Diverse Group Stock Portfolio Optimization Based on Investor Sentiment Index

Chun-Hao Chen[1], Tzung-pei Hong[2, 3], Shih-Chi Chu[2]
[1]Department of Computer Science and Information Engineering Tamkang University, Taipei, Taiwan
[2]Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan
[3]Department of Computer Science and Engineering, National Sun Yat-sen University, Taiwan
chchen@mail.tku.edu.tw, tphong@nuk.edu.tw, piliseric@gmail.com

***Abstract*—Financial data analysis has always been an interesting issue and stock portfolio optimization is one of its most popular research topics. Recently, research on investor sentiment is also very popular in financial data analysis, and some scholars claim that it has impact on the stock market. In this paper, the proposed approach thus uses two investor sentiment indices and utilizes the grouping genetic algorithm to obtain a diverse group stock portfolio. Based on the fitness function defined in the previous approach, the adopted fitness function adds a new factor called risk of investor sentiment to evaluate the chromosomes for finding an appropriate diverse group stock portfolio. At last, experiments were conducted on a real stock dataset to verify the effectiveness of the proposed approach.**

Keywords—investor sentiment index, trading strategy, grouping genetic algorithms, group stock portfolio, maximally diverse grouping problem.

I. INTRODUCTION

Since there were financial transactions in human economic activities, financial data analysis has been an issue of focus. In 1952, the theory of portfolio proposed by Markowitz laid the foundation for current investment. The stock's expected earnings and risk are used for the model called the mean-variance (M-V) model to acquire a stock portfolio [10]. Lots of algorithms have then been proposed for obtaining portfolios, or to deal with different issues in this research field [1, 2, 4, 6, 7, 8, 9, 12, 13, 15, 16].

However, providing only a stock portfolio to investors is not practical for real applications. Investors may face a situation in which not all suggested stocks are suitable for their plans. Hence, the problem is shifted from finding a stock portfolio to obtaining a set of stock portfolios. For this problem, in the previous approach, Chen *et al.* proposed an algorithm for obtaining a diverse group stock portfolio (DGSP) by the grouping genetic algorithm (GGA) [5]. A DGSP is composed of a set of stock groups, where each stock group has a set of stocks. Though a given DGSP, various stock portfolios can be provided to investors for making investment plans.

But, there is still a problem needs to be solved. That is, if an event such as the September 11 attacks (911) in New York occurs, the stock market will be significantly impacted by this sudden incident. Investors will doubt whether they should trade the stocks in this unstable situation. In the previous approach [5], the obtained stock portfolio does not provide investors with the opportunity to transact.

According to the existing literatures [3, 11, 14], they indicated that investor sentiment indices actually have the ability to reveal the relationship between the stock market and investors' sentiment that may also be used to predict future rewards of company stocks. Therefore, taking two investor sentiment indices into consideration, based on the previous approach [5], the proposed approach is designed for not only obtaining a DGSP but also providing the trading timing to investors. We first design rules based on two sentiment indices, Buy-Sell Imbalance (BSI) and the proportion of Day-Trades (DT), to decide the buying and selling points of all the input stocks. These stock price data with buying and selling points are then sent to the proposed approach for extracting a DGSP. The fitness function which is composed of the risk of investor sentiment (RIS) and the other existing factors in the previous approach is employed to evaluate the chromosomes, where each chromosome is a possible DGSP. After evolution, the chromosome with the best fitness value is output as the derived DGSP. Experimental results were made on a real dataset to show the proposed approach is effective.

II. RELATED WORK

Since how to obtain a suitable stock portfolio for investors is an interesting problem, so far, many algorithms have been proposed to deal with it in terms of different point of views [1, 2, 4, 6, 7, 8, 9, 12, 13, 15, 16]. For example, Sun and Liu established the empirical cross-correlation matrices to improve portfolio optimization by combining the Pearson's correlation coefficients (PCC) method and the detrended cross-correlation analysis (DCCA) method [12]. Yu *et al*. then developed two CVaR-based robust portfolio models. The first one was the worst-case conditional value-at-risk (WCVaR) model, and the second one was the relatively robust conditional value-at-risk (RRCVaR) model. They found that when the required return was fixed, the RRCVaR model brought higher returns, lower trading costs and higher portfolio diversity than the WCVaR model [16]. Baralis *et al*. presented an itemset-based approach to automatically identify promising sets of high-yield but diversified stocks to investors [1]. They investigated the usage of itemsets to generate appropriate stock portfolios and recommend them to investors from historical stock data. Thakur *et al*. then used the fuzzy Delphi method to identify the critical factors of investment from input data. They then used these critical factors and historical data to rank the stocks by the Dempster– Shafer evidence theory [13].

