Detecting High Risk Taxpayers Using Apriori Algorithm

Jayraj Patil[1], Payal Patil[2], Smita Patil[3], GovindRao Mettu[4]
Student[1,2,3],Faculty[4]
Department of Computer Engineering,
Pillai HOC College of Engineering and Technology, Rasayani, India

***Abstract*-Taxpayers have an Identification number, a reference number issued by a government to its citizens.A taxpayer is a person or organization (such as a company) subject to a tax on income. Risk refers to a set of events that lead to loss but risk from the tax perspective refers to the taxpayers' behaviors that may lead to negligence from the public property by the taxpayers due to tax evasion. Such actions cause unusual volatilities in the amounts envisaged in the government budgeting. The fiscal and financial transactions outside the scope of the precautionary bound and failure to achieve the expected revenues of the country.One of the most important types of tax risks is concealing the information on buying, selling and contracts that in case of being uncovered in the financial sector it leads to the issuance of amendments to taxpayers.The main purpose of this paper is to analyze, design and implement a system to extract high risk taxpayers and provide a model to forecast the amount of tax assessment notification of the taxpayers for the coming years so that it would play the role of the assistance system for the tax experts to issue the assessment notifications with realistic amounts during the assessment and tax audit to prevent major errors in the tax assessment.In order to test the proposed algorithm, some real data hasbeen collected from a tax office. In the implementation of highrisktaxpayer detection system the high-risk taxpayers aredetected by five methods: Amendment approach, Volatility approach, Diagnostic approach, Statement approach, Colorful method also used Apriori Algorithm for detection of high risk taxpayers analysis. If the taxpayer has an income higher than the taxable incomeand their definite form amounts do not comply with the amountspredicted by the regression, he is identified as Grade 1 high risk.**

Keywords: - Tax, AprioriAlgorithem,Detecting Taxpayer

I. INTRODUCTION

Around the world today, tax authorities are experiencing growing pressure to collect extra tax revenues, to discover underreporting taxpayers, and predict the irregular behavior of non-paying taxpayers. Most tax authorities require to collect tax data from a number of independent sources and perform data matching and checking with other sources to find cases ofnon-compliance. As a result, tax evasion detection performance has been rather limited in the absence of information technology tools. Tax evasion is mostly performed by the taxpayers to reduce tax liability and this illegal action is usually performed to misrepresent the financial facts to government and tax authorities by providing false tax reporting, such as declaring less income, less profit and more or exaggerated costs. Tax evasion measurement for each country represents the tax gap in that country. Gary Becker the winner of Nobel Prize and economist in 1972 presented an economic model for tax evasion and stated that tax evasion can be considered as the main source of reduced government tax revenue and tax evasion level depends on the possibility of incorrect diagnosis by tax officials and level of criminal penalties and criminal law of that country. They have stated that tax evasion is directly associated with tax rates, unemployment rates, the public revenue level and discontent with the government. In order to extract colorful taxpayers in this paper, a data model is built for each taxpayer. The model has been used as an assistant system for experts in the tax offices and obtained results were close to real results. By the mechanics of determining the tax payable. When a company receives a false invoice, it simulates a purchase that never existed, thus increasing its tax credit fraudulently and decreasing VAT payment. Also, there is a decrease of payment in the income tax due to increased costs and expenditures declared. The falsity of the document may be material if the physical elements that make up the invoice have been adulterated, or ideological when the materiality of the document is not altered, but the operations recorded in it are adulterated or nonexistent. The latter is more complex and difficult to detect because it involves fictitious transactions in which an audit is required to examine the ales books and corrections, or cross referencing the information with suppliers. Moreover, these cases are more expensive for SII, as they require a greater amount of time.

II.PROBLEM STATEMENT

Some real data has been collected from a tax office. In the implementation of highrisk taxpayer detection system the high-risk taxpayers are detected by five methods. As shown in Figure the input for all three first methods is the same and different outputs are obtained depending on the applied method and techniques and also the outputs of the four different methods are used as aggregated for the colorful approach. According to the tax data for a 5-year period between 2010 and 2015 the following output is obtained.

