Detecting High Risk Taxpayers Using Data Mining Techniques

Mehdi Samee Rad
Islamic Azad University,
Rasht Branch Faculty of Engineering
Rasht, Iran
SameeRad@iauRasht.ac.ir

Asadollah Shahbahrami
Department of Computer Engineering,
Faculty of Engineering, University of Guilan
Rasht, Iran
Shahbahrami@guilan.ac.ir

***Abstract*— 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. But if it is uncovered in the due course, it leads to the non-fulfillment of tax collection and thus negative financial waves at the national level and ultimately leads to detrimental financial impact to the financial framework of states and countries. 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. To extract high risk taxpayers using the variance and the mean standard deviation the suspicious financial behavior is detected and then the previously supervised data that exist in the tax base as amendment forms are used to classify the taxpayers and also the job coefficient field is used and high risk occupations are identified and classified. One of the strongest and best practices in this field is the use of statistical and financial calculations in time domain. The main feature is the amount of taxable income based on which the purchasing, sales, revenue and profit can be calculated. By studying the volatilities and noise detection in the amounts paid by taxpayers during the past years and also creating linear regression analysis it has been possible to discover the risk levels and forecast of the tax assessment notification for the coming years. Also using this technique tax assessment notification error tolerance of the previous years is obtained. Finally the high risk taxpayer detection system known as HTS is provided with the best and quickest manner possible.**

Keywords*— *Data mining, task risk assessment, tax fraud detection, machine learning, artificial neural networks, big data.

I. INTRODUCTION

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.

II. RELATED WORKS

The methods applied in different sources that have studied auditing the high-risk taxpayers include classification and correlation and association rules [2], Bayesian networks [1] 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 three-layer neural network model [28]. Also Dorronsoro, Ginel, Sanchez and Cruz (1997) used neural networks to detect fraud [29]. Shen, Tong and Deng (2007) applied Decision Tree, Logistic Regression and Neural Network methods for data classification for fraud detection [30]. 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 [31]. Gadi, Wang and Lago (2008) employed Neural Network, Naive Bayes, Bayesian Network, Artificial Immune and Decision Tree for fraud detection [32]. Finally, Guo and Yangli (2008) used neural networks to detect fraud [33]. By studying these methods the colorful approach presented in this article is improved in terms of performance.

III. PROPOSED METHOD

A. Identification of high-risk taxpayers

Bart Baesens [27] et al have proposed a proper exploration conceptual model based on the factors that enable financial fraud in different jobs. The model is depicted in Figure 1.

978-1-5090-5820-4/16/$31.00 ©2016 IEEE

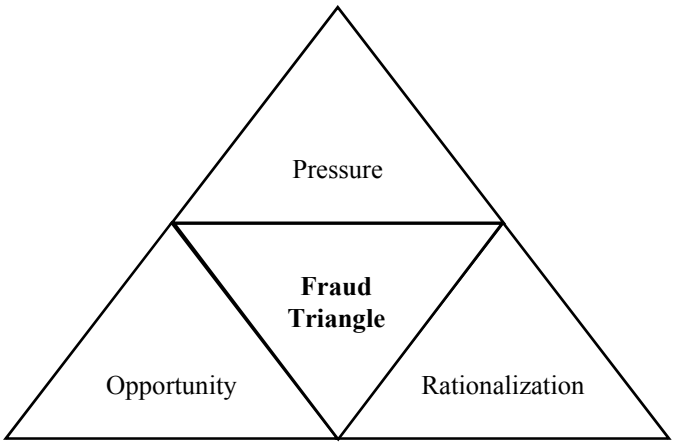

Figure 1: fraud triangle [27]

In order to test the proposed algorithm, some real data has been collected from a tax office. In the implementation of high-risk taxpayer detection system the high-risk taxpayers are detected by five methods. As shown in Figure 2 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.

B. Risk assessment by volatility

Based on tax experts' opinion the previous years' definite tax forms are extracted and refined (five years) then the assessed notifications are also studied. Then the following factors are deducted from the refined list:

- The ones that had the assessed and definite form in the final year are excluded
- The ones that had the definite form in the final year are excluded
- The ones that have been exempt from tax payment are excluded
- The ones that had no activity according to the center are excluded
- Priority of the income cases that are hereditary are reduced
- The noise and volatility of the assessment forms are calculated

The taxpayers that attempted to shift location are intelligently detected. Since Iranian Tax Administration Office's jobs system is locations based, it is necessary to detect a taxpayer in different areas and classes to analyze his tax behavior record. After collecting the result, the outputs are testes individually based on reviewing the records of high risk taxpayers case by case by tax officers and the output accuracy of high risk taxpayers' detection is analyzed 100% and the potential problems are resolved. Among 32,954 taxpayers handled by the risk assessment, 95 cases are detected as volatility high risk taxpayers that lead to the loss of public property. Also a section is developed as "regression forecast" inside the risk assessment system that enables the tax officers to forecast assessment notification amount.

