Available online at http://ijim.srbiau.ac.ir/
Int. J. Industrial Mathematics (ISSN 2008-5621)
Vol. 12, No. 1, 2020 Article ID IJIM-1052, 15 pages
Research Article

A Hybrid DEA Based CHAID and Imperialist Competitive Algorithm for Stock Selection

F. Faezy Razi *†

Received Date: 2017-04-25 Revised Date: 2018-08-25 Accepted Date: 2019-05-18

Abstract

This paper proposes a new framework for the formation of an optimal stock portfolio. The paper will argue that how an optimal stock portfolio is designed through the proposed approach compared with previous methods. In this paper, the investment portfolio is formed based on the data mining algorithm of CHAID on the basis of the risk status criteria. In the next step, the second investment portfolio is created based on the decision rules extracted by the DEA-BCC model. The final portfolio is created through a two-objective mathematical programming model based on the Imperialist Competitive algorithm. The proposed methodology is applied on a case study in the Tehran Stock Exchange. The results of the CHAID algorithm implementation based on the risk output field showed that all candidate stocks do not fall in one class and that is why it is necessary that each class of candidate stocks must be evaluated independently of other classes. The result of the Imperialist Competitive algorithm in small and large scale based on the Taguchi method showed that the studied stocks are calibrated with the used method. Unlike other models of stock portfolio selection, this paper first classifies the Stocks through the CHAID algorithm. The classified stocks in each class are evaluated independently of other classes through the DEA-BCC model. After narrowing the search space, the optimal portfolio is selected through the Imperialist Competitive algorithm.

Keywords : Data mining; Classification; DEA Based CHAID; Imperialist Competitive Algorithm; Stock Selection.

1 Introduction

One of the main decision making issues in decision theory is the problem of multi criteria decision of stocks portfolio Selection [39]. In this decision problem, the decision maker is trying to create an optimal portfolio for stock selection [27]. The high volume of traded shares in stock exchange changes the decision problem to condition of NP-Hard optimization problems [24]. Given the diversity of studied stocks in these kinds of studies and the multiplicity of the parameters governing the decision problem, the use of data mining techniques provides better analysis for the decision maker [32]. Therefore, basically in order to deal with the problem of selecting a portfolio of stocks systematically, the combination of data mining techniques with multiple criteria decision analysis models is necessary. The major combined studies carried out in this

*Corresponding author. f.faezi@semnaniau.ac.ir, Tel:+98(233)3654040.
†Department of Industrial Management, Semnan Branch, Islamic Azad University, Semnan, Iran.

area include: Grey Based KOHONEN for technology selection problem [8] and project portfolio selection problem,Grey Based Fuzzy C-means for the selection problem of oil projects [9] and DEA-CCR Based K-Means problem for selection of maintenance activities [30].Through the data mining techniques, stocks traded on the stock exchange are clustered or classified [11]. If it is not possible to define a target field for the studied data, it is necessary to use clustering methods such as KOHONEN, K-means, C-means and Two Step Clustering [18]. In situations where it is possible to define the target field for the decision maker, the application of classification methods such as C&R Tree, CHAID, C5 and QUEST seems reasonable [22]. Using multiple criteria decision analysis methods, the studied stocks are analyzed in one of the choice, sort, rank, elimination, design and description framework [17]. In this paper, in order to form a portfolio of stocks, based on the CHAID algorithm that is a well-known model in the classification topic and data mining, initially the decision options were classified. The basis of creation of a class was risk status of the output field. Then, through DEA-BCC mathematical programming model and based on the ranking philosophy, the decision options were ranked. In this way, an initial portfolio of studied stocks is created. For the evaluation and selection of the final portfolio of stocks, a two objective mathematical programming model based on design philosophy was used. The studied binary Pareto composition was obtained through the Imperialist Competitive Algorithm. This paper will continue as follows: The second part reviews the studies carried out in the field of stock selection. CHAID algorithm was described in part 3. Part 4 discusses the DEA-BCC model. In the fifth part, the Imperialist Competitive Algorithm was studied. The principles of stock portfolio selection are presented in section six. Parts 7 and 8 include the case study and sensitivity analysis. In the nine sections, Conclusion was presented.

2 Literature review

Wong and Cheung [37] studied the prediction and selection of stocks in the stock market of Hong Kong. Three main instruments were used for prediction and selection of stocks: fundamental analysis, technical analysis and portfolio analysis. The results indicate that the studied population was based on fundamental and technical analyses and relied less on portfolio analysis [37]. Based on the nonlinear integer programming idea, Gnanendran and Sundarraj [10] designed a backpack model for stocks selection. Based on the idea of group decision making and fuzzy sets theory and revision in Chen's method, Tiryaki and Ahlatcioglu [35] provided investors with a comprehensive approach of stock selection. In this study, each criterion is described by triangular fuzzy numbers [35]. Based on fuzzy theory, Tiryaki and Ahlatcioglu [34] proposed the Analytic Hierarchy Process to select stock portfolios. In this study, a scenario was created for the problem of weighting and ranking to select a stock portfolio through Analytic Hierarchy Process. Lai et al. [21] used time series method to predict the stock price on the stock exchange. Then, using the fuzzy theory, fuzzy decision tree was extracted for the studied stocks and the stocks in each class were studied and analyzed [21]. Huang and Jane [15] used the combined technique of moving average autoregressive exogenous prediction model and grey systems theory and rough set theory to predict and select the stock portfolio. In this study, the collected data were predicted by average autoregressive exogenous prediction model. Based on the gray theory, the studied data were clustered using K-Means Algorithm. Then, using Rough Set classification of the most appropriate combination of stocks, they were classified. In their study of the problem of stock selection, Hwang and Park [16] concentrated on the information received by the managers in stock exchange. Market timing is the major factor considered in their study [16]. Based on the idea of machine learning, Yu et al. [40] proposed Support Vector Machine method for classification of stocks. Support Vector Machine classification method used in this study had a high efficiency in performance when the studied data were nonlinear [40]. Castellano and Cerqueti [3] proposed the problem of portfolio selection based on the mean variance concept. In this study, based on the concept of pure jump processes, the issue of dynamics of stock selection