II. Related Work

The methods applied in different sources that have studied auditing the high-risk taxpayers include classification and correlation and association rules [1], Bayesian networks [2] and support vector machines, genetic programming, logistic regression and the likelihood neural networks to identify fraud and risk. The fuzzy and probabilistic networks are also widely used in the forecast. Ghosh and Reilly (1994) offered a threelayer neural network model [3]. Also Dorronsoro, Ginel, Sanchez and Cruz (1997) used neural networks to detect fraud [4]. Shen, Tong and Deng (2007) applied Decision Tree, Logistic Regression and Neural Network methods for data classification for fraud detection [5]. Moreover Quah and Sriganesh (2007) applied neural networks clustering capabilities. Self-Organizing Map (SOM) method is a method based on neural network that uses learning techniques without experimental data [6]. Gadi, Wang and Lago (2008) employed Neural Network, Naive Bayes, Bayesian Network, Artificial Immune and Decision Tree for fraud detection [7]. Finally, Guo and Yangli (2008) used neural networks to detect fraud .By studying these methods the colorful approach presented in this article is improved in terms of performance.

III. Implemented system

Association rule generation is usually split up into two separate steps:

1. First, minimum support is applied to find all frequent itemsetsin a database.

2. Second, these frequent itemsets and the minimum confidence constraint are used to form rules.

While the second step is straight forward, the first step needs more attention.

Finding all frequent itemsets in a database is difficult since it involves searching all possible Item sets(item combinations). The set of possible itemsets is the power set over I and has size 2n− 1 (excluding the empty set which is not a valid itemset). Although the size of the powersetgrows exponentially in the number of items n in I, efficient search is possible using thedownward-closure property of support (also called anti-monotonicity) which guarantees that fora frequent itemset, all its subsets are also frequent and thus for an infrequent itemset, all itssupersets must also be infrequent. Exploiting this property, efficient algorithms (e.g., Apriori andEclat) can find all frequent itemsets.

Step1: Select Item set.

Step2: Calculate support count.

Step3: Delete Least Support Count.

Step4: Get Support count.

Step5: Select 2 Item set Up to Maximum.

Step6: Calculate All Support Count.

Step7:Calculate Confidence.

Step8: Get Result Which Is High Confidence(High Risk Taxpayer).

CALCULATING LINEAR REGRESSION ANALYSIS FOR AMOUNTS OF INCOME

x_1(year)	1390	1391	1 392	x=91
y_1(income)	100	130	190	y=140

$$\bar{b} = \frac{n\sum_{i=1}^{n}(\mathcal{X}i \ -\bar{\mathrm{X}})\,(\mathcal{y}i - \bar{\mathcal{y}})}{\sum_{i=1}^{n} \quad (\mathcal{X}i - \bar{\mathcal{X}})^2}$$

To select interesting rules from the set of all possible rules, constraints on various measures ofsignificance and interest can be used. The best-known constraints are minimum thresholds on support andTo select interesting rules from the set of all possible rules, constraints on various measures ofTo select interesting rules from the set of all possible rules, constraints on various measures ofsignificance and interest can be used. The best-known constraints are minimum thresholds on support andconfidence.

Support

The support supp(X) of an itemsetX is defined as the proportion of transactions in the data set whichcontain the itemset.
Supp(X)= no. of transactions which contain the itemset X / total no. of transactions

Confidence

The confidenceof a rule is defined:$\mathrm{con}f(\mathcal{X} \rightarrow \mathcal{Y}) = \mathrm{Supp}(\mathcal{X} \cup \mathcal{Y}) \,/\, (\mathcal{X})$

Lift

The lift of a rule is defined as:$lift(\mathcal{X} \rightarrow \mathcal{Y}) = \frac{\mathrm{Supp}(\mathcal{X} \cup \mathcal{Y})}{\mathrm{Supp}(\mathcal{Y}) * \mathrm{Supp}(\mathcal{X})}$

