Print ISSN: 1738-3110 / Online ISSN 2093-7717
doi: 10.13106/jds.2014.vol12.no3.33.

# Prediction of Auditor Selection Using a Combination of PSO Algorithm and CART in Iran

Mahdi Salehi*, Sharifeh Kamalahmadi**, Mostafa Bahrami***

Received: November 28, 2013. Revised: February 02, 2014. Accepted: March 17, 2014.

## Abstract

**Purpose** - The purpose of this study was to predict the selection of independent auditors in the companies listed on the Tehran Stock Exchange (TSE) using a combination of PSO algorithm and CART. This study involves applied research.

**Design, approach and methodology** - The population consisted of all the companies listed on TSE during the period 2005-2010, and the sample included 576 data specimens from 95 companies during six consecutive years. The independent variables in the study were the financial ratios of the sample companies, which were analyzed using two data mining techniques, namely, PSO algorithm and CART.

**Results** - The results of this study showed that among the analyzed variables, total assets, current assets, audit fee, working capital, current ratio, debt ratio, solvency ratio, turnover, and capital were predictors of independent auditor selection.

**Conclusion** - The current study is practically the first to focus on this topic in the specific context of Iran. In this regard, the study may be valuable for application in developing countries.

**Keywords:** Independent Auditor Selection, PSO Algorithm, CART.

**JEL Classifications:** G23, G24, G32.

## 1. Introduction

The increasing dominance of security trading led to the prosperity of the capital market and involvement of small investors in this market. Governments adopted rules for regulating the management of joint-stock companies and the relationship between stockholders. Adoption of such rules and organization of security trading through stock exchanges were other factors that led to the maturity of joint-stock companies and the increasing number of a class of investors who neither directly participated in the management of companies nor were inclined to do so. A management that either consisted of large shareholders or was selected by them administered joint stock companies. The ongoing evolution in capital ownership and management in industrial countries led to the emergence of a new group of experienced, professional managers who, in spite of their full managerial authority, had little share in the capital of their businesses. This was an instantiation of the idea of separation of ownership and management. The new capital system expanded accounting duties and necessitated reports that could inform shareholders of the management of their capital. Meanwhile, the financial and credit markets expanded, and banks and credit institutions that provided a large part of the credits required for the current operations and investments of economic entities were informed about the financial statements of other companies and economic entities. Thus, the extent of accounting and financial reporting operations further expanded. The increasing number of joint-stock companies and development of capital and money markets promoted the goal of accounting from responding to the needs of a small number of capital owners to providing accounting information to managers and beneficiaries. This social role of accounting could not be played merely be the accountants employed by institutions, since the direct employment relationship forced them to accept the views of the management. In addition, these accountants were in favor of the management of the institution and they could not provide unbiased and reliable financial information that would meet the needs of various users who often had conflicting rights and interests. The solution to this problem was the selection of specialized auditors in the general meeting of shareholders who examined documents and statements of the economic entity, discovered any fraud and abuse, and provided unbiased opinions about the financial statements presented by the institution. The cooperation of experienced and specialized auditors with their display of honesty, truthfulness, and expertise formed the first professional accounting system in the late 19th century England. In the early 20th century, professional account-

* Corresponding Author, Assistant Professor of Accounting, Ferdowsi University of Mashhad, Iran, E-mail: mehdi.salehi@um.ac.ir.
** M.A in Accounting, Khorasan Science and Research Branch, Islamic Azad University.
*** PhD. Student of Banking, Marmara University, Istanbul, Turkey, E-mail: mostafa.bahrami66@yahoo.com.

ing systems were created in other countries that not only were socially accepted, but also had legal authority (Meigs et al., 1992).

This brief review shows that auditing, as a profession, expertise, and field of study, has a short history that goes back to a century ago. However, it has kept pace with the fast developments of the last century and has quickly adapted itself to contemporary conditions. Auditing is now regarded as a specialized branch of knowledge whose principles, methods, and performance has been examined in many articles and books (Meigs et al., 1992).

## 2. Theoretical issues

The theoretical issues are subdivided into 3 sections as followings:

### 2.1. Independent auditor

An independent auditor is a person who, upon the request of shareholders, beneficiaries, and/or legal authorities, examines the accounts of an entity and reports the results to the natural or legal person that has hired them (Meigs et al., 1992). The main reason for independent auditing is assurance and playing this role places auditors in a unique position. Attestation of financial statements is to ensure the reliability of these reports. It consists of two stages: (1) collection of evidence, and (2) presentation of the audit report.

