A modular Takagi-Sugeno-Kang (TSK) system based on a modified hybrid soft clustering for stock selection

Somayeh Mousavi

Department of Industrial Engineering, Meybod University, Meybod, Iran, e-mail address: mousavi@meybod.ac.ir
Yahyazadeh Blvd., Khorramshahr Blvd., Meybod, Yazd, Iran
Tel: +98(35)3321-2412; Mobile: +98-913-152-7828; Fax: +98(35)3235-3004

Akbar Esfahanipour*

* Corresponding Author

Department of Industrial Engineering and Management Systems, Amirkabir University of Technology, Tehran, Iran, e-mail address: esfahaa@aut.ac.ir
424 Hafez ave. Tehran, Iran 15914
Tel: +98(21)6454-5300; Mobile: +98-912-347-9906; Fax:+98(21)6695-4569

Mohammad Hossein Fazel Zarandi

Department of Industrial Engineering and Management Systems, Amirkabir University of Technology, Tehran, Iran, e-mail address: zarandi@aut.ac.ir
424 Hafez ave. Tehran, Iran 15914
Tel: +98(21)6454-5300; Fax:+98(21)6695-4569

A modular Takagi-Sugeno-Kang (TSK) system based on a modified hybrid soft clustering for stock selection

This study presents a new hybrid intelligent system with ensemble learning for stock selection using the fundamental information of companies. The system uses the selected financial ratios of each company as the input variables and ranks the candidate stocks. Due to the different characteristics of the companies from different activity sectors, modular system for stock selection may show a better performance in comparison with an individual system. Here, a hybrid soft clustering algorithm is proposed to eliminate the noise and partition the input data set into more homogeneous overlapped subsets. The proposed clustering algorithm benefits from the strengths of the fuzzy, possibilistic and rough clustering to develop a modular system. An individual Takagi-Sugeno-Kang (TSK) system is extracted from each subset using an artificial neural network and genetic algorithm. To integrate the outputs of the individual TSK systems, a new weighted ensemble strategy is proposed. The performance of the proposed system is evaluated among 150 companies listed on Tehran Stock Exchange (TSE) regarding information coefficient, classification accuracy and appreciation in stock price. The experimental results show that the proposed modular TSK system significantly outperforms the single TSK system as well as the other ensemble models using different decomposition and combination strategies.

Keywords: intelligent modular systems; ensemble learning; hybrid rough-fuzzy clustering; TSK fuzzy rule-based system; stock selection; Tehran Stock Exchange (TSE).

1. Introduction

The fund allocation problem involves two stages, the asset selection, and the asset allocation. In the first stage, the objective is to select some attractive and valuable assets as the potential candidates for portfolio composition. In the second stage, the objective is to determine portfolio weights of the selected assets to achieve a series of risk-return considerations [1]. Similarly, in the stock portfolio management, one should select a universe of stocks before running a stock portfolio optimization model to determine optimal portfolio weights [2]. Those selected stocks have the best chances of capital appreciation in a long or intermediate time horizon.

There are two approaches widely used by academicians and market professionals for decision making in stock exchange: technical analysis and fundamental analysis. The fundamental analysis involves a detailed study of a company's financial status using the financial ratios and other fundamentals of the company to predict the future stock price movements. In the fundamental analysis, the main concern is the company's financial health. Traders often use this approach to predict the stock price over a long-term investment horizon. On the other hand, the technical analysis uses the data related to the past behavior of a stock price and volume data series to forecast the future. Traders use this approach for short-term investment horizons. They deal with the stock timing and not the company's financial health [3-7]. Since the fundamentals have a stronger relationship to the price movement in the longer horizons, the stock selection stage should be designed based on the fundamental analysis.

This study tries to develop a stock selection system based on the fundamental analysis. The system uses the selected financial ratios and fundamental data of each company as the input variables and ranks the candidate stocks based on the fundamental data. However, the future performance of the companies may follow different patterns due to different fundamental characteristics as well as activity sectors of the companies [8, 9]. For example, the companies with high inventory turnover usually have lower current ratio than the companies in other activity sectors. A low current ratio is not always an indicator of poor liquidity performance and should be compared with current ratios of the other companies with the similar inventory turnover. Therefore, it seems a modular system for stock selection may show a better performance in comparison with an individual system [10]. Furthermore, based on the principle of divide and conquer, the complexity of the whole data space is reduced by modularity, which leads to some more homogeneous data spaces [11].

In general, three main steps should be done to develop a modular system with ensembles. In the first step, the training data set is partitioned into some smaller data regions. In the second step, an individual learner is developed for each data region, separately. In the third step, the outputs of the individual learners are combined to determine the final output of the modular system using an ensemble strategy. In this study, a hybrid rough-fuzzy noise rejection clustering algorithm is proposed to determine the overlapping data regions of the modular system and remove the noise data, simultaneously. Then, an individual Takagi-Sugeno-Kang (TSK) fuzzy rule-based system is developed for stock selection in each data region separately. Finally, the outputs of the individual systems are combined to derive the ultimate result using our proposed ensemble strategy.

The main purpose of this paper is to construct an accurate and interpretable stock selection system for portfolio managers. The system ranks the universe of stocks and selects a set of stocks that are likely to have the

best chances of capital appreciation in the subsequent period. For this purpose, we propose an ensemble learning model to develop a modular stock selection system based on our proposed clustering. Here we describe the novelties of this study from two perspectives. The first one regards the applied method to modularize the system and the ensemble strategy. This study proposes a hybrid rough-fuzzy noise rejection clustering algorithm to partition a noisy data set into some overlapping partitions without noise and outliers. The proposed model also develops a new weighted ensemble strategy to aggregate the outputs of the modules. The second one relates to the development of a modular system for stock selection problem through the ranking of the stocks. From the best of the authors' knowledge, this is the first study that develops a TSK system for stock selection problem which applies an ensemble learning model.

The rest of the paper is organized as follows. The next section provides a brief review of the previous works related to our study. The third section explains the proposed algorithm for hybrid rough-fuzzy noise-rejection clustering. Section 4 presents a comprehensive description of the proposed ensemble learning model. The fifth section describes the implementation of the proposed model to develop a modular system for stock selection on Tehran Stock Exchange as well as the computational results. Finally, section 6 reports some concluding remarks.

2. Related Works

According to the principle of divide and conquer, a complex computational learning can be simplified by dividing the learning task among some experts and then combining the solutions of the experts, which is said to constitute a committee machine [10]. Modular systems are one of the architectural types of committee machines with local accuracy perspective [12]. In the modular systems, separate learners are developed and applied to different regions of the problem domain. The regions of interest are determined first, and an individual learner is then developed for each region. With this architecture, an important issue is the identification of the best regions to be considered. The regions can be defined based on expert opinion [13] or using purely mechanical means like clustering [14].

2.1. *Related works on clustering algorithms*

Many researchers have used hard clustering to determine the data regions of a modular system [15-17]. However, the boundaries between the adjacent data regions may be unclear, and a data instance may not completely belong to only one cluster. Therefore, hard clustering is too restrictive in partitioning. The soft clustering algorithms using fuzzy and rough set theories is less restrictive than the hard clustering by permitting

an instance to belong to multiple clusters. While the fuzzy clustering can efficiently handle the overlapping clusters [18], it can be too descriptive with potentially a list of possible memberships for an individual object [19]. In this way, the data set may not be partitioned into some homogenous data spaces with less complexity. Therefore, most of the previous studies related to the modular systems development with fuzzy clustering applied the traditional fuzzy clustering in which each sample is assigned to only one cluster based on the maximum membership [20]. Rough clustering, as another soft clustering approach, allows an object to belong to multiple clusters. In rough clustering, representation of the clusters is based on the lower and upper bounds using rough set theoretic properties. Lingras and West [21] believe that the lower and upper bound representation of a cluster is more concise than the detailed and descriptive list of membership values. They suggest rough c-means clustering as the first rough clustering algorithm. In this algorithm, the clusters are represented by the crisp lower and upper approximations. The lower approximation contains objects that are members of the cluster with certainty (probability = 1), while the upper approximation contains objects that are members of the cluster with non-zero probability (probability > 0). However, the rough clusters don't determine the similarity and closeness of the instances to the cluster prototypes [22].

In the last decade, intensive works have been done for the hybridization of rough and fuzzy clustering to integrate the advantages of both fuzzy sets and rough sets [18, 22-25]. Hybrid rough-fuzzy c-means clustering is proposed by Mitra et al. in 2006 for the first time [22]. In their proposed clustering, each cluster consists of a fuzzy lower approximation and a fuzzy boundary. A hybrid clustering with the crisp lower approximation and the fuzzy boundaries is proposed by Maji and Pal [24]. Furthermore, a rough-fuzzy possibilistic c-means clustering is extended to make the previous hybrid clustering [24] robust in the presence of noise and outliers [18]. Many research fields benefit from the application of the hybrid rough-fuzzy clustering such as bioinformatics and medical imaging [26, 27] and text-graphics segmentation [28]. Recently, the ensemble-based rough-fuzzy clustering is extended for categorical data with different dissimilarity measures [29, 30]. However, the mentioned hybrid clustering algorithms suffer from some weak points that are explained in section 3. This study proposes a new hybrid clustering algorithm which tries to overcome the weakness of these algorithms for developing a modular system. The proposed algorithm benefits the strengths of the fuzzy, possibilistic and rough clustering approaches, while lacks their weak points for developing a modular system.

2.2. *Related works on ensemble strategies*

The literature on ensemble learning has shown that the ensemble can outperform single predictors in many cases [31-33]. Additionally, ensembles of expert systems have already been successfully applied in financial

forecasting [34-43]. There are different ensemble strategies in the literature, including simple majority vote, simple averaging, weighted averaging, reliability based strategies, Bayesian methods and stacking [44, 45]. Among the mentioned ensemble strategies, the simple majority vote and the simple averaging of the baseline classifiers have shown relatively poor performance in different fields including financial time series prediction [12, 40]. On the other hand, weighted averaging and stacking strategies have been applied with great success over the last few years [13, 41, 46-49]. In stacking, a high-level base learner is developed to combine the lower level base learners, while in the weighted averaging the base learners are combined with different weights. In these ensemble approaches, the integration of the base learners is done using a weighted least squares algorithm [46, 48, 49], generalized regression neural networks [47], particle swarm optimization [41] and genetic fuzzy systems [13]. Lv et al. [48] showed that using the fuzzy memberships to different data partitions improves the accuracy of the ensemble model. Considering the results of the mentioned studies and our proposed clustering method, in this work we introduce a new weighted ensemble that uses the rough fuzzy memberships of the data partitions.