was taken into consideration. The results were analyzed by the Monte Carlo simulation method [3]. Zhang et al. [41] used causal feature selection method to predict the studied data of stock market. The study was conducted in Shanghai Stock Exchanges and used principal component analysis and Classification and Regression Tree methods to classify the stocks [41]. Shen et al. [33] used VIKOR DANP model for stock portfolio selection. The main basis of this study is fundamental analysis. The major criteria of this study included the following criteria: Earning & Cash Flow Profitability, Naive Extrapolation and Accounting Conservatism [33]. This study was carried out in line with the studies of [30, 41]. In the study of [41], only candidate stocks are classified by the C&R Tree algorithm, and no analysis is provided for ranking. In the study of [30], although ranking is done, classification is done based on the C5 algorithm. In the C5 algorithm, classification is based on the output field with Nominal scale. Obviously, this scale, in comparison with the Ordinal scale used in this study through the CHAID algorithm, offers less appropriate basis for risk analysis of the studied stocks. In this study, after the use of CHAID algorithm and DEA-BCC model, the narrowed search space is studied through the Imperialist Competitive algorithm. However, in previous studies, the search space narrowed by the C5 algorithm is analyzed through the Genetic algorithm and Firefly algorithm.

3 Chi-squared Automatic Interaction Detection (CHAID) algorithm

One of the most important factors in the complexity of predicting models produced by machine learning algorithms is the number of prediction variables [36]. In order to avoid the complexity of the model, some researchers reduce the number of predictor variables and only use the more important variables in the production of models [36]. Since there are different types of predictor variables and each plays a different role in predicting the outcomes, therefore, it is better to use all of them in creation of prediction models. CHAID algorithm is capable to implement all variables in creation of the prediction model. At high volumes, statistical data are not free of missing values. These values have a major impact on the performance of numerous machine learning algorithms. CHAID algorithm is one of the few algorithms that act appropriately in the face of missing values [25]. The tree produced by this algorithm is not necessarily a binary tree. This is one of the important characteristics of this algorithm. Therefore, the possibility of understanding and recognition of models increases for experts and shows more flexibility in application of model in important decision makings [4]. About the implementation of CHAID decision tree, it should be noted that this algorithm is a modeling technique used to study the relationships between a dependent variable and many independent variables. Predictor variables can be qualitative or quantitative. This method used Chi-Square analysis to investigate the role of qualitative independent variables and used variance analysis methods to investigate the role of independent variables. Based on the P-Value, this algorithm selects the effective variables for predicting output variable [12]. About the development method of prediction model and evaluation of its effectiveness, it can be said first using a technique, the data set should be separated to individual subsets to create and test the models. To reduce the modeling Bias, the application of K-Fold validation method is recommended for this technique [26].Law inference algorithms have some differences that are important to users. Below are the differences between these algorithms [14].

- Type of output: C5.0, QUEST and list use the symbolic output field decision (a field among organized fields). CHAID and Classification and Regression Tree are capable of producing symbolic and numerical outputs and predict the binary result decision.

- Type of categorization: When the data set is divided into subgroups recursively, Classification and Regression Tree and QUEST will only support categorization into two sub-groups (training sub-group and the test group) while CHAID, C5.0 and list support the decision of division into more than two

sub-groups (training sub-group, the test sub-group and the validation sub-group).

- The rapid growth of the tree and pruning: the three algorithms of QUEST, C5.0 and C&R Tree are fast growing trees; back pruning should be used for them that is a known effective method. But they have different pruning criteria. C5.0 includes correctness (the highest accuracy on the training sample) and universality (the results are generalized to other data).

- Results: The set of rules can be easily interpreted according to the complex decision trees. The decision tree provides a unique classification for each data record, while more than a rule may be applied among the set of rules. When one data record provides a number of laws, the first law is assigned to the desired record.

Due to the differences mentioned above, the reasons for selection of CHAID algorithm can be stated as follows: The output produced by this algorithm is symbolic, but it is not binary that due to under investigation sample data, it becomes important.

4 Imperialist Competitive Algorithm

General optimization problem exists in almost every field of science, engineering and commerce [28]. So far, great efforts have been made to solve general optimization problems. The main challenge of general optimization is that the problems which are to be optimized may have many local optimizations. Many evolutionary algorithms have been proposed so far to solve the general optimization problem [2]. In the evolutionary algorithms proposed so far, the optimal solution of the optimization problem is found by modeling the natural evolution process. This is performed through evolution of a population of candidate solutions similar to biological evolution processes that can be adapted to environmental changes [6]. The Genetics algorithm [13], Particle Swarm Optimization algorithm [7] and Simulated Annealing algorithm [20] are meta-heuristic optimization algorithms. Recently, a new algorithm called Imperialist Competitive Algorithm was proposed by Atashpaz-Gargari and Lucas [1] that is not inspired by a natural phenomenon but it is inspired by a Social-Human phenomenon. The imperialist competitive algorithm is a new algorithm in evolutionary computations founded on Socio Political evolution of human. Like other evolutionary algorithms, this algorithm also starts with a random initial population all of which are called a nation. Some of the best members are selected as colonialists and the rest of members are considered as colonial populations. By considering the function f(x) in optimization problems, x is found such that its corresponding cost becomes optimal (usually minimum). In an N_{var} dimension optimization problem, a country is an $N_{var} \times 1$ array. The array is defined as equation 4.1.

$$country = [p_1, p_2, p_3, ..., p_{N_{var}}] \tag{4.1}$$

By evaluation of function f for the variables $(p_1, p_2, p_3, ..., p_{N_{var}})$ in equation (4.2) the costs of a country is presented.

$$cost = f(country) = f(p_1, p_2, p_3, ..., p_{N_{var}}) \tag{4.2}$$

In imperialist competitive algorithm, N_{coutry} initial states are created and N_{imp} of the best members of this population (the countries with the lowest cost function) are selected as the colonialists. The N_{col} rest of the countries form colonies that each belongs to an empire. The colonialist's countries apply the absorption policy along different aspects of optimization to attract the colonies toward themselves. According to equation 1, using their power, the colonialists attract the colonies toward themselves. The total power of the empire is determined by calculating the strengths of its two constituent parts i.e. the colonialist power plus a percentage of the average power of its colonies determined based on equation (4.3) [39].

$$T.C_{.n} = C(I) + \xi mC(COMn) \tag{4.3}$$

The colonial country moves x units along the line connecting the colonial country to the colonialist and it is drawn into the new position. In figure 1, d shows the distance between colonialist and

