IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 15, 2019, accepted September 28, 2019, date of publication October 23, 2019, date of current version November 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2949055

An Effective Approach for the Diverse Group Stock Portfolio Optimization Using Grouping Genetic Algorithm

CHUN-HAO CHEN[1], CHENG-YU LU[1], TZUNG-PEI HONG[2,3], JERRY CHUN-WEI LIN[4], AND MATTEO GAETA[5], (Senior Member, IEEE)

[1]Department of Computer Science and Information Engineering, Tamkang University, Taipei 251, Taiwan
[2]Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung 811, Taiwan
[3]Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung 804, Taiwan
[4]Department of Computing, Mathematics, and Physics, Western Norway University of Applied Sciences, 5063 Bergen, Norway
[5]Dipartimento di Ingegneria dell'Informazione ed Elettrica e Matematica Applicata, University of Salerno, 84084 Fisciano, Italy

Corresponding author: Tzung-Pei Hong (tphong@nuk.edu.tw)

This work was supported by the Ministry of Science and Technology of China under Grants MOST 106-2221-E-032-049-MY2, MOST 107-2218-E-390-004 and MOST 108-2221-E-032-037.

ABSTRACT Finding useful portfolios that could be a portfolio of trading strategy or a stock portfolio from financial datasets is always an attractive research topic due to the nature of financial markets. Because investors always want an approach that can continually provide various portfolios, the issue of group stock portfolio optimization (GSPO) has been raised and the algorithms to obtain a group stock portfolio (GSP) have also been described in the past. A GSP divides the whole set of stocks into several stock groups, and the stocks in a group are exchangeable in investment. Thus, when investors are not satisfied with a suggested stock, they can select another stock from the same group to replace the original one. However, the industry diversity of stocks within a group is not regarded in the existing literatures. In this paper, an algorithm for dealing with the diverse group stock portfolio optimization (DGSPO) is proposed to obtain a diverse group stock portfolio (DGSP). The proposed algorithm is based on the group genetic algorithm with the chromosome representation and the fitness function designed for the purpose of finding a good DGSP. Especially, a factor called group diversity is designed to diversify stocks from different industries and is considered in the fitness evaluation. Another factor considers cash dividend is also applied to keep companies with good quality in the portfolio for increasing the profit of a DGSP. Two real financial datasets are used in the experiments to verify the effectiveness of the proposed approach.

INDEX TERMS Stock portfolio, group stock portfolio, grouping genetic algorithm, grouping problem, group diversity.

I. INTRODUCTION

Portfolio optimization is a very important research topic in investment. In a portfolio, various assets can be considered, e.g., stocks, futures, and options [2], [25], [32], [36], [39]. Given a set of assets, effectively deriving appropriate portfolios with good profits and low risks is crucial. In the past, the mean-variance (M-V) model is commonly utilized to deal with the portfolio optimization [30], [31]. Given a set of assets and an expected return, the model can find the weights of the assets that can minimize risk. Since the obtained portfolios depended on the desired expected returns, the problem was transformed into an optimization problem and optimization techniques could be adopted to solve it. For example, some of them utilized genetic algorithms (GA) [12], [24] and some of them used multi-objective genetic algorithms (MOGA) [4], [26], [27], [29]. In addition, taking a fuzzy set into consideration, other algorithms were proposed for solving fuzzy portfolio optimization problems [1], [5], [28]. Furthermore, hybrid approaches have also been designed for optimizing portfolios [15], [21], [22]. For instance, Elhachloufi et al. used classification and GA for stock portfolio optimization [9]. Gupta et al. employed support vector machines and GA for asset portfolio

The associate editor coordinating the review of this manuscript and approving it for publication was Philippe Fournier-Viger.

TABLE 1. The two GSPs A and B.

GSP *A* (Low diversity)		GSP *B* (High diversity)	
G_1:	{1338, 2206, 2227}	G_1:	{1338, 2317, 2325}
G_2:	{2303, 2311, 2325}	G_2:	{2303, 2206, 2354}
G_3:	{2312, 2317, 2354}	G_3:	{2312, 2311, 2227}

TABLE 2. Stock information.

Date / Company	2013/01/02	2013/12/31	2014/01/02	2014/12/31
		Automobile		
2227	77.2	103	105	93.8
1338	18	48.85	47.9	28.5
2206	232	422	432.5	333
		Semiconductor		
2303	11.8	12.35	12.3	14.75
2311	26.15	27.7	27.7	38.1
2325	31.1	35.6	35.65	47.95
		Other Electronic		
2312	6.52	10.95	11.05	14.7
2317	88.7	80.1	80.4	87.9
2354	90.9	69.6	70.2	85.3

optimization [14]. Considering financial and ethical considerations, hybrid optimization models were introduced for portfolio selection in [15].

In the past, a GA-based approach which considered an investor's objective and subjective requests has been proposed to derive stock portfolios [6]. Two sets of criteria, subjective and objective, were designed to evaluate the goodness of a chromosome, which represented a candidate stock portfolio. The subjective criteria included the investment capital penalty (ICP) and the portfolio penalty (PP), and the objective criteria included return on investment (ROI) and value at risk (VaR). For a derived portfolio, investors sometimes don't like a specific stock due to some personal reasons and want to replace it. It will be very time-consuming if the undesired stock is removed and the program is re-run to get another portfolio to investors. Another approach was then designed to divide the stocks into groups and find a group stock portfolio (GSP) using the grouping genetic algorithm (GGA) [9]. In a GSP, the stocks in the same group can be substituted to each other. In other words, investors may thus have high flexibility in choosing stocks according to the suggested GSP. Thus, when investors are not satisfied with a suggested stock, they can select a substitute stock from the same group to replace the original one. Because the stock group is considered, group balance and portfolio satisfaction are employed to design the fitness function to evaluate each individual. Finding a GSP is a kind of grouping problems [17], [18].

However, the diversity of groups which is an important property of the grouping problem [20], [37] is not considered in the previous approach [9]. To describe the importance of diversity of groups, two GSPs *A* and *B* show in Table 1 are used to explain it. Stock information, including industries and stock prices of stocks, is shown in Table 2.

Table 1 shows that each of the given GSPs has three groups and every group has three stocks. From Table 2, we can see that the three stocks, 1338, 2206, and 2227, in G_1 of GSP A belong to automobile, and stocks in G_2 and G_3 belong to semiconductor and other electronic. In this case, we say that the diversity of the group in GSP A is low because stocks in each stock group have the same industries, and the problem will arise when using the real datasets from 2013/01/02 to 2013/12/31 and from 2014/01/02 to 2014/12/31 for training and testing. From Table 1, we can observe that although the returns of stocks belonging to automobile are positive in the training data, they are negative in the testing data since automobile worsens in 2014. For instance, the returns of stock 2206 are 0.82 and −0.23 on the training and testing datasets. In other words, if the diversity of a group in a GSP is low, investors do not have opportunities to avoid high-risk stocks even though other stocks in the same group can be replaced. On the contrary, if the diversity of a group in a GSP is high, investors can easily replace a suggested stock by a stock from a different industry in the same group. Take the GSP B as an example. When investors think the prospects of automobile are bad, they then can select stocks 2317, 2303 and 2312 from groups G_1, G_2 and G_3 as a stock portfolio to avoid potential risks.

In this paper, we thus adopt the grouping genetic algorithm to develop an optimization mechanism for finding a good diverse group stock portfolio (DGSP). To encode a possible DGSP, the grouping, stock and stock portfolio parts are used as those in [9]. In chromosome evaluation, not only the existing factors, including portfolio satisfaction, group balance, price balance and unit balance, are used to get a good DGSP, but also two new factors, the stability and diversity factors, are designed to increase return and ability to avoid risk. The stability factor is developed based on the cash dividends of stocks. The diversity factor is created to measure the diversity of stock groups. Using these factors, two fitness functions are presented and employed for chromosome evaluation. The three genetic operations are used to generate offspring, including crossover, mutation, and inversion. Finally, two real financial datasets consisting of 30 and 31 stocks were verified to show the effectiveness of the proposed approach.

The rest of this paper is organized as follows. The background knowledge and problem definition is given in Section 3. The related work is described in Section 4. The elements of the proposed approach are stated in details in Section 4. The proposed algorithm for optimizing a DGSP is stated in Section 5. Extensive experiments on the two real datasets are discussed in Section 6. Finally, conclusions and future work are given in Section 7.

II. BACKGROUND KNOWLEDGE AND PROBLEM DEFINITION

In this section, the maximally diverse grouping problem and the grouping genetic algorithm are introduced in Sections II.A and II.B, respectively. The definition of the DGSP optimization problem is given in Section II.C.

A. MAXIMALLY DIVERSE GROUPING PROBLEM

Given n objects, the grouping problem is to divide them into K clusters with the condition that the group sizes are as equal

as possible [17], [18]. A variant of the grouping problem is the maximally diverse grouping problem (MDGP), in which the objects in a group have the maximal diversity [20], [37]. Mathematically, the MDGP can be represented below [37]:

$$\max \sum_{g=1}^{G} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} x_{ig} x_{jg},$$

$$\text{subject to } \sum_{g=1}^{G} x_{ig} = 1, \quad i = 1, 2, \ldots n$$

$$a_g \leq \sum_{i=1}^{n} x_{ig} \leq b_g, \quad g = 1, 2, \ldots, G$$

$$x_{ig} \in \{0, 1\}, \quad i = 1, 2, \ldots, n, \ g = 1, 2, \ldots, G,$$

where d_{ij} denotes the difference between two objects i and j. Besides, if stock i is in group g, then x_{ig} is 1; otherwise, x_{ig} is 0. The two symbols, a_g and b_g, represent the minimum and maximum group sizes. Thus, the diversity of a group can be measured as the total of the distance between each pair of objects in a group. Because MDGP is NP–hard, some heuristic approaches have consequently been developed [3], [19], [34]. Examples include the variable neighborhood search [3], the hybrid genetic algorithm [19], and the artificial bee colony algorithm [34] among others.

