International Journal of Accounting Information Systems 25 (2017) 1–17

Contents lists available at ScienceDirect

International Journal of Accounting Information Systems

journal homepage: www.elsevier.com/locate/accinf

Detecting corporate tax evasion using a hybrid intelligent system: A case study of Iran

Eghbal Rahimikia[a,*], Shapour Mohammadi[b], Teymur Rahmani[c], Mehdi Ghazanfari[a]

[a] School of Industrial Engineering, Iran University of Science & Technology, Tehran 16846-13114, Iran
[b] Faculty of Management, University of Tehran, Tehran 14155-6311, Iran
[c] Faculty of Economics, University of Tehran, Tehran 14155-6445, Iran

ARTICLE INFO

Article history:
Received 21 September 2015
Received in revised form 25 November 2016
Accepted 18 December 2016
Available online xxxx

Keywords:
Corporate tax evasion detection
Data mining
Hybrid intelligent system
Support vector machine
Neural network
Harmony search

ABSTRACT

This paper concentrates on the effectiveness of using a hybrid intelligent system that combines multilayer perceptron (MLP) neural network, support vector machine (SVM), and logistic regression (LR) classification models with harmony search (HS) optimization algorithm to detect corporate tax evasion for the Iranian National Tax Administration (INTA). In this research, the role of optimization algorithm is to search and find the optimal classification model parameters and financial variables combination. Our proposed system finds optimal structure of the classification model based on the characteristics of the imported dataset. This system has been tested on the data from the food and textile sectors using an iterative structure of 10-fold cross-validation involving 2451 and 2053 test set samples from the tax returns of a two-year period and 1118 and 906 samples as out-of-sample using the tax returns of the consequent year. The results from out-of-sample data show that MLP neural network in combination with HS optimization algorithm outperforms other combinations with 90.07% and 82.45% accuracy, 85.48% and 84.85% sensitivity, and 90.34% and 82.26% specificity, respectively in the food and textile sectors. In addition, there is also a difference between the selected models and obtained accuracies based on the test data and out-of-sample data in both sectors and selected financial variables of every sector.

1. Introduction

Tax returns contain useful information for detecting tax evasion. However, much of this information is voluminous and complex. Therefore, it is necessary to use mathematical and statistical models to analyze this information. Because of the costs associated with auditing individual companies' tax returns, finding models that accurately identify misstated returns is important. Consequently, the goal for this data analysis is to help tax auditors and tax authorities to detect companies with a high probability of misstated activities in order to give these suspect companies a more detailed audit. By using these models, tax authorities can significantly increase tax income and decrease the human resource costs associated with manually auditing tax returns.

* Corresponding author.
E-mail address: erahimi@ut.ac.ir (E. Rahimikia).

http://dx.doi.org/10.1016/j.accinf.2016.12.002
1467-0895/

Due to voluminous and complex tax data sets, tax authorities in many countries have started adopting new systems based on data mining (DM) and artificial intelligence (AI) to help them to detect misstatements in financial statements. Because of this, academic research in this area (specially published research reports for tax evasion detection systems) is lagging behind and the need for more attention to this topic by the academia is especially felt more than ever. Generally, tax evasion detection, used by the tax authorities to detect tax returns suspected of tax evasions, is somewhat similar to financial statement fraud detection.

Tax administrations have to deal with a variety of risks, such as the risk of non-compliance, the risk of tax evasion and the risk of insolvency by the taxpayer. Compliance risk management allows us to deal with these risks by looking at the behavior of taxpayers. For these reasons, intelligent predictive models to identify tax returns for additional examination can be used as a tool to increase the accuracy and efficiency of auditing.

The Iranian National Tax Administration (INTA) is a governmental institution and is established under the supervision of Iran's ministry of economic affairs and finance. INTA provides full support for administering tax reforming plans and executing tax collection legal procedures in an efficient manner. In the future, its responsibilities will also extend to the monitoring of tax law enforcements and regulations and the creation of a proper basis to achieve tax objectives.

In this paper, we present a new hybrid intelligent system to detect corporate tax evasion in INTA. Our focus is to compare the effectiveness of using a system that combines the multilayer perceptron (MLP) neural network, support vector machine (SVM), and logistic regression (LR) classification models with the harmony search (HS) optimization algorithm to detect corporate tax evasion separately in two sectors. The HS optimization algorithm optimizes the selection of financial variables in addition to model parameters; so an optimized model with acceptable accuracy in terms of a combination of optimal financial variables and model parameters can be developed.

The hybrid model was developed and tested with data from the food and textile sectors. This paper is among the first research based on actual data with a large sample set of tax returns from different sectors operating in Iran. To gain a good level of confidence in our results, the output from our model was evaluated against out-of-sample data (in addition to test data and total data). Furthermore, cross-validation was repeated 8 times using parallel computing to improve the reliability of the results. Most of the articles published in this or related areas mainly focus on binary results instead of probability results. However, the binary results are not really helpful in practice. Therefore, all classification models presented in this proposed system generate the probability of membership in each of the two classes and for reporting purposes, the outputs are sorted based on the obtained probabilities. This allows auditing of companies with higher probabilities to be done with more scrutiny. Furthermore, comparing the selected financial variables based on a number of repetitions in 10 system runs for the food and textile sectors is helpful to answer the question of whether tax evasion behavior patterns are different across sectors.

This paper is structured as follows: First intelligent systems for fraud detection is presented in Section 2, then tax evasion detection algorithms that were used are described in Section 3, the hybrid intelligent system structure for the corporate tax evasion detection is explained next in Section 4, method is presented in Section 5 followed by the results and discussion in Section 6 and finally Section 7 provides conclusions and future research directions of this work.

2. Intelligent systems for fraud detection

The current published research mainly focuses on financial statement fraud detection. Relatively little published academic work is available on tax evasion detection. This is due to the limited access to datasets at the national level. It should be noted that in terms of modeling frameworks and applied methods, tax evasion detection and financial statement fraud detection have a lot in common.

Krieger (1996) mentions an example from the USA's Internal Revenue Service (IRS) which uses neural networks and polynomial regression to pinpoint potential tax non-compliance cases. Their tools focused on feature classification and extraction, descriptive statistics, polynomial networks and clustering algorithms. Through these tools, the database variables were reduced from 150 to 8. The accuracy of the polynomial network was confirmed by other neural networks (backpropagation neural network (BPN), learning vector quantization (LVQ) and self-organizing map (SOM)). Fischthal (1998) registered a patent in the USA for detecting fraud using a neural network. The architecture of this system involves first employing a conceptual clustering technique to generate a collection of classes from historical data. Neural networks are provided for each class created by the clustering step and the networks are trained using the same historical data. Yu et al. (2003) present a system architecture for the fraudulent tax declaration detection problem of Chinese commercial enterprises that includes communication with domain experts, choice of the core data mining algorithm, design guidelines of the data mining system architecture and incorporation of domain experts' knowledge. Their decision tree implementation results achieved an 85–90 % accuracy rate. Gupta and Nagadevara (2007) implement 8 models based on different combinations of DT, Discriminant model and LR in India Value-Added Tax (VAT) from 2003 to 2004 and a sample containing 402 dealers. They found that all models developed through data mining techniques were better than random selection. Wu et al. (2012) design a VAT evasion detection model in Taiwan by association rule mining (IBM DBMiner 2.0 tool) from 2003 to 2004. The results show that the designed model enhanced the detection of tax evasion and therefore can be employed to effectively reduce or minimize losses from VAT evasion. Hsu et al. (2015) present a case study of a pilot project that was developed to evaluate the use of data mining in audit selection for the Minnesota Department of Revenue (DOR) for Sales and Use Tax. Researchers used a combination of C4.5 DT, Naïve Bayes, MLP neural network, SVM and other techniques to build classification models using 10943 samples from 2004–2006 as training

dataset and 2007 as test (and validation) datasets. The feature selection of this research is based on trial and error and discussion with domain experts. The results of the pilot project showed that the data mining based approach achieved an increase of 63.1% in effectiveness.