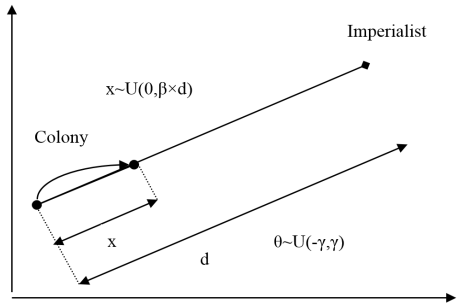

Figure 1: The movement of colonies toward the colonialist.

the colony. x is a uniformly distributed random number defined in equation (4.4) where β is a number greater than one and close to two.

$$x \sim U(0, \beta \times d) \tag{4.4}$$

A good choice would be $\beta = 2$. Also, the angle of movement is considered as uniform distribution in equation (4.5).

$$\theta \sim U(-\gamma, \gamma) \tag{4.5}$$

With a possible deviation in imperialist competitive algorithm, the colony moves toward the path of assimilation by the colonialist. This deviation angle is shown by θ that θ is chosen randomly with uniform distribution. During the movement of colonies toward the colonialist country, some of these colonies may reach to a better situation than the colonialist. In this case, the colonialist and the colony change their positions with each other. For modeling this competition, given the total cost of empire, first the probability of takeover of colonies by each empire is calculated as equation (4.6).

$$N.T.C_{.n} = T.C_{.n} - max\{T.C_{.i}\} \tag{4.6}$$

where $T.C_{.n}$ is the total cost of the n^{th} empire and $N.T.C_{.n}$ is the total normalized cost of that empire and the possible takeover of the colony by the empire is calculated as equation (4.7) [19].

$$P_{p_n} = |\frac{N.T.C_{.n}}{\sum_{i=1}^{N_{imp}} N.T.C_{.i}}| \tag{4.7}$$

5 Data envelopment analysis

Efficiency measurement because of its importance in assessing the performance of a company or organization has always been of researcher's interest. Using a method like the efficiency measurement methods in engineering topics, Farrell measured the efficiency of a manufacturing unit in 1975 [5]. The case that Farrell used for the measurement of efficiency included an input and an output. Farrell used his model to estimate the efficiency of the U.S. Agricultural Sector compared to other countries. However, he was not successful in presenting a method that incorporated multiple inputs and outputs. Charnes, Cooper and Rhodes developed the Farrell's viewpoint and presented a model that was able to measure the efficiency with multiple inputs and multiple outputs [29]. In their viewpoints, the efficiency of each decision making unit is equal to the ratio of total weighted outputs to total weighted inputs. In this expression, E_k is the efficiency of k^{th} unit under investigation. y_{rk} is the amount of r^{th} output for k^{th} decision making unit and x_{ik} is the amount of i^{th} output for k^{th} decision making unit. u_r is the weight of r^{th} output. v_i is the weight of i^{th} output. s is the number of outputs and m is the number of inputs of decision making units. Charnes, Cooper and Rhodes used this measuring technique of efficiency to present a new model. The purpose of the model was measuring and comparing the relative efficiency of organizational units with multiple similar inputs and outputs.

$$E_k = \frac{\sum_{r=1}^{s} u_r y_{rk}}{\sum_{i=1}^{m} v_i x_{ik}} \tag{5.8}$$

In model (5.9), the efficiency of the unit under investigation (K^{th} unit) is presented with the CCR model. By solving the model for the studied unit, the relative efficiency of this unit and the optimal weights to reach this efficiency are obtained. The first limitation of this model ensures that the maximum value of efficiency of decision making units is one and the next limitations ensure non negative weights for inputs and outputs. To obtain the efficiency of all decision making units, a

unique model must be solved for each unit.

$$
\begin{aligned}
Max \quad & E_k = \frac{\sum_{r=1}^{s} u_r y_{rk}}{\sum_{i=1}^{m} v_i x_{ik}} \\
\text{subject to:} & \\
& \frac{\sum_{r=1}^{s} u_r y_{rk}}{\sum_{i=1}^{m} v_i x_{ik}}, \qquad k = 1, \ldots, n, \\
& u_r \geq 0, \qquad r = 1, \ldots, s, \\
& v_i \geq 0, \qquad i = 1, \ldots, m.
\end{aligned} \tag{5.9}
$$

One of the features of data envelopment analysis model is its returns to scale structure. Returns to scale can be constant or variable. Constant returns to scale means that an increase in input amount leads to proportional increase in the amount of output. In variable returns, the increase in output is more or less than the increases ratio in the input. CCR models are among the models with constant returns to scale. Constant returns to scale models are useful when all units operate at an optimal scale. While evaluating the efficiency of the units, if incomplete conditions and space of competition impose restrictions on investment, it leads to inactivity of the unit in optimal scale [23]. In 1984, Banker, Charnes and Cooper made some changes in the CCR model to present a new model called BCC. This model is of data envelopment analysis models types that assesses the relative efficiency of units with variable returns to scale [31]. Models with constant returns to scale are more limiting than

$$
\begin{aligned}
& Max \quad E_0 = \textstyle\sum_{r=1}^{s} u_r.y_{r0} + u_0 \\
& \text{subject to:} \\
& \textstyle\sum_{i=1}^{m} v_i.x_{i0} = 1, \\
& \textstyle\sum_{r=1}^{s} u_r.y_{rk} - \sum_{i=1}^{m} v_i.x_{ik} + u_0 \leq 0, \forall k, \\
& u_r, v_i \geq 0, r = 1, \ldots, s, i = 1, ..., m. \\
& \text{W is free.}
\end{aligned} \tag{5.10}
$$

In model (5.10), x_{ij} and y_{rj} represent the j^{th} inputs and outputs of decision making units and v_i and u_r are the weights of inputs and outputs. Therefore, in the above model x_{i0} and y_{rj} are DMU_0 inputs and outputs. Also the sign of u_0 can determine returns to scale for each unit.

