2017 14th Web Information Systems and Applications Conference

Research on Portfolio Optimization Based on Affinity Propagation and Genetic Algorithm

Chong Liu*, Wenyan Gan, Yutian Chen
College of Command Information System
PLA University of Science and Technology,
Nanjing 210007, China

***Abstract*—Portfolio optimization refers to the reasonable allocation of assets to achieve the investment objectives. At present, the investment environment depression was due to sluggish economic conditions. For investors, they expect to find a balance between return and risk in a complex investment environment. In order to solve this problem, this paper proposes a portfolio optimization algorithm named Portfolio Optimization based on Affinity propagation and Genetic algorithm, as also called POGA, which based on affinity propagation and genetic algorithm. Firstly, the affinity propagation algorithm is used to construct a candidate set of portfolio based on the correlation analysis of the stock time series. Secondly, using the Sharpe-ratio as the Optimization objective function, the genetic algorithm is used to solve an optimal portfolio strategy with higher-return and lower-risk. Finally, the experimental result of real-world stock data show that a portfolio with higher return and lower risk will be selected.**

Keywords—Portfolio optimization ; Cluster analysis ; affinity algorithm ; genetic algorithm

I. Introduction

As we all know, the stock market is called the financial "barometer" [1, 2]. At present, the financial situation remains fragile and many funding markets remain impaired. Financial conditions continue to pose a downside risk to the outlook for growth. For investors, they envisioned optimizing asset allocation achieve higher return through proper diversification. Portfolio optimization [3, 4] refers to the process of construct a portfolio based on established target return and risk tolerances. Markowitz once said that "an effective portfolio is not just a list of excellent stocks, but a balance of protection and opportunity for investors in all situations" [5, 6]. Many other studies have been devoted to consider several aspects of portfolio optimization both from a theoretical and from an applied point of view [7]. Here we focus our attention on the role of Cluster analysis in portfolio optimization.

The traditional stock block is mainly divided according to industry, concept and region. The purpose of this division is to integrate stocks so as to facilitate the analysis of investors. But, there are too many subjective factors which are not suitable for investment portfolio. Clustering is an unsupervised learning method that seeks natural clustering structure of data. Therefore, it is more objective stock cluster obtained by clustering method. Meanwhile, clustering can narrow the portfolio candidate sets and eliminate stock price synchronicity.

In this paper, an algorithm of portfolio optimization based on affinity propagation and genetic algorithm is proposed. Firstly, the algorithm constructs the correlation matrix according to the stock time series, which is used further to generate the candidate sets of portfolio. Then, the genetic algorithm based on the Sharpe-ratio objective function is used to solve an optimal portfolio strategy with high-return and low-risk. Finally, the experimental results of real-world stock data demonstrate the effectiveness and efficiency of our algorithm.

II. Relatited Reseach Work

In the portfolio optimization domain, popular methods include Markowitz Mean-variance model [5], Data-driven clustering methods, and so on.

A. Mean - variance model

In 1952, Markowitz first used mathematical statistics to quantify the returns and risks, proposing the mean-variance model [5]. In the mean-variance model, Markowitz makes the following assumptions about the portfolio environment:

(1) Rational person hypothesis: Assuming that all investors are rational, they seek the most return.

(2) Efficient markets hypothesis: Under the complete market information, investors can obtain all useful information based on the stock price.

(3) Correlation hypothesis: In the stock market, the relationship between any two stocks is measured by the correlation coefficient.

(4) Variance represents the portfolio risk: Investors quantify risk as variance and tend to choose low-risk portfolios.

Based on the above assumptions, the Markowitz mean-variance model can be described as follows:

Assume that the holding period return on stocks is R = (R_1, R_2 ,…, R_n)^T, Where R_i is the expected stocks for the i-th risk stock, i=1,…, n. The covariance matrix of stocks return is $S = \left(\delta_{ij}\right)_{n\times n}$,where $\delta_{ij} = cov\left(R_i, R_j\right)$. stocks allocation weight is w=(w_1,…, w_n)^T, where w_i represents the allocation weight on the i-th risk stock. The variance of portfolio return represented as follows:

$$\delta^2 = w^T S w = \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{n} w_i w_j \delta_{ij} \tag{1}$$

*Corresponding Author

978-1-5386-4806-3/17 $31.00 © 2017 IEEE
DOI 10.1109/WISA.2017.9

$$s.t. \sum w_i = 1$$

The target return for the portfolio can be expressed as follows:

$$R_p = w^T R \tag{2}$$

Markowitz argues that reasonable investment philosophy should include two points: 1）Choose the highest return when all portfolios are at the same variance. 2）Choose the lowest variance when all portfolios are at the same return. But the above model does not apply to large-scale trading market portfolio strategy due to the lack of effective optimization methods and high computational complexity.