Characteristics such as independence, impartiality, and professional responsibility are as important to auditors as technical expertise (Meigs et al., 1992). Among these characteristics, independence is of utmost importance, that is, the auditor must not be affiliated with the company being audited. The role of the independent auditor is to ensure that financial statements optimally support the interests of all groups without bias. The independent auditor provides services other than assessment of financial statements. These services can be categorized into two general groups:

1. Other assurance services: auditor's report on special purpose, overview of historical financial statements, performing agreed-upon procedures regarding financial information, and examination of prospective financial information;
2. Compiling financial information (Nourvash et al., 2011).

### 2.2. Decision tree (CART)

Decision tree is the most efficient and most widely used methods of inductive inference (Mitchell, 1997; Alpaydin, 2010). Decision tree learning is a method of learning with observations. ID3 algorithm is an algorithm used to generate a decision tree. It uses divide-and-conquer strategy for generating the tree and uses entropy for classifying the data. The goal of ID3 algorithm is to reduce confusion in nodes and to produce smaller trees. CART, which stands for Classification and Regression Trees, is the enhanced version of ID3 (Quinlan, 1993). This algorithm is able to classify continuous and noisy data. In terms of attributes, CART functions similar to ID3 and if the value of attributes is discrete, the data is arranged in an ascending order. Then, it calculates information gain for all the splitting possibilities in order and selected the value of the attribute with the highest gain as the splitting threshold and the information gain of the attribute. The best characteristic of this algorithm is that it prunes the generated tree. Usually if the possibility of a leaf falling from the tree is less than a threshold as compared to the neighboring leaves, that leaf is omitted or combined with the neighboring leaves. The purpose of pruning is to make the tree smaller, to avoid over fitting, to remove noise, and to reduce the effect of missing data that helped generate the tree. CART uses the Gini index instead of entropy (Breiman et al., 1984).

### 2.3. Particle swarm optimization (PSO)

PSO is an algorithm that simulates the social behavior of a flock of birds. This algorithm was designed to discover the underlying rules that enabled birds to fly synchronously and often changing direction suddenly, scattering, and regrouping. PSO optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. In PSO the candidate solutions, or particles, fly across the search-space. Movement within the search-space is a function of the particle's own experience and the experience of the particle's neighbors or the experience of the whole swarm. Therefore, the position of the population of particles affects the searching pattern of a particle. For each particle, a position and a velocity are defined and respectively modeled by a space vector and a velocity vector. PSO is based on the principle that at any given time each particle adjusts its position in the search-space according to its best previous position and the best position in its neighborhood (Carvalho and Ludermir, 2006).

Considering the above introduction, the present research tries to examine whether it is possible to predict selection of independent auditors in the companies listed in Tehran Stock Exchange based on a combination of PSO and CART.

## 3. Review of Literature

Chaney et al. (2004) carried out a research on self-selection of auditors and audit pricing in private firms. Using OLS regression, they showed that auditees, when not compelled by the market, choose the lowest-cost auditor available. Chaney and colleagues also used the data from prediction of self-selection of auditors in their analysis of audit pricing.

Velury et al. (2003) carried out a research the role of institutional ownership in the selection of industry specialist auditors.

They employed a two-stage least square regression model to examine the relationship between audit quality and the level of institutional investment. Their findings suggested that institutional owners influence management's choice of audit quality and encourage management to improve reporting quality by employing auditors that provide a higher quality service.

Kane and Velury (2004) studied the role of institutional ownership in the market for auditing services. Using logistic regression, they found that firms with high level of institutional ownership are more likely to select a Big auditor. They also found a positive association between Size and Debt and the selection of a Big auditor.

Cravens et al. (1994) examined the factors that affect the auditor selection process. They used univariate statistics and standardized Z score to perform the comparisons. The results indicated that there are differences between clients of Big and non-Big auditors, but also that there are differences among clients of different big auditors.

Kirkos et al. (2007) applied data mining methodologies to auditor selection. They used decision tree, neural networks, and K-nearest neighbor algorithm to develop models that could predict auditor choice. Using a 10-fold cross-validation test, the accuracy of each of the methods was calculated as follows: 81.97% for decision tree, 78.45% for neural networks, and 73.21% for k-NN. The results of this study suggested that firms with high debt opt for audit quality.

Kirkos et al. (2008) examined auditor selection using C4.5 Decision Tree, MLP Neural Networks, and Support Vector Machine. The sample of this research included 338 UK and Irish firms. The accuracy rates according to a 10-fold cross-validation test were 83.73%, 75.44% and 79.29% for the Decision Tree, Neural Networks, and Support Vector Machine methods respectively. The results of C4.5 and SVM models suggested that debt level is a driving factor in the choice of auditors and companies with high debt structure tend to choose a big auditor.