6 A Framework for the Creation of a Portfolio of Stocks with Hybrid DEA-BCC Based CHAID and Imperialist Competitive Algorithm

This part of the paper presents a comprehensive and new framework for the creation of stock portfolio based on data mining approach and multiple criteria decision analysis. As observed in Figure 2, in this approach, first the stocks in the stock exchange are classified using CHAID algorithm, then each class is ranked by DEA-BCC and hence the initial portfolio is formed. The final portfolio is obtained by designing a binary two objective mathematical programming model that minimizes the stock risks and maximizes the rank of each share. The Pareto solution of the related mathematical model was obtained through imperialist competitive algorithm. The framework

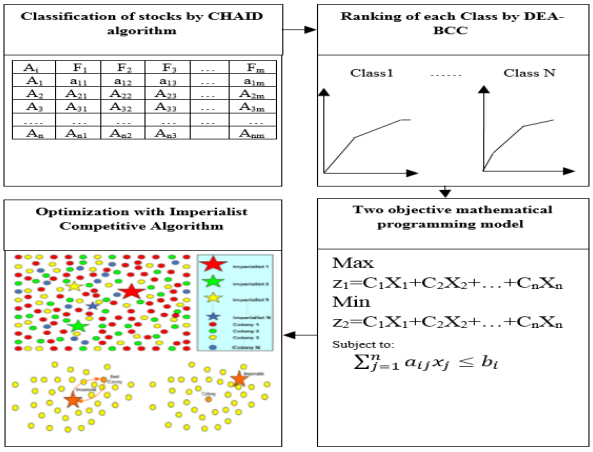

Figure 2: The framework for selection of stocks portfolio.

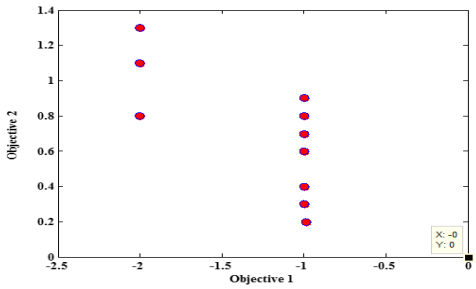

Figure 3: Risk and rank Pareto solution.

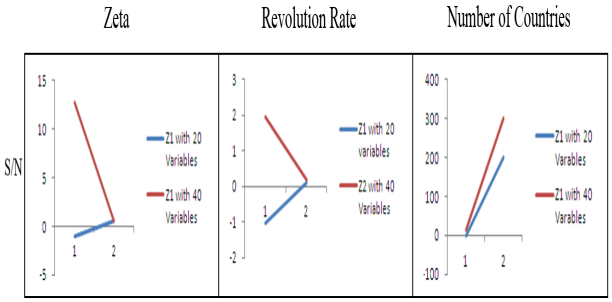

Figure 4: Taguchi ratio of the first objective function for small-scale problems (20 to 40 variables).

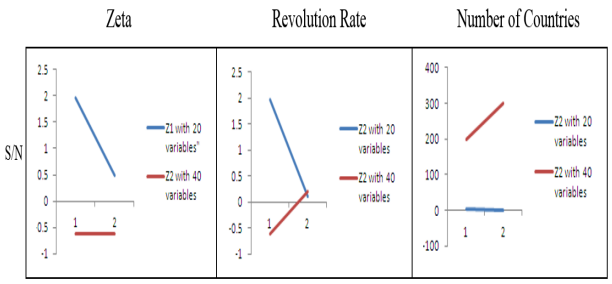

Figure 5: Taguchi ratio of the Second objective function for small-scale problems (20 to 40 variables).

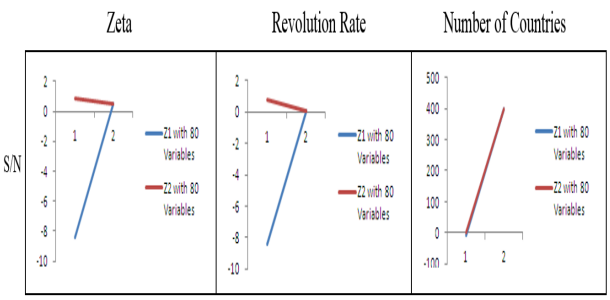

Figure 6: Taguchi ratio of the first objective function for large-scale problems (80 to 100 variables).

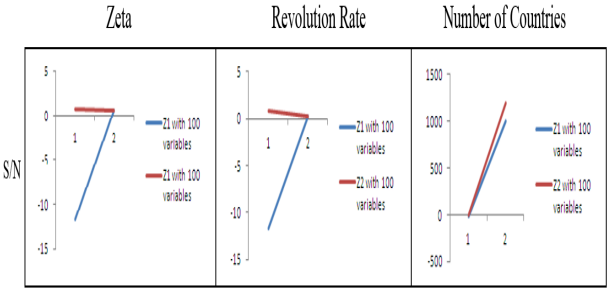

Figure 7: Taguchi ratio of the second objective function for large-scale problems (80 to 100 variables)

depicted in Figure 2 contains the main following steps;

- **Classification:** At this stage, the candidate stocks are classified based on the risks labels.
- **Ranking:** At this stage, it is ranked through DEA-BCC mathematical programming model.
- **Designing multi-objective binary programming model:** At this stage, a multi-objective binary programming model is designed that minimizes the risks per share and maximizes the ranking of each share. Variables and parameters of this model are summarized in Table 1.

In Table 4, the Beta risk coefficient per share, price per share and the expected return are presented.

Therefore, the two-objective binary mathematical programming model of this paper is presented as follows;

$$Max\ z_1 = \sum_{i=1}^{n}\sum_{j=1}^{m} Rank_{DEA-bcc_i} X_j$$

$$Max\ z_2 = \sum_{i=1}^{n}\sum_{j=1}^{m} \beta_i X_j,$$

subject to:

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{m} ROE_i X_j \geq R,\\
&\sum_{i=1}^{n}\sum_{j=1}^{m} V_i X_j \leq B,\\
&\sum_{j=1}^{m} X_j = 1\\
&\sum_{j=1}^{m} X_j = 0\\
&X_j \in \{0,1\}
\end{aligned} \tag{6.11}$$

In Table 1, j represents the number of candidate stocks and the formation of stock portfolio is considered for small scale problems with 20-40 vari-

Table 1: Variables and parameters of the mathematical model of stocks portfolio selection.