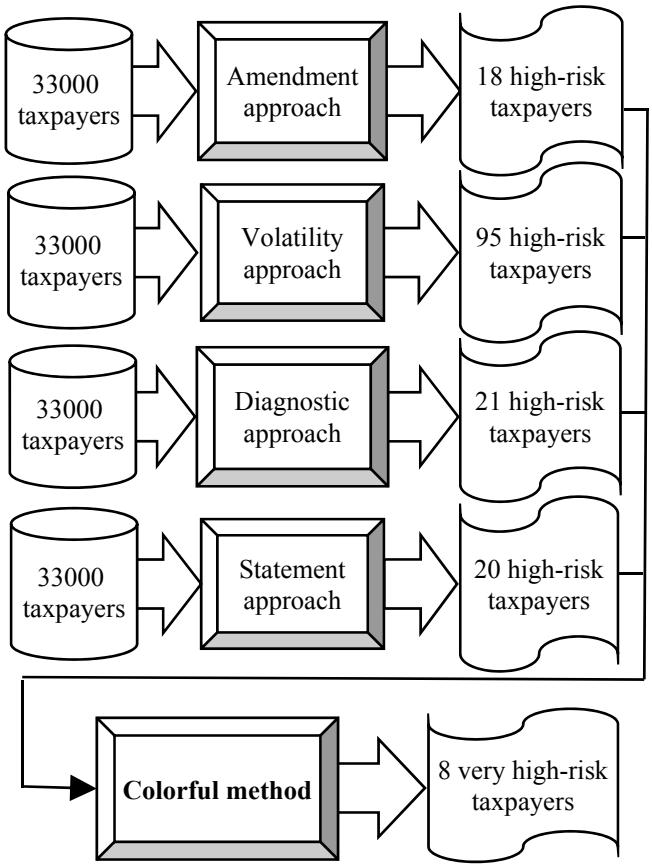

Figure 2: Implementing five algorithms on tax information

C. Risk assessment by upper limit of self-reported amounts by tax payers

Department of risk assessment using upper limit of self-reported amounts needed to run a query on the high volume of tax data and using the parallelism, parallel processing and Loop Inter Changing techniques the program speed is increased to more than 90 times and also the use of served memory is optimized and had a significant impact on correct system performance.

D. Risk assessment by mean probability method

According to the formula (1) if the tax data as $X_1, X_2, \ldots, X_n$ have the job coefficient $W_1, W_2, \ldots, W_n$, to calculate the weighted arithmetic mean with job factors effectiveness:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \qquad (1)$$

In which the weight for each job is equal with the job factor and X_i is the amount of taxable income. The equation 1 presents the weighted arithmetic mean of taxable income.

If the taxable income of all taxpayers is positive, it's essential to use the geometric mean that is presented in equation (2).

$$G = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{2}} \qquad (2)$$

The algorithm used to calculate and identify taxpayers' risk is as follows:

- Classifying the definite forms and removing the taxpayers with amendments.
- Obtaining the performance of supervisors without the effect of amendment form amount.
- Calculating the geometric mean for each taxpayer's taxable income.
- Calculating the mean taxable income minus the income assessed by the previous forms.
- Calculating the first part of the variance which is the power of each taxpayer.
- Obtaining standard deviations for each taxpayer.
- Specifying the location of the taxable income of each taxpayer in the range of standard deviation and tax risk assessment.
- Adding the acquired reputation and taxpayers' specifications per year.
- Summing the taxpayers' risk based on the case, scope and taxpayers' number.
- Testing the output accuracy.

In this the amount and distance of the definite taxable income per year of the mean and standard deviation as well as the place of standard deviation are included, i.e. as shown in Figure 3 standard deviation's precautionary bound is calculated by geometric and weighted means and bounding it by standard deviation and the model is formed.

