Applied Soft Computing 61 (2017) 593–602

Contents lists available at ScienceDirect

Applied Soft Computing

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

Full length article

Using genetic algorithm to support clustering-based portfolio optimization by investor information

Donghyun Cheong [a], Young Min Kim [a], Hyun Woo Byun [a], Kyong Joo Oh [a,*], Tae Yoon Kim [b]

[a] Department of Industrial Engineering, Yonsei University, 50, Yonsei-ro, Seodaemun-gu, Seoul 03722, South Korea
[b] Department of Statistics, Keimyung University, Daegu 704-701, South Korea

ARTICLE INFO

Article history:
Received 23 July 2015
Received in revised form 18 May 2017
Accepted 21 August 2017
Available online 31 August 2017

Keywords:
Cluster analysis
Genetic algorithm
Investor information
Portfolio
Stock market

ABSTRACT

A clustering-based portfolio optimization scheme that employs a genetic algorithm (GA) based on investor information for active portfolio management is presented. Whereas numerous studies have investigated trading behaviors, investor performance, and portfolio investment strategies, few works have developed investment strategies based on investor information. This study is conducted in two phases. First, a basket of portfolio (i.e., a collection of stocks held in individual portfolios) is developed through a cluster analysis of investor information. A GA is then employed to optimize the weights of the selected stocks. And the optimized portfolio is rebalanced to get excess return. It is concluded that the proposed multistage portfolio optimization scheme for active portfolio management generates superior results than previously proposed methods for the Korean stock market.

1. Introduction

In financial markets, three types of investors – foreign (institutional), (domestic) institutional, and individual investors – engage in constant competition. Recent studies regarding the trading performance of these three types of investors (or investment information) have concluded that foreign and institutional investors perform better than individual investors [1–9]. This stronger performance of foreign and institutional investors may be attributable to their access to optimal portfolios. Therefore, it appears evident that emulating foreign and institutional investor portfolio strategies can lead to better performance. However, there is no active portfolio management imitating such strategies employed during portfolio construction. Furthermore, portfolio investment strategies based on *volume* of all three types of investors have benefited because volume levels alone serve as a significant indicator in stock markets [10].

Stock selection strategies strongly affect portfolio performance. By allocating capital to a portfolio of stocks, investors strive to maximize their return while minimizing their risk. Modern portfolio theory, which was developed by Markowitz [11,12], is an efficient portfolio theory that is based on the mean–variance relationship. In the mean–variance portfolio framework, diversified portfolio development secures the greatest possible expected return for a given degree of risk tolerance. However, it is difficult to select stocks and assign relative weights to a portfolio. Nanda et al. [13] asserted that the complexities of diversified stock selection can be mitigated by clustering stock data and found the *k*-means method to perform best among cluster analysis methods in the context of the Indian stock market. El hachloufi et al. [14] present an approach based on the classification and genetic algorithms by Value at Risk(VaR) to obtain an optimal stock portfolio, which leads to a financial gain surplus in terms of cost and taxes reduction.

This study presents a clustering-based portfolio optimization scheme that uses a genetic algorithm (GA) based on investor information, i.e., the three types of investors – foreign, institutional, and individual investors – for active portfolio management. A cluster analysis (in this study, *k* denotes the number of clusters) is conducted to select stocks for each type of investor, and a GA is employed to optimize the weights for each selected stock, cluster and type of investor. The proposed *clustering-based portfolio optimization scheme* is employed for the Korea Composite Stock Price Index (KOSPI) 200 and actively traded constructing portfolios to get excess return. Of the 200 companies included in the KOSPI 200, the top 90 companies (or stocks) in terms of market capitalization are selected for portfolio construction. The suggested clustering-based

* Corresponding author.
E-mail addresses: dhcheong@yonsei.ac.kr (D. Cheong), wealth38@gmail.com (Y.M. Kim), next1219@nate.com (H.W. Byun), johanoh@yonsei.ac.kr (K.J. Oh), tykim@kmu.ac.kr (T.Y. Kim).

http://dx.doi.org/10.1016/j.asoc.2017.08.042
1568-4946/

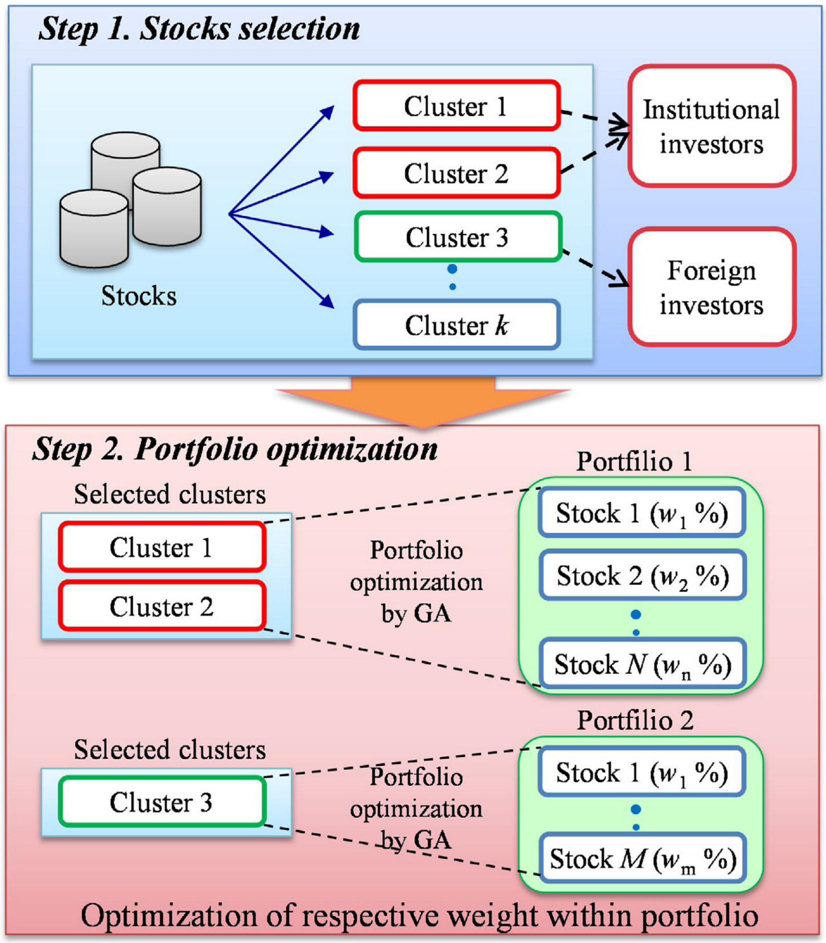

Fig. 1. Architecture of the clustering-based portfolio optimization model.

portfolio optimization scheme generates higher returns than those yielded by the KOSPI 200 market portfolio buy–and–hold strategy.