variables and parameters of the model	Description
X_j Binary variable $j = 1, ..., m$	If the Stock j is selected, it is equal to 1, otherwise, it is 0.
β_i	Risk the i^{th} Stock
$\text{Rank}_{DEA-bcc_i}$	Efficiency obtained from DEA-BCC for the i^{th} Stock
R	Short term return rate Value
ROE_i	Return On Equity the i^{th} Stock
V_i	value the i^{th} Stock
B	Budget Capacity
Research objectives;	
$Max\ z_1 = \sum_{i=1}^{n} \sum_{j=1}^{m} Rank_{DEA-bcc_i} X_j$	1-maximizing the rank of each portfolio
$Max\ z_2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \beta_i X_j$	2-minimizing the risk of each portfolio
Constraints;	
$\sum_{i=1}^{n} \sum_{j=1}^{m} ROE_i X_j \geq R$	1. Return On Equity Constraint
$\sum_{i=1}^{n} \sum_{j=1}^{m} V_i X_j \leq B$	2. Budget Constraint
$\sum_{j=1}^{m} X_j = 1$	3. Mutually Exclusive Stocks Constraint
$\sum_{j=1}^{m} X_j = 0$	4. Dependent Stocks Constraint

Table 2: Factors studied in stock selection problem by the problem scale.

Factor	Levels	
Number of Stocks	Small:20-40	Large:80-100
Number of Rank per Class	Small:20-40	Large:80-100
Number of Risk per Class	Small:20-40	Large:80-100
Number of Constraint	4-6-8	

ables and large scale problems with 80-100 variables. However, in the case study section, input data is displayed only for 20 variables. The problem constraints also respectively increased to 4, 6 and 8. Thus Model 11 is studied in small and large-scales. Table 2 shows the described factors:

- **Optimization:** At this stage, the Pareto solution of the model 11 with multi objective algorithm is analyzed using the imperialist competitive algorithm.

7 Case Study

In this part of the study, a case study is used to describe the DEA-BCC Based CHAID technique and imperialist competitive algorithm to select the portfolio of stocks. The study data presented in Table 3 are related to the stock market of Tehran Stock Exchange in 2013. The study indicators include *Price-Dividend Ratio (P/D), Price-Earnings Ratio (P/E), Price-To-Sales Ratio, Return on Equity (ROE), Return on Working Capital and Profit to Sales Ratio.* All indicators have a beneficial or the more, the better nature and they are of the Continuous Measure type.

Table 3: The study input data.

Stock No,	P/D Ratio	P/E Ratio	Price-to-Sales Ratio	Return On Equity	Return On Working Capital	Profit to Sales Ratio	Risk Status
1	1.27	10.46	5.3	13.45	14.31	54.59	2.00
2	2.54	2.89	0.23	8.65	14.23	5.63	1.00
3	2.46	-13.56	0.5	-194.9	30.38	-21.54	2.00
4	-22.8	5.98	1.31	34.89	23.7	11.06	3.00
5	6.75	-25.91	2.25	55.25	52.2	-48.22	3.00
6	1.83	-13.06	0.9	-89.85	-58.4	-32.39	3.00
7	3.62	-24.79	1.68	-40.37	-150.692	-123.37	2.00
8	0.75	0.33	0.15	-61.15	-1.08	0.38	3.00
9	1.19	-6.03	0.21	-24.94	60.71	-7.08	2.00
10	1.79	-0.79	1.38	-2.33	80.65	45	3.00
11	1.15	8.87	1.09	-32.41	-50.41	23	3.00
12	0.35	20.77	2.1	26.25	154.64	38.85	2.00
13	2.61	2.92	1.25	26.58	-30.78	19.05	4.00
14	1.47	14.67	0.33	47.38	-1,557.78	22.07	2.00
15	1.08	20.86	0.42	44.77	63.26	34.33	2.00
16	2.41	11.97	5.06	18.45	34.59	20.37	3.00
17	1.34	-11.2	2.72	-45.29	108.26	-14.37	1.00
18	1.08	7.39	2.52	22.58	68.04	41.52	2.00
19	1.25	3.36	2.31	26.79	-16.34	4.02	3.00
20	1.15	7.61	4.09	58.13	-53.68	13.54	4.00

Table 4: Beta risk coefficient per share, price per share and the expected return.

stock	1	2	3	4	5	6	7	8	9	10
risk	0.2	0.8	0.6	0.9	0.4	0.5	0.3	0.8	0.4	0.3
price	1	1.5	1.2	1.8	1.1	1.4	1.3	1.4	1.6	1.8
return	0.2	0.15	0.23	0.22	0.21	0.14	0.2	0.22	0.24	0.21
stock	11	12	13	14	15	16	17	18	19	20
risk	0.1	0.4	0.7	0.6	0.8	0.4	0.3	0.7	0.3	0.7
price	1.6	1.5	1.2	1.4	1.2	1.3	1.9	1.3	1.4	1.5
return	0.23	0.24	0.21	0.25	0.22	0.2	0.19	0.18	0.17	0.22

The last column of Table 3 is associated with the risk status that range from very low value (1) to low value (2), high value (3) and very high value (4). This scale is an Ordinal Measure.

Meanwhile, the total budget of institute is 15.6. It is expected that the dividends reach 2.3 monetary units for the total investment. After running the CHAID algorithm with IBM Modeler 14.2 software, the decision rules were derived as follows; Rule 1: If Return on working capital > -1.078 and Return on working capital ≤ 14.312 then Risk is 1.

Rule 2: If Return on working capital > 80.648 then risk is 1.

Rule 3: If Return on working capital ≤ -150.692 then risk is 2.

Rule 4: If Return on working capital > 14.312 and Return on working capital ≤ 80.648 then Risk is 2.

Rule 5: If Return on working capital> -150.692 and Return on working capital ≤ -1.078 then Risk is 3.

Table 5: Positive values of the input data.

Stock No	I(1)	I(2)	O(1)	O(2)	O(3)	O(4)
1	31.27	40.46	5.3	213.45	1614.31	204.59
2	32.54	32.89	0.23	208.65	1614.23	155.63
3	32.46	16.44	0.5	5.1	1630.38	128.46
4	7.2	35.98	1.31	234.89	1623.7	161.06
5	36.75	4.09	2.25	255.25	1652.2	101.78
6	31.83	16.94	0.9	110.15	1541.6	117.61
7	33.62	5.21	1.68	159.63	1449.308	26.63
8	30.75	30.33	0.15	138.85	1598.92	150.38
9	31.19	23.97	0.21	175.06	1660.71	142.92
10	31.79	29.21	1.38	197.67	1680.65	195
11	31.15	38.87	1.09	167.59	1549.59	173
12	30.35	50.77	2.1	226.25	1754.64	188.85
13	32.61	32.92	1.25	226.58	1569.22	169.05
14	31.47	44.67	0.33	247.38	42.22	172.07
15	31.08	50.86	0.42	244.77	1663.26	184.33
16	32.41	41.97	5.06	218.45	1634.59	170.37
17	31.34	18.8	2.72	154.71	1708.26	135.63
18	31.08	37.39	2.52	222.58	1668.04	191.52
19	31.25	33.36	2.31	226.79	1583.66	154.02
20	31.15	37.61	4.09	258.13	1546.32	163.54

Table 6: Beta risk coefficient per share, price per share and the expected return.