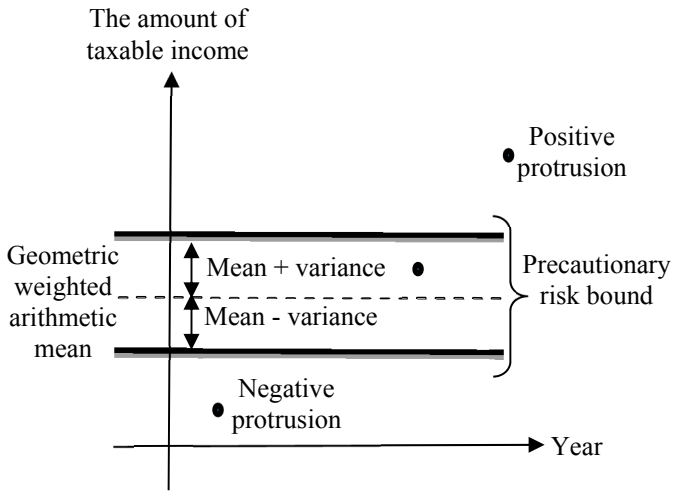

Figure 3: Calculating the mean and variance of probable taxpayer revenues

Taxpayers whose income has negative protrusion from the variance and standard deviation means i.e. they have a reduction in the declared tax and they are introduced as diverted high risks from the mean criterion.

Example: For the taxpayer with the income tax (in million Rls) based on Table (**1**) and equations (**3**), (**4**) and (**5**):

TABLE 1: CALCULATING THE VARIANCE AND STANDARD DEVIATION OF THE MEAN FOR THE TAXPAYER'S TAXABLE INCOME IN DIFFERENT YEARS

x_i **(year)**	1390	3191	1392	$\delta^2 = 1400$
y_i **(income)**	100	130	190	$\bar{x}$**= 140 (million Rls)**

By calculating the mean, SD and variance:

$$\text{Mean} \quad \bar{x} = \frac{420}{3} = \mathbf{140} \qquad (\mathbf{3})$$

$$SD \quad \delta = \sqrt{1400} = \mathbf{38} \qquad (\mathbf{4})$$

$$\text{Variance } \delta^2 = \frac{(100-140)^2 + (130-140)^2 + (190-140)^2}{3} = \mathbf{1400} \qquad (\mathbf{5})$$

E. Amount based high risks based on regression forecast

If the taxpayer has an income higher than the taxable income and their definite form amounts do not comply with the amounts predicted by the regression, he is identified as Grade 1 high risk. However after regression calculation, these individuals are divided into three groups based on the amount of taxable income and both purposes of "classification based on the income" and "noise assessment in the forecasted amount by regression" are met at the same time.

Regression calculation is done as $\hat{a} = \hat{y} - \bar{x}\hat{b}$ and $y_i = a + bx_i + \varepsilon_i \;\; i = 1, \ldots, n$. This technique is one of the most powerful methods by which it is possible to forecast the future tax revenues and perform risk assessment where Y is the dependent variable y, X is the independent variable, ε is he model error and a and b are constant values. $\hat{b}$ is calculated based on the equation 6 and to solve the linear regression fitness of the taxpayer with the taxable incomes based on Table 2:

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (6)$$

TABLE 2: CALCULATING LINEAR REGRESSION ANALYSIS FOR AMOUNTS OF INCOME

x_i **(year)**	13**90**	13**91**	13**92**	$\bar{x}$**=91**
y_i **(income)**	100	130	190	$\bar{y}$**=140**

By calculating the above amounts and according to the equation: b=45 and a=-3955. Therefore y_i=-3955+45x. Thus the forecasted amount for different years will be based on Table 3.

TABLE 3: CALCULATING THE AMOUNT OF TAXABLE INCOME WITH LINEAR REGRESSION

Performance year (Persian calendar)	13**93**	13**91**	13**90**
The amount of linear regression forecast (million Rials)	**230**	**140**	**95**

The algorithm that is used for regression calculation and risk assessment is summarized as follows:

- Forming the regression formula $Y_i=a+bX_i+E_i$
- Selecting the last amendment form and the largest definite taxable income.
- Obtaining the taxable income for each taxpayer.
- Calculating the mean taxable income of each taxpayer.
- Calculating the regression numerator and denominator.
- Calculating Y_i and regression numerator and denominator for all taxpayers.
- Calculating regression model fitness $Y_i=a+b*X_i$ per taxpayer.
- Those with a high value of the mean regression forecast for different years are amount based high-risks.
- Those with a high difference between definite form amount and the amount forecasted by the regression for that year are assessment based high-risks.
- Adding the acquired reputation and taxpayers' specifications to the output.
- Calculating the reduced amount compared to the regression forecast in all years and sorting by protrusion of the scope of regression.
- Those with a high mean regression forecasted amount are amount based high-risks.
- Testing the accuracy of the output by the tax experts.