In all the mentioned related works, Krieger (1996), Yu et al. (2003), Gupta and Nagadevara (2007), Wu et al. (2012) and Hsu et al. (2015) studied tax evasion detection in VAT and sales and use tax without using hybrid models. Therefore, the need for modeling tax evasion behavior in other main sources of tax revenue (such as corporate tax) by new and hybrid data mining models is evident which is the focus of this paper.

3. Tax evasion detection algorithms

In subsection 3.1, MLP neural network, in subsection 3.2, SVM, in subsection 3.3, LR, and finally in subsection 3.4, HS optimization algorithm will be described. As mentioned in Phua et al. (2010), SVM, MLP, and LR are the most popular models in fraud detection and other related classification cases. Therefore, we compare the effectiveness of using a system that combines MLP neural network, SVM, and LR classification models with the HS optimization algorithm. The optimization algorithm optimizes the selection of financial variables (inputs) in addition to classification model parameters; so an optimized model will be developed which has the highest detection accuracy.

3.1. Multilayer perceptron (MLP) neural network

Fig. 1 shows the main components of MLP neural network. Each hidden layer neuron receives an input ($I_1, I_2, I_3, \ldots, I_n$), multiplies it by a weight and sends each output to all other neurons of the hidden layer. Every neuron in a hidden layer calculates the output value based on an activation function, input values and the neuron's bias as shown in Fig. 2. The activation function can take many forms, such as the step function, linear function, log-sigmoid function, tan-sigmoid function, or softmax function. Neural networks use different algorithms for training. The purpose of a neural network training algorithm is to find optimized weight and bias values to simulate output(s) behavior based on input(s) characteristics. Resulting outputs from the hidden layer are sent to the output layer. Neurons in this layer use the same procedure as of the hidden layer (with similar or different activation functions) to calculate the final output(s) ($O_1, \ldots, O_n$) of the neural network. The MLP neural network can have different number of hidden layers and different number of neurons in every layer; however in most cases having one or two layers is suitable. Although a MLP neural network with many layers can represent deep circuits, training deep networks has always been challenging and empirical studies have found that deep networks generally performed not better, and often worse than neural networks with one or two hidden layers (Bengio et al., 2007) .

The scaled conjugate gradient algorithm is one of the most effective algorithms for pattern discovery and classification. This algorithm can train any network as long as its weight, net input, and activation functions have derivative functions. Backpropagation is used to calculate derivatives of performance with respect to the weight and bias variables. The scaled conjugate gradient algorithm is based on conjugate directions, but this algorithm does not perform a line search at each iteration (Møller, 1993) .

The general practice is to first divide the data into three datasets. The training dataset is used to adjust the weights and biases by computing the gradient and updating the network weights and biases, test dataset is used for testing the final trained neural network in each epoch in order to obtain the actual predictive performance of the network, and the validation dataset is used to minimize overfitting. The error on the validation set is monitored during the training process. The validation error normally decreases during the training phase, as does the training set error. However, when the network begins to overfit, the

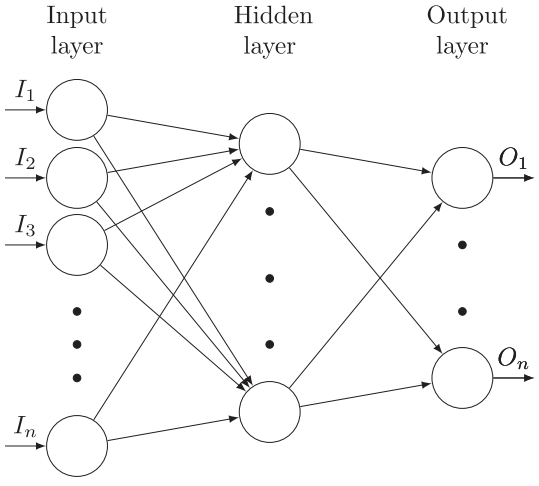

Fig. 1. Neural network structure.

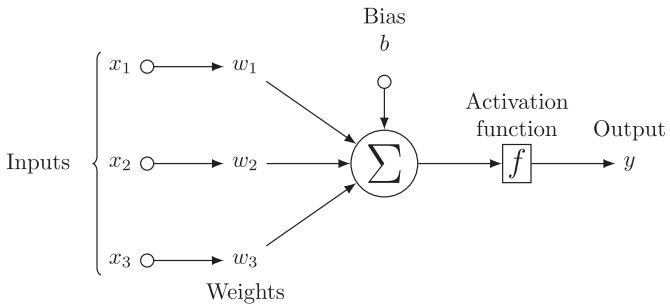

Fig. 2. Neuron structure.

error on the validation set typically begins to rise. The network weights and biases are saved when validation set error is at the minimum. Furthermore, early stopping or bayesian regularization is used to increase the generalization capability of a neural network (i.e., enhance prediction accuracy for unseen data).

3.2. Support vector machine (SVM)

The SVM was proposed by Vapnik in 1995 (Cortes and Vapnik, 1995) and was derived from structural risk minimization (SRM). SVM uses the pre-processing strategy in learning by mapping input space X to a high-dimensional feature space, F (Seo, 2007) . Suppose that we have a sample set $S_t = \{(x_i, y_i | x_i \in R^P), y_i \in \{-1, 1\}, i = 1, 2, 3, \ldots, n\}$. x_i is the input vector, n is the dimension of the input vector, and y_i is the output vector. Based on linear separability, SVM classifies these samples by solving the quadratic programming (QP) problem presented in Eq. (1) to find the optimal separator hyper-plane:

$$\begin{aligned} &\min_{w,b,\xi} \quad \frac{1}{2} w^t w + C \sum_{i=1}^{N} \xi_i \\ &\text{s.t.} \quad y_i(w\phi(x_i + b)) + \xi_i - 1 \\ &\xi \geq 0. \end{aligned} \tag{1}$$

where $w\phi(x_i)$ is a mapping function which maps training data into a high dimensional space, C is the penalty for error term, w is the weight vector, and ξ_i is the slack variable (Vapnik, 1995) . This optimization problem can be solved by the lagrangian method. The result obtained from the optimization problem is a hyper-plane. Finally we can use Eq. (2) to decide about every sample class:

$$f(y) = \text{sign} \quad \left(\sum_{i=1}^{N} y_i a_i K(x, x_i) + b \right) \tag{2}$$

where a_i is the parameter and $K(x_i, y_i) = \Theta(x_i)^T(x_j)$ is the kernel function. Kernel functions consider only inner product rather than high dimensional feature space so we have an easier implementation in this structure. The kernel function can take many forms, such as the linear, polynomial, log-sigmoid, radial basis, and sigmoid functions (Hsu et al., 2003) .

3.3. Logistic regression (LR)

Logistic regression (LR) is similar to linear regression but the response variable is discrete (Cox, 1958) . The LR model for p independent variables is as follows:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}} \tag{3}$$

where $P(Y = 1)$ is the probability of presence of tax evasion and $\beta_0, \beta_1, \ldots, \beta_p$ are regression coefficients. The natural logarithm of the ratio of $P(Y = 1)$ to $1 - P(Y = 1)$ gives a linear model as follows:

$$g(x) = ln\left(\frac{P(Y = 1)}{1 - P(Y = 1)} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{4}$$

where $g(x)$ has many of the properties of a linear regression model (Hosmer Jr and Lemeshow, 2004) .