B. GROUPING GENETIC ALGORITHM

By enhancing the genetic algorithms (GAs) which can provide an almost optimal solution within a limited time, the grouping genetic algorithm (GGA) is presented for solving the grouping problem which is attempted to divide elements into groups with a cost function [17]. Using the GGA for solving the grouping problems, Brown et al. also indicated that the GGA was better than the GAs on various datasets [6].

In the following, the main differences of GA and GGA that are the encoding schema and genetic operations are introduced. A solution in the GGA is encoded by the grouping and object parts. For instance, a possible chromosome could be "ACDBBC: ABCD". Before the semicolon, the string "ACDBBC" is the object part, and after the semicolon, the string "ABCD" is the group part. From the two parts, we can know that there are six objects and four groups. The chromosome shows that six objects should be divided into four groups. In this example, object o_1, objects o_4 and o_5, objects o_2 and o_6, and object o_3 belong to group 'A', 'B', 'C', and 'D', respectively. As to the genetic operations in the GGA, three operations are used to form new offspring, including crossover, mutation and inversions. In crossover operation, it requires a chromosome arbitrarily selected as a base chromosome. Some groups from another chromosome are then added into the base chromosome. After that, the duplicate objects are eliminated from the newly formed chromosome. The mutation operator performed only on the object part. It reassigns an object into another group randomly. The last genetic operator is the inversion operator which is intended to assist the crossover operator in having distinct group combinations to exchange between two parents.

C. DGSP OPTIMIZATION PROBLEM

Before we define the DGSP optimization problem, definition of GSP is stated firstly.

Definition 1 (Group Stock Portfolio, or GSP): Given a set of stocks $S = \{s_1, s_2, \ldots, s_n\}$ *and a group number K, the GSP is a partition of S with K stock groups. That is,* $GSP = \{G_1, G_2, \ldots, G_n\}$, *where each* G_i *is a subsetof S,* $G_1 \cup G_2 \cup \ldots \cup G_K = S$, $G_i \neq \phi$ *and* $\forall i \neq j$, $G_i \cap G_j = \phi$.

Based on *Definition 1*, considering the diversity of stock group, the diverse group stock portfolio optimization problem can be defined as follows.

Definition 2. (Diverse Group Stock Portfolio Optimization, or DGSPO): Given a set of stocks $S = \{s_1, s_2, \ldots, s_n\}$ *with related information, including stock price series, cash dividends, industry type, and a group number K, the DGSPO problem is to optimize the weights of stock groups and stocks should belong to which groups of a DGSP so that the conditions can be reached: (1) The return and risk of stock portfolios that can be maximized and minimized; (2) The diversity of stock group can be maximized.*

To get a good DGSP, other factors could also be used. For example, to make sure various stock portfolios can be provided by a DGSP, stock groups have similar number of stocks should be promised. Thus, the group balance can be considered together with existing criteria to optimize a DGSP. Hence, the aim of this paper is attempted to design an algorithm for solving the DGSPO problem.

III. RELATED WORK

As portfolio selection is an optimization issue, metaheuristics for portfolio optimization have been introduced in [23], [35]. The approaches like evolutionary algorithms, the Tabu search and ant colony optimization approaches in portfolio optimization are very effective. Below are briefly discussed some portfolio optimization techniques for the M-V model's parameter learning to acquire portfolios through evolutionary algorithms [1], [3], [4], [12], [24], [26]–[29].

Chang et al. proposed a GA-based method for portfolio optimization problems. They considered different risk measures, including semi-variance, mean absolute deviation and variance with skewness [12]. Hoklie et al. presented another evolutionary approach to solve the problem with the expected returns and risks of stocks [24]. Given a portfolio, the presented approach encoded weights of stocks into a chromosome. According to the weights, each chromosome is evaluated by the ratio of the expected return and risk of the portfolio, where the expected return of a stock is the mean profit during a certain period and the downside value of the variance of a stock is used as identified risk. Taking fuzzy theory into consideration, some approaches were also presented to optimize fuzzy portfolios using GA [1], [28]. For instance, considering portfolio liquidity and fuzzy theorem, Barak et al. designed an approach for obtaining a

portfolio using GA based on the presented fuzzy portfolio mean-variance-skewness model with the cardinality constraint [1]. In that model, trapezoidal fuzzy membership functions were adopted to represent the return of an asset, and the fuzzy credibility theory was used to derive the turnover rate of an asset. In the presented approach, the asset numbers and proportions of assets are encoded into the chromosome. The fuzzy variables that are used to represent return of assets are employed to calculate fitness values of chromosomes.

For multi-objective genetic algorithms (MOGA), some approaches were designed for portfolio optimization [4], [26], [27], [29]. For example, Li et al. designed a multi-objective genetic algorithm to select portfolios with fuzzy random returns [29]. Three criteria including return, risk and liquidity were investigated. They proposed the constrained multi-objective portfolio selection model to obtain a compromised portfolio strategy. Chromosomes that satisfied the constraints of a problem were randomly generated. Each chromosome was then evaluated using the regret values. Genetic operators were used to discover new chromosomes. After evolution, the best chromosome is considered as a compromise solution. Lwin et al. also proposed a multi-objective portfolio optimization approach, called MODEwAwL, with four constraints [27]. The encoding scheme consists of two vectors that are utilized to represent whether assets are included in the portfolio and the proportions of capital invested in assets. The maximum return and minimum risk are used as two objective functions to evaluate fitness of chromosomes to obtain non-dominated solutions.

There are also hybrid methods for portfolio optimization [5], [8], [15], [21], [22]. For example, Bermúdez et al. presented a genetic algorithm to deal with a fuzzy multi-objective portfolio selection problem for selecting efficient portfolios [5]. In that approach, fuzzy quantities were used to represent the uncertainty of the returns, and investor's aversion to risk is represented using a downside risk function. Since the goal is to find efficient portfolios, each chromosome indicates a possible portfolio in the population. According to the expected returns on and risks of chromosomes, a frontier of solutions is built. The evolution process is repeated until a good upper-boundary frontier is found. Chen et al. combined domain-driven mining framework and proposed an algorithm for finding an actionable stock portfolio according to the given subjective and objective criteria using GA [8]. Hachloufi et al. presented an approach called MinVaRMaxVaL for the selection of optimal actions portfolio using GA and VaR [15]. The goal of that algorithm is to improve the optimal choice in the sense of having the highest return and low risk. Gupta et al. then adopted the support vector machines and the real-coded genetic algorithm for asset portfolio optimization [21]. It first used the support vector machines to classify assets into less risky, high-yield and liquid assets. In accordance with user preferences, desired portfolios were then found from the classes using GA. Gupta et al. then proposed a three-stage framework to select portfolios by considering ethical and financial factors [22]. The ethical performance and financial quality scores of each asset are first calculated using the analytical hierarchy process technique and fuzzy decision making. They also designed three hybrid models to find suitable portfolios.

For group stock portfolio (GSP) optimization, Chen et al. extended the grouping genetic algorithm [9] to optimize a GSP. The representation of a chromosome includes three parts that are the grouping part, stock part and stock portfolio part. Each chromosome is evaluated by group balance and portfolio satisfaction. When the optimal GSP is derived at the end of the evolution process, different stock portfolios can be generated from the groups, with one stock chosen from one group. The approach could provide investors more flexibility in decision. To increase the similarity of stock groups in a GSP, a series-based GSP optimization algorithm was proposed in [11]. In that approach, the stock price series are firstly transformed into symbolic series. Then, the distance of the symbolic series in groups is used to evaluate the similarity of stock groups and designed as a part of factors in the fitness function. To speed up the evolution process, a map-reduce-based approach to optimize a GSP was presented in [7]. The chromosome representation contains a mapper number, a group number, a stock part and a portfolio part. The mapper number is utilized to divide chromosomes in a population into subsets and deliver to respective mappers. The fitness evaluation and genetic operations are executed in the reducers. The optimization process is repeated until reaching the terminal conditions. In addition, by using island-based genetic algorithms, the island-based group stock portfolio optimization approach which consisted of the evolution and migration phases was designed [10]. In the evolution phase, the grouping genetic algorithm is used to optimize GSPs. Then, in the migration phase, the best chromosomes in islands are selected and updated randomly to other islands to enhance its searching ability. Hence, the main difference between the proposed approach and the mentioned approaches is the diversity of stock group is considered in the designed approach.

IV. ELEMENTS OF THE PROPOSED ALGORITHM

The four main components of the proposed approach are presented in this section. They are encoding scheme in Section IV.A, initial population in Section IV.B, genetic operations in Section IV.C, and fitness and selection in Section IV.D.

A. ENCODING SCHEME

Assume there are n stocks to be chosen from. They are represented as $S = \{s_1, s_2, \ldots, s_n\}$. Given a parameter K for the group number, the proposed approach here is to obtain K stock groups as a group stock portfolio with subjective and objective criteria. Fig. 1 shows the encoding scheme of a chromosome C_q as in [9].

In Fig. 1, there are three parts in the chromosome representation: grouping, stock, and stock portfolio. In the stock part, the stocks belonging to the same group G_i are expected

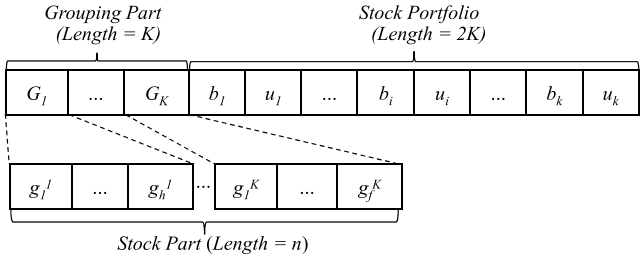


FIGURE 1. Encoding scheme for a chromosome C_q.