2.3. *Stock selection based on the fundamental analysis*

Fundamental analysis has been widely used for stock selection in the stock portfolio management. The stock selection models based on Fundamental analysis include, PROMETHEE decision making model [50], generalized data envelopment analysis model [1], multiple attributes decision making (MADM) model [51], Probit and Tobit based models [52] and also a continuous time model for active portfolio management [53]. However, soft computing models seem to be more appropriate for modeling the noisy, nonlinear and complex behavior of the stock markets [54-58]. In the literature, there are some studies on the soft computing methods such as artificial neural networks [3, 59], evolutionary algorithms [60-63], support vector machines [64] and fuzzy logic [65-68] for stock selection problem. This study proposes a hybrid genetic fuzzy system to select the best stocks for considering in the portfolio composition. The TSK fuzzy rule-based systems have shown good capability for modeling the nonlinear dynamic systems in many fields including the short-term stock trend prediction [69-71]. In this paper, we intend to evaluate the performance of TSK systems in the stock ranking and selection problem over the longer investment horizons. The structure and parameter identification phases of the TSK systems are done using artificial neural networks (ANN) and genetic algorithms (GA), respectively. Table 1 compares this study with the previous researches for stock selection based on the fundamental analysis using soft computing methods.

Please insert Table 1 about here.

3. The Developed Hybrid Rough-Fuzzy Noise-Rejection Clustering Algorithm

The hybrid rough-fuzzy clustering incorporates fuzzy membership value in the rough clustering framework. Rough-fuzzy c-means (RFCM) algorithm was proposed by Mitra et al. [22] for the first time. This algorithm partitions a set of *N* objects $X = \{x_1, \dots, x_j, \dots, x_N\}$ into c rough clusters (U_i) with a fuzzy lower approximation and a fuzzy boundary by minimizing the objective function J_{RFCM} as (1) subject to $\sum_{i=1}^{c} u_{ij} = 1$ for all $j = 1,2,\dots,N$.

$$J_{RFCM} = \sum_{i=1}^{c} JC_i$$

$$JC_i = \begin{cases} w_{low} \times A_i + w_{bound} \times B_i, & if\ \underline{A}U_i \neq \emptyset, BU_i \neq \emptyset \\ A_i, & if\ \underline{A}U_i \neq \emptyset, BU_i = \emptyset \\ B_i, & if\ \underline{A}U_i = \emptyset, BU_i \neq \emptyset \end{cases}$$

$$A_i = \sum_{x_j \in \underline{A}U_i} u_{ij}^m \|x_j - v_i\|^2$$

$$B_i = \sum_{x_j \in BU_i} u_{ij}^m \|x_j - v_i\|^2. \quad (1)$$

where v_i is the center of cluster U_i, $\|.\|$ is the distance norm, u_{ij} is the membership of x_j to cluster U_i and $1 \leq m < \infty$ is the fuzzifier in the fuzzy set theory. $\underline{A}U_i$ and $\overline{A}U_i$ are the lower and the upper approximations of U_i and $BU_i = \overline{A}U_i - \underline{A}U_i$ denotes the boundary region of the rough cluster U_i. The terms A_i and B_i represent the weighted within-groups sum of squared errors for the lower approximation and boundary of rough clusters, respectively. The parameters w_{low} and w_{bound} are the relative importance of the lower approximation and the boundary regions.

According to the definitions of lower approximation and boundary of rough sets by Pawlak, if an object is the member of a cluster's lower approximation, it definitely belongs to that cluster and it cannot be a member of its boundary or the other clusters [72].

$$IF\ x_j \in \underline{A}U_i\ THEN\ x_j \notin BU_k, \forall k\ AND\ x_j \notin \underline{A}U_k, \forall k \neq i$$

Maji and Pal [24] claimed that based on these definitions, the objects in the lower approximation should have a similar influence on only their corresponding cluster regardless of their similarity with their corresponding clusters and the other clusters. They proposed a hybrid rough-fuzzy clustering with the crisp lower approximation and the fuzzy boundaries. In this case, they reduced A_i to (2).

$$A_i = \sum_{x_j \in \underline{A}U_i} \|x_j - v_i\|^2. \quad (2)$$

They incorporated possibilistic c-means (PCM) into their previous model to develop a more robust clustering algorithm in the presence of noise and outliers [18]. In their proposed rough-fuzzy possibilistic c-means algorithm, they calculated A_i similar to their previous work and changed B_i as (3).

$$B_i = \sum_{x_j \in BU_i} \{a(\mu_{ij})^{m_1} + b(v_{ij})^{m_2}\} \|x_j - v_i\|^2 + \eta_i \sum_{x_j \in BU_i} (1 - v_{ij})^{m_2} . \quad (3)$$

where μ_{ij} is the probabilistic membership of x_j to U_i as that in FCM and v_{ij} is the possibilistic membership as in the PCM. The constants a and b determine the relative importance of the probabilistic and possibilistic memberships, respectively. The objective function of their proposed clustering algorithm is minimized when,

$$\mu_{ij} = \left(\sum_{k=1}^{c} \left(\frac{\|x_j - v_i\|^2}{\|x_j - v_k\|^2} \right)^{\frac{2}{m_1 - 1}} \right)^{-1} . \quad (4)$$

$$v_{ij} = \frac{1}{1 + \left(\frac{b\|x_j - v_i\|}{\eta_i} \right)^{\frac{1}{m_2 - 1}}} . \quad (5)$$

Also, a rough possibilistic type 2 fuzzy c-means (RPT2FCM) clustering algorithm is proposed by Sarkar et al. [73]. The RPT2FCM algorithm is so similar to the rough-fuzzy possibilistic c-means algorithm proposed by Maji and Pal [18]. The only difference is in the probabilistic memberships. In [73], the probabilistic memberships of Equation 3 are type 2 fuzzy membership values to handle some other various subtle uncertainties in the overlapping areas.

Although the lower approximation members of rough clusters definitely belong to their corresponding clusters, it is unreasonable to impose the same weight for all objects of a lower approximation [74]. We believe that different objects of a lower approximation should have different weights based on the proximity to their corresponding cluster prototypes regardless of the other prototypes. This is consistent with the above definition of lower approximation by Pawlak [72]. In this paper, we propose a hybrid rough-fuzzy noise rejection clustering (RFNRC) algorithm to resolve some of the drawbacks of the previous hybrid c-means algorithms. It incorporates the fuzzy noise rejection clustering (FNRC) [75] into the rough c-means (RCM) framework. The FNRC utilizes FCM and PCM to introduce a more robust clustering algorithm in the presence of noise and outliers.

In our proposed algorithm, different objects have different weights in determining their prototypes similar to [22]. However, unlike their work, the weights only depend on the distance of the objects from their corresponding prototypes which is more consistent with the rough set theory. The main steps of the proposed

RFNRC algorithm are as follows. The steps 1-3 are designed to define the suitable weighting exponent, the number of clusters and the initial cluster centers as the preprocessing steps of RFNRC. The fourth step determines the fuzzy clusters. The steps 5-7 are related to noise rejection from data set. In the 8th step, the PCM membership values are calculated, and the steps 9-10 determine the rough clusters.

(1) Define the suitable weighting exponent (m). The weighting exponent should be selected far from its both extremes to guarantee that the cluster validity index in the next step indicates the optimum number of clusters. According to [76], the suitable weight exponent is a value that makes the trace of the fuzzy total scatter matrix (S_T) equal to $z/2$.

$$S_T = \sum_{j=1}^{N}\left(\sum_{i=1}^{c}\left(u_{ij}\right)^m\right)\left(x_j - \bar{v}\right)(x_j - \bar{v})^T . \tag{6}$$

$$z = \text{trace}\left(\sum_{j=1}^{N}\left[\left(x_j - \frac{1}{N}\sum_{j=1}^{N} x_j\right)\left(x_j - \frac{1}{N}\sum_{j=1}^{N} x_j\right)^T\right]\right). \tag{7}$$

where $\bar{v}$ is the fuzzy total mean vector of the dataset considering the FCM based membership values.

(2) Determine the optimum number of clusters (C) through the original FCM so that the cluster validity index (8) is minimized. This index determines the optimum number of clusters that maximizes within clusters compactness and between clusters separation.

$$S_{cs} = \sum_{j=1}^{N}\sum_{i=1}^{c}\left(u_{ij}\right)^m\left(\left\|x_j - v_i\right\|^2 - \|v_i - \bar{v}\|^2\right) . \tag{8}$$

(3) Assign the initial cluster centers by the agglomerative hierarchical clustering algorithm. This clustering algorithm puts each of the n data instances in an individual cluster. Then, two or more clusters are merged using a matrix of dissimilarities, until the required number of clusters (C) are reached. This step prevents our proposed algorithm to converge to a local extreme.

(4) Identify the initial fuzzy cluster prototypes using the original FCM.

(a) Compute the matrix of membership degrees:

$$u_{ij} = \left(\sum_{k=1}^{C}\left(\frac{\left\|x_j - v_i\right\|^2}{\left\|x_j - v_k\right\|^2}\right)^{\frac{2}{m-1}}\right)^{-1} . \tag{9}$$

(b) Update the cluster centers:

$$v_i = \frac{\sum_{j=1}^{N}\left(u_{ij}\right)^m x_j}{\sum_{j=1}^{N}\left(u_{ij}\right)^m} . \tag{10}$$

(c) Repeat (a) and (b) until reaching convergence in cluster centers, i.e., $v_i^{(t)} - v_i^{(t-1)} < \varepsilon$, where t is the iteration number of the FCM algorithm.

(5) Calculate the resolution parameter of PCM (η_i). The value of η_i determines the distance that the membership value of a point in the cluster i becomes 0.5 and is chosen based on the desired "bandwidth" of the possibilistic membership distribution for each cluster [77]. It is assumed that the data in each cluster follows a Gaussian distribution and $\|x_j - v_i\|/\sigma_i$ has a chi-square distribution, with degrees of freedom equivalent to the number of features in each data instance. Therefore, the resolution parameter can be calculated as [75]:

$$\eta_i = \frac{\underset{x_j \in U_i}{median\,(\|x_j - v_i\|)}}{\chi^2_{0.5}} . \tag{11}$$

where χ^2 is the chi-square value.

(6) Calculate the cutoff distance ($u_{FC}cut^2$) to detect the noise and outliers [75].

$$u_{FC}cut^2 = \eta_i \chi^2_{\hat{z}} . \tag{12}$$

where $\hat{z}$ is the percentage of inliers in the data. The number of outliers is estimated based on W index.

$$W_j = \sum_{i=1}^{C} \|x_j - v_i\| . \tag{13}$$

This index sums the distance of the data instance to all cluster centers. The data instances with large values of W_j are considered as outliers. The threshold for outliers depends on the upper and lower bounds of the data and is selected according to the trace of W index. This step is designed to find the outliers, i.e., the data instances which are too far from all cluster centers.