## 4. Research methodology

The present research is an applied study. Based on the data collected from TSE, the research hypothesis is tested and the results are generalized to the entire population. The method of the research is a combination of PSO and CART.

### 4.1. CART decision tree

CART uses the Gini index instead of entropy (Breiman, 1984), which is defined as follows:

$$\mathrm{I}_{\mathrm{gini}} = 1 - \sum_{j} p(c_j)^2 \quad (1)$$

Where $p(c_j)$ is the ratio of the data belonging to class $c$. First this algorithm calculates the Gini coefficient for all the attributes of the primary data. Then, the information gain of each of the attributes is calculated from:

$$\mathrm{Gain}\,(\mathrm{A}) = \mathrm{I}_{\mathrm{gini}} - \mathrm{I}_{\mathrm{res}_{\mathrm{gini}}}(\mathrm{A}) \quad (2)$$

$\mathrm{I}_{\mathrm{res}_{\mathrm{gini}}}(\mathrm{A})$ is calculated in equation (3) where $\mathrm{I}_{\mathrm{res}_{\mathrm{gini}}}$ denotes the residual entropy in classes due to using attribute $\mathrm{A}$ which can be calculated as the sum of split probabilities. Then, the attribute with the highest gain is selected as the splitting attribute.

$$\mathrm{I}_{\mathrm{res}_{\mathrm{gini}}}(\mathrm{A}) = \sum_{j} \left( p(a) \times (1 - \sum_{j} p(c_j|a)^2 \right) \quad (3)$$

Where $a$ is the branch created by selecting attribute $\mathrm{A}$ as the splitting attribute.

### 4.2. Particle swarm optimization (PSO)

The PSO algorithm initializes a certain number of particles with a random position and velocity. These particles iteratively move in the N-dimensional problem space in order to search for new possible options by calculating optimization as a criterion. The space dimension of the problem is equal to the number of parameters in the optimization equation. Each particle maintains a memory of its previous best position and the best position of all the particles. Through experience, the particles decide where to move in their next turn. The particles iteratively move in the N-dimensional problem space until all the particles converge to a point in the search-space (Gupta et al., 2003).

1. The number of particles in the swarm ($\mathrm{S}$): PSO initiates a random number of particles. The size of $\mathrm{S}$ differs depending on the problem. Each solution is represented by an array of bits. $\mathrm{S} = [\mathrm{S}_{\mathrm{ij}}]$ denotes the initial population, where $i$ is the number of particles ($i = 1,2,\ldots$) and $j$ is the number of bits ($j = 1,2,\ldots$).

2. Position ($X$): The position of a particle is represented by a matrix. The number of rows in matrix $X$ is twice the rows in matrix $S$.

3. Velocity: Each particle starts with a random velocity which is updated based on equation (4):

$$V_i^{t+1} = W_{iter} \otimes V_i^t \oplus (c_1 * r_1) \otimes (p_{besti} \oslash X_i^t) \oplus (c_2 * r_2) \otimes (p_{gbest} \oslash X_i^t) \quad (4)$$

4. Objective function: The function that must be minimized.

5. Velocity-position relationship: The velocity of each particle changes with respect to $p_{besti}$(best previous position of particle $i$) and $p_{gbest}$ (global best position).

In equation (4), $c_1$ and $c_2$ are learning coefficients, $r_1$ and $r_2$ are random numbers in the interval $[0,1]$, and $W_{iter}$ denotes the inertia weight. At the end of the $t$-th iteration, the position of each particle ($X_i^t$) is updated using equation (5) and the new position of the particles is obtained (Soke and Bingul, 2007):

$$X_i^{t+1} = X_i^t + V_i^{t+1} \quad (5)$$

### 4.3. Population and sample

The population of the research consists of all the companies listed TSE during the period 2005-2010. The sample includes 576 data from 95 companies listed in TSE during the same period. First, a list was created of all the companies that were constantly active in TSE during the period 2005-2010. Investment companies, financial brokerage companies, inactive companies, and companies with incomplete information about audit fees were excluded from the sampling. The final sample included 102 data that were audited by public audit firms and 474 data that were audited by private audit firms.

### 4.4. Purpose

The purpose of the present research is to predict selection of independent auditors in the companies listed in Tehran Stock Exchange based on certain components (turnover, total assets, current assets, etc.) using a combination of PSO algorithm and CART.