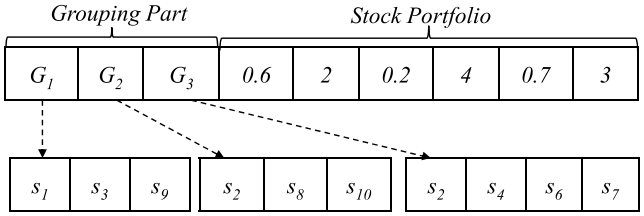


FIGURE 2. An example of an encoding scheme.

to have similar characteristics. Only one stock in each group is chosen. All the chosen stocks thus form a stock portfolio. Therefore, a stock portfolio will include K stocks. The second part is the stock portfolio part, in which each group G_i includes two genes b_i and u_i. b_i is a real value to decide whether a stock in group G_i is selected and purchased. If the value of b_i is larger than or equal to a threshold λ, then a stock in G_i is selected and purchased; otherwise, no stock from G_i is selected. The other gene, u_i, represents the amount to be purchased for s_i. One unit purchased means one thousand shares. Hence, the lengths of the three parts in a chromosome are K, $2K$ and n, respectively. Below, the encoding scheme is illustrated by a simple example.

Example 1: Suppose that ten stocks, $s_1, s_2, s_3, \ldots, s_{10}$, are considered in an investment. If the number of group K is 3, a chromosome C_1 is encoded as shown in Fig. 2.

Fig. 2 shows that the grouping part consists of three groups, G_1, G_2 and G_3, and numbers of stocks are 3, 3 and 4, respectively. Since the values of b_1 and b_3 are larger than 0.5, two stocks, one from G_1 and the other from G_3, are chosen to compose a candidate portfolio. Besides, their purchased units are 2 and 3. In this example, it expresses that twelve stock portfolios ($= 3 \times 4$) can be provided by the chromosome.

B. INITIAL POPULATION

In a genetic process, the final results will depend on the initial population. In this paper, we use the cash dividend yields of stocks to generate initial chromosomes since they are reported as an efficient investment strategy [16], [40]. An example is first given below to describe what the cash dividend yield is. Assume there are two companies, Chunghwa Telecom (CHT) and Nan Ya Plastics Corporation (NPC). Table 3 shows the cash dividends per share, stock prices and cash dividend yields of the two companies in 2011 to 2014.

TABLE 3. The dividend data of CHT and NPC in 2011 to 2014.

	2011	2012	2013	2014
CASH DIVIDENDS OF CHT (PER SHARE)	5.46	5.35	4.53	4.86
STOCK PRICE OF CHT	100.00	94.5	93.10	94.00
CASH DIVIDEND YIELD OF CHT	5.46%	5.66%	4.86%	5.71%
CASH DIVIDENDS OF NPC (PER SHARE)	2.10	0.30	1.90	2.30
STOCK PRICE OF NPC	60.10	56.00	68.9	65.5
CASH DIVIDEND YIELD OF NPC	3.49%	0.53%	2.75%	3.51%

TABLE 4. Ratio of the average cash dividend yield of each group over all groups.

G_1	G_2	...	G_i	...	G_{K-1}	G_K
$\frac{avgCD_1}{\sum_{a=1}^{K} avgCD_a}$	$\frac{avgCD_2}{\sum_{a=1}^{K} avgCD_a}$	...	$\frac{avgCD_i}{\sum_{a=1}^{K} avgCD_a}$	...	$\frac{avgCD_{K-1}}{\sum_{a=1}^{K} avgCD_a}$	$\frac{avgCD_K}{\sum_{a=1}^{K} avgCD_K}$

From Table 3, the cash dividends of CHT are NT$ 5.46, 5.35, 4.53 and 4.86, in 2011, 2012, 2013 and 2014. Their cash dividend yields are calculated by cash dividends over prices, which are 5.46%, 5.66%, 4.86% and 5.71%, respectively. Similarly, the cash dividend yields of NPC are 3.49%, 0.53%, 2.75% and 3.51%. Comparing the cash dividend yields of the two companies, CHT is thought of as more stable than NPC because the former has a better cash dividend yield than the latter.

Given n stocks, we may use existing clustering techniques, e.g., k-NN and k-means, to form K groups by their cash dividend yields. After the clusters are formed, the average cash dividend yield of stocks ($avgCD_i$) in each group G_i is calculated. The ratio of the value in one group over all the groups is then calculated as the probability for the group to be selected. Table 4 shows the results.

As a result, a group with a larger average cash dividend yield will have a larger probability for its stocks to be picked up. The strategy will thus generate a better initial population.

C. GENETIC OPERATIONS

Three genetic operations are used in the genetic process. They are crossover, mutation and inversion. The genetic operations are the same as those that were used in our previous approach [9]. Below, they are described briefly.

1) CROSSOVER

The two crossover operators are adopted here for gene exchange on the grouping and stock portfolio parts. For the grouping part, the one-point crossover operator acts on the grouping part as the GGA did [17]. For the part of stock portfolios, a chromosome *CA* is first chosen at random as a base chromosome, and then some groups from another chromosome *CB* are inserted into it to form *CA'*. At last, the redundant stocks and groups are removed from the newly

generated chromosome *CA'*. For the stock part, because the one-point crossover operator made on the grouping part will also change the stock part, the proposed approach does not apply a crossover on the stock part.

2) MUTATION

The one-point mutation operator is adopted here for mutating genes on the two parts of the stock and the stock portfolio. For the part of the stock, two groups, G_i and G_j, in a chromosome are randomly selected, with both their stock numbers larger than 1. A stock is then randomly selected from group G_i and moved into another chosen group G_j. For part of the stock portfolio, a gene is first selected at random. As mentioned before, there are two genes, b_i and u_i, in a group in the stock portfolio part. If the selected gene lies in b_i in the stock portfolio part, its value is generated and reassigned from [0, 0.5] to [0.5, 1] or [0.5, 1] to [0, 0.5]. When the selected gene lies in u_i, a random value is generated to replace the old one from the interval of [1, *maxUnit*].

3) INVERSION

The purpose of using the inversion operator is to allow the crossover operation to produce different group combinations to exchange between two parents. Different types of strategies may be utilized to achieve the goal. In this paper, the rearrangement operator is used. For example, originally, assume that there are two groups: G_1 has stocks s_1, s_3, s_9, and G_3 has stocks s_2, s_4, s_6, s_7. If G_1 and G_3 are exchanged, then stocks in the two groups are also exchanged. As a result, when the crossover operator is conducted, it increases the probability of getting various chromosomes.

D. FITNESS EVALUATION

The quality of a chromosome is measured by a fitness function. Parent chromosomes may also be randomly selected to mate based on their fitness values. Because the proposed algorithm is to optimize a diverse group stock portfolio, designing a sophisticated fitness function for evaluating chromosomes effectively is needed. Here, the enhanced fitness functions based on that used in [9] are designed to obtain a good GSP. The fitness function adopted in the previous approach [9] is shown as follows:

$$f(C_q) = GB(C_q)^{\alpha} * PS(C_q), \tag{1}$$

where $GB(C_q)$ denotes the group balance and $PS(C_q)$ denotes portfolio satisfaction. The former is employed to balance the numbers of stocks of groups in a chromosome, and the latter is used for evaluating the profit satisfaction and user's request satisfaction of a chromosome. The parameter α reflects the weight between the two factors. The group balance is defined and explained below:

$$GB(C_q) = \sum_{i=1}^{K} -\frac{|G_i|}{N} \log \frac{|G_i|}{N}, \tag{2}$$

where $|G_i|$ represents the size of the i-th group and K is the number of groups. A large group balance value is better. The portfolio satisfaction is defined and explained below (3):

$$PS(C_q) = \sum_{p=1}^{NC} subPS(SP_p)/NC, \tag{3}$$

where NC is the number of stock portfolios that can be generated from the chromosome C_q, and $SubPS(SP_p)$ is the p-th portfolio's portfolio satisfaction which can be calculated by formula (4):

$$subPS(SP_p) = \frac{ROI(SP_p)}{suitability(SP_p)}. \tag{4}$$

In Formula 4, *ROI(SPP)* and *suitability*(SP_P) are the profit and the suitability of the stock portfolio SP_p. $ROI(SP_P)$ is defined as follows:

$$ROI(SP_p) = \sum_{i=1}^{n} \left[(SP_s^{\{i\}} - SP_b^{\{i\}}) * u_i + Div^{(i)} * u_i + u_i * Risk_i \right], \tag{5}$$

where u_i is the purchased units of s_i, $SP_s^{(i)}$ and $SP_b^{(i)}$ are the sale price and purchase cost of s_i, and $Div^{(i)}$ and $Risk_i$ are cash dividend and risk of s_i. Note that the risk calculation is based on the historical simulation (HS) [31]. The *suitability* (SP_P) is stated in formula (6).

$$suitability(SP_P) = ICP(SP_P) + PP(SP_P)^{\beta}, \tag{6}$$

where ICP(SP_P) and PP(SP_P) are the investment capital penalty and the portfolio penalty of SP_p, and β is the parameter for balancing the influence of ICP and PP. ICP(SP_p) is intended to evaluate the degree of satisfaction of SP_P's investment capital to the maximum investment capital predefined, which is shown in formula (7):

$$ICP(SP_p) = \begin{cases} \dfrac{\max Inves}{Cap_p}, & if\ Cap_p \le \max Inves \\ \dfrac{Cap_p}{\max Inves}, & if\ \max Inves < Cap_p \end{cases} \tag{7}$$

where *maxInves* and Cap_P are the predefined maximum investment and investment capital of SP_P, respectively. $PP(SP_p)$ is used to assess the degree of satisfaction of the amount of stocks bought in SP_p over the predefined maximum amount of stocks bought, which is defined in formula (8):

$$PP(SP_p) = \begin{cases} \dfrac{numCom_p}{numCom}, & if\ numCom \le numCom_p \\ \dfrac{numCom}{numCom_p}, & if\ numCom_p < numCom \end{cases} \tag{8}$$

where $numCom_P$ and *numCom* are the number of purchased stocks of the stock portfolio SP_P and the predefined maximum number of purchased stocks, respectively. Hence, when a chromosome has a high portfolio satisfaction value, it means the portfolios formed from the chromosome could reach a good profit under the given criteria.