(7) Remove the noise data and outliers. If $\|x_j - v_i\| > u_{FC}cut^2$, then the data instance is recognized as noise, and it takes a zero membership to the cluster.

(8) Compute the membership matrix of the remaining data using PCM membership value [77]:

$$u_{ij} = \frac{1}{1 + \left(\frac{\|x_j - v_i\|}{\eta_i}\right)^{\frac{1}{m-1}}} . \tag{14}$$

(9) Assign each data instance to the lower approximation of a cluster or the boundaries of multiple clusters, by the following procedure:

(a) Assign the instance (x_j) to the upper bound of the cluster k ($\overline{A}U_k$), based on the maximum membership.

$$x_j \in \overline{A}U_k \,, k = \arg_i \max_{i=1,2,\ldots,C}(u_{ij}) \,. \quad (15)$$

The instance should be assigned to the upper bound and boundary of two or more clusters in the case of ties in the maximum membership.

(b) For each cluster i, i=1, 2, ...,C, $i \neq k$, IF $u_{kj} - u_{ij} < \delta$ Then $x_j \in \overline{A}U_i$ and $x_j \in BU_i$,

where δ is a small threshold value that determines the overlapping degree of the adjacent clusters.

(c) IF $x_j \notin \overline{A}U_i,\ i = 1,2,\ldots,C, i \neq k$ Then $x_j \in \underline{A}U_k$ Else $x_j \in BU_k$.

(10) Compute the new cluster centers as Equation 16.

$$v_i = \begin{cases} w_{low} \times C_1 + w_{bound} \times D_1, & if\ \underline{A}U_i \neq \emptyset, BU_i \neq \emptyset \\ C_1, & if\ \underline{A}U_i \neq \emptyset, BU_i = \emptyset \\ D_1, & if\ \underline{A}U_i = \emptyset, BU_i \neq \emptyset \end{cases}$$

$$C_1 = \frac{\sum_{x_j \in (\underline{A}U_i)} u_{ij}^m x_j}{\sum_{x_j \in (\underline{A}U_i)} u_{ij}^m}$$

$$D_1 = \frac{\sum_{x_j \in (BU_i)} u_{ij}^m x_j}{\sum_{x_j \in (BU_i)} u_{ij}^m} \,. \quad (16)$$

According to the rough set theory, the objects of the lower approximation should have much more influence to their cluster prototypes. Therefore, w_{low} should be much more than w_{bound}, i.e., $0 < w_{bound} < w_{low} < 1, w_{low} + w_{bound} = 1$.

(11) Repeat steps 8-10 until convergence, i.e., there are no more new assignments.

The strengths of our proposed RFNRC algorithm are in six aspects. First, from the compatibility with the centroid point of view, possibilistic memberships of PCM correspond more closely to the notion of typicality [77]. Unlike FCM, there is no constraint on the memberships of PCM (i.e., $\sum_{i=1}^{C} \mu_{ij} = 1, \forall j$). Therefore, the prototypes of PCM are attracted toward dense regions in the feature space, regardless of the locations of the other prototypes. Second, using FCM and agglomerative hierarchical clustering at the first steps avoids the problem of the coincident clusters of PCM. The FCM and PCM algorithms have been previously integrated to avoid the problems of noise sensitivity of the FCM and the coincident clusters of PCM [18]. Third, the steps 5-7 of the proposed algorithm makes the algorithm more robust in the presence of noise and outliers. The powerful ability of these steps in noise rejection has been confirmed by Melek et al. [75]. Our proposed algorithm removes the noise and outliers from the cluster members and, therefore the outliers don't affect the learning process of the modules in the modular system. Fourth, the c-means algorithms with random initial centers

always converge to a local extreme [78]. Our proposed RFNRC algorithm has overcome this problem using agglomerative hierarchical clustering algorithm. Fifth, it uses two preliminary steps to identify the suitable weight exponent and the optimum number of clusters which makes our algorithm more efficient. Above all, the previous hybrid clustering algorithms allow only two overlapping clusters [18, 22, 24]. However, three or more overlapping clusters are possible in clustering of real data sets. Step 9 (parts *a* &*b*) of our proposed RFNRC handles multiple overlapping clusters as well as two overlaps.

4. The Proposed Ensemble Learning Model for Stock Selection

This section describes the general architecture of our proposed model to develop a modular TSK system with weighted ensemble strategy for stock selection based on fundamental analysis. In this approach, the system uses fundamental data of companies to predict the future behavior of its stock in the following year and assigns a score to the stocks. Then the system ranks the universe of stocks based on the assigned scores. Figure 1 shows the overall framework of the proposed ensemble learning model.

Please insert Fig. 1 about here.

This framework starts with collecting fundamental information of companies along with the data preprocessing stage to treat the missing variables. We also consider data normalization as one of the necessary data transformations in the forecasting problems [16]. This study applies the min-max normalization method to obtain a database with all feature's values falling in range of 0 and 1. Due to a large number of fundamental variables as the input variables, the most influential subset of variables are selected by stepwise regression analysis. Stepwise regression analysis has been used successfully for variable selection in the stock market forecasting [69-71, 79]. This technique either adds the variables onward or removes the variables backward to find the best combination of independent variables for forecasting the dependent variable. The following sections describe the other steps of the proposed model in more detail.

4.1. *Data partitioning using the developed hybrid rough-fuzzy noise-rejection clustering*

Due to the variability of fundamental properties of the companies within different activity sectors, we believe that the data partitioning is a necessary task to have the more homogenous subsets to develop a modular system for stock selection. On the other hand, the fundamental information of the companies generally has some outliers. Therefore, the training data set is partitioned into multiple overlapping clusters using the proposed rough-fuzzy noise-rejection clustering algorithm as described in section 3. The proposed algorithm has some

advantages for our application. First, because of its noise-rejection property, it can handle the outliers within the fundamental data set. Second, it benefits from the fuzzy, possibilistic and rough clustering to resolve the ambiguity in assigning the objects to the modules. The training patterns of the modules are unique in their lower approximation, and they are similar to patterns of one or more modules in their overlapping or boundary regions. Third, this algorithm represents the cluster members using their rough-fuzzy memberships. The third property is useful for designing of an efficient ensemble strategy as described in section 4.3.

4.2. *Generating TSK fuzzy rule-based systems for stock selection*

In TSK fuzzy rule-based systems, the knowledge base includes multiple fuzzy rules with crisp functions as the consequent. For the first order TSK system with two input variables, the rules are in the form of:

$$IF\ L_1\ is\ FS_{i1}\ AND\ L_2\ is\ FS_{i2}\ THEN\ y_i = a_{i0} + a_{i1}L_1 + a_{i2}L_2$$

where L_1 and L_2 are the linguistic variables, FS_{i1} and FS_{i2} are their corresponding fuzzy sets and a_{i0}, a_{i1}, a_{i2} are the parameters of the system. The system inferences using crisp reasoning. The crisp inference of the system is determined as the weighted average of the individual rule inferences using (17).

$$\hat{y} = \sum_{i=1}^{r} DOF_i y_i \Big/ \sum_{i=1}^{r} DOF_i \ . \tag{17}$$

where, r is the number of the rules and DOF_i is the degree of firing of the ith rule, i.e., the rule's condition memberships aggregated by a t-norm operator as (18).

$$DOF_i = \mu_{FS_{i1}}(L_1) \wedge \mu_{FS_{i2}}(L_2) \ . \tag{18}$$

where $\wedge$ is a t-norm operator and $\mu_A(x)$ is the membership degree of x in the fuzzy set A.

In our proposed framework, a unique TSK system is independently developed for each data partition as an individual learner. Each data partition includes the lower approximation and the boundary of the rough cluster determined by our proposed RFNRC algorithm. The structure and parameter identification of the TSK systems are described in the following.

We design the TSK rules in canonical form, where the combination of all the input variables takes place using the conjunction operator. The structure of the TSK rules in the premises is determined by the grid partitioning. The quality of the TSK system heavily depends on the partitioning of the input space [80]. We determine the fuzzy sets for the input variables using the modified Adeli-Hung algorithm (AHA) [71]. This algorithm includes two stages. The first stage involves clustering of the data instances with a topology and

weight change neural network, known as Adeli-Hung clustering. The second stage comprises assigning the membership functions to the input space [81]. These stages are clearly explained in [82]. In the modified version of AHA, the input variables are partitioned individually, and the membership functions are determined for individual variables [71]. The modified AHA assigns symmetric triangular fuzzy sets to the input variables, where the fuzzy sets of the adjacent labels overlap to some extent, and their vertex points do not cross. These properties make the system more transparent [83].

The parameters of the linear conclusion of the TSK rules are determined using the genetic algorithm. Here, GA is used to learn the TSK system because of its flexible and powerful search capability [84], especially in the case of fuzzy systems learning [85, 86]. This evolutionary algorithm has been successfully applied for both phases of fuzzy modeling, namely structure and parameter identification [80, 87]. Among the two common approaches for genetic learning of the rule-based systems, i.e., Pittsburgh and Michigan, we apply the Pittsburgh approach to learn the TSK fuzzy systems. In the Pittsburgh approach, one individual encodes the whole rule base of the system. Figure 2 shows the encoding scheme of the TSK consequents parameters as a GA chromosome.

Please insert Fig. 2 about here.