### 4.5. Research Hypothesis

According to the purpose of the study, the following hypothesis is postulated in the study:

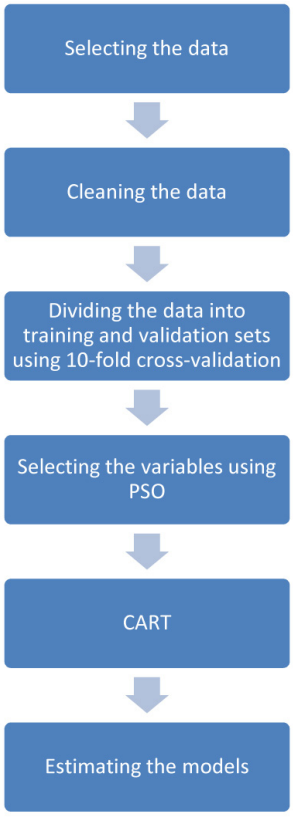
H1: Certain components (operating income, non-audit fees, etc.) affect the prediction of independent auditor selection using PSO algorithm and CART.

## 5. Data analysis

MATLAB 7.6 software was used to implement the algorithms. MATLAB is one of the best mathematical software with extensive applications in many fields of study. The toolboxes of MATLAB facilitate working with the software. In the present research, a combination of particle swarm optimization (PSO) and categorization and regression trees (CART) was used for hypothesis testing.

### 5.1. Design and implementation

This section provides an overview of the design and implementation procedure of the mixed PSO-CART model. There are six steps in the proposed method: (1)selecting the data, (2) cleaning the data, (3) dividing the data into training and validation sets, (4) selecting the variables that affect prediction of independent auditor selection using PSO for each model, (5) training the models, and (6) estimating the trained models with the data that has not been seen by the algorithms.




<Figure 1> The proposed procedure for predicting independent auditor selection using the mixed PSO-CART model

The following sections describe the different steps of the proposed procedure.

### 5.2. Selecting the data

The first step in any data mining procedure is data selection. The financial data of 95 companies were collected for a six-year period (2005-2010) with 16 independent variables from the website of Iran's Securities and Exchange Organization[1]. There was a dependent variable (Audit Firm) for all the companies. From the 95 studied companies, private audit firms audited 78 companies and public audit firms audited 17 companies.

<Table 1> List of the selected independent variables

| No. | Variable |
|---|---|
| 1 | Current debts |
| 2 | Accounts Receivable |
| 3 | Audit Fee |
| 4 | Non-Audit Fee |
| 5 | Current Assets |
| 6 | Degree of Leverage |
| 7 | Capital |
| 8 | Working Capital |

1) www.codal.ir

| | |
|---|---|
| 9 | Operating Capital |
| 10 | Gross Profit |
| 11 | Total Debts |
| 12 | Total Assets |
| 13 | Turnover |
| 14 | Debt Ratio |
| 15 | Solvency Ratio |
| 16 | Current Ratio |

## 5.3. Cleaning the data

The second step involves cleaning the data. In this step, the data with incomplete or incalculable information about the independent variables are omitted. Considering the previous studies and the research of Kirkos et al. (2008), 39 financial ratios were extracted. First, analysis of variance was used to estimate the significance of each variable. This method estimates the variation of the values for each class. If the $p$-values of the variables were less than 0.005;those variables would be selected as independent variables of the research (Kirkos et al., 2008). Overall, 16 variables were selected using this method.

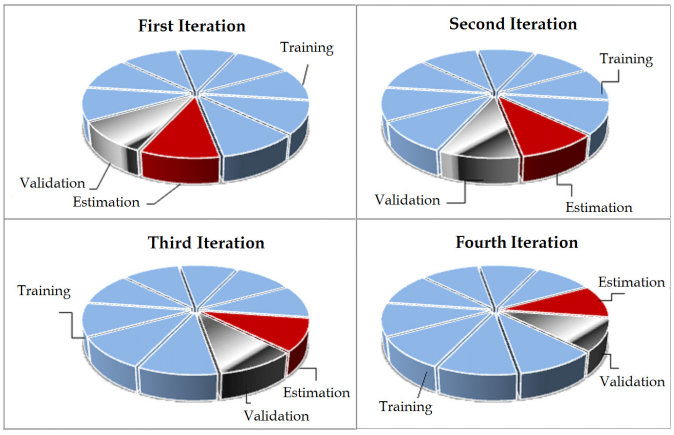
To prepare the data for model training and validation, first each of the variables is normalized using equation (6) so that the effect of large values is reduced.

$$\tilde{S}_i = \frac{(S_i - S_{min})}{S_{max} - S_{min}}, \quad i = 1, \dots, 576 \tag{6}$$

Where $S_{min}$ and $S_{max}$ are the minimum and maximum value of the variable, $\tilde{S}_i$ is the normalized value of $S_i$, and $i$ is the number of the variable.