B. Portfolio Optimization Based on Clustering

Recently, the clustering-based optimization method is a hot research direction. Clustering is an unsupervised learning method for finding the natural clustering structure of data, which can produce the general description of the data in the problem and construct the data clustering structure automatically. For the portfolio study, the clustering method has the following advantages:1) it can eliminate the stock price synchronicity [8].2) it can reduce the computational complexity of the combinatorial optimization. 3) it can enhance the portfolio's ability to resist risks. In recent years, clustering methods commonly used in portfolio optimization include K-means, K-mediods and hierarchical clustering (single-linkage, average-linkage, complete-linkage).

V Tola and F Lillo et al [9]. introduce single-linkage and average-linkage clustering methods to compare with common methods, the experimental results show that the clustering methods can make the forecast risk of the portfolio close to the actual situation. Weidi Wu et al [10]. improved the mean-variance model by the CVaR and used the K-means clustering algorithm to calculate the portfolio return and the optimal investment weights under different conditions. Xiaoxue Cen, Jiangtao Qin et al [11]. introduce the idea of hierarchical clustering to improve the K-means algorithm and apply it to the trend of stock price fluctuation. Victoria and Payam et al [12]. used K-means, K-mediods and hierarchical clustering to analyze the portfolio using the daily CRSP stock index and they use visualization techniques to analyze the impact of the sensitivity of different clustering methods on the visualization of portfolio risk.

In addition to the clustering approach, a portfolio approach based on community division is also explored. Huiping Wang [13] used the GN algorithm to discover the largest community and pick the portfolio, the experimental results show that the portfolio after "denoising" has a stronger ability to resist risks. Ming Yuan [14] proposes a method of building stock market network based on measuring weighted multi-fractal similarity among CSI 300 index stocks through multi-scale curve.

In this paper, we introduce the affinity propagation algorithm to realize the stock clustering and construct the portfolio candidate set. We realize the automatic discovery of the portfolio strategy by optimizing an objective function based on Markowitz mean-variance model and Sharpe-ratio.

III. Basic concepts and related techniques

Before describing the algorithm, we first introduced the basic concepts related to this article.

A. Basic concepts

Definitions 1: Stock return. Assuming that $p_i(t)$ is the closing price of stock i on the t-th trading day, at time interval Δt, the return of stock i is

$$r_i(t) = \frac{p_i(t+\Delta t) - p_i(t)}{p_i(t)} \tag{3}$$

Definitions 2: Return correlation matrix. Assuming that $\{r_1,\ldots,r_n\}$ represents the return sequence of n stocks, the return correlation coefficient for any stock i and j is

$$\rho_{ij} = \frac{cov(r_i, r_j)}{\sigma_{r_i}\sigma_{r_j}} \tag{4}$$

The return correlation matrix is an n-order symmetric matrix with ρ_{ij} as an element, which is denoted by S

$$S = \begin{bmatrix} \rho_{11} & \cdots & \rho_{1j} & \cdots & \rho_{1n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{i1} & \cdots & \rho_{ij} & \cdots & \rho_{in} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{n1} & \cdots & \rho_{nj} & \cdots & \rho_{nn} \end{bmatrix} \tag{5}$$

Where r_i and r_j represent the return of stock i and j, respectively. $\rho_{ij} \in [-1,1]$. When $\rho_{ij} \in [-1,0)$, it means that the return of stock i and j are negative correlation, i.e. i stock rose, while j stock fell . when $\rho_{ij} \in (0,1]$, it means that the return of stock i and j are positive correlation, i.e. i stock rose while j stock also rose . when $\rho_{ij} = 0$, it means that the return of stock i and j are completely irrelevant.