3.4. Harmony search (HS)

The HS heuristic optimization algorithm is inspired by the improvisation process of musicians (Geem, 2009) . Solutions of the optimization process correspond to the musicians and the harmony of the notes generated by a musician corresponds to the fitness of the solution (Lee and Geem, 2005) .

HS consists of three main steps based on Algorithm 1:

Algorithm 1. Harmony search algorithm.

```
begin
    Define objective function F(x), x = (x_1, x_2, ..., x_d)^T;
    Define harmony memory accepting rate (r_accept);
    Define pitch adjusting rate (r_pa) and other parameters;
    Generate harmony memory (HM) with random harmonies;
    while (t < max number of iterations) do
        while (i ≤ number of variables) do
            if (rand < r_accept) then
                Choose a value from HM for the variable i;
                if (rand < r_pa) then
                    Adjust the value by adding certain amount;
                end
            else
                Choose a random value;
            end
        end
        Accept the new harmony (solution) if better;
    end
    Find current best solution;
end
```

1. **Initialization:** Algorithm parameters are defined and the harmony memory is initialized by filling it up with random solutions; each harmony is evaluated using an objective function.
2. **Harmony improvisation:** The three rules of the algorithm (memory consideration, pitch adjustment, and random selection) are used to generate new harmony vector.
   (a) **Create a new solution:** Create a new solution randomly (probability of $(1 - r_{accept})$) by (1) or an existing solution in harmony memory (probability of (r_{accept})) by (2).
   (b) **Adjustment:** Modification of elements of the new harmony with probability of (r_{pa}).
   (c) **Evaluation:** Evaluation of the new harmony using objective function.
3. **Selection:** After activating the termination criterion, the best harmony of the harmony memory is selected in this phase.

4. Corporate tax evasion detection hybrid intelligent system structure

The main process of the proposed 'Corporate tax evasion detection hybrid intelligent system' is illustrated in Fig. 3. This section, examines parts of this process separately.

4.1. Data consolidation and cleaning

Databases of INTA are hosted on different servers. All data used in this study were extracted from the Integrated Tax System (ITS) of INTA and all returns were submitted electronically in this system. We consolidated these databases using a consolidation module based on different characteristics of taxpayers including sector name, registration number, place of registration, and International Standard Industrial Classification (ISIC) code. After building a consolidated database, corrupt, inaccurate, and blank records were removed from the new database in the cleaning phase ('Data consolidation and cleaning' phase in Fig. 3).

4.2. Financial variables computation and tax evasion criterion

Based on the most commonly used financial variables in related studies (Bellovary et al., 2007; Brigham and Ehrhardt, 2013; Kumar and Ravi, 2007; Lin et al., 2015; Persons, 2011) , feedback from tax authorities and experts, and data availability, 21 financial variables were collected to develop the system ('Financial variables computation' phase in Fig. 3). Since articles and reports relevant to tax evasion detection are very limited, we rely on the above-mentioned financial statement fraud detection articles for variable selection. The obtained financial variables are presented in Table 1.

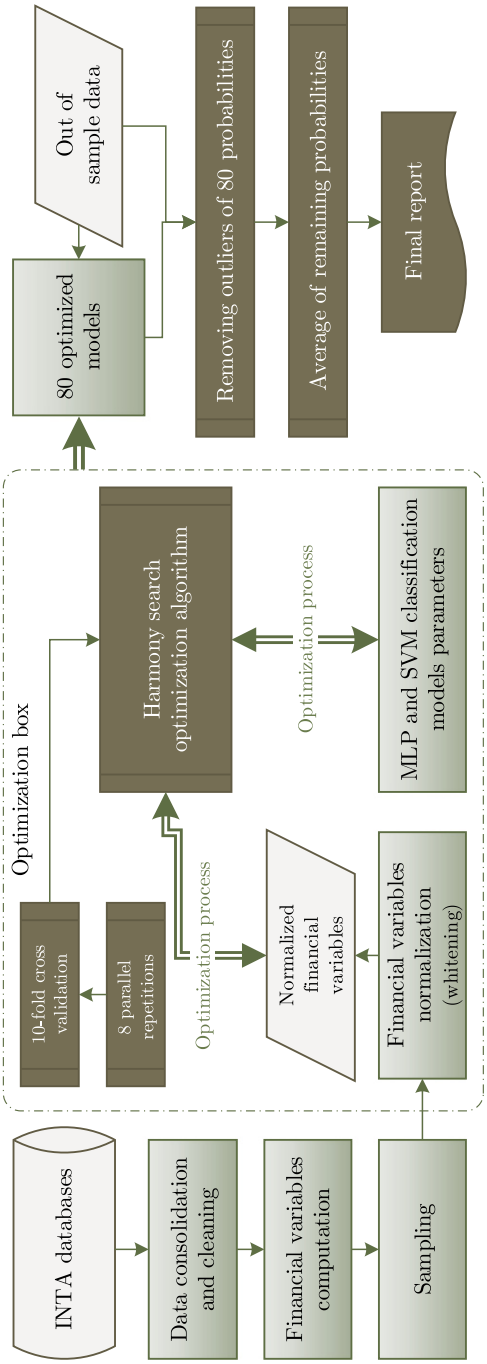

Fig. 3. Process of corporate tax evasion detection hybrid intelligent system.

Table 1
Financial variables.

1	EBIT[1]/Total assets	12	Fixed assets/Total assets
2	Net sales/Total assets	13	Inventory/Total assets
3	Net income/Total assets (ROA)	14	Total liabilities/Shareholders' equity
4	Net income/Shareholders' equity (ROE)	15	Total liabilities
5	Retained earnings/Total assets	16	Accounts receivables
6	Total liabilities/Total assets (debt ratio)	17	Accounts receivables/Total assets
7	Current assets/Total assets	18	Gross profit/Total assets
8	Quick assets/Total assets	19	Total liabilities/(TL+SE[2])
9	Total assets	20	Gross profit
10	Total assets/Shareholders' equity	21	Net profit
11	Operating income/Total assets		

[1] Earnings before interest and taxes.
[2] Total liabilities+Shareholders equity.

Our proposed model is a binary classification model and outputs of training, testing and validation datasets are labeled as 1 (misstated companies) and 0 (not-misstated companies). We use the Eq. (5) formula for classifying companies as misstated and not-misstated samples:

$$E = 1 - \frac{1}{1 + \frac{(R-P)}{mean(R-P)}} \tag{5}$$

where P is declared Earnings Before Tax (EBT), R is audited EBT, and E is the calculated output of tax evasion criterion for every imported sample ranging from 0 to 1. Based on this formula, higher E values are equivalent to larger companies with higher degree of misstatement considering the fact that $mean(R-P)$ is the average of differences between the declared and audited EBTs for a specific sector. Finally, a value ranging from 0 to 1 is considered as the threshold here. Companies with E values greater than this threshold classify as misstated and companies with E values smaller than this threshold classify as not-misstated companies. Tax authorities and experts of INTA selected 0.7 for this threshold but other values can be chosen.