Stock	1	2	3	4	5	6	7	8	9	10
Class	1	1	4	4	4	5	3	5	4	2
Efficiency	1	0.99	0.95	1	1	1	1	1	1	1

Stock	11	12	13	14	15	16	17	18	19	20
Class	5	2	5	3	4	4	2	4	5	5
Efficiency	1	1	1	1	1	1	1	1	1	1

As specified in the decision rules, the number of classes relating to the research data is equal to 5 classes. The stocks 1 and 2 fall in the first class, the stocks 10, 17 and 12 fall in the second class, the stocks 14 and 7 fall in the third class, the stocks 3, 4, 5, 9, 15, 16 and 18 fall in the fourth class and the stocks 6, 8, 11, 13, 19 and 20 fall in the fifth class. Since Table 3 contains negative values, the translation invariant concept was used to solve the DEA-BCC model. The transforma-

Table 7: The main settings for application of ICA Algorithm.

Parameters	Value
Number of Imperialist	10
Number of Population	200
Max of Decades	700
Beta	0.4

Table 8: Average Pareto combination of risk and rank.

Solution	1	2	3	4	5	6	7	8	9	10
Rank	-2.001	-2	-2	-1.001	-1	-1	-1	-1	-1	-0.99
Risk	1.301	1.1	0.8	0.901	0.8	0.7	0.6	0.4	0.3	0.2

Table 9: Types of Taguchi functions for calibration.

Performance characteristic/metric	S/N ratio formula	Description of formula parameters
Smaller the better	$S/N = -10log[\frac{1}{n}\sum_{i=1}^{n} OF_i^2]$	n=number of observation (signals) OF=Objective Function
Nominal is best	Mean and Variance, $S/N = -10log(S^2)$ Variance only, $S/N = 10log(\frac{\bar{OF}}{S})^2$	OF=Average of Observation of Objective Function, S=Standard Deviation of n observation
Larger the better	$S/N = -10log[\frac{1}{n}\sum_{i=1}^{n} \frac{1}{OF_i^2}]$	n=number of observation (signals) OF=Objective Function

Table 10: Settings for sensitivity analysis of the model with Taguchi method for small-scale problems.

Factors	Value
Zeta	0.5-0.6
Revolution Rate	0.1-0.2
Number of Imperialist	10-20
Number of (Countries-Iteration)	(200,700)-(300-1000)

tion result is provided in Table 5.

Based on the translation invariant concept in Table 5, +30 was added to the first input values, +30 was also added to the second input values, +200 was added to the second output values, +1600 was added to the third output values and +150 was added to the fourth output values. The first phase of the input oriented DEA-BCC model for the sixth unit which is located in the fifth class is shown in 12. It should be noted that in this study, P/E and P/D were input criteria and other parameters were considered as output criteria.

$$\begin{aligned} &Min\ y_0 = \theta, \\ &\text{S.t:} \end{aligned} \tag{7.12}$$

Table 11: Settings for sensitivity analysis of the model with Taguchi method for large-scale problems.

Factors	Value
Zeta	0.5-0.6
Revolution Rate	0.1-0.2
Number of Imperialist	10-20
Number of (Countries-Iteration)	(400,1000)-(500-1200)

$$
\begin{aligned}
&1.83\lambda_1 + 0.75\lambda_2 + 1.15\lambda_3 \\
&+2.61\lambda_4 + 1.25\lambda_5 + 1.15\lambda_6 \geq 1.83, \\
&-13.06\lambda_1 + 0.33\lambda_2 + 8.87\lambda_3 + 2.92\lambda_4 \\
&+3.36\lambda_5 + 7.61\lambda_6 \geq -13.06, \\
&-0.90\theta + 0.90\lambda_1 + 0.15\lambda_2 + 1.09\lambda_3 \\
&+1.25\lambda_4 + 2.31\lambda_5 + 4.09\lambda_6 \leq 0, \\
&89.85\theta - 89.85\lambda_1 - 61.15\lambda_2 - 32.41\lambda_3 \\
&+26.58\lambda_4 + 26.79\lambda_5 + 58.13\lambda_6 \leq 0, \\
&58.40\theta - 58.40\lambda_1 - 1.08\lambda_2 - 50.41\lambda_3 \\
&-30.75\lambda_4 - 16.34\lambda_5 - 53.68\lambda_6 \leq 0, \\
&32.39\theta - 32.39\lambda_1 + 0.38\lambda_2 + 23\lambda_3 \\
&+19.05\lambda_4 + 4.02\lambda_5 + 13.54\lambda_6 \leq 0, \\
&\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 + \lambda_6 = 1, \\
&\lambda_j \geq 0, \theta \text{ is free.}
\end{aligned}
$$