The contribution of this study as follows:

- Investor information of individual stock is used to develop investment strategy. Few studies have developed investment strategy based on investor information. Clustering technique selects relatively more invested stocks by specific investor than the other stocks. Using investor information of individual stock enables to classify and select promising stocks.
- Clustering for portfolio construction and GA for portfolio optimization are adopted sequentially for portfolio emulation strategy. The proposed multistage investment strategy shows that imitating investor behaviors and patterns can be used to make the investment strategy and enhances the portfolio performance.
- Investment tendency of investor is identified. This study shows that institutional and foreign investors tend to participate via long-term (six-month) investments in the KOSPI stock market and institutional investors tend to consider return and risk simultaneously, whereas foreign investors tend to consider portfolio risk only.

The remainder of this paper organized as follows. A literature review is presented in Section 2. Section 3 describes the details of the proposed model. The empirical experiments are described in Section 4. Finally, Section 5 presents concluding remarks.

2. Literature review

2.1. Portfolio management

Portfolio management involves selecting stocks to include in a portfolio, deciding how much capital to allocate to each stock, and determining when it is necessary to rebalance the portfolio. Throughout the portfolio management process, it is necessary to consider the investor's risk tolerance level because some investors are more willing to accept a greater degree of risk in exchange for potentially greater return.

Two different portfolio management strategies can be employed, depending on the investor's degree of risk tolerance [15]. When a passive strategy is employed, the investor tracks the market index. This approach is related to random walk theory [16], which states that it is impossible to predict future prices or to guarantee a return greater than the market return. In short, the passive management strategy strictly involves establishing a well–diversified portfolio without attempting to identify stocks of high expected return. The passive management strategy is typically executed using a buy–and–hold approach. In contrast, in active management, an investor constructs a portfolio to outperform a benchmark index by buying stocks of high expected return and selling stocks of low expected return. Investors who use active management strategies are generally more willing to accept higher risk. Moreover, depending on the investor's degree of risk tolerance, active management strategies differ. This study applies

Table 1
Variables used in the cluster analysis (From *t* day to *T* day).

Variable	Formula
Return	$Return = \frac{Price\ \ of\ \ stock\ \ (T) - Price\ \ of\ \ stock\ \ (t)}{Price\ \ of\ \ stock\ \ (t)}$
Institutional investor trading volume proportion (TVP)	$TVP\ \ of\ \ institutional\ \ invesetor = \frac{\sum_{i=t}^{T} Net\ \ trading\ \ volume\ \ of\ \ institutional\ \ investor\ \ (i)}{\sum_{i=t}^{T} trading\ \ volume\ \ (i)}$
Foreign investor TVP	$TVP\ \ of\ \ foreign\ \ invesetor = \frac{\sum_{i=t}^{T} Net\ \ trading\ \ volume\ \ of\ \ foreign\ \ investor\ (i)}{\sum_{i=t}^{T} trading\ \ volume\ (i)}$
Individual investor TVP	$TVP\ \ of\ \ individual\ \ invesetor = \frac{\sum_{i=t}^{T} Net\ \ trading\ \ volume\ \ of\ \ individual\ \ investor\ (i)}{\sum_{i=t}^{T} trading\ \ volume\ (i)}$
Volume ratio (VR)	$VR = \frac{\sum_{i=t}^{T} (up\ \ volume\ \ (i) + 0.5 \times steady\ \ volume\ \ (i))}{\sum_{i=t}^{T} (down\ \ volume\ \ (i) + 0.5 \times steady\ \ volume\ (i))}$

an active management strategy by formulating a clustering-based portfolio optimization scheme based on a GA.

2.2. Investor information

The different trading behaviors of each type of investor create differences in their trading performance. Froot and Ramadorai [3] studied international portfolio flows in several countries and found that foreign investors, who have an information advantage, predict future equity returns relatively well. Grinblatt and Keloharju [2] found that foreign investors exhibit superior performance in the Finnish stock market. Kamesaka et al. [4] and Bae et al. [8] also found that in the Japanese market, foreign investors exhibit strong market-timing abilities and thus achieve better performance. However, other studies have presented conflicting results. Brennan and Cao [1] and Dvorak [6] concluded that foreign investors are less informed than domestic investors and thus exhibit inferior performance. Choe et al. [7] also found that because of poor market-timing tendencies, foreign investors perform poorly. Agarwal et al. [9] reported that foreign investors' inferior performance is attributable to their aggressiveness rather than to their information disadvantage or poor market timing.

Institutional investors appear to have informational advantages relative to other investor types, which cause institutional investors to have superior performance. Barber et al. [5] found that Taiwanese institutional investors profit from other investors that exhibit an information disadvantage. The authors also found the performance of individual investors to be poor in the Taiwanese stock market. Kamesaka et al. [4] reported that individual investors exhibit poor market timing and performance in the Japanese market. Thus, foreign and institutional investor trading data are examined in this study. The model presented in this study disregards individual investor data because previous studies have found such individuals to exhibit relatively poor trading performance.

2.3. K-means clustering

Clustering is often employed in multivariate data analysis [17]. Cluster analysis involves the use of an unsupervised learning algorithm that generates clusters. The purpose of cluster analysis is to obtain a meaningful series of N variables that are categorized based on their characteristics [18,19]. Clustering involves grouping objects with similar characteristics. The degree of dissimilarity between objects is often measured based on the Euclidean, Manhattan, or Minkowski distances. Cluster analyses have been employed in numerous research areas to address classification problems. In the field of finance, Yu and Wang [20] presented an approach that clusters data via *k*-means clustering to divide stocks into different categories based on their financial data. Zhou et al. [21] conducted a cluster analysis to examine stock investments. Investors can use a synthetic evaluation index system that measures the similarities between stocks to predict stock price fluctuations. In this study, we employ *k*-means clustering to construct a portfolio basket. The Euclidean distance is used as the *k*-means clustering measure.

2.4. Genetic algorithms

GAs are a stochastic optimization technique developed by Holland [22] that are based on the principles of Darwinian evolution. GAs employ an evolutionary process of selection, crossover and mutation [23–29] that, in principle, yields the optimal solution to a problem. Because of the lack of restrictive assumptions regarding the solution space, GAs are an effective technique for solving optimization problems. Via selection, crossover and mutation, the population converges to one high-fitness chromosome [30]. The

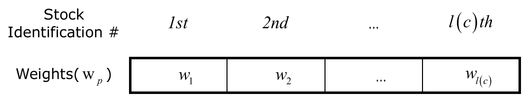

Fig. 2. The example of Chromosome.

fitness value of a chromosome denotes how strong the chromosome is. GAs have been employed in various research disciplines for several years [31,32] and have been demonstrated to be effective for financial applications. Rivera et al. surveyed the number of approaches proposed for financial applications and found that genetic algorithms are the most used approach for financial applications [33]. The technique was used by Bauer [34] to generate trading rules, by Colin [35] to identify optimal lengths of moving average crossover strategies and by Deboeck [36] to form a neural network trading system. GAs have also proven quite effective for portfolio formulation [37,38]. Chen et al. reported that GAs show promising results in financial applications and GAs are effective on the investment strategy portfolio problem [39]. In this study, GA is employed to optimize the weights ascribed to each stock of a portfolio. To find a possible solution, each chromosome, which denotes the weight of an individual stock within the portfolio, is optimized.