The GA chromosomes are evaluated by information coefficient (IC) as the fitness function. IC is a performance measure used for evaluating the forecasting skill of financial analysts. It is an appropriate fitness measure for ranking and classifying the stocks in the investment universe [60, 88]. Information coefficient measures the Spearman correlation of the ranking that the model assigns to the stocks and their actual rankings in the following period.

$$IC_t = \frac{\sum_{j=1}^{S} MRank_{j,t} * RRank_{j,t} - (\sum_{j=1}^{S} MRank_{j,t} * \sum_{j=1}^{S} RRank_{j,t})/n_t}{\sqrt{\left(\sum_{j=1}^{S} MRank_{j,t}^2 - (\sum_{j=1}^{S} MRank_{j,t})^2/n_t\right) * \left(\sum_{j=1}^{S} RRank_{j,t}^2 - (\sum_{j=1}^{S} RRank_{j,t})^2/n_t\right)}},\ t = 1, 2, \dots, T. \tag{19}$$

where $MRank_{j,t}$ is the rank of stock j which is predicted by the model at time t, $RRank_{j,t}$ is the ranking of the realized return of stock j at time t, and T is the number of periods in the training set. Our model evaluates the GA chromosomes by the average information coefficient of the proposed system as (20).

$$fitness function: \max\{\sum_{t=1}^{T} IC_t/T\}. \tag{20}$$

4.3. *Aggregating the TSK systems based on the proposed weighted ensemble strategy*

Our proposed ensemble learning model develops an individual TSK system for each partition of the training instances. Similar to the training instances, a new instance may belong to multiple partitions with different

membership values. Therefore, our model combines the outputs of the corresponding TSK systems using an ensemble strategy to reach the final score of the new instance. In our proposed modular system, we design a new weighted ensemble strategy to combine the outputs of the TSK systems. In this design, the weight of each module is different for each instance and depends on the proximity of the instance concerning the prototype of module. It uses both rough and possibilistic-fuzzy memberships to calculate the relative importance of the modules as their weights. The ensemble weight of the module *i* for the instance $x_j (EW_{ij})$ is determined by (21).

$$EW_{ij} = \begin{cases} \dfrac{\underline{w}(u_{ij})}{\underline{w}(u_{ij}) + \sum_{k|x_j \notin \overline{A}U_k} w(u_{kj})} & if\ x_j \in \underline{A}U_i \\ \dfrac{\overline{w}(u_{ij})}{\sum_{k|x_j \in BU_k} \overline{w}(u_{kj}) + \sum_{k|x_j \notin \overline{A}U_k} w(u_{kj})} & if\ x_j \in BU_i \\ \dfrac{w(u_{ij})}{\sum_{k|x_j \in \underline{A}U_k} \underline{w}(u_{kj}) + \sum_{k|x_j \in BU_k} \overline{w}(u_{kj}) + \sum_{k|x_j \notin \overline{A}U_k} w(u_{kj})} & if\ x_j \notin \overline{A}U_i \end{cases} \quad (21)$$

where, $\underline{w}, \overline{w}$ and w correspond to the relative importance of the modules regarding the rough partitions. $\underline{w}$ is the weight of the module if x_j belongs to its lower approximation. $\overline{w}$ is the weight of the module if x_j belongs to its boundary. And w is the weight of the module if x_j does not belong to its partition. Since the instances out of the upper approximation don't contribute in the module training, w is too small relative to $\underline{w}$ and $\overline{w}$ ($0 < w \ll \overline{w} < \underline{w} < 1$). The relative importance of the modules depends on the possibilistic-fuzzy membership of the instance to its partition (u_{ij}), as well.

Figure 3 shows a typical partitioning of the data set in a two-dimensional space. The thickness of the arrows corresponds to the ensemble weight of the arrowhead's module. In this figure, $x_1 \in \underline{A}U_3$ and the ensemble weight of the third module is very high ($EW_{31} = \underline{w}(u_{31})/(\underline{w}(u_{31}) + w(u_{11} + u_{21}))$). This instance does not belong to the other partitions ($x_1 \notin \overline{A}U_1, x_1 \notin \overline{A}U_2$), therefore the weights which are assigned to their corresponding modules are very low ($EW_{11} = w(u_{11})/(\underline{w}(u_{31}) + w(u_{11} + u_{21}))$, $EW_{21} = w(u_{21})/(\underline{w}(u_{31}) + w(u_{11} + u_{21}))$). Another instance, x_2, belongs to the boundaries of U_1 and U_2. Therefore, the ensemble weights of the first and the second modules are relatively high based on their possibilistic-fuzzy membership values ($EW_{11} = \overline{w}(u_{11})/(\overline{w}(u_{11} + u_{21}) + w(u_{31}))$, $EW_{21} = (\overline{w}(u_{21})/\overline{w}(u_{11} + u_{21}) + w(u_{31}))$). However, the weight of the third module is too low ($EW_{31} = (w(u_{31})/\overline{w}(u_{11} + u_{21}) + w(u_{31}))$).

Please insert Fig. 3 about here.

Consequently, the final score of the stock with fundamental properties represented by x_j is determined by (22).

$$Score_{x_j} = \sum_{i=1}^{C}(EW_{ij} * \hat{y}_{ij}) \,. \tag{22}$$

where $\hat{y}_{ij}$ is the output of module i for the instance x_j. Finally, the system selects the stocks according to their ranked scores at time *t*.

5. Experimental Results

We implemented our proposed ensemble learning model to develop a modular TSK system for stock selection among 150 companies with different activity sectors listed on Tehran Stock Exchange (TSE). This section describes the collected data at first. Then, it reports the implementation results of our proposed model step by step. Finally, it analyses the performance of the developed modular system and its comparison results.

5.1. *Data*

In this work, our data comprise of the fundamental data of 150 Iranian companies listed on TSE during 24 fiscal years to develop a stock selection system for investing in this market. These companies are the most liquid companies according to six liquidity measures including the number of traded shares, the value of traded shares, the number of trading days, the number of trades, the average number of shares issued and the company's value in a fiscal year. The liquidity measures are aggregated by harmonic mean (Equation 23) to find the most liquid companies on TSE.

$$M_j = N / \sum_{i=1}^{N} I_{ij}. \tag{23}$$

where, M_j is the *j*th company's score, *N* is the number of indices and I_{ij} is the value of the *i*th index for the *j*th company. The companies are selected from the most liquid companies that were active before 2008 from different activity sectors [89]. We selected thirty-six financial ratios as the potential fundamental variables in five categories of the profitability, activity, liquidity, leverage and valuation ratios [1, 3, 59, 62, 65, 90]. Table 2 reports a list of the selected financial ratios. We calculated the financial ratios of the Iranian companies listed on TSE using the financial statements information of the companies.

Please insert Table 2 about here.

The historical data includes the above mentioned financial ratios and the dividend and split-adjusted close price for the start and end of the companies' fiscal year from March 20, 1991 to March 19, 2014. We used the

extending window approach to define the training and testing periods. This approach uses all the available historical data to train the system and then applies the trained system on the next period immediately after the last training period [91]. In this work, we designed five series of experiments to evaluate the evolved system for stock selection on five fiscal years, March 20, 2009, to March 19, 2014. Table 3 shows the training and the testing periods of the five series of experiments.

Please insert Table 3 about here.

5.2. *Implementation of our proposed ensemble learning model for stock selection*

This section reports the implementation results of our proposed ensemble learning model step by step.

Step 1. This step involves the data preprocessing and the variable selection. First, the fundamental data were studied to handle the missing data and some outliers. Then, the data were normalized. Finally, the most effective financial ratios were selected using the stepwise regression analysis. We used the statistical software, IBM SPSS statistics 20 for setting up the regression forecasting model. We considered the thirty-six mentioned financial ratios as the independent variables and the rate of return as the dependent variable. We set the probabilities of F statistics to enter and remove a variable at 0.05 and 0.1, respectively. Due to our experiments, five financial ratios were chosen as the input variables of the TSK systems. The selected ratios are the return on asset (ROA), total asset turnover, equity ratio, dividend yield and book value per share.

Step 2. We implemented the proposed hybrid rough-fuzzy noise-rejection clustering algorithm to decompose the training data set into several overlapping subsets (i.e., clusters). At first, the suitable weighting exponent was selected as m=2.5, which gives a value for the trace of the total scatter matrix equal to $z/2$ (Figure 4). Then, the optimum number of clusters was identified using the cluster validity index. According to Figure 5, the rate of reduction in the cluster validity index is very high till 4 clusters, and the index gradually decreases till 7 clusters. We set the number of clusters at C=4 that ensures almost minimum cluster validity index as well as sufficient training instances to learn the individual TSK systems. The noise data are those with large values of the noise-rejection index (W) as shown in Figure 6. In our experiments, the threshold was selected as 2 for calculating the cutoff distance and removing the noise. Then, the FNRC membership degrees were calculated, and the rough-fuzzy data partitions were determined through the iterative procedure of the proposed algorithm. In this algorithm, the threshold δ represents the size of granules of rough-fuzzy clusters. The threshold could be determined as the median or the mean of the difference of the highest and the second highest memberships of all the instances to the specified clusters [19, 73]. However, when the distribution of the membership differences is

skewed, the median would work better than the mean. Therefore, we have used the following definition for δ assignment:

$$\delta = Median_{j=1,2,\ldots,N}(u_{ij} - u_{kj}) . \quad (24)$$

where u_{ij} and u_{kj} are the two highest memberships of instance x_j. Based on this definition, the threshold was set to $\delta = 0.22$. Consequently, the modular system included four modules with about 600 samples to train the individual TSK systems.

Please insert Fig. 4 about here.

Please insert Fig. 5 about here.

Please insert Fig. 6 about here.

Step 3. This step resulted in generating the individual TSK systems for each data region, separately. The structure of the TSK rules' antecedents was specified through AHA. In our design, three fuzzy sets were allocated to each input variable, interpreted by linguistic labels of low, medium and high. The parameters of the TSK rules' consequents were learned using GA with Pittsburgh approach. We set the parameters of GA as shown in Table 4. Crossover and mutation rates and reproduction size are selected based on some primary experiments. The population size is selected as 100 which showed a better performance in multiple runs. Since no significant improvements were observed after generation 150, the number of generations is set 200. Four TSK systems were learned for the four data partitions on the five time frames. Table 5 reports the performance of the TSK systems in terms of IC on training data subsets. According to this table, the individual TSK systems could learn a stock selection system with the average IC value of 33.03% in the case of their corresponding training subsets.

Please insert Table 4 about here.

Please insert Table 5 about here.

Step 4. This step involved aggregation of the outputs of all of the evolved individual TSK systems to reach an overall stock ranking. The modules' outputs were aggregated using the proposed weighted ensemble strategy, and then the stocks were ranked according to their scores. According to the preliminary experiments, we set the $\underline{w}, \overline{w}$ and w to 0.8, 0.6 and 0.2, respectively. The next section reports the performance of our proposed ensemble learning model on the five testing periods and the comparison results.

5.3. *Performance evaluation of our proposed ensemble learning model*

The performance evaluation of our proposed RFNRC-TSK-MBE (rough-fuzzy noise rejection clustering based modular TSK system with membership based ensemble) is based on information coefficient for stock selection on TSE. For comparison purposes, we established three alternative models: single TSK, FNRC-TSK-AveE (FNRC based TSK system with averaging ensemble) and RFNRC-TSK-AveE (RFNRC based TSK system with averaging ensemble). We designed the single TSK model based on the first and the third steps of our proposed RFNRC-TSK-MBE. However, the third step in the single TSK system uses the whole training dataset instead of the training subsets. FNRC-TSK-AveE partitions the dataset using FNRC and then assigns the instances to clusters based on the maximum membership. This model develops an individual TSK for each disjoint subset and applies simple averaging as the ensemble strategy. The RFNRC-TSK-AveE uses the same decomposition and sub-modeling method with RFNRC-TSK-MBE but aggregates the modules by simple averaging ensemble strategy.