Fig. 4 illustrates the audit selection problem (Hsu et al., 2015) . The y-axis is the ratio of the average benefit (or revenue) obtained from an audit case to the average cost of an audit; the x-axis is the company rank (*lower* values of x-axis are equivalent to larger companies and *higher* values of x-axis are equivalent to smaller companies). Different areas of this curve can be selected for tax evasion investigation and modeling. Based on the current economic and taxation characteristics of Iran and the opinion of tax experts and authorities, we selected the area labeled with a star (large companies with higher amounts of tax evasion) to investigate. As mentioned, according to Eq. (5), samples with higher values of E are larger companies with higher amount of tax evasion and samples with smaller values of E are smaller companies with lower amount of tax evasion.

In Iran, as in most other countries, audit costs (labor costs, organizational costs, etc.) are not significant comparing to the benefits of finding misstated companies and obtaining revenues. Therefore, the focus of this research is on larger companies with higher amounts of E because audit efforts will likely result in the recovery of larger amounts of revenue.

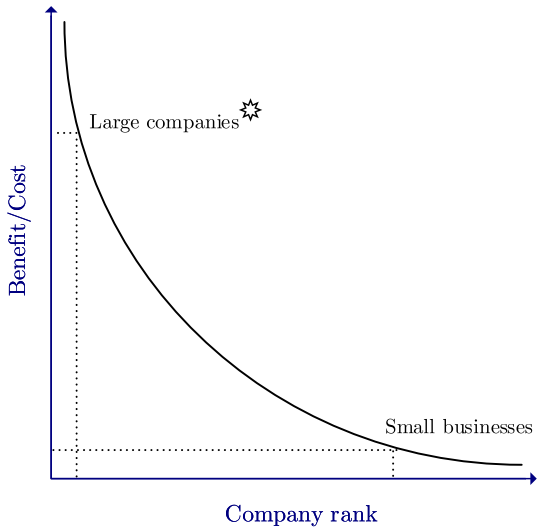

Fig. 4. The picture for the audit selection problem Hsu et al. (2015).

4.3. Sampling

One of the major challenges in regards to corporate tax evasion detection is the fact that misstated tax returns constitute a small percentage of the total tax returns. Therefore, each misstated company was matched with a not-misstated company based on size (total assets as a proxy of company size) and ISIC code ('Sampling' phase in Fig. 3).

We classify not-misstated companies in 5 groups based on the amount of total assets in a specific ISIC code. For each misstated company, a matching not-misstated company was picked from the same industry and total asset class. This results in an equal number of misstated and not-misstated companies in the sample.

4.4. Financial variables normalization

In this phase, calculated financial variables were normalized by mapping mean values to 0 and standard deviations to 1 by the following equation ('Financial variables normalization (whitening)' phase in Fig. 3):

$$N = \frac{(UN - \mu_{UN})}{\sigma_{UN}} \tag{6}$$

where UN is the financial variable before normalization, μ_{UN} is the average of UN, σ_{UN} is the standard deviation of UN, and N is the normalized financial variable.

4.5. Optimization box process

The proposed system optimizes model parameters and a combination of financial variables according to the two optimization processes in Fig. 3. It selects optimal 'Normalized financial variables' and 'MLP and SVM classification models parameters' simultaneously using the 'Harmony search optimization algorithm'. Decision variable size of the optimization algorithm is 21 (number of initial financial variables) plus the number of parameters for every classification model (2 for MLP neural network and SVM, 0 for LR). So for MLP neural network and SVM, the number of decision variables is 23 and for LR, the number of decision variables is 21. Considering the fact that the MLP neural network and SVM results are sensitive to train/test/validation indices and model parameters, and the LR model is only sensitive to train/test/validation indices, we used 8 parallel repetitions of 10-fold cross-validation in order to increase the reliability of average accuracies obtained (Demšar, 2006) ('10-fold cross-validation' and '8 parallel repetitions' in Fig. 3). Cost function of optimization algorithm is calculated based on the averaged classification error rates of 80 models (10-fold cross-validation repeated for 8 times). This repetition is due to the difference in results (output probabilities) in different runs of 10-fold cross-validation, random sample selection of this class-validation procedure, and difference in results in MLP neural network training phase in cases of using MLP neural network. The eight repetitions are based on available hardware resources and 10-fold cross-validation is a standard procedure.

In addition, before starting HS optimization algorithm, we need to set the maximum number of iterations, harmony memory size (number of harmonies in harmony memory), number of decision variables, harmony memory consideration rate (r_{accept}) (the rate of choosing a value from the harmony memory), and pitch adjustment rate (r_{pa}) (the rate of choosing a neighboring value).

4.6. Removing outliers and preparing final results

80 optimized models are found based on the best combination of financial variables and model parameters. We use these 80 models as a black box for detection of corporate tax evasion and import out-of-sample financial variables based on selected financial variables of the optimization algorithm to each of 80 models. Every model returns a probability of tax evasion for that specific company. To improve the reliability of the system, output probabilities that were greater than the mean output probability by two standard deviations or more were identified as outliers and removed from reported results. The average of all the remaining probabilities is calculated as the probability of tax evasion for a specific out-of-sample company that has the same ISIC code of the created model. It should be noted that this outlier detection procedure is applied only on the output tax evasion probabilities of the system not the input features.

5. Method

5.1. Data and software

The implemented system was evaluated against data from the food and textile sectors. The food sector dataset consists of 3097 companies (209 misstated and 2888 not-misstated companies before sampling) and the textile sector consists of 2356 companies (178 misstated and 2178 not-misstated companies before sampling). The dataset which is used for designing models (train, test, and validation datasets) is from tax returns in a time span of two years (2010 and 2011). Also, 1118 and 906 companies from tax returns of the subsequent year (2012) were used as out-of-sample of food and textile sectors. ISIC codes for the evaluated sectors are presented in Table 2. These two sectors are selected because of a higher population size compared to

Table 2
Related ISIC codes of sectors.

Sector	ISIC codes
Food	1595,1599,1599,1599,1599
Textile	1515,1519,1519,1515,1519,1515,1519

other sectors, authorities' preferences and data availability. The sample is limited to two years due to structural breaks in the last decade in the economic environment of Iran (Gujarati, 2007) .

MATLAB R2014a, R2014b, and R2015a 64-bit were used to implement IFDS[1] for INTA and for reporting purposes. All financial variables and related information were extracted directly from the INTA databases using Microsoft Excel 2013 and Microsoft SQL Server 2014. LIBSVM (Chang and Lin, 2011) (an open-source library for support vector machine) was used for implementing SVM classification model.

5.2. Classification models and optimization algorithm configuration

In the proposed system, MLP neural network can have one layer or two layers. For the first layer, the searching interval of HS optimization algorithm is [4, 25] and for the second layer, the searching interval is [0, 25]. If the second output is zero, the second layer is removed. For input and hidden layer(s), the defined activation function is tangent sigmoid and for output layer, the defined activation function is softmax. Training algorithm of MLP neural network is scaled conjugate gradient backpropagation. Maximum validation check is equal to 6, maximum epochs is equal to 1000, performance goal is equal to 0, minimum gradient is equal to $1e-06$, sigma is equal to $5e-05$, lambda is equal to $5e-07$ (default parameters of MATLAB for classification neural network), and performance function of implemented MLP neural network is cross-entropy.

The radial basis function (RBF) is kernel function of SVM. For C parameter, HS optimization algorithm searching interval is $[10^{-7}, 10^{7}]$, and searching interval of γ parameter is $[10^{-3}, 10^{3}]$. As mentioned in subsection 4.5, optimization of LR classification model is only for financial variables combination. All searching intervals of MLP and SVM models were obtained based on trial and error and one of the main tasks in development of the proposed system is to investigate different intervals of each parameter and other configurations for the proposed models.