Definitions 3: Sharpe-Ratio. The Sharpe-Ratio(SR) [15] basic idea is how much excess profits will be made to investors under the risk of stock loss. This article uses it to assess the quality of the portfolio. It is described as follows

$$SR = \frac{E\left(\sum_{i\in p} w_i R_i\right) - R_f}{\sum_{i\in p}\sum_{j\in p} w_i w_j\, cov_{ij}} \tag{6}$$

Where w_i represents the weight of stock i in the portfolio. R_i represents the return of stock i, cov_{ij} is the covariance of stock i and j. R_f is the risk-free rate of return(In this paper, $R_f = 0.041$, issued in October 2016 five-year Treasury bond rate) .

B. Affinity propogation algorithm

In 2007, a paper titled " Clustering by passing messages between data points " appeared in the Journal of Science, which proposed Affinity Propagation(AP) algorithm [16]. The basic idea of the algorithm is that it does not need to specify the initial exemplars to decide the cluster center according to the data point vote. In order to minimize the distance between the data point, the algorithm uses the Euclidean distance as the measure of the correlation, that is, the correlation between the arbitrary data point x$_i$ and x$_j$ is

$$s(i,j) = -\left\| x_i - x_j \right\|^2 \tag{7}$$

In the clustering process, the AP algorithm introduces two parameters—Responsibility and Availability—as message passing between data points. The iterative formula is as follows:

$$r_{t+1}(i,k) = \begin{cases} s(i,k) - \max\limits_{k' \neq k}\left\{ a_t(i,k') + s(i,k') \right\}, i \neq k \\ s(i,k) - \max\limits_{k' \neq k}\left\{ s(i,k') \right\}, i = k \end{cases} \tag{8}$$

$$a_{t+1}(i,k) = \begin{cases} \min\left(0, r_t(k,k) + \sum_{i' \notin \{i,k\}} \max\left\{0, r_t(i',k)\right\}\right), i \neq k \\ \sum_{i' \notin \{i,k\}} \max\left\{0, r_t(i',k)\right\}, i = k \end{cases} \tag{9}$$

Preference (p) is another parameter in the AP algorithm, it evaluates whether the data point k can be an exemplar. At the same time, in order to avoid oscillation in the iterative process, the AP algorithm introduces the attenuation coefficient λ ($\lambda \in (0.5, 1)$). The formulas are as follows:

$$r_{t+1}(i,k) \leftarrow (1-\lambda) r_{t+1}(i,k) + \lambda r_t(i,k) \tag{10}$$

$$a_{t+1}(i,k) \leftarrow (1-\lambda) a_{t+1}(i,k) + \lambda a_t(i,k) \tag{11}$$

IV. Algorithm description

Given the closing price of n stocks D={$s_1,\ldots,s_n$}, where s_i={$x_i(t_1),\ldots,x_i(t_m)$} represents the time series of the closing price of the i-th stock, $x_i(t_k)$ represents the closing price of the i-th stock at t_k. Investor usually chooses a portfolio with greater returns and less risk [17]. Hence, portfolio selection can be considered as a combinatorial optimization problem. We through the AP algorithm to obtain the candidate portfolio candidate set. The searching capability of GA has been used in this paper for the purpose of appropriately determining a portfolio in stock set. These are now described GA in detail.

a) Chromosome representation: Each chromosome is a sequence of stock codes selected from n portfolio candidate sets. For the problem space, the length of a chromosome is n, where the first positions represent the one stock, the next position represent another one, and so on.

b) Population initialization: The N stocks encoded in each chromosome are initialized to n randomly chosen stock clusters from the candidate set. This process is repeated for each of the k chromosomes in the population, where k is the size of the population.

c) Fitness computation:

$$\max_{\{w_1,\ldots,w_n\}} SR_p = \frac{E(R_p) - R_f}{\delta^2} \tag{12}$$

$$s.t. \sum w_i = 1, w_i \geq 0$$

$E(R_p)$ represents the holding period return on the portfolio. δ^2 represents the portfolio variance. R_f represents the risk-free rate of return. SR_p represents portfolio Sharpe-ratio. w_i represents the weight of stock i in the portfolio.

d) Selection: In the proportional selection strategy adopted in this article, chromosomes are chosen as criteria for their fitness in populations, that go into the mating pool for further genetic operations.