## 5.4. Dividing the data using 10-fold cross-validation

Error rate is one of the measures used for classification. Generally, we cannot make a good judgment about the performance of algorithms by comparing the errors calculated on the training data. Usually the error rate on the training data is less than the error rate on the data that has not been seen during the training process. Therefore, we cannot use training error for comparing the two algorithms. For more complex models, classifications that have more parameters have a more complex boundary. This reduces error on the training data as compared to simpler models (Alpaydin, 2010). Therefore, we need estimation data along with the training data. Validation data is also used to select the right variables for each model. Hence, each dataset is divided into three independent subsets: training data, validation data, and estimation data. Training data is used for training the model, validation data is used to examine whether the right variables are selected for each model, and estimation data is used to calculate the algorithms'error rate on the data that has not been seen. Indeed one implementation of the algorithms is not enough for a good estimation. Usually algorithms tend to have expected error rate as close to the real error rate as possible and this can be done only with repeated implementation of the training and estimation processes. Thus, a part of a given data set is set aside for final estimation and the rest of the data is used for training and validation; finally, all the three sets are changed and the model is again estimated. K-fold cross-validation is a common method for this purpose. In this method, the data set is randomly partitioned into $K$ equal parts. $K$ pairs $\{x_i, y_i\}_{i=1}^{K}$ are randomly extracted where $x_i$ is the independent variables and $y_i$ is the dependent variable related to sample $i$. Subsequently $K$iterations of training and validation are performed such that within each iteration a different fold of the data is held-out for validation while the remaining $K-1$ folds are used for learning. Usually 20% of the available data is used for validation and the remaining 80% is used for training. The training, validation, and estimation data must be large enough so that the expected error rate will be closer to the actual error rate. Meanwhile, the training, validation, and estimation data must have the least overlap with the data from other iterations so that all the data are included in training, validation, and estimation processes. It must also be noted that the ratio of the data of each class in each of the training, validation, and estimation sets must be equal to the ratio of the data of the same class in the entire data sets. For instance, if the samples of a classroom include 20% of the learning samples of the entire data sets, 20% of the data in each of the training, validation, and estimation sets must belong to that class. There are two important issues here. First, the ratio of the validation set to the learning set is small. Second, as the value of $N$ (total number of samples in the data sets) increases, we can decrease $K$, and if the value of $N$ is small, $K$ must be large enough to provide enough samples for training. If $K$equals $N$, this method changes into leave-out-one method. Figure 2 displays the first four iterations for selecting the training, validation, and estimation data sets using 10-fold cross-validation.




<Figure 2> The first four rounds of selecting training, validation, and estimation data sets with $k = 10$

In each iteration, an error rate is calculated for training and validation data and the mean of these values is considered as the error rates of these two data sets.

### 5.5. Selection of the variables that affect prediction of independent auditor selection using PSO algorithm for each model

After dividing the data into three groups, i.e. learning, validation, and evaluation sets, the variables that affect the prediction of independent auditor selection must be identified. A combination of PSO algorithm and CART is used to identify these variables. First, the 16 variables of the problem must be encoded in order to be incorporated into the mixed model. The table below presents an example of such encoding.

<Table 2> List of the selected independent variables

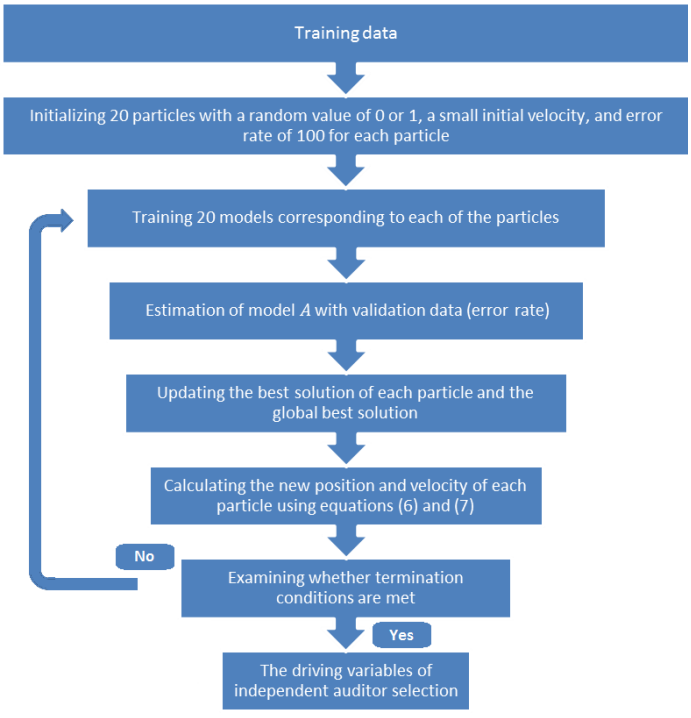
| No. | Variable | Encoding |
|---|---|---|
| 1 | Total Assets | 1 |
| 2 | Current Assets | 0 |
| 3 | Total Debts | 1 |
| 4 | Current Debts | 1 |
| 5 | Gross Profit | 0 |
| 6 | Operating Profit | 0 |
| 7 | Accounts Receivable | 0 |
| 8 | Audit Fee | 1 |
| 9 | Non-Audit Fee | 1 |
| 10 | Working Capital | 1 |
| 11 | Current Ratio | 0 |
| 12 | Debt Ratio | 1 |
| 13 | Solvency Ratio | 0 |
| 14 | Turnover | 0 |
| 15 | Leverage | 1 |
| 16 | Capital | 1 |