As mentioned in subsection 3.1, MLP neural network needs train, test, and validation datasets but in the 10-fold cross-validation used in this research, only train and test datasets were available as output indices. To overcome this problem, we select the first 10% of data for validation, the second 10% for testing, and all the remaining parts for training (equals to 80% of total sample size) in the first iteration. In the second iteration, we select the second part (the second 10%) as validation set, the third part as test set, and all the remaining parts as training set. This process continues for 10 iterations to cover the entire data. In this procedure, neural network is trained by total sample but is not tested against total sample.

HS optimization algorithm parameters are based on default proposed values of algorithm designers (Geem, 2009) . Harmony memory size and the number of new harmonies is equal to 40, fret width (bandwidth) is equal to 0.01, pitch adjustment rate is equal to 0.3, fret width damp ratio is equal to 0.999, and, finally, harmony memory consideration rate is equal to 0.9. The total number of iterations of HS optimization algorithm is set to 150.

To increase confidence in the results, the proposed system was run 10 times for every combination of models and optimization algorithm. We averaged 10 runs of output probabilities to obtain one output probability for every imported company. It should be noted that this arbitrary 10 runs of the system is not usually performed in practice. This procedure is only for increasing confidence in the results and more importantly, determining the most frequently selected input financial variables.

6. Results and discussion

In this section, optimization process comparison is examined in subsection 6.1, out-of-sample results are examined in subsection 6.2, output probabilities comparison are illustrated and reviewed in subsection 6.3, additional performance measurement metrics to compare models are examined in subsection 6.4, and, finally, selected financial variables are presented in subsection 6.5.

6.1. Optimization process comparison

Total accuracy, sensitivity, specificity, and Area Under Receiver Operating Characteristic (AUROC)[2] for test set results of the system are presented in Table 3 and confusion matrices is presented in Table 4. Based on the obtained results, SVM classification model has higher total accuracy compared to MLP and LR. For the food sector, a slightly higher accuracy for LR (compared to MLP) and for the textile sector, a slightly higher accuracy for MLP (compared to LR) was obtained. Otherwise, McNemar's statistical

[1] Intelligent Fraud Detection System.

[2] ROC is plotted by true positive rate versus false positive rate. AUROC is an unbiased estimator of the probability of correctly ranking a (event, no-event) pair (Bamber, 1975).

Table 3
Test set results.

Model	Sector	Accuracy	Sensitivity	Specificity	AUROC
MLP	Food	85.37	82.69	88.05	0.853
	Textile	84.17	83.97	84.37	0.841
SVM	Food	88.34	87.50	89.18	0.883
	Textile	87.59	86.62	88.55	0.876
LR	Food	85.56	81.50	89.59	0.855
	Textile	81.88	77.50	86.16	0.818

test is not appropriate to compare performance across models on the test data set results because these are the output of 800 models trained and tested on data that was chosen randomly in 10 runs of 8 repetitions of 10-fold cross-validation.

To compare MLP and LR classification models and considering only sensitivity, MLP classification model can detect higher misstated companies. In the proposed system and based on tax evasion criterion presented in subsection 4.2 and Fig. 4, higher sensitivity is more critical than specificity due to tendency of the system to find larger misstated companies. This is a cost-benefit situation where finding larger companies with a higher amount of tax evasion has priority over smaller companies with a lower amount of tax evasion. In addition, it should be noted that due to the lower costs of auditing compared to the benefits of discovering tax evasion, sensitivity is more critical than specificity in Iran. Considering total accuracy, sensitivity and specificity, respectively, SVM, MLP, and LR are the best classification models based on the optimization process.

Best accuracy (maximum accuracy value), average accuracy, sensitivity (detection rate of misstated companies), and specificity (detection rate of not-misstated companies) of the averaged 10 runs are presented in Figs. 5 and 6 for the food and textile sectors. In both sectors, SVM classification model outperformed other classification models based on the best and average accuracy.

As shown in Figs. 5 and 6, the SVM classification model has a higher sensitivity to random initial values of decision variables compared to MLP and LR classification models. This occurs because the SVM model has lower accuracy in the starting point of the optimization algorithm process. Also, based on the sensitivity and specificity plots, lower best accuracy of SVM classification model at the starting point originates from lower specificity rather than sensitivity.

In order to perform statistical comparisons of model accuracy, all the training, testing, and validation before sampling datasets were imported to the models. Obtained results are presented in Table 5, confusion matrices are presented in Table 6, and finally McNemar's test results are presented in Table 7.

As shown in Table 7, MLP has higher total accuracy in the food sector compared to SVM and LR, and SVM has higher total accuracy in the textile sector compared to MLP and LR and these differences are statistically significant at the 5% confidence level. However, as mentioned, in the test set results, SVM has higher total accuracy compared to other classification models in both sectors. Therefore, we examine the performance of the system using out-of-sample data in subsection 6.2.

6.2. Out-of-sample results

Total accuracy, sensitivity, specificity, and AUROC for out-of-sample results of the system are presented in Table 8. As mentioned in subsection 5.2, average output probabilities of 10 runs of the system were used to calculate results and to compare models.

We compared the obtained results using McNemar's test in Table 8. MLP has higher total accuracy compared to SVM and LR in the food sector and SVM has higher accuracy compared to MLP and LR in the textile sector and these differences are statistically significant at the 5% confidence level. Confusion matrix of MLP, SVM, and LR classification models for the food and textile sectors of out-of-sample dataset are presented in Table 10.

The differing results for the test, total, and out-of-sample data show why it is important to evaluate the models on these different data sets. The additional analyses include only the MLP and SVM models, given that the LR model consistently performs more poorly than the other two models (based on lower accuracies in Tables 5 and 8) and in three out of four tests, the difference between LR and other classification models is statistically significant (based on statistical tests in Table 7 and 9).

Table 4
Confusion matrices of test set.

		PV						
		M	NM	M	NM	M	NM	
AV	M	169	35	183	26	170	39	Food
	NM	25	181	23	186	22	187	
	M	147	28	154	24	138	40	Textile
	NM	27	148	21	157	24	154	
		MLP		SVM		LR		

† AV: Actual value, PV: Predicted value.
M: Misstated, NM: Not-misstated.