Because the cash dividend could be used to indicate that a company is stable in years when the company's cash dividends are similar, the stability factor is provided based on the cash dividend to decrease the effect of the suggested stocks that have high profits in the training stage but cause enormous losses in the test stage. Other literature also indicated that, for example, (1) Huxley reported that high-dividend yield stocks usually outperform those with small yields [16], and (2) You et al. indicated that a cash dividend yield portfolio is better than other kinds of portfolios on the Taiwan stock market [40]. In other words, the aim of the stability factor is attempted to prevent using stocks in the final DGSP with a big variance of cash dividend. The stability factor $SF(C_q)$ is shown in formula (9).

$$SF(SP_p) = 2 \times (1 + \max(NCD(s_1^p), \ldots, NCD(s_i^p), \ldots, NCD(s_m^p)), \quad (9)$$

where $NCD(s_i^p)$ is the normalized variance of cash dividends of stock s_i in stock portfolio SP_p. Assume that SP_p consists of m stocks and each stock has l cash dividends, $NCD(s_i^p)$ can be calculated by formula (10).

$$NCD(s_i^p) = \frac{varCD(s_i^p)}{h - thVarCD(S, h)}, \quad (10)$$

where $varCD(s_i^p)$ is the variance of cash dividends of stock s_i and *h-thVarCD(S, h)* is the h-th largest variance of cash dividends of the given stocks in the set S. When h is set to 1, it means that the influence of the stability factor on the fitness value is the smallest. On the contrary, if h is set to n, the influence will be the largest. If high fluctuation is not allowed for investors, it is suggested that the h should be set to a higher value. The $varCD(s_i^p)$ is calculated using formula (11).

$$varCD(s_i^p) = \frac{\sum_{b=1}^{l} (CD_b^{s_i^p} - \overline{CD}^{s_i^p})^2}{l - 1}, \quad (11)$$

where l is number of cash dividends of stock s_i, and $CD_b^{s_i}$ and $\overline{CD}^{s_i}$ are the *b-th* cash dividend and the average cash dividend of stock s_i. Hence, in this paper, the modified $m\_subPS(SP_p)$ is shown in formula (12).

$$\text{m\_}subPS(SP_p) = \frac{ROI(SP_p)}{suitability(SP_p) * SF(SP_p)}. \quad (12)$$

In addition, to avoid the unit u_i purchased for a group being too large or too small, the unit balance is presented to cause the units purchased to fall within the predefined range [*minPurchasedUnit, maxPurchasedUnit*]. The unit balance is given in formula (13).

$$UB(C_q) = \begin{cases} ubv_1, & \text{if } \sum_{i=1}^{k} U_i = K, \\ ubv_2, & \text{if } \sum_{i=1}^{k} U_i > 0, \\ 1, & otherwise, \end{cases} \quad (13)$$

where U_i represents whether the purchased unit u_i of group G_i is in the predefined range and K is number of groups. If the purchased unit is between *minPurchasedUnit* and *maxPurchasedUnit*, then U_i is 1. Otherwise, U_i is -1. When $UB(C_q)$ is ubv_1, the purchased units of all groups are inside the predefined range. If $UB(C_q)$ is ubv_2, it means some purchased units do not fall within in the predefined range. The value of ubv_2 should be smaller than ubv_1. In this paper, ubv_1 and ubv_2 are set at 1.4 and 1.15. Otherwise, $UB(C_q)$ is 1.

To make stocks in the same group have as similar a stock price as possible, the price balance is then presented, which is shown in formula (14).

$$PB(C_q) = Max(1, \sum_{i=1}^{k} \sum_{j=1}^{n} -\frac{|Sec_j|}{|G_i|} \log \frac{|Sec_j|}{|G_i|}), \quad (14)$$

where Sec_j is stock price section which is a stock price range defined by user, $|Sec_j|$ is number of stocks in j-th section and $|G_i|$ is number of stocks in group G_i. Based on $SF(C_q)$, $UB(C_q)$ and $PB(C_q)$, the enhanced fitness function $f(C_q)$ is defined in formula (15).

$$f(C_q) = \frac{GB(C_q)^{\alpha} * PS(C_q) * UB(C_q)}{PB(C_q)}. \quad (15)$$

Furthermore, the diversity factor is intended to boost the diversity of stocks in the same group in this paper in order to attain the objective of acquiring a DGSP. The diversity factor is defined as formula (11).

$$DF(C_q) = \frac{\sum_{i=1}^{K} D_i^q}{K}, \quad (16)$$

where K is the given number of groups, and D_i^q means the diversity value of the group G_i in chromosome C_q. D_i^q is calculated using formula (17):

$$D_i^q = \frac{\sum_{s_h, s_t \in G_i, and\ h \neq t} dissMatrix(s_h, s_t)}{|G_i|}, \quad (17)$$

where s_h and s_t are two stocks in group G_i, and *dissMatrix*(s_h, s_t) is used to check whether s_h and s_t are in the same category (industry). If s_h and s_t are in the same category, the value is zero. Otherwise, the value is one. Note that the dissimilarity matrix can be evaluated by various attributes. For example, the capital amount could be used to evaluate the company scale, or the debt asset ratio could be utilized to measure the financial leverage. Thus, instead of using the industries of stocks, a set of attributes, including the industries of stocks, the company scale, the debt asset ratio, etc., can be used to measure the diversity of stock groups. Based on the original fitness function stated in formula (1), by combining the stability factor and the diversity factor shown in formulas (9) and (16), the first fitness function $f_1(C_q)$ is given in formulas (18).

$$f_1(C_q) = GB(C_q)^{\alpha} * PS(C_q) * DF(C_q)^{\beta}, \quad (18)$$

where α and β are parameters used to reflect the influence of these factors. The second fitness function $f_2(C_q)$, according to

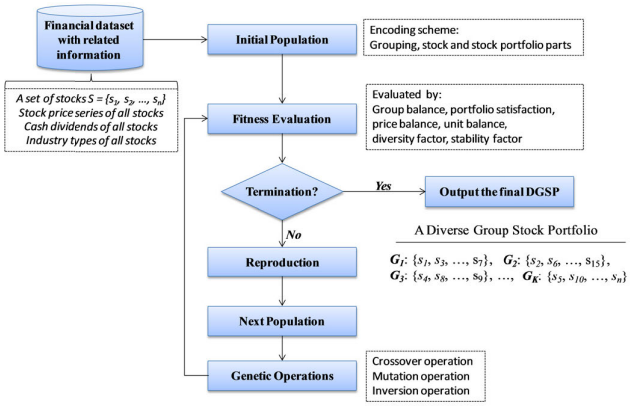


FIGURE 3. Framework of the DGSP optimization approach.

the enhanced fitness function and the diversity factor shown in formulas (15) and (16), is defined in formula (19).

$$f_2(C_q) = \frac{GB(C_q)^{\alpha} * PS(C_q) * UB(C_q) * DF(C_q)^{\beta}}{PB(C_q)}. \quad (19)$$

V. GGA-BASED OPTIMIZATION APPROACH FOR OBTAINING A DGSP

In this section, the framework of the proposed approach is stated in Section V.A. Details of the proposed algorithm is illustrated in Section.B.

A. FRAMEWORK OF PROPOSED APPROACH

The framework of the proposed approach is shown in Fig. 3.

Fig. 3 shows that the proposed approach first generates initial population using the financial dataset with related information, including a set of stocks, stock price series, cash dividends, and industry types of all stocks. For every chromosome in the population, the group, stock and stock portfolio parts are used to encode a possible DGSP. Then, the designed fitness function which composes of various criteria that are the group balance, portfolio satisfaction, price balance, unit balance, diversity factor, and stability factor is employed to evaluate the quality of every chromosome. If the termination conditions are not reached, the reproduction is executed to form next population, as well as the genetic operations that are crossover, mutation and inversion. Otherwise, the chromosome with the highest fitness value will be outputted as the final DGSP to provide investors making investment plans.

B. DETAILS OF PROPOSED ALGORITHM

The proposed algorithm for obtaining a diverse group stock portfolio by grouping genetic algorithms is described below.

The proposed GGA-based algorithm for diverse group stock portfolio optimization:

INPUT: A set of stocks $S = \{s_i \mid 1 \le i \le n\}$ with their cash dividends $CD = \{CD_i \mid 1 \le i \le n\}$ that can be used to calculate cash dividend yields $Y = \{y_i \mid 1 \le i \le n\}$, a predefined maximum investment capital *maxInves*, a predefined maximum number of purchased stocks *numCom*, a predefined maximum number of purchased units of a stock *maxUnit*, a number of groups K, parameters α, β and h, a crossover rate p_c, a mutation rate p_m, an inversion rate p_I, a population size *pSize*, and a number of generations *Gene*.

OUTPUT: A diverse group stock portfolio *DGSP*.

STEP 1: Form an initial population with *pSize* chromosomes using the procedure described in Section IV.B.

STEP 2: For each chromosome C_q, calculate its fitness value by the following sub-steps.

Sub-step 2.1: Find the value of portfolio satisfaction of C_q as follows.