Each of the bits in the above table indicates whether the independent variable is present (one) or absent (zero) in CART. Absence or presence of the variables in CART is determined by PSO algorithm. The training data is used to identify the variables that affect the prediction of independent auditor selection, and the validation data is used to estimate the variables selected in PSO algorithm. As seen in Figure 3, first, 20 particles with 16 random numbers of either 0 or 1 are initialized with a small random velocity, an error rate of 100, and a best solution value (best error rate) of 100 for each particle. It is assumed that all the particles are the worst particles possible. Then, using the training data, 20 models of $A$ are trained with an input equal to the number of variables that have a value of one ($A$ denotes the CART tree). The training data is fed to model $A$ for it to learn them. Then the error rate for each of the models of A (the particle corresponding to model A) is obtained and the best solution of each particle and the global best solution are updated. Since all the particles have a binary value, a newer version of PSO must be used for updating velocity and position (Khanesar, 2007). Using equations (6) and (7), the velocity and position of each particle is updated and the algorithm is recalled. This procedure continues until the best solution does not change over 50 generations and finally the driving variables of independent auditor selection are selected.

$$v'_{i+1} = sig(v_{i+1}) = \frac{1}{1+e^{-v_{i+1}}} \quad (6)$$

$$X'_{i+1} = \begin{cases} 1 & U < sig(v_{i+1}) \\ 0 & Otherwise \end{cases} \quad (7)$$

In equation (7), $U$ is a random number between zero and one. Once the mixed PSO-CART algorithm is converged, the best variables for predicting independent auditor selection are yielded.




<Figure 3> The process of the proposed PSO algorithm for identifying the driving variables of independent auditor selection

### 5.6. Training process and model estimation

After dividing the samples into training, validation, and estimation sets and identifying the driving variables of independent auditor selection, the CART tree is trained using the training data and then the accuracy of the model is estimated using the esti-

mation data that has not been seen by the tree. As described earlier, K-fold cross-validation method is used and the data is divided into 10 parts. Accuracy rate is used as a measure for estimating the prediction models:

$$\text{Accuracy Rate} = \frac{\text{Number of True Predictions}}{\text{Total Number of Samples}} \quad (8)$$

The closer the accuracy rate is to 100, the closer are the predictions of algorithms to reality. The procedure for calculating type I and type II errors is as follows (Walsh et al., 1969):

TP: Total number of private companies that have been truly predicted as private.

FP: Total number of public companies that have been falsely predicted as public.

TN: Total number of public companies that have been truly predicted as public.

FN: Total number of private companies that have been falsely predicted as private.

<Table 3> Type I and II errors for the estimation data of CART algorithm

| | | Actual Rate: Private | Actual Rate: Public | |
|---|---|---|---|---|
| Prediction | Private | TP | FP | Accuracy of Prediction of Praivate Companies $\frac{TP}{TP+FP}$ |
| | Public | FN | TN | Accuracy of Prediction of Public Companies $\frac{TN}{TN+FN}$ |
| | | Sensitivity $\frac{TP}{TP+FN}$ | Specificity $\frac{TN}{FP+TN}$ | |

| | |
|---|---|
| Type I Error | Specificity - 1 |
| Type II Error | Sensitivity - 1 |
| Accuracy Rate | $\frac{TP+TN}{TP+FP+TN+FN}$ |

## 5.7. CART

After selecting the core variables using PSO-CART algorithm, the training data are applied to CART and the algorithm creates a decision tree based on the data. During the training phase, the tree is allowed to fully grow and then, to avoid over fitting, the tree is pruned. The structure of the tree is saved in the computer once the tree is trained. To estimate the model, the estimation data that have not been seen by the tree are applied to it and the estimation error is calculated. Figure 4 presents the trained CART.