F. Amendment and assessment high risks (Detecting the tax experts' errors)

For taxpayers who have declared α Rls income but after presenting the declaration, there are evidences suggesting that the income is more than the declared amount, a tax amendment form is issued and the main tax and the related fines are collected from them. These typed of data are called the supervised data and they are used to determine the high risk taxpayers. As the amendment and the definite form information and the amounts listed in the definite tax form of the previous years is obtained, the linear regression amount is forecasted and also the taxpayer's mean variance SD is available, the deviation in the assessment (tax expert's error) is easily calculated. All severe noises in the amendment amount or the definite tax form is obtained in all years and the distance between the definite taxable income and the mean and standard deviation and the place in SD is calculated and by including the amendment and job factor the standard deviations of the mean high risks are detected. Finally all outputs are integrated as the high risks and the taxpayer risks are counted and colorful high risks are detected. It means that the taxpayers with the maximum tax risk are identified and presented to the tax office.

The final cumulative algorithm to detect tax risk (approved by senior tax experts in Iranian National Tax Administration):

1. Obtaining the class, scope and functions and searching in job tax bank.
2. Finding the definite forms in all years, but in the same class and scope.
3. Finding all definite forms in other classes and scopes of a job.
4. If multiple jobs are found the user is asked about the job to be regression forecasted.
5. Regression calculation for the entered year by the operator.
6. Choosing the greatest amount among the amendment forms of the taxpayer.
7. Calculating the taxable income per taxpayer.
8. Calculation of Y_i and dividing the regression to the definite taxpayer forms.
9. Calculating and fitting the regression model $Y_i=a+b*X_i$
10. Collecting the final result.

In Figure 4 a form of final calculation that is a kind of support vector machine (SVM) is observed. The problem with the mean variance and standard deviation is that they are less sensitive to the changes with sharp edges but in this method the severe noise and financial changes are detectable. Regression method acts well in coping with protruding data or the protrusion from the tax security scope. It can be observed that in the case of the error appeared in the mean standard deviation is rectified and the system presents the output without error.

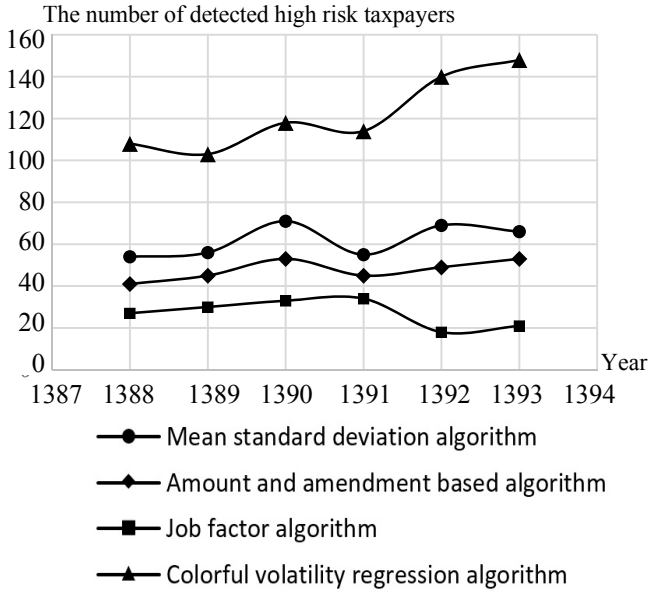

FIGURE 4: REGRESSION ALGORITHM BASED ON SUPPORT VECTOR MACHINE

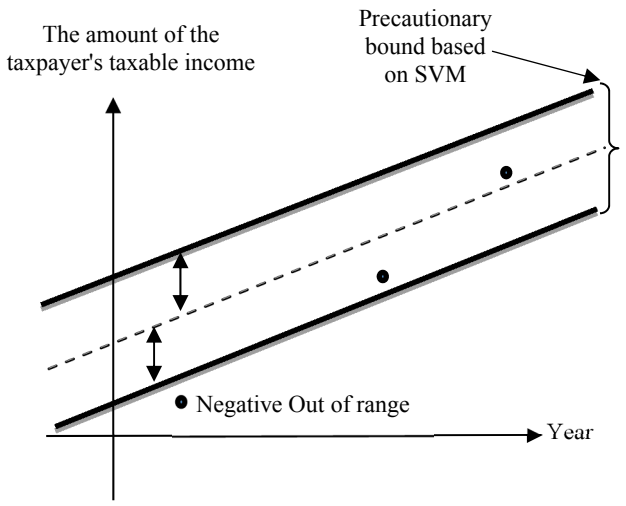

FIGURE 5: DETECTED HIGH RISK TAXPAYERS WITH COLORFUL METHOD

I. CONCLUSIONS

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.

