to train the clustering and classification models.

4.1.1 Cluster Modelling

The clustering task of segmenting tax claims is done using the WEKA simple K-Means algorithm. The need for determining the threshold values is solely to determine what patterns are discovered for each subsequent cluster models with K=2, and changing the other default parameters. This helps a lot to identify fraud suspicious segments easily. The cluster validity is a very difficult issue and subject of endless arguments since the view of good clustering is strictly related to the application domain and its specific requirements.

Table 2. Depicts the threshold values to assess the cluster result

Gross profit/Loss	Business Income /sales tax	Liquid Cash	Net Worth	Net Book Value
GP/L<=99,999 Low	PIT<=99,999 Low	LC<=99,999 Low	NW<=99,999 Low	NBV<=99,999 Low
GP/L<=499,999 Medium	PIT<=499,999 Medium	LC<=499,999 Medium	NW<=499,999 Medium	NBV<=499,999 MEDIUM
GP/L >=500,000 High	PIT>=500,000 High	LC>=500,000 High	NW>=500,000 High	NBV>=500,000 High

4.1.2 Choosing the Best Clustering model

The entire dataset output from the three different cluster experiments are available together with their segment distribution and the resulting cluster output. This enabled the domain experts to compare resulting tax claim segments from the different cluster experiments. Three clustering experiments have been carried out by the researcher using varying distance functions and seed values viz. Euclidean distance with K=2 and seed value = 10, Euclidean distance with K=2 and seed value = 100, Manhattan distance with K=2 and seed value = 1000.The value of within cluster sum of squared error is used to evaluate the goodness of clustering in the WEKA DM tool. The number of iteration the algorithm has undergone to converge. This shows the algorithm has relocated all misplaced data items in their correct classes within a few looping. The minimum valueexhibits K-Means algorithm has converged very soon.

Table 3. Sum of squared error values of the three cluster experimentations within the cluster

Experimentation	No of iterations	Sum of squared error within cluster
Euclidean distance with K=2 and seed value = 10	3	17540
Euclidean distance with K=2 and seed value = 100	4	17569
Manhattan distance with K=2 and seed value = 1000	3	19195

After in-depth analysis, the model developed in the first cluster experiment is selected as the final clustering model of this study.

4.1.3 Classification Modelling

In order to classify the records based on their values for the given cluster index, the model is trained by changing the default parameter values of the algorithms J48, and Naïve Bayes by employing the 10-fold cross validation and the percentage split classification models subsequently analyzed to measure the accuracy of the classifiers in categorizing the tax claims into specified classes.The classification accuracy of each of these models is reported and their performance is compared in classifying new instances of records. A separate test dataset is used for testing the performance of the classification models.

Selecting a better classification technique for building a model, which performs best in handling the prediction and detection of fraudulent tax claims, is one of the aims of this study.

Table 4. Accuracy of the J48 decision tree and Naïve Bayes models

Classification model	**Overall accuracy (5374)**	
	Correctly classified	**Misclassified**
Decision Tree	5309(98.79%)	65(1.21%)
Naïve Bayes	4285(79.73%)	1089(20.26%)
10-fold Cross Validation		

The result showed that J48 decision tree outperforms Naïve Bayes by 4.39% in identifying fraud suspicious tax claims. The reason for the J48 decision tree to perform better than Naïve Bayes is because of the linearity nature of the dataset. That means there is a clear demarcation point that can be defined by the algorithm to predict the class for a particular tax claim.

4.2 Evaluation of discovered Knowledge

From the decision tree developed in the aforementioned experiment, it is possible to find out a set of rules simply by traversing the decision tree and generating a rule for each leaf and making a combination of all the tests found on the path from the root to the leaf node [8]. This produces rules that are unambiguous in that it doesn't matter in what order they are executed. The following are some of the rules extracted from the decision tree.

Rule1. If Total gross income = low and Gross profit = low andInterest expense = low and Businessincome/sales= medium and Net book value = medium THEN cluster 0 (Non-fraud suspicious).

Rule2. If Total gross income = high and Gross profit = low and Interest expense = low and Business income/sales = low and Net book value = low THEN cluster 1 (Fraud suspicious).

Rule3. If Total gross income = low and Gross profit = high and Interest expense = high and Business income/sales = medium and Net book value = high THEN cluster 1 (Fraud suspicious).

Rule4. If Total gross income = medium and Gross profit = medium and Interest expense = medium and Businesincome/sales = medium and Net book value = medium THEN cluster 0 (Non-fraud suspicious).