The second phase of DEA-BCC input oriented arrayed model is presented in model 12 with description 13:

$$
\begin{aligned}
&Max\ S = s_1^+ + s_2^+ + s_3^+ + s_4^+ + s_1^- + s_2^- \\
&\text{S.t:} \\
&1.83\lambda_1 + 0.75\lambda_2 + 1.15\lambda_3 + 2.61\lambda_4 \\
&+1.25\lambda_5 + 1.15\lambda_6 - s_1^- = 1.83, \\
&-13.06\lambda_1 + 0.33\lambda_2 + 8.87\lambda_3 + 2.92\lambda_4 \\
&+3.36\lambda_5 + 7.61\lambda_6 - s_2^- = -13.06, \\
&-0.90\theta^* + 0.90\lambda_1 + 0.15\lambda_2 + 1.09\lambda_3 \\
&+1.25\lambda_4 + 2.31\lambda_5 + 4.09\lambda_6 + s_1^+ = 0, \\
&89.85\theta^* - 89.85\lambda_1 - 61.15\lambda_2 - 32.41\lambda_3 \\
&+26.58\lambda_4 + 26.79\lambda_5 + 58.13\lambda_6 + s_2^+ = 0, \\
&58.40\theta^* - 58.40\lambda_1 - 1.08\lambda_2 - 50.41\lambda_3 \\
&-30.75\lambda_4 - 16.34\lambda_5 - 53.68\lambda_6 + s_3^+ = 0, \\
&32.39\theta^* - 32.39\lambda_1 + 0.38\lambda_2 + 23\lambda_3 \\
&+19.05\lambda_4 + 4.02\lambda_5 + 13.54\lambda_6 + s_4^+ = 0, \\
&\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 + \lambda_6 = 1, \\
&\lambda_j \geq 0, s_1^+ \geq 0, s_2^+ \geq 0, s_3^+ \geq 0, \\
&s_4^+ \geq 0, s_1^- \geq 0, s_2^- \geq 0.
\end{aligned} \tag{7.13}
$$

Table 6 shows the efficiency of candidate locations in each cluster using input oriented DEA-BCC model.

For implementation of Imperialist Competitive algorithm, 30 independent tests are used. Average solution for the small scale problem with 20 variables can be seen in Table 8 and Figure 3.

Based on the model presented in this study, the binary two objective programming model is structured as follows:

$$
\begin{aligned}
Max\ Z_1 &= x_1 + 0.99x_2 + 0.95x_3 + x_4 \\
&+x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} \\
&+x_{11} + x_{12} + x_{13} + x_{14} + x_{15} \\
&+x_{16} + x_{17} + x_{18} + x_{19} + x_{20},
\end{aligned} \tag{7.14}
$$

$$
\begin{aligned}
Min\ Z_2 &= 0.2x_1 + 0.8x_2 + 0.6x_3 + 0.9x_4 \\
&+0.4x_5 + 0.5x_6 + 0.3x_7 + 0.8x_8 \\
&+0.4x_9 + 0.3x_{10} + 0.1x_{11} + 0.4x_{12} \\
&+0.7x_{13} + 0.6x_{14} + 0.8x_{15} + 0.4x_{16} \\
&+0.3x_{17} + 0.7x_{18} + 0.3x_{19} + 0.7x_{20}
\end{aligned}
$$

s.t:

$$
\begin{aligned}
&1x_1 + 1.5x_2 + 1.2x_3 + 1.8x_4 + 1.1x_5 \\
&+1.4x_6 + 1.3x_7 + 1.4x_8 + 1.6x_9 \\
&+1.8x_{10} + 1.6x_{11} + 1.5x_{12} + 1.2x_{13}
\end{aligned}
$$

$$+1.4x_{14} + 1.2x_{15} + 1.3x_{16} + 1.9x_{17}$$

$$+1.3x_{18} + 1.4x_{19} + 1.5x_{20} \leq 15.6,$$

$$0.2x_1 + 0.15x_2 + 0.23x_3 + 0.22x_4$$

$$+0.21x_5 + 0.14x_6 + 0.2x_7 + 0.22x_8$$

$$+0.24x_9 + 0.21x_{10} + 0.23x_{11} + 0.24x_{12}$$

$$+0.21x_{13} + 0.25x_{14} + 0.22x_{15} + 0.2x_{16}$$

$$+0.19x_{17} + 0.18x_{18} + 0.17x_{19}$$

$$+0.22x_{20} \leq 2.3,$$

$$x_{11} + x_{16} \leq 1,$$

$$x_4 - x_8 \leq 0,$$

$$x_i \in \{0, 1\}, i = 1, 2, ..., 20.$$

The result of application of binary two objective imperialist competitive algorithm can be seen in figure 3. In order to conduct the experiments, we implemented imperialist competitive algorithm in MATLAB R2010a run on a personal computer with a 2.3GHz up to 2.8 GHz Core i5 and 2 GB RAM memory. The main parameters of algorithm are summarized in Table 7.

8 Sensitivity Analysis

The Taguchi method is used to analyze the model sensitivity. The model control parameters are calibrated through the Taguchi Method. The basis for calibrating the control parameters in the Taguchi method is the signal to noise ratio. The term signal refers to the values of desired variables and the term noise refers to the values of unfavorable variables. The S/N ratio refers to the variance in response to the variable.According to the type of objective function, one of the functions in Table 9 is used for analysis of control parameters: Zeta, Revolution Rate and Number of Countries.

To analyze the results with the Taguchi method with small scale problems, the studied model elements are defined as seen in Table 10. Results for the first and second objective functions can be seen in figures 4 and 5.

As can be seen in figures 4 and 5, the problem is calibrated in small scale based on the Taguchi method. To analyze the results with Taguchi method with the large scale problems, the studied model elements are defined as seen in Table 11.

Results for the first and second objective functions can be seen in figures 6 and 7.

As can be seen, the problem is calibrated in large scale based on the Taguchi method.

9 Conclusion

One of the most important issues for investors in stock exchange markets is the stocks selection technique. Achieving the methods that can assist investors in selecting stocks in the stock exchange is very important. If investors act rationally in stock selection decisions, they can achieve the desired return. The important factor that can help investors to select the optimal stocks is concentration on the criteria approved by financial experts and specialists. The important point in the stock investment is that decision making is not a one dimensional process. The successful decision maker is the one who decides the issue from different aspects and jointly and simultaneously uses multiple criteria and then, while investigating different factors influencing on that choice, selects the best options based on their priorities. In order to select a portfolio of Stocks through CHAID data mining algorithm in this paper, first the studied stocks were classified. The classified stocks were ranked using the DEA-BCC model. Through a binary two objective programming model, the combination of risk and rating Pareto was analyzed based on imperialist competitive algorithm.

References

[1] E. Atashpaz-Gargari, C. Lucas, Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition, *In 2007 IEEE congress on evolutionary computation* 4 (2007) 4661-4667.

[2] T. Back, Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms, *Oxford university press*, 1996.

[3] R. Castellano, R. Cerqueti, Mean Variance portfolio selection in presence of infrequently traded stocks, *European Journal of Operational Research* 234 (2014) 442-449.

[4] C. Çiflikli, E. Kahya-Özyirmidokuz, Implementing a data mining solution for enhancing carpet manufacturing productivity, *Knowledge-Based Systems* 23 (2010) 783-788.