e) Crossover: Crossover is a probabilistic process that exchanges information between two parent chromosomes to produce two child chromosomes. In this paper, a single point crossover with a fixed crossing probability of μ_c is used. For chromosomes of length n, randomly generated a position, the two chromosome exchanges correspond to the gene.

f) Mutation: Each chromosome undergoes mutation with a fixed probability μ_m. When a mutation occurs, a new stock is selected randomly from the stock candidate sets to replace the original gene position in the chromosome.

g) Termination criterion: In this paper, a given iteration number is used to terminate the operation of the genetic algorithm. After each selection, crossover and mutation operation, the individuals with the highest fitness in the current generation are recorded. After all the iterations are completed, the individual with the highest fitness is the optimal solution of the combinatorial optimization problem.

In combination with AP and GA algorithms, the pseudo code description of the program is as follows:

Algorithm: Portfolio Optimization based on Affinity propagation and Genetic algorithm(POAG).

Input: Stock dataset D={$s_1,\ldots,s_n$}, Genetical gorithm iterations: iterate_num.

Output: Stock ID and its investment weight: portfolio = {(w_1,I_1),…,(w_k,I_k)}.

Steps:

Step 1. S = np.corrcoef(D); % Calculate the return correlati on matrix S

Step 2. [cluster_centers, cluster_n, labels] = affinity_prop agation (S); % Stock clustering is generated by using an affinit y propagation algorithm, where cluster_centers represents the cluster center; cluster_n represents the number of clusters; labe ls represents each stock clusters.

Step 3. Portfolio = Genetic_algorithm(labels,iterate_num); %Combine the stock clusters and the number of iterations to g et the preferred portfolio and weight.

Step 4. Output portfolio;

According to the description of the algorithm, step (3) using genetic algorithm [18, 19] for portfolio optimization. The chromosome coding rules are as follows: chromosome length corresponds to the number of clusters, each gene of the chromosome corresponds to a stock ID, each chromosome of any two genes from different stock clusters. The algorithmic time complexity is analyzed as follows: In step (1), Assuming that the time length of the stock closing price is m and the time complexity of the stock return sequence is $O(N*m)$, the time complexity of constructing the correlation matrix is $O(N^2*m)$. In step (2), The time overhead is mainly the iterative information transfer between A and R, the time complexity is $O\ (N^2*log\ N)$. In step (3), the algorithm continues to iterate three operations: selection, crossover, mutation, the time complexity is $O(N^2*iterate\_num)$.

Therefore, the time complexity of the integrated algorithm is $O(N^2*m) + O\ (N^2*log\ N) + O(N^2*iterate\_num) \approx O(N^2*log\ N)$ (When m $\geqslant log\ N$, the time complexity of the algorithm is $O(N^2*m)$).

TABLE I. A-SHARE MARKET 2291 STOCK CLOSING PRICE DATA CLUSTE RING RESULTS (P = -241.78)

ID	Community center node	Number of community nodes
1	600810	436
2	601999	346
3	002087	302
4	002747	265
5	600112	217
6	601216	209
7	002591	144
8	600863	95
9	603519	80
10	600282	71
11	000682	68
12	601857	58

V. Experimental results and discussion

This section uses the February 20, 2016 to February 16, 2017, a total of 250 trading days, the China A-share market 2317 stock closing price data as experimental data (Exclude stock trading for less than one year and replace the name of the stock)， Which contains 1228 shares of Shenzhen A-share market, 1063 shares of Shanghai A-share market. All programs are implemented using the software Pajek + Ananconda4.3 + python2.7, testing in a PC machine（i7-5500U CPU、8GB memory、Win7）. The POAG algorithm was used to cluster the 2291 stock's close price, and a total of 12 stocks were clustered. The clustering results are shown in TABLE I, where the largest stock clustering structure is shown in Fig.1. Then, the POAG algorithm optimizes the portfolio strategy through multiple iterations, as shown in TABLE II which is the portfolio optimization strategy and Sharpe-ratio obtained when the number of iterations is 100,300,500. Obviously, the Sharpe-ratio of the portfolio has improved markedly with the increase in the number of iterations, which means that the ability to obtain excess profits gradually increasing.