As to algorithms with investor sentiment indices, in 1998,

Neal *et al*. studied the following three investor sentiment indices including the degree of closed-end funds discount, the odd-lot ratio and the net redemptions for forecasting stock returns [11]. They found that the degree of closed-end funds discount and the net redemption were predictable for small companies, and the odd-lot ratio had no predictability for either large or small companies. In 2000, Barber *et al*. obtained the fact that there was a strong negative relationship between the IPO rate and the following year's market returns [3]. This relationship could provide a stronger market return prediction. In 2012, Wang *et al*. pointed out that there was a high correlation between the three retail sentiment variables including the buy-sell imbalance, the individual investor turnover, and the proportion of day-trades [14].

III. The Proposed Approach

In this section, details of the proposed approach are described, including data preprocessing, encoding scheme, initial population, fitness function and genetic operations are stated.

A. Data Preprocessing for Forming Desired Trainng Stock Price Sequences

For considering the sentiment factor in the proposed approach for finding a good DGSP, we first transform the input stock price sequences to the ones with sentiment considered. We design some rules based on two sentiment indices to decide the buying and selling points of all the input stocks. The two investor sentiment indices are selected from Wang and Lin's research [14]. The first investor sentiment index is Buy-Sell Imbalance (BSI_t), which is defined as follows:

$$BSI_t = \frac{BML_t - BSL_t}{VB_t + VS_t},$$

where BML_t is the balance of margin loan of the stock on the day t; BSL_t is the balance stock loan of the stock on the day t; VB_t means the retail investors' buying volume of the stock on the day t, and VS_t means the retail investors' selling volume of the stock on the day t. At time t, a higher BSI_t value means that the retail investors are more optimistic about the stock.

The second investor sentiment index is the proportion of Day-Trades (DT_t), which is defined as follows:

$$DT_t = \frac{NDT_t}{Shares_t},$$

where NDT_t is the number of day-trades and $Shares_t$ is the outstanding shares. In the stock market, the number of day-trades (NDT_t) means investors' twice trading behaviors (including buy after sell or sell after buy) by margin loan and stock loan on the same day t on the same stock. The $Shares_t$ means the amount of stock that the company offers to investors on the day t. It can be seen that there is a positive relation between DT_t and the degree of investors' interest in a stock.

Next, we state how to determine the situations of DT_t and BSI_t for a stock on the day t. DT_t has two possible values: rising and falling. The way to identify the value of DT_t is as follows:

$$DT_t Rise \leftrightarrow (DT_t - DT_{t-1} > 0), \text{ and}$$

$$DT_t Fall \leftrightarrow (\mathrm{DT_t} - DT_{t-1} < 0).$$

On the other hand, BSI_t has three possible values: low, medium and high. We use the two thresholds 10% and 90% to classify BSI_t into one of them. We then define the following trading rules:

- *Rule 1*: *On a day t, if a stock's BSI_t is Low and DT_t is rising, then buy the stock.*
- *Rule 2*: *On a day t, if a stock's BSI_t is High and DT_t is rising, then sell it.*

For the other situations, the stock is neither bought nor sold, i.e. held. In this study, we set the first buying point and the last selling point as the trading timing of a stock in its given training price sequence. Thus, we can get both the buying and selling prices on every stock in the input stock price sequence data. These data are then fed into the proposed GGA-based approach to find a good diverse group stock portfolio.

B. Encoding Scheme

When a GA-based approach is used, we first need to encode the possible solutions into chromosomes. The encoding scheme in this paper is described below. We use a set S to represent n stocks, and denote them as $\{s_1, s_2, \ldots, s_n\}$. The number of the groups is K. Then, the encoding scheme of the chromosome is defined as follows in Fig. 1.

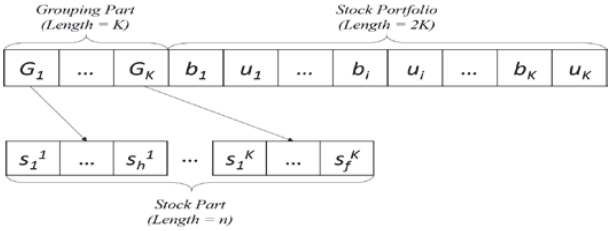

Fig. 1. The encoding scheme.