Table 6 reports the experimental results of the four models for stock selection along the five testing periods. The reported numbers are the average results of 30 independent runs. According to this table, the RFNRC-TSK-MBE shows a remarkable performance for stock selection. The single TSK system is able to rank the TSE stocks with IC of 8.5% on average, which is a good performance. However, our proposed modular systems could outperform the single system. The first modular system extended by FNRC-TSK-AveE model could improve the ranking ability of the TSK system to IC of 12.74% on average. The ability of the modular system is further improved by our proposed clustering method. It reaches to 15.24% information coefficient. Above all, our proposed ensemble method could boost the predictability of the modular system from IC of 15.24% to 18.15%, on average. Similar results are found in all experiments along the five periods. The correlation between the RFNRC-TSK-MBE model's output and the one year forward return is substantially high over each test period.

Please insert Table 6 about here.

According to the comparison results, the single TSK model showed the weakest performance for stock selection. The reason is different fundamental characteristics of the companies in different activity sectors. Therefore, a modular system works better for such a problem. Furthermore, the FNRC-TSK-AveE got a less IC in comparison with RFNRC-TSK-AveE. That is because it is difficult to distinguish a certain boundary between different clusters of data set. Moreover, RFNRC-TSK-MBE arrived at the highest IC value among all the models. The superiority of the RFNRC-TSK-MBE over the other ensemble models is because of its

decomposition and combination methods. However, it is notable that all the developed TSK systems have shown a good performance in stock selection. As a general guideline, 5% is an acceptable IC value in investment management [88]. Also, according to a research published by JP Morgan [92], managers with ICs between 0.05 and 0.15 can achieve significant risk-adjusted excess returns. Therefore, TSK system can model the stock selection problem properly. The proposed RFNRC-TSK-MBE model could reach IC value of 18.15% on average which is much more than the other stock selection models in the literature [60, 88, 93, 94]. The genetic programming model provided by Becker et al. [60] could reach to a maximum IC of 8%. Additionally, the average IC of 9% was obtained in [88] using grammatical evolution. The authors claimed that 9% is a high IC value and their models were successful at the stocks' ranking. Also, Gillam et al. [93] studied on the earnings prediction in a global stock selection model. They could improve the predictability of the model to the IC of 6%.

Finally, we carried out the statistical tests to examine whether the proposed model significantly outperforms the other three models or not. The results of student t-test are reported in Table 7. According to this table, the RFNRC-TSK-MBE significantly outperforms the other three models at 99% statistical significance level. This table shows the impact of modularization (FNRC-TSK-AveE over single TSK) is more than other factors, i.e., the clustering and ensemble methods.

Please insert Table 7 about here.

In another experiment, the profitability of our proposed model has been investigated for stock classification. In our experiment, all the stocks (described in section 5.1) have been classified into two classes. Similar to [3], we have defined class 1 as the stocks which appreciate in share price to or more than 80% within one year. The other stocks have been classified as class 2. In this design, the first class constitutes the minority of the data, while it is our interested class. We have applied the over-sampling technique to deal with the imbalanced data in training set. In over-sampling, the samples of the rare class are increased by data replication.

The classification performance of our proposed model has been examined in terms of classification accuracy. For comparison purposes, ANFIS (adaptive neuro-fuzzy inference system) has been implemented on the same data base to develop a TSK fuzzy rule-based system for stock classification. The performance of ANFIS for stock classification was previously investigated on Dow Jones Industrial Average (DJIA) market [3]. According to that research, ANFIS outperforms other neural network models, multi-layer perceptron and radial basis function, in terms of classification accuracy and time complexity. The classification accuracy of our proposed RFNRC-TSK-MBE model is compared with ANFIS model in Table 8. The ANFIS system is trained in 40 epochs, with two Gaussian membership functions for each of the five input variables. According to this

table, our proposed model outperforms ANFIS model in all the five investigated periods. Our proposed model has reached to the average classification accuracy of 75.22% in five periods, whereas ANFIS could earn classification accuracy of 65.06%, on average.

Please insert Table 8 about here.

Furthermore, the average appreciation in the stock price of the selected stocks (i.e., the stocks that are classified as class 1 by the model) is compared with the average appreciation of all investigated stocks in the subsequent year. Table 9 reports the experimental results in terms of average appreciation. Again, this table confirms the ability of our proposed model for stock selection. The RFNRC-TSK-MBE could earn the excess appreciation of 11.9%, 52.4%, 17.5%, 4.2%, 10.8% of the selected stocks over all 150 stocks in five consecutive test periods and 19.36% on average.

Please insert Table 9 about here.

6. Conclusion

This paper proposes a new ensemble learning model to develop a modular TSK system for stock selection. The proposed ensemble learning model includes four stages. The first stage involves data preprocessing and variable selection. The second stage is about the data partitioning of the training data into several overlapping regions using the proposed rough-fuzzy noise rejection clustering (RFNRC) algorithm. The proposed algorithm benefits the strengths of rough, fuzzy and possibilistic clustering, while lacks their week points for developing a modular system. The diversity of the individual learners is guaranteed by such a data partitioning algorithm. At the third stage, an individual TSK system is generated for each region. The structure and parameter identification phases of the TSK systems are done using Adeli-Hung algorithm and genetic algorithm. At the fourth stage, the outputs of the individual TSK systems are aggregated using the proposed weighted ensemble strategy based on the rough-fuzzy memberships.

Within this framework, while handling large data sets, each module may concentrate on knowledge discovery within a different region of the problem. Subsequently, all of the modules contribute to problem-solving with a degree based on the similarity of the instance with the prototypes of the modules. The similarity is measured according to the rough partitions and the possibilistic-fuzzy memberships.

We implemented our proposed ensemble learning model on 150 Iranian companies with different activity sectors listed on Tehran Stock Exchange (TSE) to develop a modular TSK system for stock selection. Based on the experimental results, our developed modular system could appropriately select the stocks with information

coefficient of 18.15% on average, which is a good performance with respect to the previous researches [60, 88, 93]. Furthermore, our proposed system significantly outperformed the single TSK system (IC=8.5%) and also other modular TSK systems, i.e., fuzzy noise rejection clustering based TSK system with averaging ensemble (IC=12.74%) as well as rough-fuzzy noise rejection clustering based TSK system with averaging ensemble (IC=15.24%). Investigating the comparison results, some conclusions can be drawn. First, modular systems outperform a single system for problems with different regions of data characteristics, like stock selection problem based on fundamental analysis. Second, the data partitioning using our proposed hybrid clustering algorithm leads to the more capable modular systems. Additionally, considering the memberships in the ensemble strategy improves the performance of the ensemble model, significantly.

Additionally, the performance of our proposed model is investigated for stock classification. According to the results, our proposed model outperforms ANFIS regarding classification accuracy (75.22% Vs. 65.06%). Based on this experiment, we can earn much more return on investment using the selected stocks by our proposed model for portfolio diversification. The selected stocks reached to 19.36% excess appreciation on average, where the average appreciation of all investigated stocks is 44.02%.

References

1. Edirisinghe, N. C. P. and Zhang, X., "Generalized DEA model of fundamental analysis and its applications to portfolio optimization". *Journal of Banking & Finance*, **31**(11), pp. 3311–3335 (2007).
2. Liu, H., Mulvey, J., and Zhao, T., "A semiparametric graphical modelling approach for large-scale equity selection". *Quantitative Finance*, **16**(7), pp. 1053–1067 (2016).
3. Quah, T. S., "DJIA stock selection assisted by neural network". *Expert Systems with Applications*, **35**, pp. 50–58 (2008).
4. Chen, Y. S. and Cheng, C. H., "Evaluating industry performance using extracted RGR rules based on feature selection and rough sets classifier". *Expert Systems with Applications*, **36**, pp. 9448–9456 (2009).
5. Esfahanipour, A. and Mousavi, S., "A genetic programming model to generate risk-adjusted technical trading rules in stock markets". *Expert Systems with Applications*, **38**(7), pp. 8438-8445 (2011).
6. Mousavi, S., Esfahanipour, A. and Fazel Zarandi, M. H., "A Novel Approach to Dynamic Portfolio Trading System Using Multitree Genetic Programming". *Knowledge-Based Systems*, **66**, pp. 68-81 (2014).
7. Shen, R.K., Yang, C. Y., Shen, V.R.L., Li, W.C. and Chen, T.S., "A Stock Market Prediction System Based on High-Level Fuzzy Petri Nets". *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **26**(5), pp. 771-808 (2018).
8. Yunusoglu, M. G. and Selim, H., "A fuzzy rule based expert system for stock evaluation and portfolio construction: An application to Istanbul Stock Exchange". *Expert Systems with Applications*, **40**, pp. 908–920 (2013).

9. Reilly, F. K. and Brown, K. C., *Investment analysis and portfolio management*, 7th ed., South-Western College Publications, (2004).
10. Haykin, S., *Neural Networks, A comprehensive foundation*, Chapter. 7, pp. 351- 391, Prentice Hall Inc., New Jersey (1999).
11. Kumar, R. S. and Arasu, G. T., "Rough Set Theory and Fuzzy Logic Based Warehousing of Heterogeneous Clinical Databases". *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **25**(3), pp. 385-408 (2017).
12. Finlay, S., "Multiple classifier architectures and their application to credit risk assessment". *European Journal of Operational Research*, **210**, pp. 368–378 (2011).
13. Melin, P., Sánchez, D. and Castillo, O., "Genetic optimization of modular neural networks with fuzzy response integration for human recognition". *Information Sciences,* **197**, pp. 1–19 (2012).
14. Kuncheva, L. I., "Switching between selection and fusion in combining classifiers: an experiment". *IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics*, **32**, pp. 146–156 (2002).
15. Lim, M. K. and Sohn, S. Y., "Cluster-based dynamic scoring model". *Expert Systems with Applications*, **32**, pp. 427–431 (2007).
16. Shahrabi, J., Hadavandi, E. and Asadi, S., "Developing a hybrid intelligent model for forecasting problems: Case study of tourism demand time series". *Knowledge-Based Systems*, **43**, pp. 112–122 (2013).
17. Alikhani, M., Nedaie, A. and Ahmadvand, A., "Presentation of clustering-classification heuristic method for improvement accuracy in classification of severity of road accidents in Iran". *Safety Science*, **60**, pp. 142–150 (2013).
18. Maji, P. and Pal, S. K., "Rough set based generalized fuzzy C-means algorithm and quantitative indices". *IEEE Trans. Syst. Man Cybern. B: Cybern*., **37** (6), pp. 1529–1540 (2007).
19. Lingras, P., Nimse, S., Darkunde, N. and Muley, A., "Soft Clustering from Crisp Clustering using Granulation for Mobile Call Mining". *IEEE International Conference on Granular Computing (GrC)*, (2011).
20. Liu, Y. H., Lin, S. H., Hsueh, Y. L. and Lee, M. J., "Automatic target defect identification for TFT-LCD array process inspection using kernel FCM-based fuzzy SVDD ensemble". *Expert Systems with Applications*, **36**, pp. 1978–1998 (2009).
21. Lingras, P. and West, C., "Interval set clustering of web users with rough k-means". Tech. Rep. 2002-002, Department of Mathematics and Computer Science, St. Mary's University, Halifax, Canada (2002).
22. Mitra, S., Banka, H. and Pedrycz, W., "Rough-fuzzy collaborative clustering". *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, **36** (4), pp. 795–805 (2006).
23. Peters, G., Crespo, F., Lingras, P. and Weber, R., "Soft clustering - fuzzy and rough approaches and their extensions and derivatives". *International Journal of Approximate Reasoning*, **54**, pp. 307–322 (2013).
24. Maji, P. and Pal, S. K., "RFCM: A hybrid clustering algorithm using rough and fuzzy sets". *Fundamenta Informaticae*, **80**(4), pp. 475–496 (2007).