Sample usage of Apriori algorithm:

A large supermarket tracks sales data by Stock-keeping unit (SKU) for each item, and thus is able to know what items are typically purchased together. Apriori is a moderately efficient way to build a list of frequent purchased item pairs from this data. Let the database of transactions consist of the sets {1,2,3,4}, {1,2,3,4,5}, {2,3,4}, {2,3,5}, {1,2,4}, {1,3,4}, {2,3,4,5}, {1,3,4,5}, {3,4,5}, {1,2,3,5}. Each number corresponds to a product such as "butter" or "water". The first step of Apriori is to count up the frequencies, called the supports, of each member item separately:

Item	Count
1	6
2	7
3	9
4	8
5	6

We can define a minimum support level to qualify as "frequent," which depends on the context. For this case, let min support = 4. Therefore, all are frequent. The next step is to generate a list of all 2-pairs of the frequent items. Had any of the above items not been frequent, they wouldn't have been included as a possible member of possible 2-item pairs. In this way, Apriori prunes the tree of all possible sets. In next step we again select only these items (now 2-pairs are items) which are frequent (the pairs written in bold text):

Item	Count
{1,2}	4
{1,3}	5
{1,4}	5
{1,5}	3
{2,3}	6
{2,4}	5
{2,5}	4
{3,4}	7
{3,5}	6
{4,5}	4

We generate the list of all 3-triples of the frequent items (by connecting frequent pair with frequent single item).

Item	Count
{1,3,4}	4
{2,3,4}	4
{2,3,5}	4
{3,4,5}	4

The algorithm will end here because the pair {2,3,4,5} generated at the next step does not have the desired support. We will now apply the same algorithm on the same set of data considering that the min support is 5. We get the following results:

Step 1:

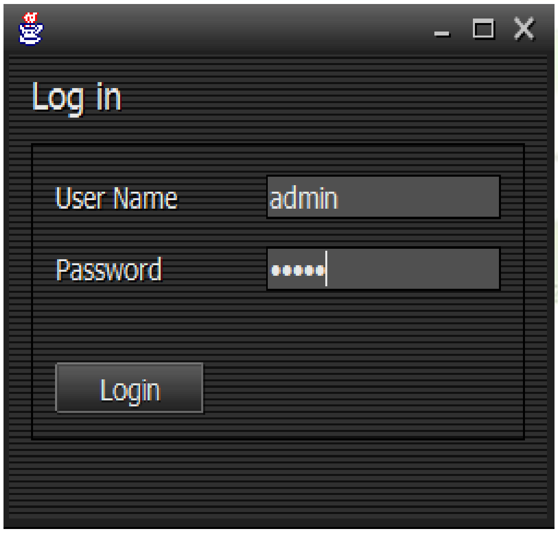
Item	count
1	6
2	7
3	9
4	8
5	6

Step 2:

Item	Count
{1,2}	4
{1,3}	5
{1,4}	5
{1,5}	3
{2,3}	6
{2,4}	5
{2,5}	4
{3,4}	7
{3,5}	6
{4,5}	4

The algorithm ends here because none of the 3-triples generated at Step 3 have de desired support.