In Fig. 1, the chromosome consists of three parts including the grouping part, stock part, and stock portfolio. The number of groups in the grouping part is K. The results of the stock grouping are expressed in stock part. The s_a^b means the stock is the a-th element in the group b. Only one stock can be selected to compose a stock portfolio. Therefore, it means that only K stocks may be selected from groups into a stock portfolio at most. In the stock portfolio, two genes are used to associate to each group. The first gene b_i is the threshold for deciding whether or not to buy the stock selected in the group G_i. The second gene u_i is the number of units purchased. If b_i is greater than or equal to 0.5, it means a stock in group G_i will be purchased u_i units as part of a stock portfolio.

C. Initial Population

The strategy of generating initial population is important, because it reflects on the final optimization result. You *et al*.

pointed out that a portfolio with cash dividend yield is better than a portfolio with types in the Taiwan stock market [17]. Therefore, the proposed approach utilizes cash dividend yields of stocks to frame the initial population. We calculate the cash dividend yields defined as that cash dividend divided by stock price. With the cash dividend yield (y_i) of n companies, we calculate the proportion of average cash dividend of each group to all groups. The proportions of the average cash dividend of each group are used as a probability of being selected in the stock portfolio. As a results, a better initial population could be formed.

D. Fitness Evaluation

In this section, the fitness function used to determine the quality of the chromosome is introduced. Taking investor sentiment into consideration, the fitness function used in the proposed approach is defined as follows:

$$f(C_q) = \frac{RIS(C_q) * PS(C_q) * GB(C_q)^{\alpha} * UB(C_q) * DF(C_q)^{\beta}}{PB(C_q)},$$

where $RIS(C_q)$ is the risk of investor sentiment of chromosome C_q and is used to represent the risk of the stock, which is defined as follows:

$$RIS(C_q) = \sum_{p=1}^{NC} subRIS(SP_p) / NC,$$

where NC is the number of stock portfolios generated from chromosome C_q, and $subRIS(SP_p)$ is the risk of investor sentiment of the p-th stock portfolio SP_p, which is defined as follows:

$$subRIS(SP_p) = \sum_{i=1}^{n} nomalDTP(s_i) * u_i,$$

where u_i is the number of purchased units of stock s_i, and $nomalDTP(s_i)$ means the normalized difference of trading price of stock s_i

The $PS(C_q)$ is the portfolio satisfaction which is used to assess the profit and requests given by users of a chromosome. The $GB(C_q)$ is the group balance for a chromosome. If a chromosome has a large group balance value, it means that numbers of stocks in groups are similar. The $UB(C_q)$ is the unit balance of a chromosome. If a chromosome has a large unit balance value, it indicates the purchased units for stocks are similar. The price balance of a chromosome $PB(C_q)$ is designed to make the price of every stock in the same group as similar as possible. The diversity factor of a chromosome $DF(C_q)$ is used to increase the diversity of stocks in the same group.

E. Genetic Operations

In the proposed approach, three genetic operations, including crossover, mutation and inversion, are applied on the population to generate new offspring. For crossover operations, in the grouping part, it randomly selects chromosome C_A to be the base chromosome, and inserts some groups from another chromosome C_B. Then, it deletes the duplicate stocks to from the new chromosome C^{new}. In the stock portfolio part, the one-point crossover is utilized to generate new offspring. For mutation operations, in the stock part, the two groups are selected randomly, where the number of stocks of each group should larger than 1. Then, the mutation operator will reassign a stock into another group. In the stock portfolio part, one-point mutation is employed. For the inversion operator, its goal is to generate various combinations of groups so that the crossover operation can get different solutions. In the proposed approach, the rearrangement is done randomly.

IV. Experimental Results

In this study, the experimental dataset contains 46 stocks that were collected from the Taiwan Economic Journal (TEJ) database from 2011/01/01 to 2015/12/31 is used to evaluate the proposed approach. The training period and testing periods used in the proposed approach and the previous approach [5] are 2011/1/1 to 2013/12/31, and 2014/1/1 to 2015/12/31, respectively. The comparison results on ten runs of two approaches in terms of three standards, including *average AvgROI*, averge *MaxROI* and average *MinROI*. The comparison results on ten runs of two approaches are shown in TABLE I.

TABLE I. Comparison results of two approaches.

	avg. AvgROI	*avg. MaxROI*	*avg. MinROI*
Proposed Approach	0.265	0.382	0.135
Previous Approach	0.169	0.59	-0.145

TABLE I shows that although the *avg. MaxROI* of the proposed approach is lower than the existing approach, the most important standard *avg. AvgROI* and the *avg. MinROI* of the proposed approach are higher than the existing approach. Thus, we can conclude that the derived DGSP with investor sentiment index can provide a more stable return than that without investor sentiment index.