3. Model specification

The proposed clustering-based portfolio optimization scheme (see Fig. 1) is conducted in two steps. Stocks are first selected via investor information k-means clustering to form the portfolio. In the second step, weights of selected stocks in the first step are optimally assigned via the GA. The performance of the clustering-based portfolio optimization scheme is also evaluated using the sliding window technique. Whenever a portfolio is reconstructed for each rebalancing point, the first and second steps are repeated to generate a new portfolio.

3.1. Step 1: stock selection

In this step, a cluster analysis is employed to select stocks of high expected return based on investor information. The five variables listed in Table 1 are used to select stocks via k–means clustering. The returns refer to the direction and strength of stocks price movement for a given period of time. The three trading volume proportions (TVPs), one for each investor type, indicate the net trading volume as a proportion of the total trading volume within a given period of time. Finally, the volume ratio (VR) (the volume indicator used in the technical analysis) is calculated as the trading volume of a stock on days when the price increases divided by the trading volume on days when the price decreases, i.e., the VR denotes the proportion of the trading volume for which the price of stocks increases. The five variables are standardized as Z-scores for the cluster analysis.

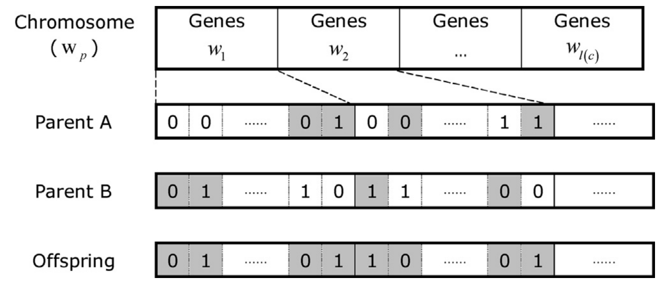

Fig. 4. Crossover in GA.

Let $\mathbf{X}$ be a $m \times n$ matrix that element x_{ij} denotes the ith variable of the jth stock ($i = 1, 2, \cdots, m$ and $j = 1, 2, \cdots, n$) and $\mathbf{x}_j$ be a $m \times 1$ column matrix, jth column of $\mathbf{X}$

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n) \tag{1}$$

$$\mathbf{x}_j = \left(x_{1j}, x_{2j}, \cdots, x_{mj}\right)^T, \quad j = 1, 2, \cdots, n \tag{2}$$

Let $\mathbf{S}$ be a disjoint set of clusters $\mathbf{s}_c$ which is a $m \times l(c)$ matrix.

$$\mathbf{S} = \left\{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_k\right\}, \quad \bigcup_{c=1}^{k} \mathbf{s}_c = \mathbf{X}, \quad \mathbf{s}_c \in \mathbf{S} \tag{3}$$

$$\mathbf{s}_c = \left(\mathbf{x}_1^c, \mathbf{x}_2^c, \cdots, \mathbf{x}_{l(c)}^c\right), \quad c = 1, 2, \cdots, k \tag{4}$$

where k is the number of clusters for k–means clustering and $l(c)$ is the number of stocks for the cluster $\mathbf{s}_c$ ($l(c) \le n$).

The main purpose of k–means clustering is to enhance the expected return of a portfolio. Thus, k–means clustering is employed to obtain a meaningful k groups among $n$stocks ($k \le n$) and to select clusters for a portfolio set $P = \left\{\mathbf{s}_c | \mathbf{s}_c \text{ is selected cluster}\right\}$ construction.

$$\mathbf{S} = arg \min_{\mathrm{S}} \sum_{c=1}^{k} \sum_{\mathrm{x}_j \in \mathrm{s}_c} |\mathrm{x}_j - \mu_c| \tag{5}$$

where μ_c is the center of $\mathbf{s}_c$.

The results of the previous studies described in Section 2.2, which focused on the trading behaviors and performance of different types of investors, indicate that foreign and institutional investors tend to exhibit better performance [4,5,8]. Thus, after k–means clustering, formula (5), clusters are arranged in descending order of mean TVP values for foreign and institutional investors.

And then, select cluster $\mathbf{s}_c$ whose ith center value μ_c^i is the highest of all clusters and a portfolio set P is constructed of stocks within chosen cluster $\mathbf{s}_c$

$$\mu_c^i = \frac{1}{l(c)} \sum_{j=1}^{l(c)} x_{ij}^c, \quad x_{ij}^c \in \mathbf{x}_j^c \in \mathbf{s}_c, \quad i = 1, 2, \cdots, m \tag{6}$$

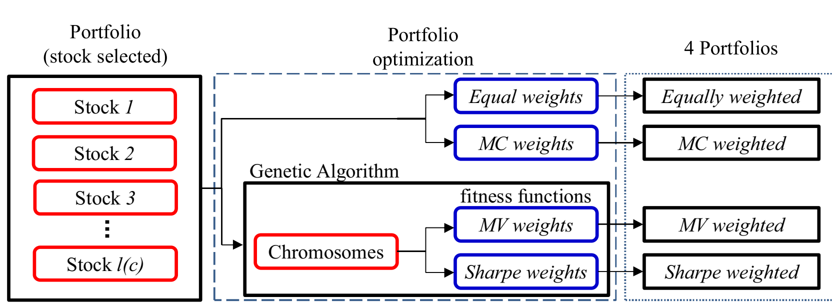

Fig. 3. Portfolio optimization by GA.

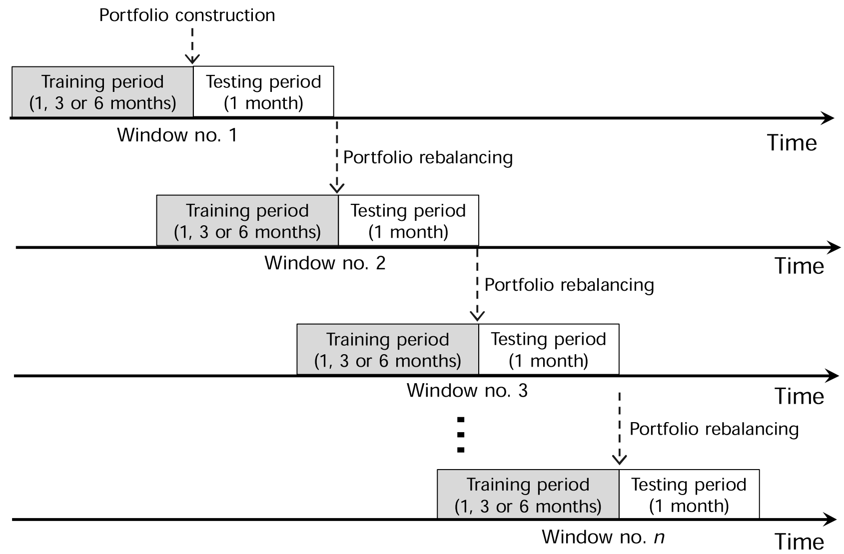

Fig. 5. Sliding window technique.

```
Proposed clustering-based portfolio optimization scheme
Parameter : Investor type, training period
While sliding window do
    Training period :
        For k varies do
            k-means clustering by changing k
            construct a portfolio with the highest mean TVP value cluster
            GA with 4 objective functions
                1) equal weights
                2) market capital weights
                3) minimum variance weights
                4) Sharpe weights
            comparing portfolio returns in accordance with k
            Find k which maximize the return of portfolio
        End
    Testing period :
        k-means clustering with chosen k in training period
        construct a portfolio with the highest mean TVP value cluster
        GA with 4 objective functions
    Return 4 kinds of portfolio's return
End
Repeat sliding window technique as changing parameters
```

Fig. 6. Proposed clustering-based portfolio optimization scheme.

$$\mathbf{x}_j^c = \left(x_{1j}^c, x_{2j}^c, \cdots, x_{mj}^c\right)^T \quad (7)$$

$$P = \left\{\mathrm{s}_c{}^*\right\} = arg \max_{\mathbf{s}_c} \mu_c^i \quad (8)$$

In other words, one or two clusters with the highest mean TVP value for foreign and institutional investor is selected for portfolio construction. The portfolio is then constructed from stocks within the chosen clusters for foreign and institutional investors.