Sub-step 2.1.1: Use the grouping part in C_q to generate possible stock portfolios and denote them as $SP = \{SP_i \mid 1 < i < |G_1| \times |G_2| \times ... \times |G_K|\}$. Each SP_i is a combination generated from the grouping part.

Sub-step 2.1.2: Evaluate the profit of each SP_i by formula (5).

Sub-step 2.1.3: Evaluate the suitability of each SP_i by formula (6).

Sub-step 2.1.4: Evaluate the stability factor of SP_i by formula (9).

Sub-step 2.1.5: Set the value of the portfolio satisfaction for each SP_i according to formula (12).

Sub-step 2.1.6: Evaluate the portfolio satisfaction for each chromosome C_q by formula (3).

Sub-step 2.2: Measure the group balance of C_q by formula (2).

Sub-step 2.3: Calculate the unit balance of C_q by formula (13) when the fitness function f_2 is selected to evaluate a chromosome.

Sub-step 2.4: Measure the price balance of C_q by formula (14) when the fitness function f_2 is selected to evaluate a chromosome.

Sub-step 2.5: Measure the diversity factor of C_q by formula (16).

Sub-step 2.6: Set the fitness value of C_q according to the selected fitness function (formulas (18) or (19)).

Step 3: Conduct the selection operation on the population to form the next population.

STEP 4: Conduct the crossover operation on the population.

STEP 5: Conduct the mutation operation on the population.

STEP 6: Conduct the inversion operation on the population.

STEP 7: Repeat Steps 2 to 6 until the stop criterion is satisfied.

STEP 8: Output the chromosome with the highest fitness value as the optimized DGSP.

Note that in Step 3, either the elitist or the roulette wheel selection strategy can be adopted as the selection mechanism. In this paper, the elitist selection strategy is used to form the next population. Because chromosome evolution may time-consuming with a large number of stocks, to speed up the evolution process, the island-based parallel grouping genetic algorithms [10], [13] or parallel genetic algorithms based on Hadoop MapReduce [7], [14] are suggested to solve

TABLE 5. Parameter setting.

Parameter	*Value*	*Parameter*	*Value*
pSize	50	*numCom*	4
p_c	0.8	*maxInves*	1 million
p_m	0.03	*maxUnit*	40
p_i	0.6	α	3
Gene	100	β	2
K	6		

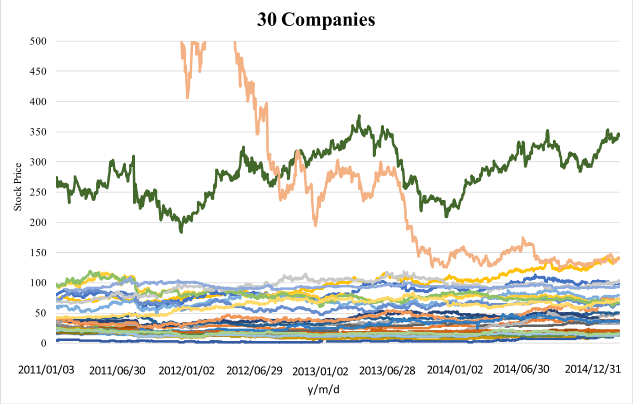


FIGURE 4. The first dataset in the experiments.

the problem. In addition, a data processing procedure could also be designed and utilized for obtaining a small subset from the hundreds of stocks in the market. To keep valuable stocks and reduce the stock size, financial indicators can be employed to reach the goal, including return on equity (ROE), the price-to-earnings (P/E) ratio, earnings per share (EPS), etc.

VI. EXPERIMENTAL RESULTS

Experiments were made to verify the proposed algorithm. In Section VI.A, the experimental datasets are stated. Then, the derived DGSPs are described in Section VI.B. The comparing proposed approach with existing approaches is given in Section VI.C. The effectiveness of the stability factor is shown in Section VI.D. The parameter setting in the experiments is listed in Table 5.

A. DATA DESCRIPTIONS

Two real datasets that have 30 and 31 stocks are used in the experiments. The first dataset contains 30 stocks that were obtained from the Taiwan Stock Exchange (TSE) from 2011/01/01 to 2014/12/31. They were chosen from six categories, namely, semiconductor, finance, computer and peripheral equipment, plastic, optoelectronic, and communication network. The first dataset are shown in Fig. 4.

From Fig. 4, we can observe one phenomenon whereby most of the stock price series are within 0 to 100 and some of them are larger than 250. The attributes of the dataset are the stock prices, the cash dividends, risk and industries of stocks. The variance of cash dividend of each stock can be calculated using its cash dividends. The risk value can be derived by

TABLE 6. Related information about the first dataset.

Category	Semiconductor	Finance	Computer and Peripheral equipment	plastic	Optoelectronic	Communication network
Number of Stocks	5	5	5	5	5	5
Avg. Buying Prices	53.38	25.08	89.79	57.16	9.25	81.82
Avg. Selling Prices	68.1	27.67	111.8	52.28	13.18	86.02
Avg. Cash Dividends	3.57	1.34	6.014	1.13	0.47	3.02
Avg. Risk Values	-3.68	-0.92	-9.04	-4.15	-1.45	-8.51
Stock Symbol	2303 2311 2325 2330 2451	2801 2809 2881 2851 2891	2301 2357 2377 2382 4938	1301 1303 1321 1325 1326	2340 2409 2426 2438 2460	2412 2419 2498 3045 4904

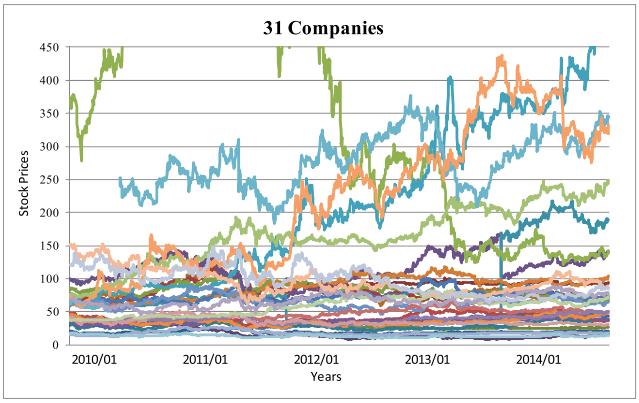


FIGURE 5. The second dataset in the experiments.

historical data simulation. Related information about the first dataset is shown in Table 6.

Table 6 shows that the number of stocks in each industry is five. According to the average cash dividends of stocks of the six industries, semiconductor is similar to communication network, and the financial industry is similar to the plastic industry as well. We can also observe that semiconductor and plastic are similar in terms of the average buying and selling prices. Computer and peripheral equipment and communication network have higher risk than other industries. Computer and peripheral equipment has the largest cash dividend.

The second dataset collects data from 2010/01/01 to 2014/12/31 from the TSE. It has 31 stocks, and its attributes are the same as the first dataset. The 31 stock price series are depicted in Fig. 5.

From Fig. 5, it can be observed that most of the stock prices are between 0 to 100, some of them are between 100 to 400, and only a few of them are larger than 400. Comparing Fig. 4 and Fig. 5, we can understand that the variance of the stocks in Fig. 5 is larger than that in Fig. 4.

TABLE 7. Initial and final best DGSPs using fitness function f_1.

Initial diverse group stock portfolio		
	Grouping and stock parts	Stock portfolio
G_1:	{2311, 2451, 2357, 2382, 1325, 2409, 2426, 2438}	0.87, 19, 0.49, 15, 0.32, 22, 0.74, 31, 0.53, 34, 0.15, 13
G_2:	{ 2330, 2801, 2891, 1303, 1321}	Fitness Value=2956.47
G_3:	{ 2851, 4938, 2460, 4904, 2412, 2419}	PortfolioSatisfaction=9.97
G_4:	{2809, 1301, 1326}	Group Balance=3.00
G_5:	{2325, 2881, 2377}	Diversity=3.314
G_6:	{2303, 2301, 2340, 2498, 3045}	
The derived diverse group stock portfolio		
	Grouping and stock parts	Stock Portfolio
G_1:	{4938, 2460, 4904, 2412, 1326}	0.94, 5, 0.54, 5, 0.14, 10, 0.20, 30, 0.53, 39, 0.89, 5
G_2:	{2330, 2801, 1303, 1321, 2409, 2419}	Fitness Value=17101.48
G_3:	{ 2311, 2451, 2891, 2357, 2340}	PortfolioSatisfaction=40.88
G_4:	{2303, 2301, 2498, 3045, 2809, 1301}	Group Balance=3.13
G_5:	{2325, 2881, 2377}	Diversity=3.689
G_6:	{2382, 1325, 2426, 2438, 2851}	

Next, we will compare the proposed approach and the previous approach [9]. We called the previous approach with its original fitness function as "Previous Approach (O)". The "Previous Approach (M)" means the stability factor is taken into consideration in the previous approach. Then, instead of the original fitness function, the enhanced fitness function, which is formula (15) without using the stability factor, is used in the previous approach to obtain GSP and named as "Previous Approach (E)". As to the proposed approach, we run two different versions based on the two fitness functions f_1 and f_2.

B. RESULTS AND ANALYSIS OF THE DERIVED DGSPs

The derived DGSP is given and analyzed in this section. Using a one-year dataset (2013) as training data, Table 7 displays the initial best DGSP and the final DGSP using the proposed algorithm with the f_1 fitness function after 100 generations.