Table 4 shows the selected components for predicting the selection of independent using the PSO-CART algorithm.

<Table 4> The list of the variables selected by PSO-CART algorithm

| | | |
|---|---|---|
| 1 | Total Assets | 1 |
| 2 | Current Assets | 1 |
| 3 | Total Debts | 0 |
| 4 | Current Debts | 0 |
| 5 | Gross Profit | 0 |
| 6 | Operating Profit | 0 |
| 7 | Accounts Receivable | 0 |
| 8 | Audit Fee | 1 |
| 9 | Non-Audit Fee | 0 |
| 10 | Working Capital | 1 |
| 11 | Current Ratio | 1 |
| 12 | Debt Ratio | 1 |
| 13 | Solvency Ratio | 1 |
| 14 | Turnover | 1 |
| 15 | Leverage | 0 |
| 16 | Capital | 1 |

Table 5 presents the mean accuracy rate of the model after implementing 10-fold cross-validation.

<Table 5> Accuracy rate of the model

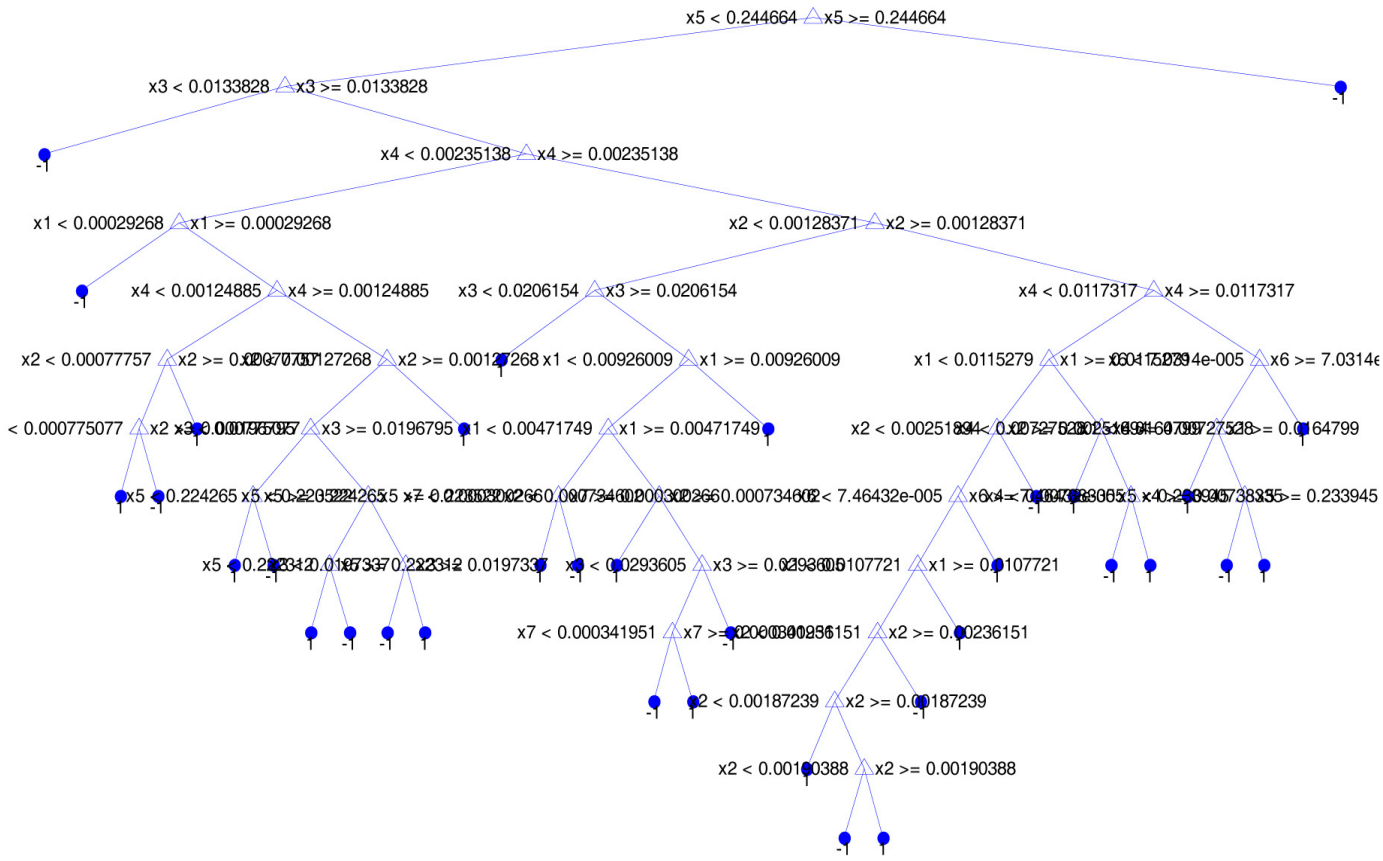
| Fold | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Accuracy Rate (%) | 100 | 88.77 | 98.25 | 94.74 | 87.77 | 85.25 | 10 | 96.49 | 96.49 | 96.49 | 94.43 |

The calculation procedure for the fourth implementation is showed in the table below.