3.2. Step 2: portfolio optimization

In this step, optimization techniques are used to minimize portfolio risk while maximizing the portfolio return.

Let w_j be the proportion invested in stocks and w be a $1 \times n$ row matrix, a corresponding vector of w_j

$$\mathrm{w} = (w_1, w_2, \cdots, w_n) \quad (9)$$

$$\sum_{j=1}^{n} w_j = 1, \quad w_j \geq 0, \quad j = 1, 2, \cdots, n \quad (10)$$

Table 2
Top 90 companies of the KOSPI 200.

No.	Company	No.	Company
1	Amore Pacific	46	LG
2	Cheil Industriees	47	LG Chem
3	Cheil Worldwide	48	LG Display
4	CJ	49	LG Electronics
5	CJ Korea Express	50	LG Household & Health Care
6	Daelim	51	LG International
7	Daewoo E&C	52	LG Uplus
8	Daewoo International	53	Lotte Chemical
9	Daewoo Securities	54	Lotte Chilsung
10	Dongkuk Steel	55	Lotte Confectionery
11	Doosan	56	Lotte shopping
12	Doosan Heavy Industries & Construction	57	LS
13	Doosan Infracore	58	LSIS
14	DSME	59	Mirae Asset Securities
15	GS	60	NCsoft
16	GS E&C	61	NHN
17	Halla Climate Control	62	Nongshim
18	Hana Financial Group	63	OCI
19	Hanwha	64	Orion
20	Hanwha Chemical	65	POSCO
21	Hyosung	66	S-1
22	Hyundai Department Store	67	Samsung
23	Hyundai Dvp	68	Samsung C&T
24	Hyundai E&C	69	Samsung Electro-Mechanics
25	Hyundai Glovis	70	Samsung Engineering
26	Hyundai Heavy Industries	71	Samsung Fine Chemicals
27	Hyundai Hysco	72	Samsung Fire & Marine Insurance
28	Hyundai Merchant Marine	73	Samsung Heavy Industries
29	Hyundai Mipo Dockyard	74	Samsung SDI
30	Hyundai Mobis	75	Samsung Securities
31	Hyundai Motor Company	76	Samsung Techwin
32	Hyundai Securities	77	Shinhan Financial Group
33	Hyundai Steel	78	Shinsegae
34	IBK	79	SK
35	Kangwon Land	80	SK Chemicals
36	KCC	81	SK Hynix
37	KEPCO	82	SK Networks
38	Kia Motors	83	SK Telecom
39	Korea Gas	84	S-Oil
40	Korea Investment Holdings	85	STX
41	Korea Zinc	86	Taihan Electric Wire
42	Korean Air	87	Woongjin Coway
43	KT	88	Woori Bank
44	KT&G	89	Woori Investment & Securities
45	Kunho Petrochemical	90	Yuhan

The vector of portfolio weights w_p is as follows

$$\mathrm{w}_p = \mathrm{w}_{\mathbf{s}_c} = \left(w_1, w_2, \cdots, w_{l(c)}\right) \tag{11}$$

$$\sum_{j=1}^{l(c)} w_j = 1, \quad w_j \geq 0, \quad j = 1, 2, \cdots, l(c) \tag{12}$$

After construction of a portfolio P by k–means clustering, determine the proportion of stocks w_p following 4 methods, respectively. Thus, GA constructs different 4 portfolios corresponding to 4 methods.

3.2.1. Equal weights

The weights of stocks w_p within portfolio P are equally invested.

$$\mathrm{w}_p = \left(w_1, w_2, \cdots, w_{l(c)}\right), \quad w_j = \frac{1}{l(c)} \tag{13}$$

3.2.2. Market capital weights

MC_j means the market capital of jth stocks and weights of stocks within the portfolio are invested as proportion of market capital.

$$\mathrm{w}_p = \left(w_1, w_2, \cdots, w_{l(c)}\right), \quad w_j = \frac{MC_j}{MC_P} \tag{14}$$

3.2.3. Minimum variance weights

In the mean–variance framework, risk is quantified by the total variance of the portfolio return. The portfolio risk is based on the portfolio variance, which is the weighted sum of the variances of individual stocks within the portfolio.

V_P be a covariance matrix of the return of portfolio R_P

Thus, the risk of portfolio σ_P is calculated as follows

$$\sigma_P{}^2 = \mathrm{w}_p \mathrm{V}_P \mathrm{w}_p{}^T \tag{15}$$

$$\mathrm{w}_p = \left(w_1{}^*, w_2{}^*, \cdots, w_{l(c)}{}^*\right) = \underset{w_1, w_2, \cdots, w_{l(c)}}{arg\ \min}\ \sigma_P{}^2 \tag{16}$$

$$\sum_{j=1}^{l(c)} w_j{}^* = 1, \quad w_j{}^* \geq 0, \quad j = 1, 2, \cdots, l(c) \tag{17}$$

The weights of stocks which minimize the risk of portfolio σ_P are determined as optimal weights.

3.2.4. Sharpe weights

The Sharpe ratio [40] measures the excess return (or risk premium) per unit of risk for an investment asset or trading strategy [41]. A portfolio with a higher Sharpe ratio provides a better return for the same unit of risk. The objective function of the Sharpe ratio is denoted as S_p and is calculated as follows:

$$S_P = \frac{R_P - R_f}{\sigma_P} \tag{18}$$

where R_f is a return of benchmark asset.

$$\mathrm{w}_p = \left(w_1{}^*, w_2{}^*, \cdots, w_{l(c)}{}^*\right) = \underset{w_1, w_2, \cdots, w_{l(c)}}{arg\ \max}\ S_P \tag{19}$$

$$\sum_{j=1}^{l(c)} w_j{}^* = 1, \quad w_j{}^* \geq 0, \quad j = 1, 2, \cdots, l(c) \tag{20}$$

The weights of stocks which maximize Sharpe ratio of portfolio S_P are determined as optimal weights. Then, a GA is used to determine the optimal weight for each selected stock within the portfolio.

Traditional optimization techniques are expected to converge towards the global optimum solution when a large number of stocks (or large $l(c)$) provides a stable surface of fitness function. Since this algorithm handles a relatively small number of stocks selected form the previous stage (or small $l(c)$), traditional optimization techniques tend to produce somewhat biased or less accurate w_p. But GA is likely to solve such dilemma by focusing on the relatively small number of stocks selected form the previous stage. Thus GA is suitable here.