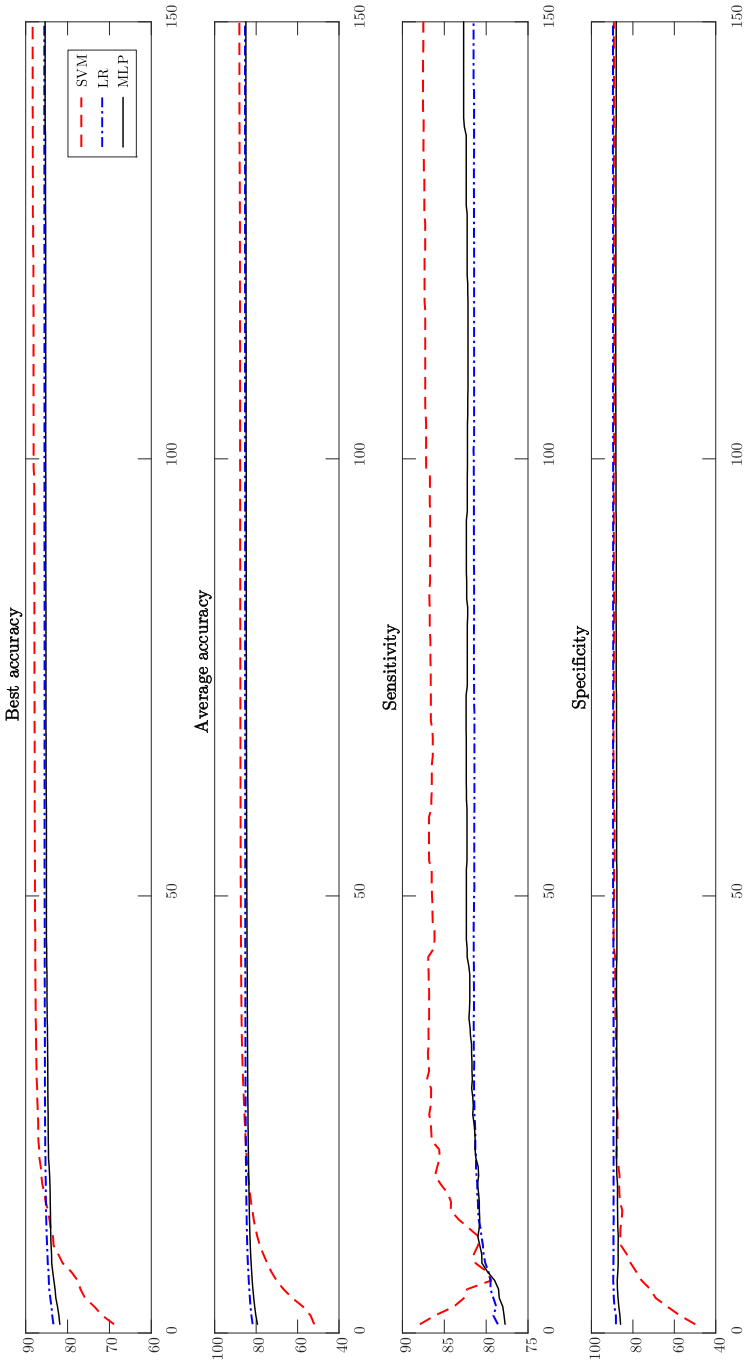

Fig. 5. Food sector optimization results.

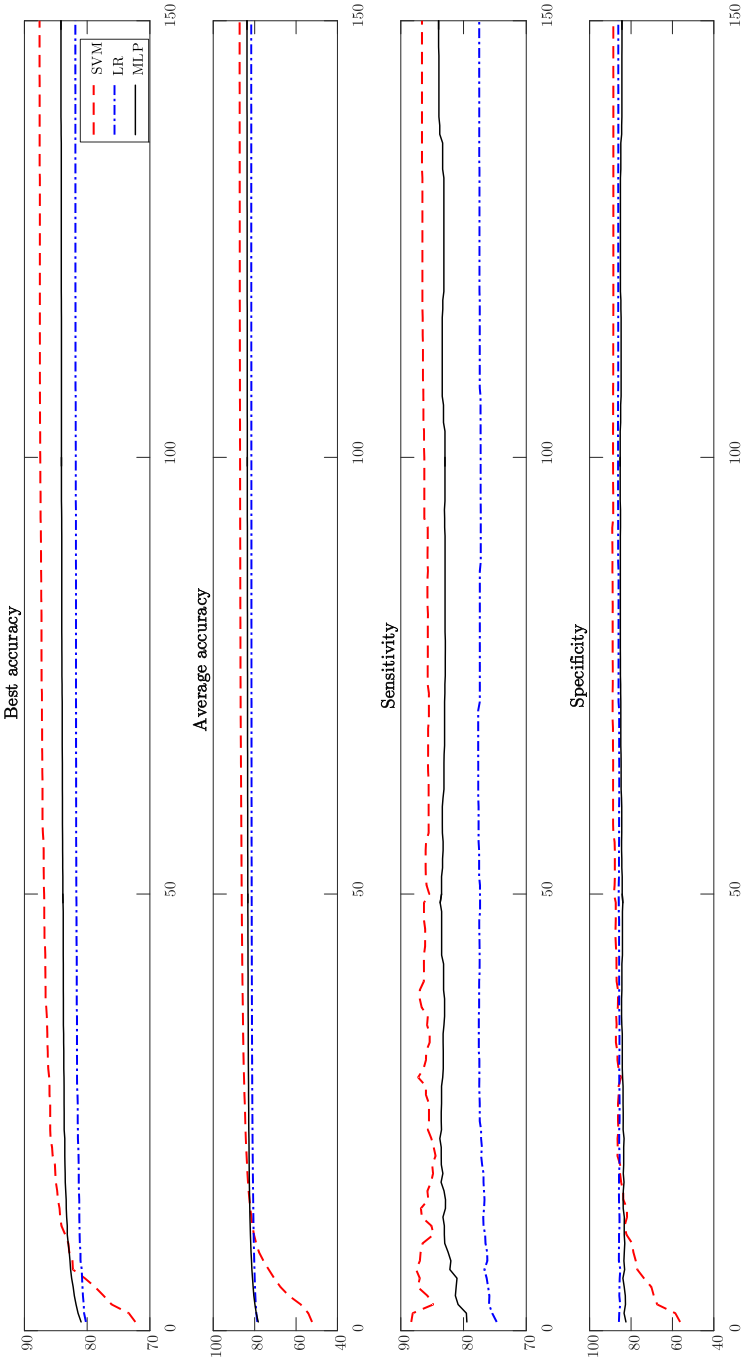

Fig. 6. Textile sector optimization results.

Table 5
Total data results.

Model	Sector	Accuracy	Sensitivity	Specificity	AUROC
MLP	Food	87.84	83.50	88.23	0.858
	Textile	80.56	83.24	80.32	0.817
SVM	Food	83.39	84.00	83.34	0.836
	Textile	85.24	84.97	85.27	0.851
LR	Food	83.02	78.00	83.47	0.807
	Textile	77.69	76.88	77.77	0.773

Table 6
Confusion matrices of total data.

		PV						
		M	NM	M	NM	M	NM	
AV	M	167	33	168	32	156	44	Food
	NM	265	1986	375	1876	372	1879	
	M	144	29	147	26	133	40	Textile
	NM	370	1510	277	1603	418	1462	
		MLP		SVM		LR		

† AV: Actual value, PV: Predicted value.
M: Misstated, NM: Not-misstated.

In subsections 6.3 and 6.4, we compare output probabilities of SVM and MLP models and examine additional performance measurement metrics in two sectors.

6.3. Output probabilities comparison

Histogram of output probabilities of out-of-sample data for MLP and SVM classification models in the food and textile sectors are presented in Figs. 7 and 8.

This study employed two-sample Kolmogorov-Smirnov test (Manoukian, 1986) to verify whether the output probabilities for any two classification models come from the same distribution or not. In this test, the null hypothesis states that output probabilities of MLP and SVM come from populations with the same distribution. In both cases of food and textile sector, at the significance level of 5%, the null hypothesis is rejected, therefore, the output probabilities of MLP and SVM come from different populations.

Furthermore, the tendency for higher values of misstatement probabilities is greater in SVM classification model compared to MLP and this effect is greater in the textile sector (63.33% of differences are positive in food sector and 70.20% of differences are positive in textile sector). As mentioned in subsections 6.1 and 6.2, MLP has higher total accuracy in the food sector and SVM has higher total accuracy in the textile sector in both total and out-of-sample data. These different results indicate that we should develop a specific classification model for all sectors. This will improve the ability of comparison between different sectors.

Table 7
McNemar's test results of total data.

McNemar chi-square *p*-value	MLP	SVM	LR
MLP	—	44.6494 0.0000	41.0227 0.0000
SVM	35.3616 0.0000	—	0.5144 0.4732
LR	8.5827 0.0033	50.7220 0.0000	—

† Dark gray: food sector, light gray: textile sector.