Table 7 shows that the number of stocks in the derived DGSP, which are 5, 6, 5, 6, 3 and 5, is better than that in the initial DGSP, which are 8, 5, 6, 3, 3 and 5. The result means that the derived groups are balanced. As to the diversity of DGSP, we can see that the three stocks (2409, 2426, 2438) belong to optoelectronic in group G_1 and the three stocks (4904, 2412, 2419) belong to communication network in group G_3 in the initial DGSP. But in the derived DGSP, they are divided into different groups. The stock symbols 2409 and 2419 are moved to groups G_2. From the derived DGSP and its diversity value, it can be concluded that the diversity of groups obtained by the proposed approach is increased. Besides, a total of 450 ($= 5 \times 6 \times 3 \times 5$) stock portfolios can be generated from the four chosen groups and suggested to investors.

Then, experiments were conducted to compare the DGSPs derived by fitness function f_1 with those derived by fitness function f_2. Table 8 shows the results.

The difference in the fitness functions f_1 and f_2 is that when f_2 is used in the proposed approach, it means the stock prices and purchased units of a stock will be considered during the evolution process. From Table 8, we can verify that the unit and price balances of the DGSP derived by f_2, 1.4 and 1.027, are both better than those derived by f_1. However, the portfolio satisfaction of the derived DGSP has decreased. In other words, there is a trade-off between portfolio satisfaction and unit and price balances. To verify them more clearly, the series of stock prices of the DGSPs obtained by f_1 and f_2 are shown in Fig. 6 and Fig. 7.

TABLE 8. The DGSPs derived using fitness functions f_1 and f_2.

The diverse group stock portfolio derived by f_1		
	Group and stock parts	Stock Portfolio
G_1:	{4938, 2460, 4904, 2412, 1326}	0.94, 5, 0.54, 5, 0.14, 10, 0.20, 30, 0.53, 39, 0.89, 5
G_2:	{2330, 2801, 1303, 1321, 2409, 2419}	Fitness Value=17101.48
G_3:	{ 2311, 2451, 2891, 2357, 2340}	PortfolioSatisfaction=40.88
G_4:	{2303, 2301, 2498, 3045, 2809, 1301}	Group Balance=3.13
G_5:	{2325, 2881, 2377}	UB = 1.00, PB =1.838
G_6:	{2382, 1325, 2426, 2438, 2851}	Diversity=3.689
The diverse group stock portfolio derived by f_2		
	Group and stock parts	Stock Portfolio
G_1:	{2325, 2377, 2340, 2851}	0.96, 30, 0.28, 23, 0.69, 30, 0.42, 17, 0.51, 10, 0.91, 10
G_2:	{2330, 2357, 1301, 2498, 4904}	Fitness Value=18339.37
G_3:	{2801, 2301, 1326, 1321, 2426}	PortfolioSatisfaction=30.93
G_4:	{451, 2382, 1303, 3045, 2412}	Group Balance=3.19,
G_5:	{2311, 2891, 2809, 2409, 2460}	UB = 1.40, PB =1.027
G_6:	{2303, 2881, 4938, 1325, 2419, 2438}	Diversity = 3.667

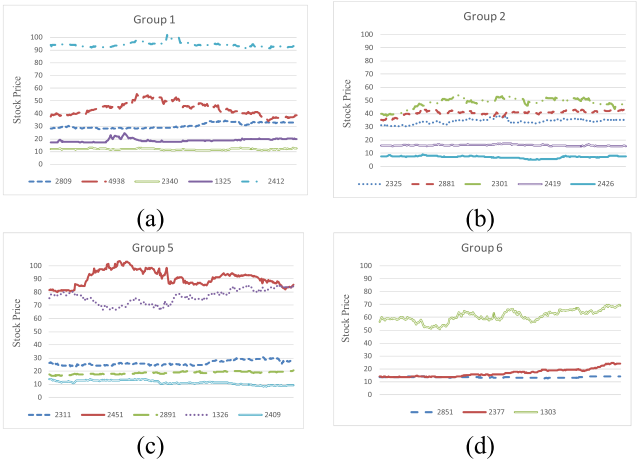


FIGURE 6. The series of stock prices of the diverse GSP by f_1.

Comparing the results in Fig. 6 and Fig. 7, we can observe that the stock price series in Fig. 7 have a higher similarity than that in Fig. 6. For instance, the prices of the two stock symbols 2451 and 1326 are higher than those of the other three stocks in Fig. 6(c) as well as stock symbol 1303 in Fig. 6(d). From Fig. 7, we can also observe that the stock price series in groups are similar. For instance, the prices of the stock series in Fig. 7(c) are within 10 to 35.

C. COMPARING PROPOSED APPROACH WITH EXISTING APPROACHES

To show the diversity of the derived DGSPs more clearly, experiments were then made to show the average diversity

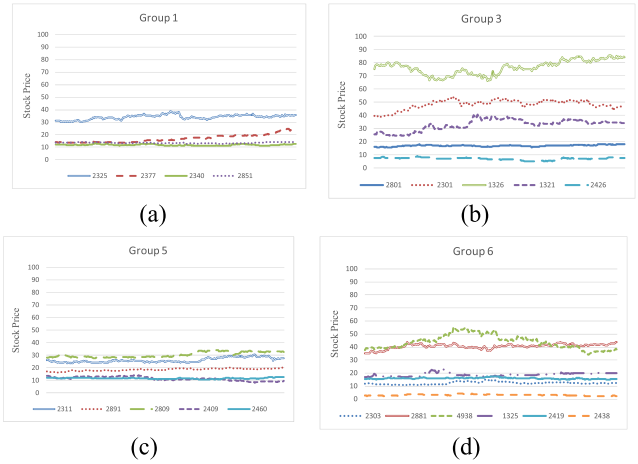


FIGURE 7. The series of stock prices of the diverse GSP by f_2.

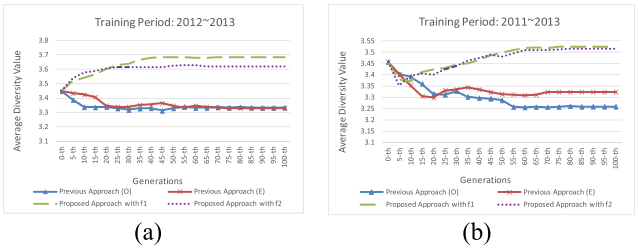


FIGURE 8. The average diversity values of the previous and proposed approaches.

values of the derived DGSPs by the previous approach and proposed approach with fitness functions f_1 and f_2 using two-year and three-year datasets as training data. Fig. 8 shows the results.

Fig. 8 shows the average diversity values of the proposed approach with f_1 and f_2 are better than those of "Previous approach (M) " and "Previous approach (E) ". The results show that the DGSPs derived by the proposed approach do actually increase the diversity of groups when compared with existing approaches.

Experiments were then conducted to verify the effectiveness of the derived DGSPs. The proposed approach was compared with the previous one and the benchmark in terms of returns and diversity of groups. The benchmark represents the best solution and is obtained by a brute-force way by trying all combinations of stock portfolios. The average ROI, maximum ROI and minimum ROI of the stock portfolios were compared. The results were averaged by ten runs on the two-year training and one-year testing datasets, shown in Tables 9.

It can be easily observed from Table 9 that the average ROI by the proposed approach and the previous one are both around 45% and 25% on the training and testing datasets, which are better than 12% and 18% on all the combinations in the benchmark. Comparing the proposed approach with the previous one in terms of returns, they show that the returns of both approaches are similar. When the diversities of the approaches are compared, we can find that the proposed

TABLE 9. Returns and diversity of the derived DGSPs on the first dataset.

$h = 3$		Avg. ROI	Avg. Max. ROI	Avg. Min. ROI	Avg. Diversity
2012~2013	Previous Approach (O)	0.493	0.828	0.25	3.342
	Previous Approach (E)	0.435	0.78	0.187	3.331
	Proposed Approach (f_1)	0.451	0.804	0.196	3.684
	Proposed Approach (f_2)	0.429	0.738	0.202	3.62
	Benchmark	0.128	0.825	-0.675	
2014	Previous Approach (O)	0.257	0.56	-0.038	
	Previous Approach (E)	0.275	0.605	-0.016	
	Proposed Approach (f_1)	0.259	0.646	-0.021	
	Proposed Approach (f_2)	0.228	0.53	-0.045	
	Benchmark	0.183	1.033	-0.153	

TABLE 10. Returns and diversity of the obtained DGSPs on the second dataset.

$h = 3$		Avg. ROI	Avg. Max. ROI	Avg. Min. ROI	Avg. Diversity
2012~2013	Previous Approach (O)	1.639	2.259	0.992	3.889
	Previous Approach (E)	1.429	2.17	0.697	3.762
	Proposed Approach (f_1)	0.995	1.606	0.379	4.007
	Proposed Approach (f_2)	0.944	1.524	0.379	3.997
	Benchmark	0.317	1.861	-0.639	
2014	Previous Approach (O)	-0.027	0.117	-0.158	
	Previous Approach (E)	0.059	0.284	-0.157	
	Proposed Approach (f_1)	0.164	0.302	0.006	
	Proposed Approach (f_2)	0.188	0.315	0.046	
	Benchmark	0.108	0.446	-0.163	

approach has better diversity of DGSPs than the previous one. Thus, the proposed approach can not only achieve almost the same returns as the previous one but also has a better diversity of groups than the previous one. Then, experimental results were conducted to compare the two approaches on returns for the second dataset. The results are summarized in Table 10.

From Table 10, in the training phase, although the previous approach with the original fitness function has the highest average ROI, the average ROI is −0.027 in the testing phase. Using the enhanced fitness function, the average ROI of the previous approach is increased, which is 0.059 in the

TABLE 11. Returns of the derived DGSPs on the first dataset with different h as training data.