<Table 6> Type I and II errors for the estimation data of CART algorithm

| | | Actual Rate: Private | Actual Rate: Public | |
|---|---|---|---|---|
| Prediction | Private | 74 | 1 | 97.92% |
| | Public | 2 | 7 | 77.87% |
| | | 95.92% | 87.50% | |
| | | Sensitivity | Specificity | |

| | |
|---|---|
| Type I Error | 12.50% |
| Type II Error | 4.08% |
| Accuracy Rate | 94.74% |




<Figure 4> CART Tree

## 6. Conclusion

The present research used a combination of PSO algorithm can CART for hypothesis testing. The results showed that from the 16 studied variables, 9 variables (Total Assets, Current Assets, Audit Fee, Working Capital, Current Ratio, Current Debt, Solvency Ratio, Turnover, and Capital) affect the prediction of independent auditor selection. Moreover, the prediction rate of the mixed model (PSO-CART) is 94.43%, the type I error of the model is 12.50%, and the type II error of the model is 4.08%. Therefore, it can be argued that the mixed model is appropriate for predicting independent auditor selection due to having a high accuracy rate and low type I error rate.

## References


Alpaydin, E. (2010). *Introduction to Machine Learning* (2nd Edition). London, England: MIT Press.

Breiman, L.,Friedman, J., Stone C.J., & Olshen R.A. (1984). *Classification and Regression Trees* (1st Edition). New York: Chapman and Hall/CRC.

Carvalho, M., & Ludermir, T.B. (2006). An analysis of PSO hybrid algorithms for feed-forward neural networks training. Proceedings of the Ninth Brazilian Symposium on Neural Networks (SBRN'06).

Chaney, P., Jeter, D., &Shivakumar, L. (2004). Self-selection of auditors and audit pricing in private firms. *The Accounting Review,* 79(1), 51-72.

Cravens, K.S., Flagg, J.C., & Glover, H.D. (1994). A comparison of client characteristics by auditor attributes: Implications for the auditor selection process. *Managerial Auditing Journal,* 9(3), 27-36.

Gupta, M.M., Jin, L., & Homma, N. (2003). *Static and Dynamic Neural Networks: From Fundamentals to Advanced Theory* (1st Edition). New Jersey: Wiley-IEEE Press.

Kane, G.D., & Velury, U. (2004). The role of institutional ownership in the market for auditing services: An empirical investigation. *Journal of Business Research,* 57(9), 976-983.

Khanesar, M.A., Teshnehlab, M., & Shoorehdeli, M.A. (2007). A novel binary particle swarm optimization. Proceedings of the Mediterranean Conference of MED'07 (pp1-6). Control and Automation.

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications,* 32(4),

995-1003.

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2008). Support vector machines, decision trees and neural networks for auditor selection. *Journal of Computational Methods in Science and Engineering,* 8(3), 213-224.

Meigs, W.B., Whittington, O.R., Pany, K., & Meigs, R.F. (1992). *Principles of Auditing* (10th Edition). Boston: Irwin Professional Publishing.

Mitchell, T.M. (1997). *Machine Learning.* New York, NY : McGraw-Hill.

Nourvash, I., Mehrani, S., Karami, G.R., & Shahbazi, M. (2011). *A Comprehensive Review of Auditing.* Tehran, Iran: Negahe Danesh.

Quinlan, J.R. (1993). *C4.5:Programs for Machine Learning* (1st Edition). San Francisco: Morgan Kaufmann.

Soke, A., & Bingul, Z. (2007). Applications of discrete PSO algorithm to two-dimensional non-guillotine rectangular packing problems. Proceedings of GEM (pp. 127-132), GEM.

Velury,U., Reisch, J.T., & O'Reilly, D.M. (2003). Institutional ownership and the selection of industry specialist auditors. *Review of Qualitative Finance and Accounting,* 21, 35-48.

Walsh, T.J., Garden, J.W., & Gallagher, B. (1969). Obliteration of retinal venous pulsations during elevation of cerebrospinal-fluid pressure. American *Journal of Ophthalmology,* 67(6), 954-956.