Rule5. If Total gross income = low and Gross profit = low and Interest expense = high and Business income/sales = high and Net book value = high THEN cluster 1 (Fraud suspicious).

Rule6. If Total gross income = low and Gross profit = low and Interest expense = low and Business income/sales = low and Net book value = low THEN cluster 0 (Non-fraud suspicious).

Rule7. If Total gross income = high and Gross profit = high and Interest expense = high and Business income/sales= high and Net book value = high THEN cluster 0 (Non-fraud suspicious).

Rule8. If Total gross income = high and Gross profit = high and Interest expense = low and Business income/sales= low and Net book value = low THEN cluster 1 (Fraud suspicious).

The rules that are presented above indicate the possible conditions in which a tax claim record could be classified in each of the fraud and non-fraud suspicious classes. Five of the total fourteen variables are used for constructing the decision tree model. These attributes are claim total gross income, gross profit, Business income/sales, Interest expense, and net book value which are basis for building the decision tree. From these, the generated decision tree has shown that the total gross income and net book value is the most determinant variable, which is the top splitting variable of the model.

5 CONCLUSION

In this research, an attempt has been made to apply the DM technology in support of detecting and predicting fraudulent tax claims. This research is mainly conducted for an academic purpose.

The data used in this research has been gathered from the BTRA office and SNNPR revenue authority office database. The study was then conducted in two sub phases. First clustering using the K-Means clustering algorithm for segmenting the data into the target classes of Fraud

suspicious and Non-Fraud suspicious. This result of the model complies with the authority's assumption that different rated claims are more of fraud suspicious claims.

Cluster records are then submitted for the classification module for model building using the J48 decision tree algorithm. By changing the training test options and the default parameter values of the algorithm, different decision tree models have been created. The model developed with the 10-fold cross validation with the default parameter values has shown a better classification accuracy of 98.79% on the training dataset, with the Total gross income as a splitting variable. This model is then valuated with a separate test dataset and scored an accuracy of 98.65% of classifying new tax datasets as fraud and non-fraud suspicious claims.

The results of this study have also shown that DM technology particularly the K-Means clustering and the J48 decision tree classification technique are well applicable in the efforts of tax fraud detection. However, the results of this study are found promising to be applied to address practical problems of tax fraud. This research work can contribute a lot towards a comprehensive study in this area in the future, in the context of our country.

REFERENCES

[1] Hand, D. 2006. "Principles of data mining". Prentice-Hall of India, New Delhi, India.
[2] Piatetsky. "From Data Mining to Knowledge Discovery in Databases" American Association for Artificial Intelligence, USA.
[3] Chapman, P Clinton, J Kerber, R Khabaza, T Reinartz, T Shearer, C and Wirth, R. (2000). "CRISP- DM 1.0: Step-by-Step data mining guide", SPSS Inc., USA.
[4] Organization for Economic Co-operation and Development, 2004. "Compliance Risk Management: Managing and Improving Tax compliance". OECD Press. New York, USA.
[5] SAS Institute Inc. 2003. "Finding the solution to data mining: Exploring the features and components of enterprise miner", Release 4.1, SAS Institute Inc.
[6] Palshikar, G. 2002. Data Analysis Techniques for Fraud Detection.
[7] Singh, Y. and Chauhan, A. 2009. Neural networks in data mining. India: Journal of theoretical and applied Information Technology. pp. 37-42.
[8] Han, J. and Kamber, M. 2006. Data Mining: Concepts and Techniques, 2nd ed. Morgan Kaufman publishers, San Francisco.
[9] Kantardzic, M., 2002. Data mining: Concepts, models, methods, and algorithm. Wiley IEEE Press.
[10] Gillett, P. R., & Uddin, N. (2005). CFO intentions of fraudulent financial reporting. Auditing: A Journal of Practice & Theory, 24(1), 55-75.
[11] Deshpande, S. P. and Thakare, V.M., 2010. Data mining system and applications: A review.
[12] Cao. L, 2007. Fraud detection using Data mining. Wiley IEEE Press, London
[13] "Application of Data Mining Technology to Support Fraud Protection, he Case of Ethiopian Revenue and Custom Authority"
[14] Qui, M. Davis, S. and Ikem, F. 2010. Evaluation of clustering techniques in data mining tools. Issues in Information Systems, Vol. 5(1), pp. 254-160.
[15] Pham, DT. Dimov, SS. And Nguyen, CD. 2005. Selection of K in K-means clustering, Journal of Mechanical Engineering Science, Vol. 219, pp. 103-119.
[16] Koh, C and Gervais, G. 2010. Fraud detection using data mining techniques: Applications in the motor insurance industry's. Singapore