With regard to portfolio optimization, the weights of stocks w_p within portfolio P are represented by chromosome as binary strings. The chromosome is consists of a number of genes which indicate the capital to be allocated to each stock.

Table 3
Institutional investor results for 09/01/2007-05/30/2014.

Train period	Measure	*Equal*	*MC*	*MV*	*Sharpe*
One month	Annual return (%)	1.95	8.52	12.42	14.05
	Sharpe ratio	−0.05	0.18	0.35	0.39
Three months	Annual return (%)	5.87	−0.67	7.81	10.93
	Sharpe ratio	0.09	−0.17	0.19	0.28
Six months	Annual return (%)	9.66	9.87	14.83	24.12
	Sharpe ratio	0.25	0.26	0.50	0.70

Table 4
Comparison GA with Conventional optimization for 09/01/2007-12/31/2008.

Train period	*MV*		*Sharpe*	
	GA	Conventional optimization	GA	Conventional optimization
Sep-07	−17.14	−17.14	−6.73	−6.73
Oct-07	−15.67	−15.67	−16.48	−16.48
Nov-07	−1.58	−1.58	−0.67	−0.99
Dec-07	−11.70	−11.70	−12.23	−12.23
Jan-08	−7.47	−7.47	−1.29	−1.29
Feb-08	1.41	1.41	−1.34	−1.34
Mar-08	15.63	15.63	13.40	13.42
Apr-08	2.67	−3.86	−3.35	−3.35
May-08	−9.63	−9.63	−9.75	−9.75
Jun-08	−1.16	−1.16	−1.93	−1.93
Jul-08	4.06	4.06	7.21	7.21
Aug-08	−7.77	−7.77	−8.67	−8.67
Sep-08	−7.32	−6.67	−11.57	−11.57
Oct-08	5.12	5.12	−9.10	−9.10
Nov-08	8.14	8.14	17.75	17.75
Dec-08	−15.62	−15.62	−15.62	−15.62
Annual return	−38.23	−40.33	−39.21	−39.32

- Initialization: GA randomly generates initial set of solutions called population of chromosomes. Chromosomes, shown in Fig. 2, represent a set of weights of stocks w_p assigned to portfolio *P*. The position of genes in a chromosome is used to represent identification of stock and its value is used to represent the weights of stock.

Chromosomes evolve into new generation of population through the process of selection, crossover and mutation. Such evolutionary process, simple GA replaces entire chromosomes in existing population. But steady-state approach of GA, used in this paper, generates only one new chromosome and replaces existing chromosome at a time.

- Evaluation: After initialization, population is evaluated using fitness functions above 4 methods – *Equal weights*, *Market capital weights*, *Minimum variance weights* and *Sharpe weights*. The fitness functions measures the degree of goodness of the chromosomes which is representing the weights of stocks w_p. The fitness function of *Minimum variance weights* and *Sharpe weights* evaluates chromosomes respectively. Otherwise *Equal weights* and *Market capital weights* portfolios are constructed basic arithmetic technique. Thus, GA optimizes *Minimum variance weights* and *Sharpe weights* portfolios only. After portfolio optimization step, 4 different optimized portfolios are constructed corresponding with 4 objective functions which represent the different weight of stocks each other(see Fig. 3).
- Selection: To create new chromosome, called offspring, two parent chromosomes in a portfolio are selected with a rank-based approach. In general selection approach, chromosomes are selected by proportion to its fitness value. Whereas in rank-based approach, the population is sorted according to its fitness value and rank is assigned to each individual chromosomes. Then, parent chromosomes are selected by rank of its fitness. The rank-based approach overcomes the scaling problems of fitness value which cause highly fitted chromosome to dominate the evolution process.
- Crossover: Crossover is a unique feature of GA. The parent chromosomes in a portfolio are mixed to generate offspring. General single or double point crossover cuts and combines two parent chromosomes at certain point. Every string beyond one point in either chromosomes or between two points is swapped between the two parent chromosomes. The uniform crossover selects each gene randomly from one parent and copies into the other parent in the same order(see Fig. 4). The uniform crossover is considered more exploratory crossover approach and better at preserving schema than traditional crossover approach. In this paper, the uniform crossover is applied.
- Mutation: Point mutation alters several genes in chromosome by generating random number between 0 and 1 every individual gene and flipping the gene value if random number is less than the specified mutation rate. Mutation prevents solution from getting caught in local optimum.

Through this process, existing generation changes into new generation. This process is repeated until optimum solutions are found. Finally, 4 different optimized portfolios are constructed correspond with 4 objective functions.

3.3. Sliding window technique

A sliding window technique is applied to portfolio rebalancing and performance evaluation. Each window is divided into two periods: training and testing periods (see Fig. 5). The portfolio is constructed by applying the proposed clustering-based portfolio optimization scheme. In training period, various portfolios are constructed in accordance with *k*. And most profitable portfolio's *k* is selected for stock selection. In testing period, *k*-means clustering is employed with chosen *k* in training period for stock selection and weights of stocks is optimized by GA. After optimized portfolio is constructed, the portfolio is traded and the performance of the portfolio is evaluated during the testing period (see Fig. 6).

The portfolio is then rebalanced at the end of each testing period via clustering-based portfolio optimization scheme testing during the following training period. This process is repeated for the whole dataset. The performance of the clustering-based portfolio optimization scheme is also evaluated by adjusting the time interval of the training period.

4. Empirical results

Daily stock data obtained from the Korea Securities Computing Corporation (KOSCOM) were used for the empirical experiments. The top 90 companies of the KOSPI 200 in terms of market capitalization were sampled for portfolio construction (see Table 2).

The dataset covers the period from 04/01/2007 to 05/01/2014. To evaluate the performance of the proposed model, we also measured the performance of other portfolio management weighting strategies. The training periods spanned one, three or six months, and the performance evaluation testing period was fixed at one

Table 5
Foreign investor results for 09/01/2007-05/30/2014.