V. Conclusion and Future Work

In this paper, we have proposed an approach for obtaining a DGSP by the grouping genetic algorithm with investor sentiment index. We add the concept of investor sentiment to the previous approach, and provide the appropriate opportunity for trading stocks to a stock portfolio. In the experiment, we show the performance of the proposed approach by comparing it with the previous approach. The result shows that the proposed approach can have a better return than the previous approach in terms of *avg. AvgROI* and *avg. MinROI*, which means that the derived DGSP with investor sentiment index can provide a more stable return than the previous approach. In the future, more investor sentiment indices will be considered to improve the efficiency of the proposed approach.

References

[1] E. Baralis, L. Cagliero and P. Garza, “Planning stock portfolios by means of weighted frequent itemsets,” *Expert Systems with Applications*, Vol. 86, No. 15, pp. 1-

17, 2017.
[2] J. Bermúdez, J. Segura and E. Vercher, “A multi-objective genetic algorithm for cardinality constrained fuzzy portfolio selection,” *Fuzzy Sets and Systems*, Vol. 188, pp. 16-26, 2012.
[3] M. Baker and J. Wurgler, “The equity share in new issues and aggregate stock returns,” *Journal of Finance*, Vol. 55, No. 5, pp. 2219-2257, 2000.
[4] C. H. Chen and C. Y. Hsieh, “Mining actionable stock portfolio by genetic algorithms,” *Journal of Information Science and Engineering*, Vol. 32, No. 6, pp. 1657-1678, 2016.
[5] C. H. Chen, C. Y. Lu, T. P. Hong and J. H. Su, “Using grouping genetic algorithm to mine diverse group stock portfolio,” *The IEEE Congress on Evolutionary Computation*, pp. 4734-4738, 2016.
[6] M. Escobar, S. Ferrandoa and A. Rubtsov, “Portfolio choice with stochastic interest rates and learning about stock return predictability,” *International Review of Economics & Finance*, Vol. 41, pp. 347-370, 2016.
[7] Z. G. M. Elhachlofi and F. Hamza, “Stocks portfolio optimization using classification and genetic algorithms,” *Applied Mathematical Sciences*, Vol. 6, pp. 4673-4684, 2012.
[8] P. Gupta, M. K. Mehlawat and G. Mittal, “Asset portfolio optimization using support vector machines and real-coded genetic algorithm,” *Journal of Global Optimization*, Vol. 53, pp. 297-315, 2012.
[9] P. Gupta, M. K. Mehlawat and A. Saxena, “Hybrid optimization models of portfolio selection involving financial and ethical considerations,” *Knowledge-Based Systems*, Vol. 37, pp. 318-337, 2012.
[10] H. M. Markowitz, "Harry Markowitz: Selected Works," *World Scientific Publishing Company*, 2009.
[11] R. Neal and S. M. Wheatley, “Do measures of investor sentiment predict returns,” *Journal of Financial and Quantitative Analysis*, Vol. 33, No. 4, pp. 523-547, 1998.
[12] X. Sun and Z. Liu, “Optimal portfolio strategy with cross-correlation matrix composed by DCCA coefficients: Evidence from the Chinese stock market,” *Physica A: Statistical Mechanics and its Applications*, Vol. 444, pp. 667-679, 2016.
[13] G. S. M. Thakur, R. Bhattacharyya and S. Sarkar, “Stock portfolio selection using Dempster–Shafer evidence theory,” *Journal of King Saud University - Computer and Information Sciences*, pp. 1-13, 2016.
[14] J. Y. Wang and Y. C. Lin, “Individual Investor Sentiment and Stock Return－Evidence from Taiwan,” *Journal of China University of Science and Technology*, Vol. 50, pp. 147-167, 2012.
[15] E. Wah, Y. Mei and B. W. Wah, “Portfolio optimization through data conditioning and aggregation,” *IEEE International Conference on Tools with Artificial Inte elligence,* pp. 253-260, 2011.
[16] J. R.Yu, W. J. P. Chiou and R.T Liu, “Incorporating transaction costs, weighting management, and floating required return in robust portfolios,” *Computers & Industrial Engineering*, Vol. 109, pp. 48-58, 2017.
[17] C. F. You, S. H. Lin and H. F. Hsiao, “Dividend yield investment strategies in the Taiwan stock market,” *Investment Management and Financial Innovations*, Vol. 7, No. 2, pp. 189-199, 2010.