[5] W. W. Cooper, Data envelopment analysis, *Encyclopedia of Operations Research and Management Science* (2001) 183-191.

[6] K. Deb, Multi-objective optimization using evolutionary algorithms, *John Wiley & Sons*, 2001.

[7] R. Eberhart, J. Kennedy, A new optimizer using particle swarm theory, *InMHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science* 6 (1995) 39-43.

[8] A. T. Eshlaghy, F. F. Razi, A hybrid grey-based KOHONEN and genetic algorithm to integrated technology selection, *International Journal of Industrial and Systems Engineering* 20 (2015) 323-342.

[9] F. Faezy Razi, A. T. Eshlaghy, J. Nazemi, M. Alborzi, A. Pourebrahimi, A hybrid grey based KOHONEN model and biogeography-based optimization for project portfolio selection, *Journal of Applied Mathematics* (2014).

[10] K. Gnanendran, J. K. Ho, R. P. Sundarraj, Stock selection heuristics for interdependent items, *European Journal of Operational Research* 145 (2003) 585-605.

[11] F. Gorunescu, Data Mining: Concepts, models and techniques, *Springer Science & Business Media*, 2011.

[12] J. Han, J. Pei, M. Kamber, Data mining: concepts and techniques, *Elsevier*, 2011.

[13] J. H. Holland, Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence, *MIT press*, 1992.

[14] H. C. Huang, T. K. Lin, P. W. Ngui, Analysing a mental health survey by chi-squared automatic interaction detection, *Annals of The Academy of Medicine, Singapore* 22 (1993) 332-337.

[15] K. Y. Huang, C.-J. Jane, A hybrid model for stock market forecasting and portfolio selection based on ARX, grey system and RS theories, *Expert systems with applications* 36 (2009) 5387-5392.

[16] S. Hwang, J. Park, Performance Evaluation with Information on Portfolio Compositions, *AsiaPacific Journal of Financial Studies* 40 (2011) 710-730.

[17] A. Ishizaka, P. Nemery, Assigning machines to incomparable maintenance strategies with ELECTRE-SORT, *Omega* 47 (2014) 45-59.

[18] M. Kantardzic, Data mining: concepts, models, methods, and algorithms, *John Wiley & Sons*, 2011.

[19] A. Kaveh, Advances in metaheuristic algorithms for optimal design of structures, *Springer*, 2014.

[20] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, Optimization by simulated annealing, *science* 220 (1983) 671-680.

[21] R. K. Lai, C.-Y. Fan, W.-H. Huang,P.-C. Chang, Evolving and clustering fuzzy decision tree for financial time series data forecasting, *Expert Systems with Applications* 36 (2009) 3761-3773.

[22] D. T. Larose, C. D. Larose, Discovering knowledge in data: an introduction to data mining, *John Wiley & Sons*, 2014.

[23] C. A. K. Lovell, J. T. Pastor, Units invariant and translation invariant DEA models, *Operations research letters* 18 (1995) 147-151.

[24] D. Maringer, Portfolio management with heuristic optimization, *Springer*, 2005.

[25] J. A. McCarty, M. Hastak, Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression, *Journal of business research* 60 (2007) 656-662.

[26] D. L. Olson, D. Delen, Advanced data mining techniques, *Springer Science, Business Media*, 2008.

[27] J.-L. Prigent, Portfolio optimization and performance analysis, *CRC Press*, 2007.

[28] R. V. Rao, Decision making in the manufacturing environment: using graph theory and fuzzy multiple attribute decision making methods, *Springer Science & Business Media*, 2007.

[29] S. C. Ray, Data Envelopment Analysis: Theory and Techniques for Economics and Operations Research, *Cambridge university press*, 2004.

[30] F. F. Razi, A. T. Eshlaghy, J. Nazemi, M. Alborzi, A. Poorebrahimi, A hybrid grey-based fuzzy C-means and multiple objective genetic algorithms for project portfolio selection, *International Journal of Industrial and Systems Engineering* 21 (2015) 154-179.

[31] L. M. Seiford, J. Zhu, Modeling undesirable factors in efficiency evaluation, *European Journal of Operational Research* 142 (2002) 16-20.

[32] J. Shadbolt, J. G. Taylor, Neural Networks and the Financial Markets: Bpredicting, Combining, and Portfolio Optimisation, *Springer*, 2002.

[33] K.-Y. Shen, M.-R. Yan, G.-H. Tzeng, Combining VIKOR-DANP model for glamor stock selection and stock performance improvement, *Knowledge-Based Systems* 58 (2014) 86-97.

[34] F. Tiryaki, B. Ahlatcioglu, Fuzzy portfolio selection using fuzzy analytic hierarchy process, *Information Sciences* 179 (2009) 53-69.

[35] F. Tiryaki, M. Ahlatcioglu, Fuzzy stock selection using a new fuzzy ranking and weighting algorithm, *Applied Mathematics and Computation* 170 (2005) 144-157.

[36] M. van Diepen, P. H. Franses, Evaluating chi-squared automatic interaction detection, *Information Systems* 31 (2006) 814-831.

[37] M. C. Wong, Y. L. Cheung, The practice of investment management in Hong Kong: market forecasting and stock selection, *Omega* 27 (1999) 451-465.

[38] P. Xidonas, G. Mavrotas, T. Krintas, J. Psarras, C. Zopounidis, Multicriteria portfolio management, *Springer*, 2012.

[39] B. Xing, W.-J. Gao, Introduction to Computational Intelligence, *In Innovative Computational Intelligence: A Rough Guide to 134 Clever Algorithms* 5 (2014) 3-17. Springer, Cham.

[40] H. Yu, R. Chen, G. Zhang, A SVM stock selection model within PCA, *Procedia computer science* 31 (2014) 406-412.

[41] X. Zhang, Y. Hu, K. Xie, S. Wang, E. W. T. Ngai, M. Liu, A causal feature selection algorithm for stock prediction modeling, *Neurocomputing* 142 (2014) 48-59.

Farshad Faezy Razi has got PhD degree from Islamic Azad University Science and Research Branch in 2014 he has been member of the faculty in Islamic Azad University Semnan branch since 2000. Main research interest include: Data Mining, Data Envelopment Analysis, Multi criteria decision Making.