Training Period: 2013		Avg. ROI	Avg. Max. ROI	Avg. Min. ROI
	Previous Approach (O)	0.493	0.828	0.25
	Benchmark	0.042	0.452	-0.498
$h = 1$	Previous Approach (M)	0.341	0.666	0.155
	Proposed Approach (f_1)	0.324	0.662	0.104
$h = 3$	Previous Approach (M)	0.371	0.694	0.185
	Proposed Approach (f_1)	0.296	0.635	0.101
$h = 5$	Previous Approach (M)	0.321	0.663	0.12
	Proposed Approach (f_1)	0.328	0.657	0.125
$h = 7$	Previous Approach (M)	0.205	0.459	-0.026
	Proposed Approach (f_1)	0.197	0.372	-0.016
Testing Period: 2014		Avg. ROI	Avg. Max. ROI	Avg. Min. ROI
	Previous Approach (O)	0.257	0.56	-0.038
	Benchmark	0.183	1.033	-0.153
$h = 1$	Previous Approach (M)	0.301	0.621	0.072
	Proposed Approach (f_1)	0.266	0.67	0
$h = 3$	Previous Approach (M)	0.262	0.597	0.007
	Proposed Approach (f_1)	0.229	0.554	-0.015
$h = 5$	Previous Approach (M)	0.244	0.646	-0.035
	Proposed Approach (f_1)	0.287	0.576	0.091
$h = 7$	Previous Approach (M)	0.206	0.52	-0.017
	Proposed Approach (f_1)	0.147	0.362	-0.038

TABLE 12. Returns of the derived DGSPs on the second dataset with different h as training data.

Training Period: 2013		Avg. ROI	Avg. Max. ROI	Avg. Min. ROI
	Previous Approach (O)	0.6	0.843	0.304
	Benchmark	0.132	0.782	-0.446
$h = 1$	Previous Approach (M)	0.572	0.853	0.297
	Proposed Approach (f_1)	0.55	0.857	0.243
$h = 3$	Previous Approach (M)	0.435	0.633	0.224
	Proposed Approach (f_1)	0.49	0.761	0.236
$h = 5$	Previous Approach (M)	0.437	0.656	0.228
	Proposed Approach (f_1)	0.439	0.602	0.257
$h = 7$	Previous Approach (M)	0.217	0.443	-0.065
	Proposed Approach (f_1)	0.223	0.402	-0.02
Testing Period: 2014		Avg. ROI	Avg. Max. ROI	Avg. Min. ROI
	Previous Approach (O)	-0.009	0.186	-0.156
	Benchmark	0.108	0.446	-0.163
$h = 1$	Previous Approach (M)	0.048	0.232	-0.16
	Proposed Approach (f_1)	0.033	0.225	-0.158
$h = 3$	Previous Approach (M)	0.139	0.258	0.026
	Proposed Approach (f_1)	0.121	0.267	-0.049
$h = 5$	Previous Approach (M)	0.164	0.276	0.056
	Proposed Approach (f_1)	0.181	0.26	0.12
$h = 7$	Previous Approach (M)	0.129	0.365	-0.083
	Proposed Approach (f_1)	0.122	0.278	-0.031

testing phase. In the proposed approach, we can find that the average ROIs of the optimized DGSPs by utilizing the fitness function f_1 or f_2 are obviously better than those of the previous approach and benchmark. These results indicate that the proposed approach, considering diversity and stability factors, can derive better DGSPs than the previous approach.

D. EFFECTIVENESS OF THE STABILITY FACTOR

Experiments were performed in this section for testing the stability factor with different h values on the two datasets in terms of average ROI. Tables 11 and 12 state the results for the first and second datasets, respectively.

From Table 11, it can be seen that the average ROI values of the proposed and the previous approaches with different h values are better than those of the benchmark on training data. All of them are larger than 0.197. On the testing data, most of the average ROI values of the proposed and the previous approaches are also better than those of the benchmark. Comparing the proposed approach with the previous one, it can be observed that the proposed approach may have average ROI values similar to those of the previous approach. When h is set to 5, the average ROI value of the proposed approach is better than that of the previous one. Table 11 shows that when the stocks in the given dataset have small variance, the proposed approach (f_1), the previous approach (O) and the previous approach (M) can reach good average ROI values.

In Table 12, the average ROI values of both the proposed and previous approaches are better than the benchmark in the training data, but the previous approach (O), which has

the highest return in the training data, has a negative return in the testing data. But, if the stability factor is considered, the average ROI values of the previous approach (M) with different h are larger than 0.048 at least. In the proposed approach (f_1), we can also see that the average ROI values are 0.033, 0.121, 0.181 and 0.122 for h values that were set to 1, 3, 5 and 7, respectively. And, when h was set to 5, the proposed approach has the highest average ROI. As a result, comparing Tables 11 and 12, we can conclude that when the variance of stock prices is high, the stability factor can increase the average ROI values.

VII. CONCLUSION AND FUTURE WORK

Many portfolio optimization methods with distinct criteria have been suggested over the past decades to address distinct kinds of portfolio optimization problems. Previously, an algorithm has been proposed to find a GSP using the GGA in order to provide a more effective way for investment. Through the GSP, various stock portfolios can be suggested. In other words, when investors are not satisfied with a suggested stock, it is possible to select a substitute stock from the same stock group to replace the original. Taking into account the diversity of stock industries, the GSPO problem becomes the DGSPO problem. An algorithm has been proposed for solve DGSPO problem to obtain a DGSP using the GGA with the enhanced fitness functions in this paper. Experimental findings on two real datasets also showed the effectiveness of the presented approach. The main contributions of this work can be summarized as follows:

(1) In order to provide a more effective way for investment, by considering the diversity of stock industries, we have proposed an algorithm to obtain a diverse group stock portfolio using the grouping genetic algorithm.

(2) To reach the goal, the two enhanced fitness functions have been designed by adding the stability factor and the diversity factor. The stability factor is intended to decrease the probability of the suggested stock suffering enormous loss based on the cash dividend in testing. The diversity factor is used to enhance a group's diversity of stock industries.

(3) Experimental results conducted on two real datasets indicate that the proposed approach can not only has similar or even better returns than the previous approach, but also improve the diversity of the derived DGSPs, especially when stock prices fluctuate heavily.

In the future, we will continually assess the suggested approach through bigger databases and investigate the outcomes using distinct parameter configurations as well as the distinct methods of optimization.

ACKNOWLEDGMENT

This work was supported by the Ministry of Science and Technology of China under Grants MOST 106-2221-E-032-049-MY2, MOST 107-2218-E-390-004 and MOST 108-2221-E-032-037. This article was presented at the 2016 IEEE Congress on Evolutionary Computation.

REFERENCES

[1] S. Barak, M. Abessi, and M. Modarres, "Fuzzy turnover rate chance constraints portfolio model," *Eur. J. Oper. Res.*, vol. 228, no. 1, pp. 141–147, 2013.

[2] M. H. Babaei, M. Hamidi, E. Jahani, and H. Pasha, "A new approach to solve an extended portfolio selection problem," in *Proc. Int. Conf. Ind. Eng. Oper. Manage.*, 2012, pp. 1954–1960.

[3] J. Brimberg, N. Mladenović, and D. Urošević, "Solving the maximally diverse grouping problem by skewed general variable neighborhood search," *Inf. Sci.*, vol. 295, pp. 650–675, Feb. 2015.

[4] V. Bevilacqua, V. Pacelli, and S. Saladino, "A novel multi objective genetic algorithm for the portfolio optimization," in *Advanced Intelligent Computing*. Berlin, Germany: Springer, 2012, pp. 186–193.

[5] J. D. Bermúdez, J. V. Segura, and E. Vercher, "A multi-objective genetic algorithm for cardinality constrained fuzzy portfolio selection," *Fuzzy Sets Syst.*, vol. 188, pp. 16–26, Feb. 2012.

[6] E. C. Brown and R. T. Sumichrast, "Evaluating performance advantages of grouping genetic algorithms," *Eng. Appl. Artif. Intell.*, vol. 18, no. 1, pp. 1–12, 2005.

[7] C.-H. Chen, C.-C. Chen, and Y. Nojima, "An efficient and effective approach for mining a group stock portfolio using MapReduce," *Intell. Data Anal.*, vol. 21, no. S1, pp. 217–232, 2017.

[8] C. H. Chen and C. Y. Hsieh, "Actionable stock portfolio mining by using genetic algorithms," *J. Inf. Sci. Eng.*, vol. 32, no. 6, pp. 1657–1678, 2016.

[9] C.-H. Chen, C.-B. Lin, and C.-C. Chen, "Mining group stock portfolio by using grouping genetic algorithms," in *Proc. IEEE Congr. Evol. Comput.*, May 2015, pp. 738–743.

[10] C.-H. Chen, W.-Y. Shen, T.-P. Hong, and J.-H. Su, "An island-based algorithm for group stock portfolio optimization," in *Proc. Joint 17th World Congr. Int. Fuzzy Syst. Assoc., 9th Int. Conf. Soft Comput. Intell. Syst.*, Jun. 2017, pp. 1–5.

[11] C.-H. Chen and C.-H. Yu, "A series-based group stock portfolio optimization approach using the grouping genetic algorithm with symbolic aggregate approximations," *Knowl.-Based Syst.*, vol. 125, pp. 146–163, Jun. 2017.

[12] T.-J. Chang, S.-C. Yang, and K.-J. Chang, "Portfolio optimization problems in different risk measures using genetic algorithm," *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10529–10537, Sep. 2009.

[13] T. Dokeroglu and A. Cosar, "Optimization of one-dimensional Bin Packing Problem with island parallel grouping genetic algorithms," *Comput. Ind. Eng.*, vol. 75, pp. 176–186, Sep. 2014.

[14] L. D. Geronimo, F. Ferrucci, A. Murolo, and F. Sarro, "A parallel genetic algorithm based on Hadoop MapReduce for the automatic generation of JUnit test suites," in *Proc. IEEE Int. Conf. Softw. Test., Verification Validation*, Apr. 2012, pp. 785–793.