Train period	Measure	*Equal*	*MC*	*MV*	*Sharpe*
One month	Annual return (%)	9.32	8.66	10.25	10.24
	Sharpe ratio	0.25	0.22	0.30	0.27
Three months	Annual return (%)	4.33	5.07	6.61	6.63
	Sharpe ratio	0.04	0.06	0.14	0.13
Six months	Annual return (%)	4.80	4.10	13.16	8.10
	Sharpe ratio	0.06	0.03	0.51	0.24

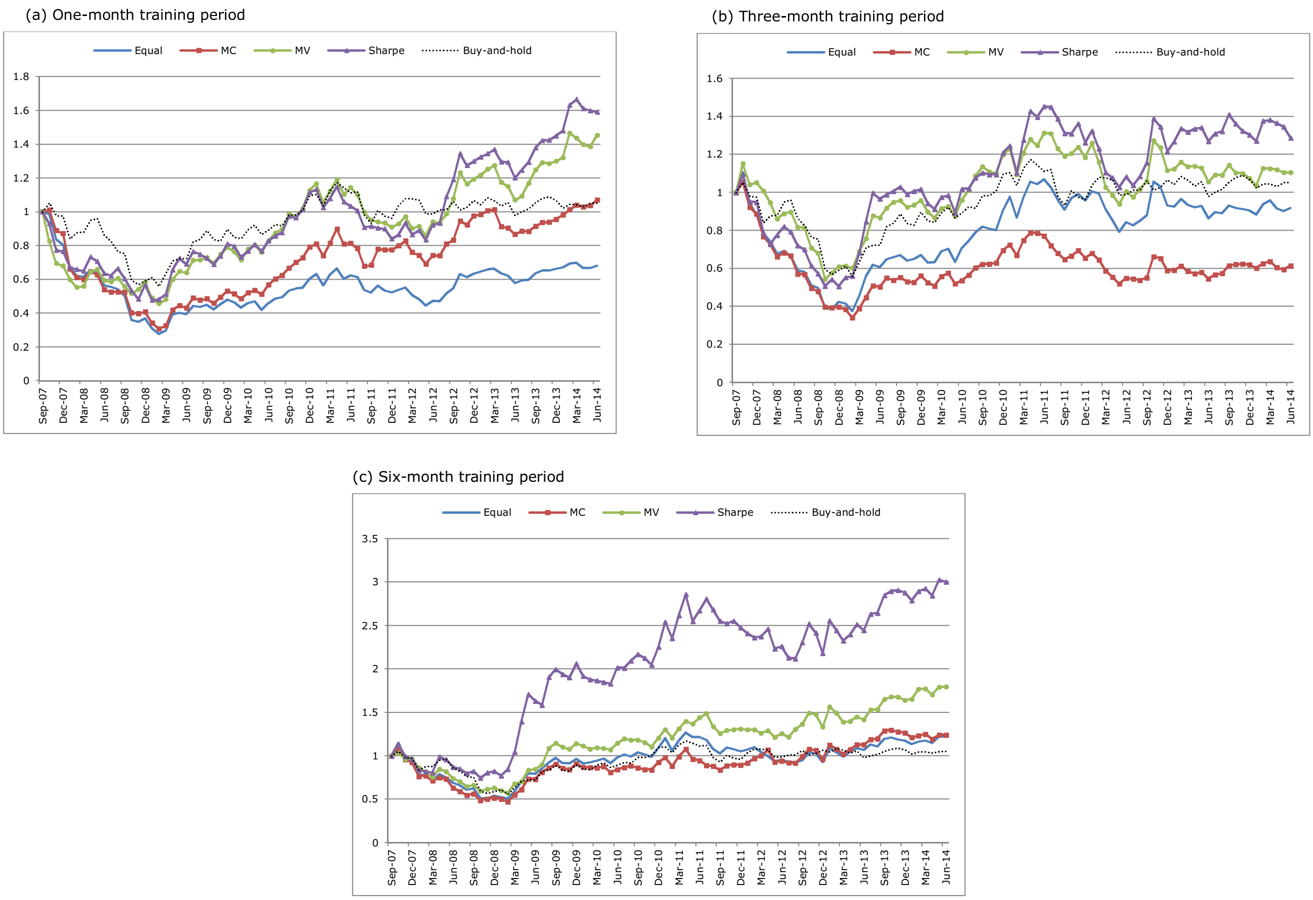

Fig. 7. Flow of cumulative returns of institutional investors. (Testing dataset: 09/01/2007-05/30/2014)

month. Thus, 81 testing sets were used in total (09/01/2007-05/30/2014), and the portfolio was rebalanced each month using investor data (or data regarding investor types) for one, three or six months. In short, 2 kinds of parameters – 2 investor types and 3 training periods – with 4 GA objective functions make 24 testing results as constructing 1944 portfolios.

When using GA for portfolio optimization, GA parameters are very important because GA parameters affect the scope of the search during the evolutionary process. Through preliminary experiments, population size, crossover rate and mutation rate were selected as 200, 0.5 and 0.06, respectively. The number of generations was fixed at 500 and GA was stopped when improvement change is less than 1% in 1000 consecutive trials.

To compare performance levels, the return of the proposed model was annualized. Annual return rates were calculated by computing the ratio of the current portfolio value to the initial portfolio value after subtracting transaction costs. We considered transaction costs (0.35%) equal to the actual value charged by the Korean stock exchange. The Sharpe ratio was also calculated to measure the risk-adjusted performance. The Sharpe ratio quantifies how well a portfolio return compensates investors for risks taken. A portfolio with a higher Sharpe ratio provides a higher expected return for the same level of risk. Thus, the risk-adjusted performance of portfolios can be compared based on their Sharpe ratios. The certificate of deposit (CD) interest rate was used as a risk-free rate or benchmark asset to measure the Sharpe ratio.

Tables 3 and 4 summarize the performance of the proposed model by investor type for the dataset. The terms *Equal*, *MC*, *MV* and *Sharpe* denote the performance of the equally-weighted portfolio, market capital-weighted portfolio, minimum variance-weighted portfolio and Sharpe-weighted portfolio, respectively. For the examined dataset, the annual return rate of the KOSPI 200 market portfolio buy–and–hold strategy was 2.90%, and the Sharpe ratio was −0.02.

Table 3 presents the results of the proposed model for portfolios constructed of stocks traded by institutional investors. These stocks are among the top cluster(s) that were selected in descending order of their mean TVP values for the training period. Thus, if the performance of a portfolio was the highest for a given training period, we employed that portfolio's parameters in the following testing period. As a result, Sharpe-weighed portfolios tended to perform better than other weighted portfolios (*Equal*, *MC*, and *MV*). The portfolio that performed best was the Sharpe-weighted, institutional-investor-based portfolio that employed a six-month training period. The annual return of this portfolio was 24.12%, and the Sharpe ratio was 0.70 (see Table 2). Furthermore, the performance of portfolios optimized via GA was superior to that of the KOSPI 200 market portfolio. Regarding the Sharpe ratio, the Sharpe-weighted portfolio significantly outperformed the other portfolios and the KOSPI 200 market portfolio. The outstanding Sharpe ratio of the Sharpe-weighted portfolio is attributable to the high excess return and low standard deviation of this portfolio. Therefore, the Sharpe ratio could be successfully maximized via the GA, i.e., the GA played an important role in improving the performance.

To verify the effectiveness of GA, conventional deterministic optimization method is used to optimize portfolios period from 09/01/2007 to 12/31/2008. Table 4 presents the performance of institutional-investor-based portfolio that employed a one-month training period.