IV. RESULTS



Log in

User Name: admin

Password: •••••

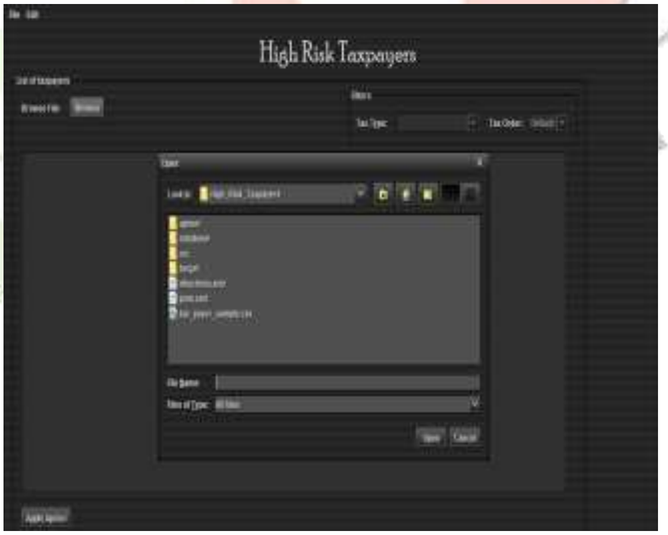
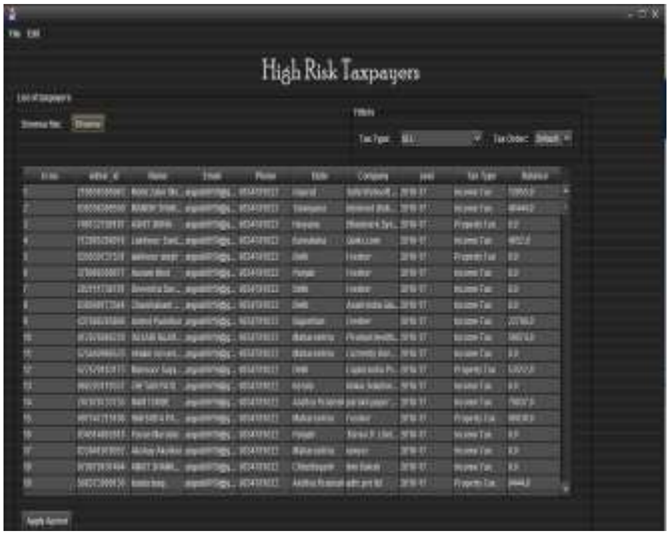
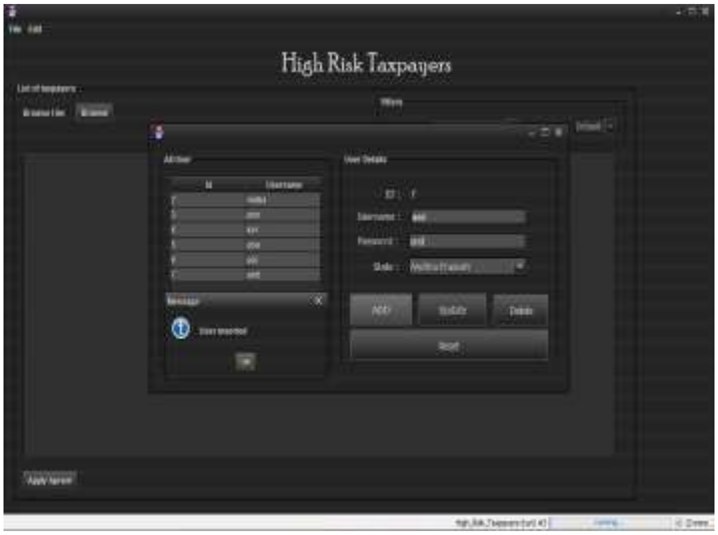