25. Hu, J., Li, T., Luo, C., Fujita, H. and Yang, Y., "Incremental fuzzy cluster ensemble learning based on rough set theory". *Knowledge-Based Systems*, **132**, pp. 144-155 (2017).
26. Mitra, S. and Barman, B., "Rough-Fuzzy Clustering: An Application to Medical Imagery". in *Rough Sets and Knowledge Technology, RSKT 2008.* Lecture Notes in Computer Science, vol. 5009, eds. Wang G., Li T., Grzymala-Busse J.W., Miao D., Skowron A. and Yao Y., pp. 300-307, Springer, Berlin, Heidelberg (2008).
27. Maji, P. and Pal, S. K., *Rough-Fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging*, Wiley-IEEE Press (2012).
28. Maji, P. and Roy, S., "Rough-fuzzy clustering and multiresolution image analysis for text-graphics segmentation". *Applied Soft Computing*, **30**, pp. 705–721 (2015).
29. Saha, I., Sarkar, J. P. and Maulik, U., "Ensemble based rough fuzzy clustering for categorical data". *Knowledge-Based Systems*, **77**, pp. 114-127 (2015).
30. Saha, I., Sarkar, J. P. and Maulik, U., "Integrated Rough Fuzzy Clustering for Categorical data Analysis". *Fuzzy Sets and Systems*, **361**, pp. 1-32 (2019).
31. Chandra, A. and Yao, X., "Evolving hybrid ensembles of learning machines for better generalization". *Neurocomputing*, **69**, pp. 686-700 (2006).
32. Rodriguez, J. J, Kuncheva, L. I., Alonso, C. J., "Rotation forest: A new classifier ensemble method". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**, pp. 1619-1630 (2006).
33. Masnadi-Shirazi, H. and Vasconcelos, N., "Cost-sensitive boosting". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**, pp. 294-309 (2011).
34. Abdullah, M. and Ganapathy, V., "Neural network ensemble for financial trend prediction". *In TENCON Proceedings IEEE*, pp. 157–161 (2000).
35. Chun, S. H. and Park, Y. J., "Dynamic adaptive ensemble case-based reasoning: Application to stock market prediction". *Expert Systems with Applications*, **28**, pp. 435–443 (2005).
36. Chen, Y., Yang, B., Abraham, A., "Flexible neural trees ensemble for stock index modeling". *Neurocomputing*, **70**, pp. 697–703 (2007).
37. Creamer, G. and Freund, Y., "Automated trading with boosting and expert weighting". *Quantitative Finance*, **4**, pp. 401–420 (2010).
38. Tsai, C. F., Lin, Y. C., Yen, D. C. and Chen, Y. M., "Predicting stock returns by classifier ensembles". *Applied Soft Computing*, **11**, pp. 2452–2459 (2011).
39. Xiao, Y., Xiao, J., Lu, F. and Wang, S., "Ensemble ANNs-PSO-GA approach for day a head stock e-exchange prices forecasting". *International Journal of Computational Intelligence Systems*, **6**, pp. 96–114 (2013).
40. Booth, A., Gerding, E. and McGroarty, F., "Automated trading with performance weighted random forests and seasonality". *Expert Systems with Applications*, **41**, pp. 3651–3661 (2014).

41. Pulido, M., Melin, P. and Castillo, O., "Particle swarm optimization of ensemble neural networks with fuzzy aggregation for time series prediction of the Mexican Stock Exchange". *Information Sciences*, **280**, pp. 188-204 (2014).
42. Suzuki, T. and Ohkura, Y., "Financial technical indicator based on chaotic bagging predictors for adaptive stock selection in Japanese and American markets". *Physica A: Statistical Mechanics and its Applications*, **442**, pp. 50–66 (2016).
43. Holopainen, M. and Sarlin, P., "Toward robust early-warning models: a horse race, ensembles and model uncertainty". *Quantitative Finance*, **17**(12), pp. 1933-1963 (2017).
44. Kuncheva, L. I., Bezdek, J. C. and Duin, R. P. W., "Decision templates for multiple classifier fusion: an experimental comparison". *Pattern Recognition*, **34**, pp. 299–314 (2001).
45. Kuncheva, L. I., *Combining Pattern Classifiers, Methods and Algorithms*. Wiley, Hoboken, New Jersey (2004).
46. Zhou, L., Lai, K. K. and Yu, L., "Least squares support vector machines ensemble models for credit scoring". *Expert Systems with Applications*, **37**, pp. 127-133 (2010).
47. Gheyas, I. A. and Smith, L. S., "A novel neural network ensemble architecture for time series forecasting". Neurocomputing, **74**(18), pp. 3855-3864 (2011).
48. Lv, Y., Liu, J., Yang, T. and Zeng, T., "A novel least squares support vector machine ensemble model for NOx emission prediction of a coal-fired boiler", *Energy*, **55**, pp. 319-329 (2013).
49. Liu, Y., Li, C. and Gao, Z., "A novel unified correlation model using ensemble support vector regression for prediction of flooding velocity in randomly packed towers". *Journal of Industrial and Engineering Chemistry*, **20**, pp. 1109–1118 (2014).
50. Albadvi, S., Chaharsooghi, K. and Esfahanipour, A., "Decision making in stock trading: An application of PROMETHEE". *European Journal of Operational Research*, **177**, pp. 673–683 (2007).
51. Shen, K. Y., Yan, M. R. and Tzeng, G. H., "Combining VIKOR-DANP model for glamor stock selection and stock performance improvement". *Knowledge-Based Systems*, **58**, pp. 86–97 (2014).
52. Eckbo, B. E., Makaew, T., Thorburn, K. S., "Are stock-financed takeovers opportunistic?". *Journal of Financial Economics*, **128**(3), pp. 443-465 (2018).
53. Zhang, H. and Yan, C., "Modelling fundamental analysis in portfolio selection". *Quantitative Finance*, **18**(8), pp. 1315-1326 (2018).
54. Chen, Y. S. and Cheng, C. H., "Forecasting PGR of the financial industry using a rough sets classifier based on attribute-granularity". *Knowledge and Information Systems*, **25**, pp. 57–79 (2010).
55. Esfahanipour, A., Goodarzi, M. and Jahanbin, R., "Analysis and forecasting of IPO underpricing". *Neural Computing and Applications*, **27**, pp. 651–658 (2016).
56. Tan, Z., Yan, Z., Zhu, G., "Stock selection with random forest: An exploitation of excess return in the Chinese stock market". *Heliyon*, **5 (8)**, (2019).
57. Yang, F., Chen, Z., Li, J., Tang, L., "A novel hybrid stock selection method with stock prediction", *Applied Soft Computing*, **80**, pp. 820-831 (2019).

58. Babazadeh, H., Esfahanipour, A. "A novel multi period mean-VaR portfolio optimization model considering practical constraints and transaction cost", *Journal of Computational and Applied Mathematics*, **361**, pp. 313-342 (2019).
59. Vanstone, B., Finnie, G. and Hahn, T., "Creating trading systems with fundamental variables and neural networks: The Aby case study". *Mathematics and Computers in Simulation*, **86**, pp. 78–91 (2012).
60. Becker, L. Y., Fei, P. and Lester, A. M., "Stock Selection – An Innovative Application of Genetic Programming Methodology", In *Genetic Programming Theory and Practice IV,* edited by R. Riolo, T. Soule, B. Worzel, pp. 315-334, Springer-Verlag, US (2007).
61. Parque, V., Mabu, S. and Hirasawa, K., "Evolving Asset Selection using Genetic Network Programming". *IEEJ transactions on electrical and electronic engineering*, **7**, pp. 174–182 (2012).
62. Ince, H., "Short term stock selection with case-based reasoning technique". *Applied Soft Computing*, **22**, pp. 205–212 (2014).
63. Silva, A., Neves, R. and Horta, N., "A Hybrid Approach to Portfolio Composition based on Fundamental and Technical Indicators". *Expert Systems with Applications*, **42** (4), pp. 2036-2048 (2015).
64. Yu, H., Chen, R. and Zhang, G., "A SVM Stock Selection Model within PCA". *Procedia Computer Science*, **31**, pp. 406-412 (2014).
65. Huang C. F., Chang C. H., Chang B. R. and Cheng D. W., "A Study of a Hybrid Evolutionary Fuzzy Model for Stock Selection". *IEEE International Conference on Fuzzy Systems*, Taipei, Taiwan (2011).
66. Shen, K. Y. and Tzeng, G. H., "Combined soft computing model for value stock selection based on fundamental analysis". *Applied Soft Computing*, **37**, pp. 142–155 (2015).
67. Sang, X., Zhou, Y., Yu, X., "An uncertain possibility-probability information fusion method under interval type-2 fuzzy environment and its application in stock selection", *Information Sciences*, **504**, pp. 546-560 (2019).
68. Thakur G.S.M., Bhattacharyya, R., Sarkar, S., "Stock portfolio selection using Dempster–Shafer evidence theory", *Journal of King Saud University - Computer and Information Sciences*, **30 (2)**, pp. 223-235 (2018).
69. Chang, P. C. and Liu, C. H., "A TSK type fuzzy rule based system for stock price prediction". *Expert systems with Applications*, **34**(1), pp. 135-144 (2008).
70. Esfahanipour, A. and Aghamiri, W., "Adapted neuro-fuzzy inference system on indirect approach TSK fuzzy rule base for stock market analysis". *Expert systems with Applications*, **37**(7), pp. 4742-4748 (2010).
71. Mousavi, S., Esfahanipour, A. and Fazel Zarandi, M. H., "MGP-INTACTSKY: Multitree Genetic Programming-based learning of INTerpertable and Accurate TSK sYstems for dynamic portfolio trading". *Applied soft computing*, **34**, pp. 449- 462 (2015).
72. Pawlak, Z. *Rough Sets: Theoretical Aspects of Reasoning About Data*. Springer, Netherlands (1991).
73. Sarkar, J. P., Saha, I. and Maulik, U., "Rough Possibilistic Type-2 Fuzzy C-Means clustering for MR brain image segmentation". *Applied Soft Computing*, **46**, pp. 527–536 (2016).