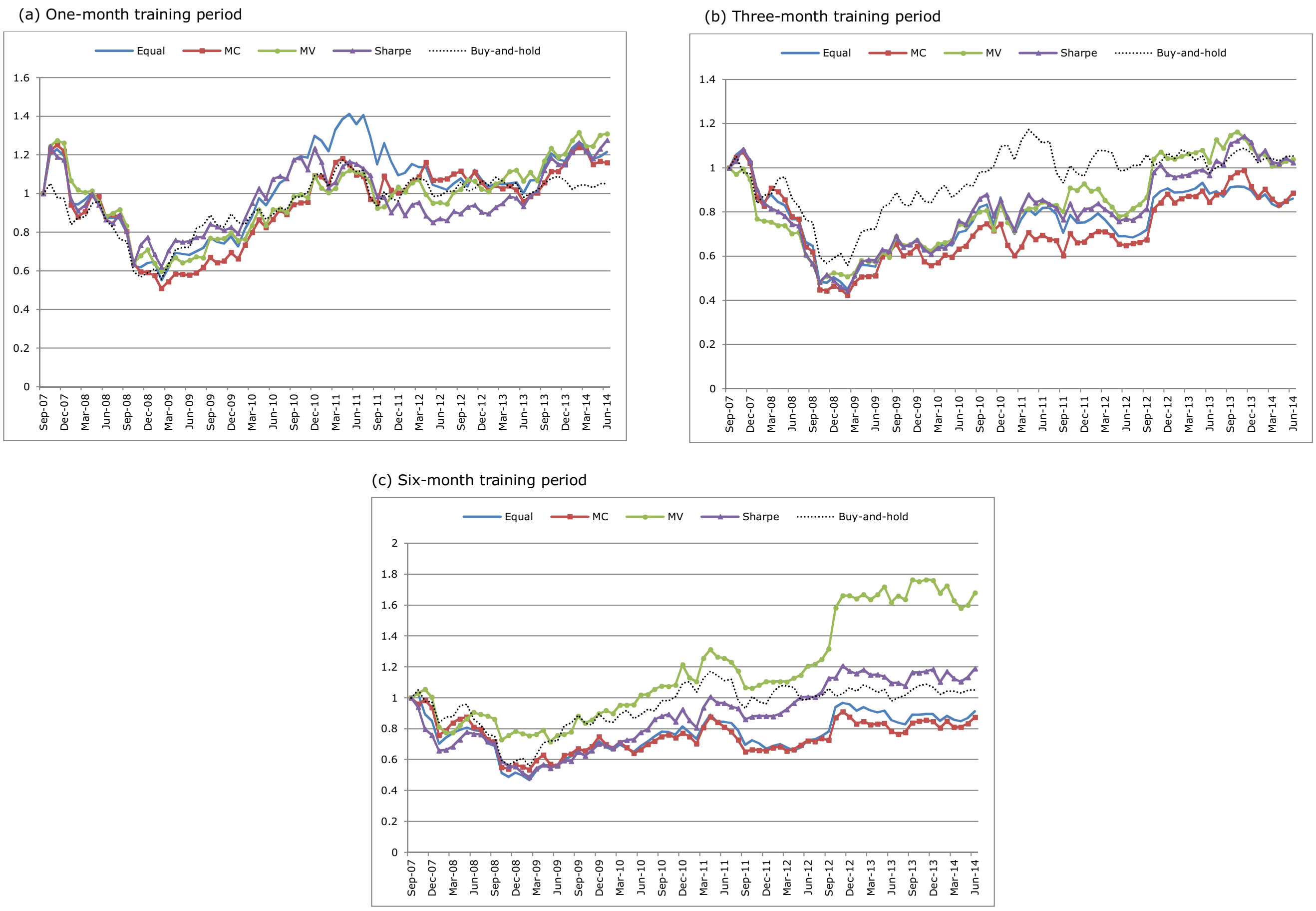

Fig. 8. Flow of cumulative returns of foreign investors. (Testing dataset: 09/01/2007-05/30/2014)

The annual return of minimum variance-weighted and Sharpe-weighted portfolio optimized by GA is −38.23% and −39.21%. In comparison, the annual return of portfolio optimized by conventional deterministic optimization method is −40.33% and −39.32%. The optimization results of 20 portfolios out of 256 portfolios in training period are different as optimization technique varies and such difference of optimization results make a different return, shaded gray in Table 4, in testing period. The most optimization results of GA show better results than conventional deterministic optimization method.

As discussed earlier. GA is used here because we are dealing with a relatively small number of stocks selected form the previous stage (small $l(c)$). Thus the GA successfully works to find the weight of optimal portfolio that maximizes the portfolio return while minimizing the portfolio risk. Consequently, the GA leads to superior results in capital allocation for our portfolio investment strategy.

Table 5 presents the results of the proposed model for portfolios constructed of stocks traded by foreign investors. These stocks were also among the top cluster(s) selected in descending order in terms of mean TVP value for the training period. In this case, the six-month training period minimum variance-weighted portfolio performed best. This minimum variance-weighted portfolio also outperformed other portfolios and the KOSPI 200 market portfolio, with the exception of portfolios constructed over a three-month training period. Foreign investors who implement the clustering-based portfolio optimization scheme for a minimum variance-weighted portfolio should thus use a six-month training period.

The monthly cumulative returns of all of the portfolios from 09/01/2007 to 05/30/2014 are plotted and compared with the KOSPI 200 buy-and-hold strategy in Figs. 7 and 8, which present the cumulative returns for the overall testing period by investor type (institutional and foreign investors) based on various training periods (i.e., one, three, or six months). The cumulative return of the buy-and–hold strategy was 1.09.

As Shown in Figs. 7 and 8, portfolios that were constructed of stocks of the selected cluster(s) and employed a six-month training period yielded better returns for both institutional and foreign investors. After the global economy was shaken by the United States subprime mortgage crisis in 2008, the cumulative returns of the proposed model (see Figs. 7 (c) and 8 (c)) decreased less than those of the KOSPI 200 buy-and-hold strategy. Therefore, the proposed model is valid and would likely generate results superior to those of the KOSPI 200 buy-and-hold approach. These results indicate that the clustering analysis was properly employed for portfolio construction and that both institutional and foreign investors tend to participate via long-term (six-month) investments in the KOSPI stock market.

5. Conclusion

This paper presents a clustering-based portfolio optimization scheme based on a GA of investor information (or investor type). A cluster analysis is conducted to construct a portfolio by selecting stocks which are more invested than the other stocks by institutional or foreign investors. After a portfolio is formed, a GA is employed to allocate portfolio stock weights as minimizing portfolio risk while maximizing the portfolio return in the mean–variance framework. Consequently, the proposed model affects portfolio management performance positively (i.e., the return, risk and Sharpe ratio).

When constructing a portfolio based on investor information, long-term (six-month) investor information performs well for institutional and foreign investors. The results also indicate that long-term information is superior to short-term (one-month) and medium-term (three-month) information in the Korean stock market. This study also examines how different weighting strategies using a GA affect portfolio performance. The results of the proposed model, in which a portfolio is constructed of stocks of the best selected cluster(s) and optimized weights (maximized Sharpe ratio), indicate that institutional investors perform best. However, the minimum variance-weighted portfolio strategy is suitable for foreign investors. It must be noted that in the Korean stock market, institutional investors anticipate higher returns and lower risk, whereas foreign investors only consider portfolio risk.