Table 8
Out-of-sample results.

Model	Sector	Accuracy	Sensitivity	Specificity	AUROC
MLP	Food	90.07	85.48	90.34	0.879
	Textile	82.45	84.85	82.26	0.835
SVM	Food	87.47	90.32	87.31	0.888
	Textile	84.65	74.24	85.48	0.7986
LR	Food	86.76	77.42	87.31	0.823
	Textile	79.13	78.79	79.17	0.789

Table 9
McNemar's test results of out-of-sample.

McNemar chi-square *p*-value	MLP	SVM	LR
MLP	—	10.6055 0.0011	9.5211 0.0020
SVM	14.7162 0.0001	—	0.3181 0.5727
LR	7.1129 0.0076	24.3951 0.0000	—

† Dark gray: food sector, light gray: textile sector.

Table 10
Confusion matrix of out-of-sample.

		PV						
		M	NM	M	NM	M	NM	
AV	M	53	9	56	6	48	14	Food
	NM	102	954	134	922	134	922	
	M	56	10	49	17	52	14	Textile
	NM	149	691	122	718	175	665	
		MLP		SVM		LR		

† AV: Actual value, PV: Predicted value.
M: Misstated, NM: Not-misstated.

In selecting MLP or SVM classification models, we should also focus on different aspects of the system architecture. One of these aspects is the optimization duration. Based on the obtained results, in larger sample sizes, SVM model is slower (i.e., has a higher system runtime) than MLP in the training phase. If we consider the duration of optimization process and resource limitations, using the MLP classification model is more reliable in the long term. In addition, it is important to consider metrics that concentrate more on detection of misstated companies than detection of not-misstated companies to select the best classification model.

Table 8 shows that for the food sector, total accuracy is higher for MLP but AUROC is higher for SVM. In the textile sector, total accuracy for SVM is higher but AUROC is higher for MLP. Total accuracy depends on the ability of the classifier to rank patterns, but also on its ability to select a threshold in the ranking used to assign patterns to the positive class if above the threshold. However, AUROC measures the classifier's ability to ranking a set of patterns according to the degree to which they belong to the positive class. For the food sector, MLP selects the threshold well, but ranks samples with lower performance and for the textile sector, MLP ranks samples well, but selects the threshold with lower performance. Based on the above results, there is a need to consider overall accuracy versus the AUROC in additional sectors for future research.

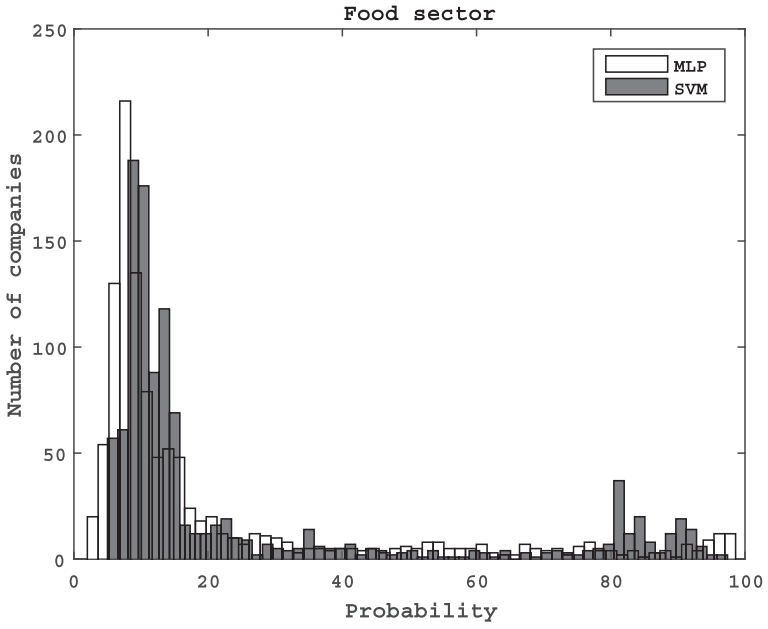

Fig. 7. Food sector output probabilities histograms.

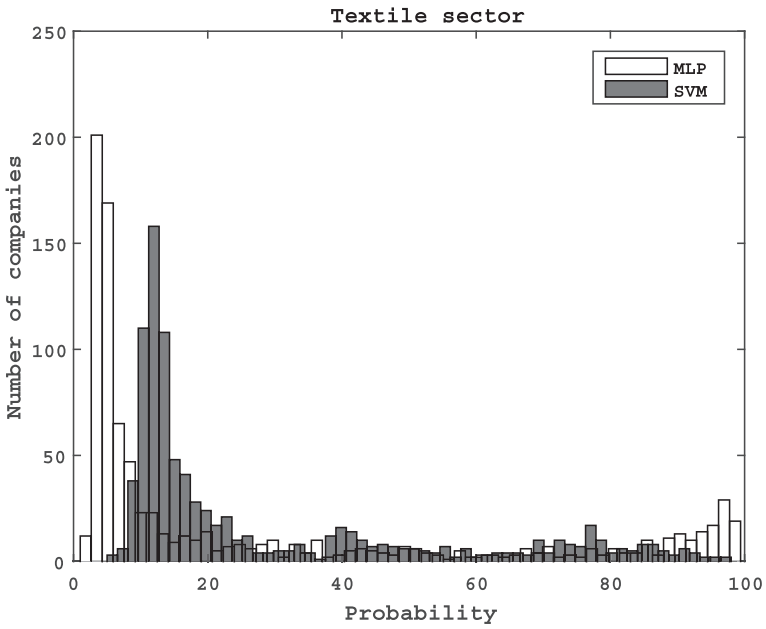

Fig. 8. Textile sector output probabilities histograms.

6.4. Additional performance measurement metrics to compare models

F-measure (Kononenko and Kukar, 2007) , which concentrates on the detection of misstated companies and the Matthews correlation coefficient (MCC) (Matthews, 1975) which takes into account true and false positives and negatives were used to compare MLP and SVM models. The F-measure is presented in Eq. (7):

$$F_\beta = \frac{(1+\beta^2) \times TP}{(1+\beta^2) \times TP + \beta^2 \times FN + FP} \tag{7}$$

where *TP* indicates true positive rate, *FN* indicates false negative rate, and *FP* indicates false positive rate. Different values for β can be considered. Here we have chosen the commonly used value 2 for β which weights recall (also known as sensitivity) higher than precision (the proportion of number of true positives to number of true positives and false positives) in corporate tax evasion detection case. This takes into account the fact that the costs of auditing in Iran are lower than the earnings from finding misstated companies. The MCC is presented in Eq. (8):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{8}$$

where *TN* indicates true negative rate. Additional performance measurement metrics results for out-of-sample data are presented in Table 11.

In both food and textile sectors, higher values of F-measure and MCC are obtained from the MLP classification model.

6.5. Selected financial variables

The selected financial variables obtained by the MLP neural network classification model in combination with harmony search optimization algorithm are presented in Tables 12 and 13, respectively, for the food and textile sectors. The numbers in

Table 11
F-measure and MCC performance measurements for out-of-sample data.

Model	Sector	F-measure	MCC
MLP	Food	**0.6576**	**0.5022**
	Textile	**0.5970**	**0.4168**
SVM	Food	0.6393	0.4731
	Textile	0.5632	0.3966

Table 12
Food sector selected financial variables.