74. Zhang, T., Chen, L., and Ma, F., "A modified rough c-means clustering algorithm based on hybrid imbalanced measure of distance and density". *International Journal of Approximate Reasoning*, **55**(8), pp. 1805-1818 (2014).
75. Melek, W. W., Goldenberg, A. A. and Emami, M. R., "A fuzzy noise-rejection data partitioning algorithm". *International Journal of Approximate Reasoning*, **38**, pp. 1–17 (2005).
76. Emami, M. R., Turksen, I. B. and Goldenberg, A. A., "Development of a systematic methodology of fuzzy logic modeling". *IEEE Transactions on Fuzzy Systems*, **6**(3), pp. 346–361 (1998).
77. Krishnapuram, R. and Keller, J. M., "A Possibilistic Approach to Clustering". *Fuzzy Systems, IEEE Transactions on*, **1**(2), pp. 98 - 110 (1993).
78. Fazel Zarandi, M. H., Doostparast Torshizi, A., Turksen, I. B. and Rezaee, B., "A new indirect approach to the type-2 fuzzy systems modeling and design". *Information Sciences*, **232**, pp. 346-365 (2013).
79. Hadavandi, E., Shavandi, H. and Ghanbari, A., "Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting". *Knowledge-Based Systems*, **23**(8), pp. 800-808 (2010).
80. Park, B. J., Kim, W. D., Oh, S. K. and Pedrycz, W., "Fuzzy set-oriented neural networks based on fuzzy polynomial inference and dynamic genetic optimization". *Knowledge and Information Systems*, **39**, pp. 207–240 (2014).
81. Adeli, H. and Hung, S. L., *Machine learning: neural networks, genetic algorithms, and fuzzy system*s, John Wiley & Sons, New York (1994).
82. Karray, F. O. and De Silva, C. W., *Soft computing and intelligent systems design: theory, tools, and application*s, Addison-Wesley, Boston (2004).
83. Alcal´a, R., Ducange, P., Herrera, F., Lazzerini, B. and Marcelloni, F., "A multiobjective evolutionary approach to concurrently learn rule and data bases of linguistic fuzzy-rule-based systems". *Fuzzy Systems, IEEE Transactions on*, **17**(5), pp. 1106-1122 (2009).
84. Casillas, J. and Carse, B., "Special issue on Genetic Fuzzy Systems: Recent Developments and Future Directions". *Soft Computing*, **13**(5), pp. 417-418 (2009).
85. Ishibuchi, H. and Nojima, Y., "Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning". *International Journal of Approximate Reasoning*, **44**(1), pp. 4-31 (2007).
86. Herrera, F., "Genetic fuzzy systems: taxonomy, current research trends and prospects". *Evolutionary Intelligence*, **1**(1), pp. 27-46 (2008).
87. Verikas, A., Guzaitis, J., Gelzinis, A. and Bacauskiene, M., "A general framework for designing a fuzzy rule-based classifier". *Knowledge and Information Systems*, **29**, pp. 203–221 (2011).
88. McGee, R., O'Neill, M. and Brabazon, A., "The Syntax of Stock Selection: Grammatical Evolution of a Stock Picking Model". *IEEE Congress on Evolutionary Computation (CEC)*, pp. 1-8 (2010).
89. TSETMC: Tehran Securities Exchange Technology Management Co., Available online at: http://irbourse.com/NewsTag.aspx?tag=50, (accessed 15 June 2017).

90. Barak, S. and Modarres, M., "Developing an approach to evaluate stocks by forecasting effective features with data mining methods". *Expert Systems with Applications*, **42**, pp. 1325–1339 (2015).

91. Ghandar, A., Michalewicz, Z., Schmidt, M., Tô, T. D. and Zurbrugg, R., "Computational intelligence for evolving trading rules". *IEEE Transactions on Evolutionary Computation*, **13**(1), pp. 71-86 (2009).

92. Kroll, B., Trichilo, D. and Braun, J., Extending the fundamental law of investment management, JPMorgan asset management, 2005, available online at: https://www.jpmorgan.com/cm/BlobServer/Extending_the_Fundamental_Law_of_Investment_Management_.pdf?blobkey=id&blobwhere=1158630145176&blobheader=application%2Fpdf&blobheadername1=Cache-Control&blobheadervalue1=private&blobcol=urldata&blobtable=MungoBlobs.

93. Gillam, R. A., Guerard, J. B. and Cahan, R., "News volume information: Beyond earnings forecasting in a global stock selection model". *International Journal of Forecasting*, **31**(2), pp. 575-581 (2015).

94. Guerard, J. B., Markowitz, H. and Xu, G., "Earnings forecasting in a global stock selection model and efficient portfolio construction and management". *International Journal of Forecasting*, **31**(2), pp. 550-560 (2015).

95. Huang, C. F., "A hybrid stock selection model using genetic algorithms and support vector regression". *Applied Soft Computing*, **12**, pp. 807–818 (2012).

Somayeh Mousavi received her B.Sc., M.Sc. and PhD degrees in Industrial Engineering from Amirkabir University of Technology, Tehran, Iran in 2007, 2009 and 2015, respectively.

She is currently an Assistant Professor at Industrial Engineering Department, Meybod university, Meybod, Yazd, Iran. Her main research interests include financial forecasting, portfolio selection, financial risk management, fuzzy expert systems and applications of artificial intelligence and Meta-heuristics in financial markets.

Dr. Mousavi has published her research articles on financial decision making using soft computing methods in journals of *Expert Systems with Applications, Knowledge based systems and Applied soft computing*.

Akbar Esfahanipour received his B.Sc in Industrial Engineering from Amirkabir University of Technology, Tehran, Iran in 1995. His M.Sc and PhD degrees are in industrial engineering from Tarbiat Modares University, Tehran, Iran in 1998 and 2004, respectively. He is currently an Associate Professor at Industrial Engineering Department, Amirkabir University of Technology. He has worked as a senior consultant for over 15 years in his specialized field of expertise in various industries.

His research interests are in the areas of forecasting in financial markets, application of soft computing methods in financial decision making, behavioral finance, financial resiliency, and analysis of financial risks. Dr. Esfahanipour has published his research articles on financial decision making in prestigious journals such as *European Journal of Operational Research*, *Journal of Management Information Systems*, *Expert Systems with Applications, Quantitative Finance, Knowledge-Based Systems,* and *Applied Soft Computing*.

Mohammad Hossein Fazel Zarandi is Professor in Department of Industrial Engineering at Amirkabir University of Technology, Tehran, Iran, and a member of the Knowledge-Information Systems Laboratory at University of Toronto, Canada. His main research interests focus on Big Data Analytics, Artificial Intelligence,

Data Modeling, Soft Intelligent Computing, Deep Learning, Fuzzy Sets and Systems, Meta-heuristics, and Optimization.

Professor Fazel Zarandi has published over 25 books and Book Chapters, more than 300 scientific journal papers, more than 200 refereed conference papers and several technical reports in the above areas, most of which are also accessible on the web. He has taught several courses in Big Data Analytics, Data Modeling, Fuzzy Systems Engineering, Decision Support Systems, Information Systems, Artificial Intelligence and Expert Systems, Systems Analysis and Design, Scheduling, Deep Learning, Simulations, and Multi-Agent Systems, at several universities in Iran and North America.

Figure and table captions

Table 1. Position of this study among the related studies in the literature.

Reference	Learning tool	No. of fundamental variables	Handle nonlinearity	Fitness measure	Type of the system	Consider different fundamental characteristics of sectors
Becker et al. [60]	Genetic programming	65	✓	Information coefficient and spread	crisp	×
Quah [3]	Artificial neural networks	11	✓	Classification accuracy	Crisp and fuzzy	×
Huang et al. [65]	Genetic algorithm-Fuzzy	12	×	Return of top ranked stocks	Fuzzy	×
Huang [95]	support vector regression-genetic algorithms	14	✓	Cumulative return of the selected stocks	Crisp	×
Vanstone et al. [59]	Artificial neural networks	4	✓	Max. percentage change in price over next 200 days	Crisp	×
Parque et al. [61]	Genetic network programming	10	✓	Risk-adjusted return of selected stocks	Crisp	×
Ince [62]	Genetic algorithm-Case-based reasoning	7	✓	Classification accuracy	Crisp	×
Yu et al. [64]	Support Vector Machines	20	✓	Classification accuracy & the portfolio accumulated return	crisp	×
Silva et al. [63]	Genetic algorithm	10	×	Return and risk of proposed portfolio	Crisp	×
Shen & Tzeng [66]	VIKOR, DANP, DEMATEL (Decision-making trial and evaluation Laboratory)	17	✓	Classification accuracy	fuzzy	×
This study	Genetic algorithm-Artificial neural networks	36	✓	Information coefficient	TSK type fuzzy rule-based system	✓

Table 2. Financial ratios as the possible input variables

Category	Financial ratio
Profitability ratios	Percentage of net profit to sale, percentage of operating profit to sale, percentage of gross profit to sale, percentage of gross margin to sale, percentage of net profit to gross margin, return on assets (after tax), return on equity (after tax), return on working capital, working capital return percentage, fixed assets return percentage
Liquidity ratios	Current ratio, quick ratio, liquidity ratio, current assets ratio, networking capital
Activity ratios	Inventory turnover, Average payment period, inventory to working capital, current assets turn over, fixed asset turnover, total asset turnover
Leverage ratios	Debt coverage ratio, debt to total assets ratio, debt to equity ratio, fixed assets to equity ratio, long-term debt to equity ratio, current debt to equity ratio, equity ratio, interest coverage ratio
Valuation ratios	Actual Earning per Share (EPS), net dividend per share, price to EPS ratio (P/E), book value per share, dividend yield, price to book ratio(P/B), capitalization

Table 3. The training and testing periods of the five experiments

Experiment series	Training period	Testing period
Exp 1	Mar. 20, 1991- Mar. 19, 2009	Mar. 20, 2009- Mar. 19, 2010
Exp 2	Mar. 20, 1991- Mar. 19, 2010	Mar. 20, 2010- Mar. 19, 2011
Exp 3	Mar. 20, 1991- Mar. 19, 2011	Mar. 20, 2011- Mar. 19, 2012
Exp 4	Mar. 20, 1991- Mar. 19, 2012	Mar. 20, 2012- Mar. 19, 2013
Exp 5	Mar. 20, 1991- Mar. 19, 2013	Mar. 20, 2013- Mar. 19, 2014

Table 4. Parameter settings of genetic algorithm

Population size	100
Number of generations	200
Crossover rate	0.7
Mutation rate	0.3
Reproduction size	20

Table 5. Performance of the individual TSK systems on their corresponding training data subsets in terms of Information Coefficient (IC)

Training Subset	Exp 1	Exp2	Exp3	Exp4	Exp5
U_1	28.41%	28.11%	28.03%	27.67%	26.67%
U_2	40.70%	37.84%	30.64%	32.19%	30.45%
U_3	48.03%	45.45%	43.12%	41.60%	35.11%
U_4	33.01%	28.04%	31.00%	26.21%	28.35%

Notes: U_i is the ith training subset provided by the rough-fuzzy noise rejection clustering.