Figure.1 Login Page **figure.2 Selecting Database**

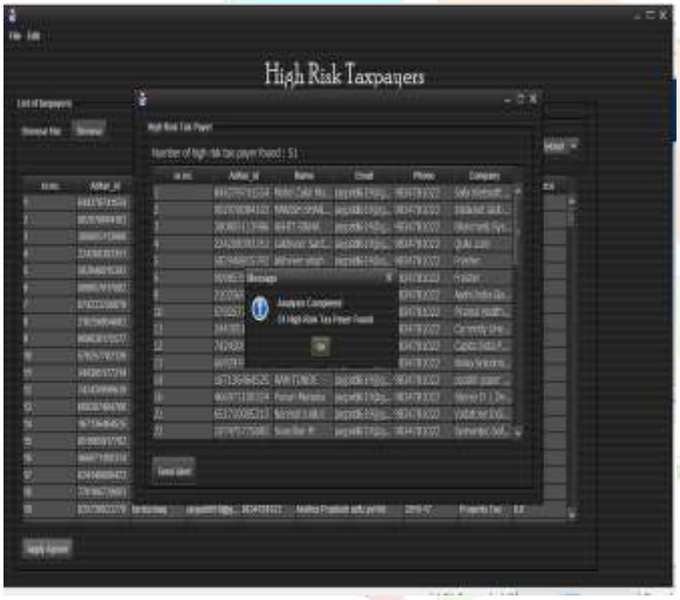




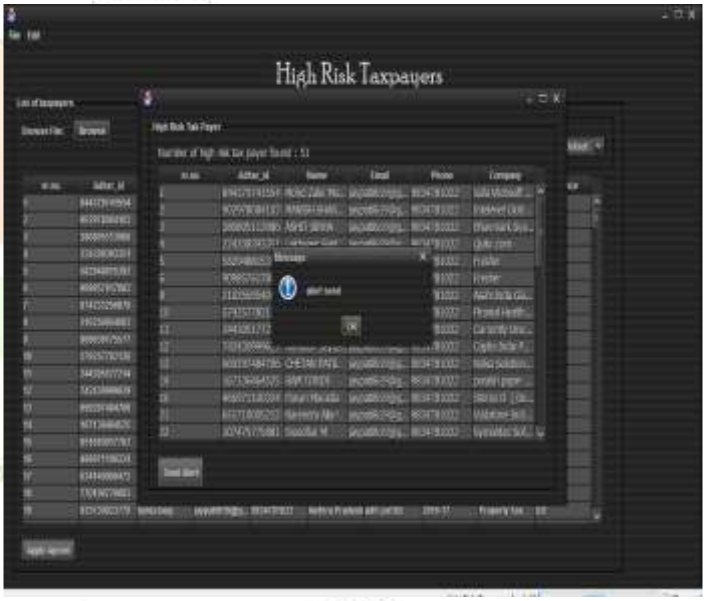
User Inserted

figure.3 Database

figure.4 Create User



Analysis Completed
51 High Risk Tax Payer Found



alert send

figure.4Detecting Of High Risk Taxpayer

figure.5 Send Alert

Conclusion

Various methods are implemented to extract the high risk taxpayers but in practice the colorful taxpayers' algorithm presented the best result compared to the mean standard deviation, job factor, amount based and amendment based methods (Figure 5). The power of this method is the combined use of regression techniques, support vector machine and prioritizing the high-income taxpayers.

eferences

[1] Mehdi SameeRad andAsadollahShahbahrami, "Detecting High Risk Taxpayers Using Data Mining Techniques",ICSPIS 2016, 14-15 Dec. 2016, Amirkabir University of Technology, Tehran, Iran.

[2] V. Ajay, D.V. Ashoka, V.N. Aradya. "Application of Data Mining Techniques for Defect Detection and Classification," In Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA), pp. 387-395.Springer,2015.

[3] P.C. González, J.D. Velásquez, "Characterization and detection of taxpayers with false invoices using data mining techniques", Expert Systems with Applications vol.40, no. 5, pp. 1427-1436, (2013).

[4] A. Ahmadi, A. Mohebbi, “Business Intelligence: data mining and optimization,” Amirkabir University Press, 2013

[5] R.S. Wu, C.S. O.U. H. Lin, S.I. Chang, DC Yen, "Using data mining technique to enhance tax evasion detection performance," Expert Systems with Applications vol.39, no. 10, pp. 8769-8777, (2012).

[6] Bond University, Central Michigan University, Deakin University; “Computational Data Mining Techniques in Automotive Insurance Fraud Detection”; Journal of Data Science 10(2012), 537-561

[7] A.S. Sabau, “Survey of Clustering based Financial Fraud Detection Research,” InformaticaEconomică vol. 16, no. 1/2012

[8] Price water house Coopers LLP (2009). "Global Economic Crime Survey". Retrieved June 29, 2011.