Financial variable	Number of repetitions
Total assets	10
Net income	10
EBIT/Total assets	8
Net sales/Total assets	8
Total liabilities	8
Gross profit/Total assets	8
Gross profit	8
Total liabilities/(TL+SE)	7

Table 13
Textile sector selected financial variables.

Financial variable	Number of repetitions
Net sales/Total assets	10
Net income/Total assets (ROA)	10
Total assets	10
Gross profit/Total assets	10
Net profit	10
EBIT/Total assets	9
Retained earnings/Total assets	9
Gross profit	5

both tables represent the number of repetitions of financial variables in 10 system runs, in cases where the variables appear more than five times. Variables are included, regardless of whether the coefficient sign is positive or negative. Overlapping selected financial variables in the food and textile sectors are highlighted in the tables.

As can be seen in Table 12 and 13, 'Net income', 'Total liabilities', and 'Total liabilities/(TL+SE)' are selected variables only for the food sector and 'Net income/Total assets (ROA)', 'Net Profit', and 'Retained earnings/Total assets' are selected variables only for the textile sector. As mentioned, most of the research published in this or related areas mainly focuses on designing a model for a mixed dataset that contains different sectors. However, here we can see there are notable differences in the selected financial variables for the two examined sectors. Based on the above results, it can be concluded that different financial variables may be associated with tax evasion behavior across each sector.

7. Conclusion and future research directions

This research presents results obtained from a system that has been developed for INTA by the authors to detect corporate tax evasion. This system combines classification models with optimization algorithm to optimize financial variables combination besides classification model parameters. We verified this system using datasets from two sectors, using 10-fold cross-validation and an iterative method of system training, testing, and validation which is presented in the previous sections. In the test phase, it was concluded that SVM outperformed other classification models based on the optimization process and SVM has higher sensitivity to random initial values of decision variables. Analyzing out-of-sample and total data sets showed that MLP performed best in the food sector and SVM performed best in the textile sector. This highlights the importance of using out-of-sample for classification models evaluation.

In the next phase, significantly different distributions of output probabilities were found for MLP and SVM, so it can be concluded that the same classification model should be used (MLP or SVM in this study) to make the results from different sectors comparable. In the last phase, we found that MLP outperformed SVM in both sectors based on F-measure and MCC. Finally, we found that there are differences between selected financial variables of the two examined sectors.

Out of sample results show that the proposed hybrid intelligent system can discover hidden patterns in tax returns to detect tax evasion before auditing with acceptable accuracy. Also, probabilities obtained from the system can work as a signal for auditors and guide them towards suspicious taxpayers that should be audited with greater scrutiny. Furthermore, the selected financial variables for every sector reveal a valuable amount of information about the structure of tax evasion in that specific sector. It can be concluded that this hybrid intelligent system does not just work as a black-box to detect suspicious taxpayers, but can be a system that identifies patterns suggestive of tax evasion.

To extend this research, other system structures (i.e., different classification models, optimization algorithms, sampling methods, etc.) can be evaluated. In addition, other sectors and longer time periods should be examined to increase the reliability of our system and results. Finally, and most importantly, this is a flexible system, so as a result, this system can be implemented

in tax infrastructure of other countries with only minimal modifications based on the tax system characteristics of each nation. One current major limitation of neural network and SVM is their 'black box' nature and in particular, the fact that it is not very clear in what direction and how much model input(s) can affect model output(s). It should be noted that the first and the foremost purpose of this research and the implemented system is to achieve higher accuracy in tax evasion detection at the nationwide level, not analysis of aforementioned effects. This is another area for future research.

Acknowledgments

This paper and the results obtained are based on a system that is developed by the authors for INTA. Thanks to Dr. Ali Askari, former Head of the Iranian National Tax Administration for his support for the IFDS project and the publication of this paper. The authors also would like to thank the support of INTA under grant number 240/7624.

References

Bamber, D., 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J. Math. Psychol. 12, 387–415.
Bellovary, J.L., Giacomino, D.E., Akers, M.D., 2007. A review of bankruptcy prediction studies: 1930 to present. J. Financ. Educ. 1–42.
Bengio, Y., LeCun, Y., et al. 2007. Scaling learning algorithms towards AI. Large Scale Kern. Mach. 34.
Brigham, E., Ehrhardt, M., 2013. Financial Management: Theory & Practice. Cengage Learning.
Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 27:1—27:27. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.
Cox, D.R., 1958. The regression analysis of binary sequences. J. R. Stat. Soc. Ser. B Methodol. 215–242.
Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30.
Fischthal, S., 1998, Oct. 13. Neural network/conceptual clustering fraud detection architecture. US Patent 5822741
Geem, Z.W., 2009. Music-inspired Harmony Search Algorithm: Theory and Applications. 191. Springer Science & Business Media.
Gujarati, D.N., 2007. Sangeetha (2007) Basic Econometrics. 110. Tata McGraw Hill Publishing Company Limited, New Delhi., pp. 451–452.
Gupta, M., Nagadevara, V., 2007. Audit Selection Strategy for Improving Tax Compliance: Application of Data Mining Techniques. Foundations of Risk-Based Audits. Proceedings of the eleventh International Conference on e-Governance, Hyderabad, India, December. pp. 28–30.
Hosmer, D.W., Jr, Lemeshow, S., 2004. Applied Logistic Regression. John Wiley & Sons.
Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. 2003. A Practical Guide to Support Vector Classification.
Hsu, K.-W., Pathak, N., Srivastava, J., Tschida, G., Bjorklund, E., 2015. Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue. Real World Data Mining Applications. Springer. pp. 221–245
Kononenko, I., Kukar, M., 2007. Machine Learning and Data Mining: Introduction to Principles and Algorithms. Horwood Publishing.
Krieger, C., 1996. Neural Networks in Data Mining.
Kumar, P.R., Ravi, V., 2007. Bankruptcy prediction in banks and firms via statistical and intelligent techniques-a review. Eur. J. Oper. Res. 180, 1–28.
Lee, K.S., Geem, Z.W., 2005. A new meta-heuristic algorithm for continuous engineering optimization: harmony search theory and practice. Comput. Methods Appl. Mech. Eng. 194, 3902–3933.
Lin, C.-C., Chiu, A.-A., Huang, S.Y., Yen, D.C., 2015. Detecting the financial statement fraud: the analysis of the differences between data mining techniques and experts judgments. Knowl.-Based Syst.
Manoukian, E.B., 1986. Mathematical Nonparametric Statistics.
Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochim. Biophys. Acta Protein Struct. Mol. Enzymol. 405, 442–451.
Møller, M.F., 1993. A scaled conjugate gradient algorithm for fast supervised learning. Neural Netw. 6, 525–533.
Persons, O.S., 2011. Using financial statement data to identify factors associated with fraudulent financial reporting. J. Appl. Bus. Res. 11, 38–46.
Phua, C., Lee, V., Smith, K., Gayler, R., 2010. A Comprehensive Survey of Data Mining-based Fraud Detection Research. arXiv preprint arXiv:1009.6119
Seo, K.-K., 2007. An application of one-class support vector machines in content-based image retrieval. Expert. Syst. Appl. 33 (2), 491–498.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc., New York, NY, USA.
Wu, R.-S., Ou, C.-S., Lin, H.-y., Chang, S.-I., Yen, D.C., 2012. Using data mining technique to enhance tax evasion detection performance. Expert. Syst. Appl. 39, 8769–8777.
Yu, F., Qin, Z., Jia, X.-L., 2003. Data Mining Application Issues in Fraudulent Tax Declaration Detection. Machine Learning and Cybernetics, 2003 International Conference on (pp. 2202–2206). IEEE volume 4.