Table 6. Performance of our proposed ensemble learning model and the other comparative models on testing periods in terms of Information Coefficient (IC)

Stock Selection system	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Average
Single TSK	6.51%	19.42%	1.81%	14.53%	0.26%	8.50%
FNRC-TSK-AveE[a]	7.51%	15.32%	15.54%	16.48%	8.86%	12.74%
RFNRC-TSK-AveE[b]	11.89%	18.06%	15.17%	17.85%	13.22%	15.24%
Our proposed system: RFNRC-TSK-MBE[c]	12.43%	26.22%	19.86%	18.06%	14.16%	18.15%

[a]FNRC-TSK-AveE: fuzzy noise rejection clustering based TSK system with averaging ensemble

[b]RFNRC-TSK-AveE: rough-fuzzy noise rejection clustering based TSK system with averaging ensemble

[c]RFNRC-TSK-MBE: rough-fuzzy noise rejection clustering based modular TSK system with membership based ensemble

Table 7.Results of the student t-test for the pair wise comparison of the stock selection systems

Stock selection system	RFNRC-TSK-AveE	FNRC-TSK-AveE	Single TSK
Our proposed system: RFNRC-TSK-MBE	0.0012 [2.91%]	0.0000 [5.41%]	0.0000 [9.64%]
RFNRC-TSK-AveE		0.0027 [2.50%]	0.0000 [6.73%]
FNRC-TSK-AveE			0.0000 [4.24%]

Notes: The table reports the p-values of tests for the pair wise dominance of systems' ICs. The difference between the IC averages of the respective systems are reported in brackets

Table 8. Classification performance of our proposed model versus ANFIS on testing periods in terms of classification* accuracy

Model	EXP1	EXP2	EXP3	EXP4	EXP5	Average
RFNRC-TSK-MBE	77.3%	80.9%	88.1%	79.5%	50.3%	75.22%
ANFIS	69.1%	72.6%	72%	61.9%	49.7%	65.06%

* In this experiment, all stocks have been classified in two classes. Class 1 represents stocks which appreciate in share price equal or more than 80% within one year. Class 2 contains all other stocks.

Table 9. Performance of our proposed RFNRC-TSK-MBE model on testing periods in terms of appreciation in selected stocks price

	EXP1	EXP2	EXP3	EXP4	EXP5	Average
Average appreciation of the selected stocks	67.2%	89.5%	19.7%	12.2%	128.3%	63.38%
Average appreciation of all 150 stocks	55.3%	37.1%	2.2%	8%	117.5%	44.02%
Excess appreciation of the selected stocks over all 150 stocks	11.9%	52.4%	17.5%	4.2%	10.8%	19.36%

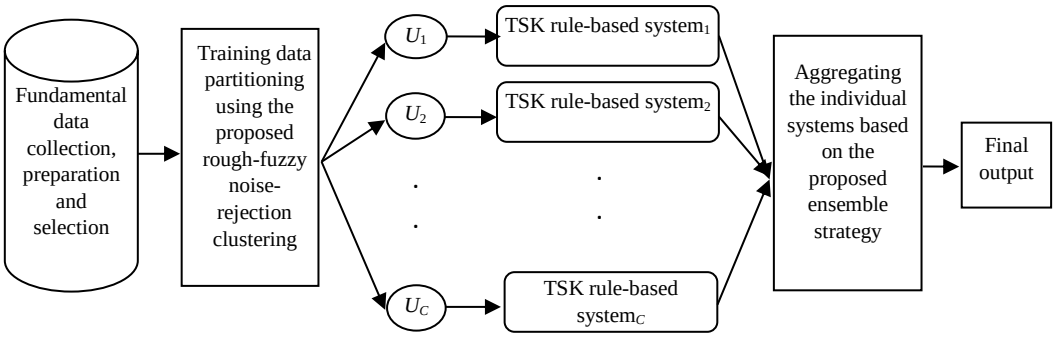

Fig. 1. The overall framework of the proposed ensemble learning model for stock selection.

a_{10}	a_{11}	….	a_{1n}	a_{20}	a_{21}	….	a_{2n}	….	a_{r0}	a_{r1}	…	a_{rn}

Fig. 2. Encoding the TSK consequent parameters as a GA chromosome.

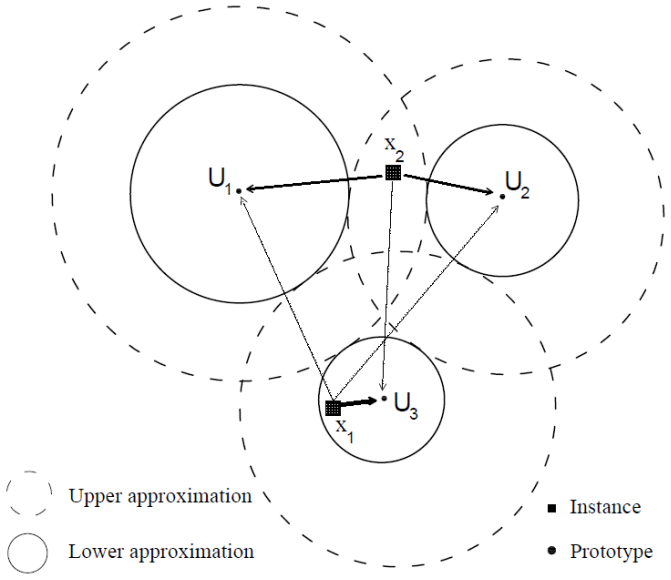

Fig. 3. The module's ensemble weights in a typical rough-fuzzy partitioning.

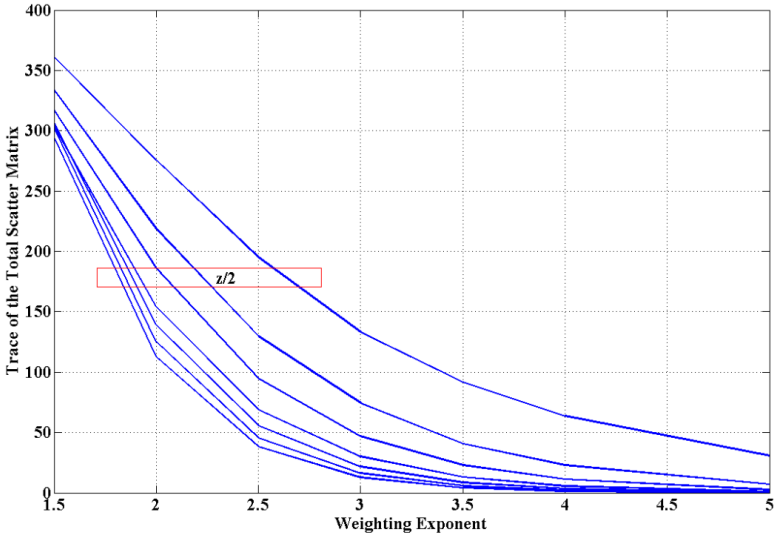

Fig. 4. Selection of the suitable weighting exponent.

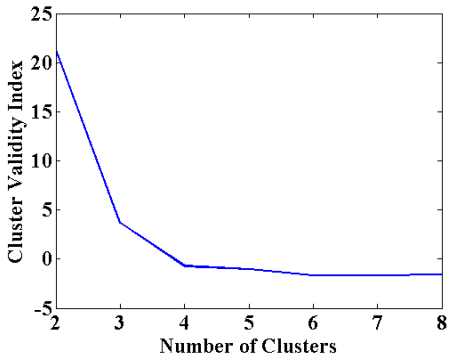

Fig. 5. Identification of the optimum number of clusters.

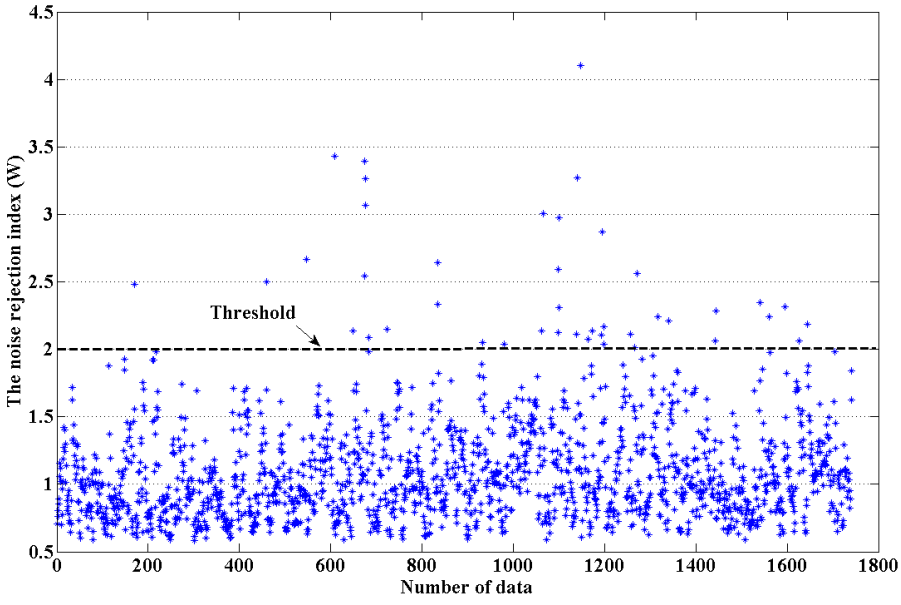

Fig. 6. Application of the noise-rejection criterion.