[15] M. El hachlou, Z. Guennoun, and F. Hamza, "Stocks portfolio optimization using classification and genetic algorithms," *Appl. Math. Sci.*, vol. 6, no. 94, pp. 4673–4684, 2012.

[16] T. H. Huxley, "The high dividend yield return advantage: An examination of empirical data associating investment in high dividend yield securities with attractive returns over long measurement periods," Tweedy, Browne, New York, NY, USA, Tech. Rep., 2007.

[17] E. Falkenauer, "A New representation and operators for genetic algorithms applied to grouping problems," *Evol. Comput.*, vol. 2, pp. 123–144, Jun. 1994.

[18] E. Falkenauer, "A hybrid grouping genetic algorithm for bin packing," *J. Heuristics*, vol. 2, pp. 5–30, Jun. 1996.

[19] Z. P. Fan, Y. Chen, J. Ma, and S. Zeng, "A hybrid genetic algorithmic approach to the maximally diverse grouping problem," *J. Oper. Res. Soc.*, vol. 62, no. 1, pp. 92–99, 2011.

[20] M. Gallego, M. Laguna, R. Martí, and A. Duarte, "Tabu search with strategic oscillation for the maximally diverse grouping problem," *J. Oper. Res. Soc.*, vol. 64, no. 5, pp. 724–734, 2013.

[21] P. Gupta, M. K. Mehlawat, and G. Mittal, "Asset portfolio optimization using support vector machines and real-coded genetic algorithm," *J. Global Optim.*, vol. 53, pp. 297–315, Jun. 2012.

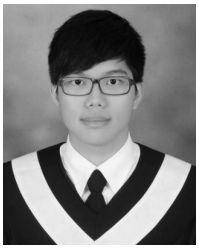
[22] P. Gupta, M. K. Mehlawat, and A. Saxena, "Hybrid optimization models of portfolio selection involving financial and ethical considerations," *Knowl.-Based Syst.*, vol. 37, pp. 318–337, Jan. 2012.

[23] L. Di Gaspero, A. Schaerf, G. Di Tollo, and A. Roli, "Hybrid metaheuristics for constrained portfolio selection problems," *J. Quant. Finance*, vol. 11, no. 10, pp. 1473–1487, 2010.

[24] Hoklie and L. R. Zuhal, "Resolving multi objective stock portfolio optimization problem using genetic algorithm," in *Proc. Int. Conf. Comput. Autom. Eng.*, Feb. 2010, pp. 40–44.

[25] R. Kumar and S. Bhattacharya, "Cooperative search using agents for cardinality constrained portfolio selection problem," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 1510–1518, Nov. 2012.
[26] P.-C. Lin, "Portfolio optimization and risk measurement based on non-dominated sorting genetic algorithm," *J. Ind. Manage. Optim.*, vol. 8, no. 3, pp. 549–564, 2012.
[27] K. Lwin, R. Qu, and G. Kendall, "A learning-guided multi-objective evolutionary algorithm for constrained portfolio optimization," *Appl. Soft Comput.*, vol. 24, pp. 757–772, Nov. 2014.
[28] Y.-J. Liu and W.-G. Zhang, "Fuzzy portfolio optimization model under real constraints," *Insurance, Math. Econ.*, vol. 53, no. 3, pp. 704–711, 2013.
[29] J. Li and J. Xu, "Multi-objective portfolio selection model with fuzzy random returns and a compromise approach-based genetic algorithm," *Inf. Sci.*, vol. 220, pp. 507–521, Jan. 2012.
[30] H. Markowitz, "Portfolio selection," *J. Finance*, vol. 7, no. 1, pp. 77–91, 1952.
[31] H. M. Markowitz, *Harry Markowitz: Selected Works*. Singapore: World Scientific, 2009.
[32] T. P. Patalia and G. R. Kulkarni, "Design of genetic algorithm for knapsack problem to perform stock portfolio selection using financial indicators," in *Proc. Int. Conf. Comput. Intell. Commun. Netw.*, Oct. 2011, pp. 289–292.
[33] A. D. Roy, "Safety first and the holding of assets," *Econometrica*, vol. 20, no. 3, pp. 431–449, Jul. 1952.
[34] F. J. Rodriguez, M. Lozano, C. García-Martínez, and J. D. González-Barrera, "An artificial bee colony algorithm for the maximally diverse grouping problem," *Inf. Sci.*, vol. 230, pp. 183–196, May 2013.
[35] G. D. Tollo and A. Roli, "Metaheuristics for the portfolio selection problem," *Int. J. Oper. Res.*, vol. 5, no. 1, pp. 13–35, 2008.
[36] I. Ucar, A. M. Ozbayoglu, and M. Ucar, "Developing a two level options trading strategy based on option pair optimization of spread strategies with evolutionary algorithms," in *Proc. IEEE Congr. Evol. Comput.*, May 2015, pp. 2526–2531.
[37] R. R. Weitz and S. Lakshminarayanan, "An empirical comparison of heuristic methods for creating maximally diverse groups," *J. Oper. Res. Soc.*, vol. 49, no. 6, pp. 635–646, 1998.
[38] E. Wah, Y. Mei, and B. W. Wah, "Portfolio optimization through data conditioning and aggregation," in *Proc. IEEE Int. Conf. Tools Artif. Intell.*, Nov. 2011, pp. 253–260.
[39] M.-E. Wu, C.-H. Wang, and W.-H. Chung, "Using trading mechanisms to investigate large futures data and their implications to market trends," *Soft Comput.*, vol. 21, no. 11, pp. 2821–2834, Jun. 2017.
[40] C. F. You, S. H. Lin, and H. F. Hsiao, "Dividend yield investment strategies in the Taiwan stock market," *Investment Manage. Financial Innov.*, vol. 7, no. 2, pp. 189–199, 2010.

CHUN-HAO CHEN received the Ph.D. degree in computer science and information engineering from National Cheng Kung University, Taiwan, in 2008. He joined as a Postdoctoral Fellow with the Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan, in 2009. From February 2010 to July 2013 and from August 2013 to July 2017, he served as an Assistant and Associate Professor with the Department of Computer Science and Information Engineering, Tankang University, respectively. He is currently a Professor with the Department of Computer Science and Information Engineering, Tamkang University, Taiwan. He has published more than 120 research articles in refereed journals and international conferences. His research interests include data mining, time series, machine learning, evolutionary algorithms, fuzzy theory, portfolio selection, trading strategy, business data analysis, and time series pattern discovery.



CHENG-YU LU received the M.S. degree in Department of Computer Science and Information Engineering at Tamkang University, Taiwan, in 2016. Currently, he is working as a system development engineer at Cimforce. His research interests include machine learning, financial data analysis, and genetic algorithms.

TZUNG-PEI HONG received the B.S. degree in chemical engineering from National Taiwan University, in 1985, and the Ph.D. degree in computer science and information engineering from National Chiao Tung University, in 1992. He served at the Department of Computer Science, Chung Hua Polytechnic Institute, from 1992 to 1994, and at the Department of Information Management, I-Shou University, from 1994 to 2001. He was an in charge of the whole computerization and library planning for the National University of Kaohsiung in preparation, from 1997 to 2000, where he served as the First Director of the Library and Computer Center, from 2000 to 2001. He also served as the Dean of Academic Affairs, from 2003 to 2006, the Administrative Vice President, from 2007 to 2008, and the Academic Vice President, in 2010. He is currently a Distinguished and the Chair Professor with the Department of Computer Science and Information Engineering and with the Department of Electrical Engineering, and the Director of the AI Research Center, National University of Kaohsiung, Taiwan. He is also a joint Professor with the Department of Computer Science and Engineering, National Sun Yat-sen University, Taiwan. He has published more than 600 research articles in international/national journals and conferences and has planned more than fifty information systems. His current research interests include knowledge engineering, data mining, soft computing, management information systems, and www applications. He was a recipient of the first National Flexible Wage Award from the Ministry of Education, Taiwan. He is also the Board Member of more than forty journals and the Program Committee Member of more than five hundred conferences.

JERRY CHUN-WEI LIN received the Ph.D. degree from the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan. He is currently an Associate Professor with the Department of Computing, Mathematics, and Physics, Western Norway University of Applied Sciences, Bergen, Norway. He has published more than 200 research articles in refereed journals and international conferences. His research interests include data mining, soft computing, artificial intelligence, social computing, multimedia and image processing, and privacy-preserving and security technologies. He is also the Project Leader of SPMF: An Open-Source Data Mining Library, which is a toolkit offering multiple types of data mining algorithms. He also serves as the Editor-in-Chief of the *International Journal of Data Science* and *Pattern Recognition*.

MATTEO GAETA received the degree in Information Science at the University of Salerno. He is a Full professor in the information processing systems area and is the Scientific Coordinator of the KnowMIS Lab - Knowledge Management and Information Systems Lab. His main research interests are complex information systems, Decision Support Systems, knowledge management systems and Computational Intelligence. He has authored over 200 scientific paper published in journals, proceedings, and books and he has planned and designed more than 10 information systems. He is editor-in-chief of the *International Journal of Information Technology*, *Communications and Convergence* (Inderscience) and associate editor of the *Journal of Ambient Intelligence and Humanized Computing* (Springer). He is also a member of the editorial board of international journals and of the program committees of many conferences. Prof. Matteo Gaeta is Senior Member of IEEE and he is a member of the Computational Intelligence Society. He gained a large experience in the implementation and design of complex information systems. He was promotor and director of Research and Development Centers, Technology Transfer and Spin-offs. He is the Scientific Coordinator and Manager of several International Research Projects. He is the Coordinator of the Working Group ex. art.14 D.M. 593/2000 for the financing of SMEs of the Italian Ministry of Instruction, University and Research (MIUR).