REFERENCES

[1] M. SameeRad, A. Shahbahrami, "High performance implementation of tax fraud detection algorithm", IEEE 2015, ISBN: 978-1-5090-0138-5, Signal Processing and Intelligent Systems Conference (SPIS).

[2] G. Miner, J. Elder, R. Nisbet, "Handbook of Statistical Analysis and Data Mining Applications", ISBN: 9780123747655, Elsevier Inc (2009).

[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] 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).

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

[6] E.W.T. Ngai, Y. Hu, Y.H. Wong, Y. Chen, X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," Decision Support Systems vol.50, no. 3, pp. 559-569, (2011).

[7] K. R. Karkera, "Building Probabilistic Graphical Models with Python," Packt Publishing 2014, ISBN:978-1-78328-900-4, Open Source community experiances distilled, www.PacktPub.com

[8] 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.

[9] M.S. Abadeh, S. Mahmoodi, M. Taherparvar, "Application Data Mining," Niyaz Danesh Press, 2012

[10] J. Shahrabi, V. Shakor; "Data mining Concepts," Metalon Press, 2007

[11] A. Ahmadi, A. Mohebbi, "Business Intelligence: data mining and optimization," Amirkabir University Press, 2013

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

[13] F. Nonyelum, "Data Mining Application in credit card fraud detection system," Journal of Engineering Science and Technology Vol. 6, (2011).

[14] C. Sakoda, A. Nagasaki, T. Itoh, M. Ise, K. Miyashita, "Visualization for Assisting Rule Definition Tasks of Credit Card Fraud Detection Systems," Journal of Data Science, 2011

[15] P. Murugavel, M. Punithavalli, "Improved Hybrid Clustering and Distance-based Technique for Outlier Removal," International Journal on Computer Science and Engineering (IJCSE)

[16] Carlo Vercellis, "Business Intelligence: Data Mining and Optimization for Decision Making," Politecnico di Milano, Italy, WILEY 2009.

[17] V. Dheepa, R. Dhanapal; "Analysis of Credit Card Fraud Detection Methods," International Journal of Recent Trends Engineering, 2009

[18] A.S. Sabau, "Survey of Clustering based Financial Fraud Detection Research," Informatica Economică vol. 16, no. 1/2012

[19] A. Sharma, P.K. Panigrahi; "A Review of Financial Accounting Fraud Detection based on Data Mining Techniques," International Journal of Computer Applications (0975 – 8887) Volume 39– No.1, February 2012

[20] C. Lin, I. Lin, C. Wu, Y. Yang and J. Roan, "The application of decision tree and artificial neural network to income tax audit: the examples of profit-seeking enterprise income tax and individual income tax in Taiwan," Journal of the Chinese Institute of Engineers Vol. 35, No. 4, June 2012

[21] E. Kumar, A. Solanki, "A Combined Mining Approach and Application in Tax Administration," International Journal of Engineering and Technology Vol.2(2), 2010, 38-44

[22] C. Phua, V. Lee, K. Smith & R. Gayler, "A Comprehensive Survey of Data Mining-based Fraud Detection Research," Monash University Prees, 2010

[23] X. Liu, D. Pan, S. Chen, "Application of Hierarchical Clustering in Tax Inspection Case-selecting," National Natural Science Foundation of China, 2010

[24] Y. Wang, "Research on Rough Sets Theory Based Tax Data Mining," International Conference on Future Information Technology and Management Engineering, 2010

[25] S. Basta, F. Fassetti, M. Guarascio, G. Manco, "High Quality True-Positive Prediction for Fiscal Fraud Detection," IEEE International Conference on Data Mining Workshops, 2009.

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

[27] B. Baesens, V. Vlasselaer, W. Verbeke. "Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques". Wiley, 2015.

[28] G. and D.L. Reilly; "Credit Card Fraud Detection with a Neural-Network"; IEEE, vol. 3, pp. 621-630, 1994.

[29] Dorronsoro, Ginel, Sanchez, Cruz; "Neural fraud detection in credit card operations"; IEEE, vol. 8, pp. 827-843, 1997.

[30] Shen, Tong, Deng; "Application of Classification Models on Credit Card Fraud Detection"; IEEE, 2007.

[31] Quah, Sriganesh; "Real-time Credit Card Fraud Detection Using Computational" Intelligence, Elsevier,2007

[32] Gadi, Wang, Lago; "Comparison with Parametric Optimization in Credit Card Fraud Detection"; IEEE; 2008.

[33] Guo, Yangli; "Neural data minign for credit card fraud detection"; IEEE, vol 7, pp 3630 – 3634, 2008.