To summarize, investment strategy based on investor information of individual stock is developed by using *k*-means clustering and GA. The results of this study indicate that proposed model improves the performance of stock selection and capital allocation for portfolio investment strategy. The GAs can be used to increase portfolio performance by allocating portfolio stock weights and that *k*-means clustering based on investor information can lead to better performance by selecting stocks in the Korean stock market. And this study also shows that investor behaviors and patterns can be used to make the investment strategy. Each type of investor constructs the portfolio according to their risk-return profile on the mean–variance framework and imitating portfolio strategies of certain type of investor can enhance the portfolio performance.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2014R1A2A1A11052670).

References

[1] M.J. Brennan, H.H. Cao, International portfolio flows, J. Finance 52 (1997) 1851–1880.
[2] M. Grinblatt, M. Keloharju, The investment behavior and performance of various investor types: a study of Finland's unique data set, J. Financ. Econ. 55 (2000) 43–67.
[3] K.A. Froot, T. Ramadorai, The Information Content of International Portfolio Flows. Working paper, Harvard University, 2001.
[4] A. Kamesaka, J.R. Nofsinger, H. Kawakita, Investment patterns and performance of investor groups in Japan, Pacific–Basin Finance J. 11 (2003) 1–22.
[5] B.M. Barber, Y.T. Lee, Y.J. Liu, T. Odean, Who Gains from Trade? Evidence from Taiwan, Working Paper, University of California, and National Chengchi University, 2004.
[6] T. Dvorak, Do domestic investors have an information advantage? Evidence from Indonesia, J. Finance 60 (2005) 817–839.
[7] H. Choe, B.C. Kho, R.M. Stulz, Do domestic investors have an edge? The trading experience of foreign investors in Korean, Rev. Financ. Stud. 18 (2005) 795–829.
[8] K.H. Bae, T. Yamada, K. Ito, How do individual institutional, and foreign investors win and lose in equity trades? Evidence from Japan, Int. Rev. Financ. 6 (2006) 129–155.
[9] S. Agarwal, S. Faircloth, C. Liu, S.G. Rhee, Why do foreign investors underperform domestic investors in trading activities? Evidence from Indonesia, J. Financ. Mark. 12 (2009) 32–53.
[10] C. Wang, S. Chin, Profitability of return and volume-based investment strategies in China's stock market, Pacific-Basin Finance. J. 12 (2004) 541–564.
[11] H. Markowitz, Portfolio selection, J. Finance 7 (1952) 77–91.
[12] H. Markowitz, Portfolio Selection: Efficient Diversification of Investments, Wiley, New York, 1959.
[13] S.R. Nanda, B. Mahanty, M.K. Tiwari, Clustering Indian stock market data for portfolio management, Expert. Syst. Appl. 37 (2010) 8793–8798.
[14] M. El hachloufi, Z. Guennoun, F. Hamza, Stocks portfolio optimization using classification and genetic algorithms, Appl. Math. Sci. 6 (2012) 4673–4684.
[15] J.L. Maginn, D.L. Tuttle, D.W. McLeavey, J.E. Pinto, Managing Investment Portfolio: A Dynamic Process, Wiley, Hoboken, NJ, 2007.
[16] B. Malkiel, A Random Walk down Wall Street, sixth ed., Norton, New York, 1973.
[17] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, London UK, 1979.
[18] E.R. Hruschka, N.F.F. Ebecken, A genetic algorithm for cluster analysis, Intell. Data. Anal 7 (2003) 15–25.
[19] V. Tola, F. Lillo, M. Gallegati, R.N. Mantegna, Cluster analysis for portfolio optimization, J. Econ. Dyn. Control. 32 (2008) 235–258.
[20] L.-A. Yu, S.-Y. Wang, Kernel principal component clustering methodology for stock categorization, Syst. Eng.-Theory Pract. 29 (12) (2009) 1–8.
[21] Z.-H. Zhou, W.-N. Chen, Z.-Y. Zhang, Application of cluster analysis in stock investment, J. Chongqing Univ. 25 (7) (2002) 122–126.
[22] J.H. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, 1975.
[23] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison–Wesley, Boston, 1989.
[24] M. Srinivas, L. Patnaik, Genetic algorithms: a survey, Computer 27 (1994) 17–26.
[25] M. Mitchell, An Introduction to Genetic Algorithms, MIT Press, Cambridge MA, 1996.
[26] T. Bäck, D.B. Fogel, Z. Michalewicz, Evolutionary Computation 1: Basic Algorithms and Operators, Institute of Physiscs Publishing (IOP), Bristol and Philadelphia, 2000.
[27] T. Bäck, D.B. Fogel, Z. Michalewicz, Evolutionary Computation 2: Advanced Algorithms and Operators, Institute of Physiscs Publishing (IOP), Bristol and Philadelphia, 2000.
[28] A. El Imrani, A. Bouroumi, M. Limouri, A. Essaïd, A coevolutionary genetic algorithm using fuzzy clustering, Intell. Data. Anal. 4 (2000) 183–193.
[29] K.A. De Jong, Evolutionary Computation: A Unified Approach, MIT Press, Cambridge, MA, 2006.
[30] H. Adeli, S.L. Hung, Machine Learning: Neural Networks, Genetic Algorithms, and Fuzzy Systems, Wiley, New York, 1995.
[31] K.J. Oh, T.Y. Kim, S. Min, Using genetic algorithm to support portfolio optimization for index fund management, Expert. Syst. Appl. 28 (2005) 371–379.
[32] K.J. Oh, T.Y. Kim, S.H. Min, H.Y. Lee, Portfolio algorithm based on portfolio beta using genetic algorithm, Expert. Syst. Appl. 30 (2006) 527–534.
[33] R.A. Rivera, M.V. Rendon, J.J.R. Ortiz, Genetic algorithms and Darwinian approaches in financial applications: a survey, Expert. Syst. Appl. 42 (2015) 7684–7697.
[34] R.J. Bauer, Genetic Algorithms and Investment Strategies, Wiley, New York, 1994.
[35] A.M. Colin, Genetic algorithms for financial modeling, in: G.J. Deboeck (Ed.), Trading on the Edge: Neural, Genetic and Fuzzy Systems for Chaotic Financial Markets, Wiley, New York, 1994, pp. 148–173.
[36] G.J. Deboeck, Using Gas to optimize a trading system, in: G.J. Deboeck (Ed.), Trading on the Edge: Neural, Genetic and Fuzzy Systems for Chaotic Financial Markets, Wiley, New York, 1994, pp. 174–188.
[37] Y. Xia, B. Liu, S. Wang, K.K. Lai, A model for portfolio selection with order of expected returns, Comput. Oper. Res. (2000) 409–422.
[38] Y. Orito, H. Yamamoto, G. Yamazaki, Index fund selections with genetic algorithms and heuristic classifications, Comput. Ind. Eng. 45 (1) (2003) 97–109.
[39] J.-S. Chen, J.-L. Hou, S.-M. Wu, Y.-W.C. -Chien, Constructing investment strategy portfolios by combination genetic algorithms, Expert. Syst. Appl. 36 (2009) 3824–3828.
[40] W.F. Sharpe, The sharpe ratio, J. Portfolio Manage. 21 (1994) 49–58.
[41] T.-C. Fu, C.-P. Chung, F.-L. Chung, Adopting genetic algorithms for technical analysis and portfolio management, Comput. Math. Appl. 66 (2013) 1743–1757.