In order to verify the effectiveness of the proposed algorithm, the stochastic combination strategy and the K-means algorithm are used as the contrast method. Using Sharpe-ratio to measure the risk of each portfolio. The results are shown in Table III. Index300_12 is from the CSI 300 constituent stocks stochastically selected 12 stocks to form a portfolio. Since the K-means algorithm needs to give the initial k value, we choose k = 41(The stock market divides all stocks into 41 segments based on the industry) and k = 12 as the initial values of the contrast algorithm, which generate the portfolio Kmeans_41 and Kmeans_12. Kmeans_41 is a portfolio composed of 41 stocks and its Sharpe-ratio is 3.551. (Kmeans_41 includes 000758, 000923, 603009 ,…, 002358, 002600, 600191) POGA_12 indicates that the POGA algorithm automatically constructs a portfolio with a combined size of 12. As shown in TABLE III, the Sharpe-ratio of the portfolio based on stock clustering is higher than that of the stochastic combination strategy. The Sharpe-ratio of Kmeans_41, which divided by each industry as a cluster, is better than K-means_12. It is indicating that the introduction of domain knowledge can improve the effectiveness of portfolio strategy based on stock clustering. The Sharpe-ratio of POAG_12 is significantly better than other contrast algorithms, and it is not need to introduce domain knowledge.

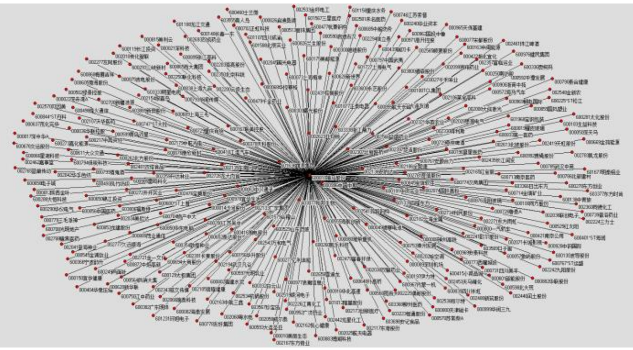

Fig. 1. The largest cluster of 436 stocks

TABLE II. THE PORTFOLIOS AT THE 100TH,300TH,500TH ITERATION

iterations	iteration =100	iterations =300	iterations =500
Portfolio	002073	002180	002716
	601789	600182	600182
	002127	002212	002212
	002695	002695	002695
	000011	603800	603800
	002292	002648	002648
	600282	002059	600638
	002611	002611	600487
	002543	600571	002031
	002073	002180	002018
	601789	600182	002716
	002127	002212	600182
SR_p	2.3093	3.2916	4.2771

Further, to evaluate the relationship between risk and return of different portfolios. The effective frontier of the above four portfolios is calculated with the least risk as the optimization target. As can be seen in Figure 2, under the same risk conditions, the return of clustering-based portfolios is much higher than the stochastic portfolio. Under the same conditions, the portfolio K-means_41, which introduces domain knowledge, has a better return than the portfolio K-means_12. While the portfolio POAG_12 return higher than other portfolios. With the increase in risk, the proportion of POAG_12's return is increasing, showing the effectiveness of the POAG algorithm.

VI. CONCLUSION

In this paper, a portfolio optimization algorithm named POGA is proposed, which uses the affinity propagation algorithm to generate the portfolio candidate sets and uses the genetic algorithm to solve an optimization objective function based on Sharpe-ratio for an optimal portfolio strategy with higher-return and lower-risk. The experimental results on real-world data sets show that POGA algorithm is able to provide more reliable portfolio in the same risks and do not need to introduce domain knowledge. The next work will discuss the impact of the stock time series length on portfolio construction and improve the efficiency of the algorithm in the actual stock analysis.

TABLE III. THE COMPARISON OF SHARPE-RATIO IN DIFFERENT ALGORITHMS INCLUDES STOCHASTIC COMBINATION, K-MEANS, POGA

Portfolio construction method	**Index300_12**	**K-means (k=12)**	**POAG_12**
Portfolio	000157	000710	002716
	000009	600764	600182
	000793	600596	002212
	000917	002352	600638
	600111	600698	002695
	600309	601028	600487
	600547	002676	603800
	600704	600260	002031
	600816	002059	002018
	600688	600241	600271
	601186	002695	002648
	601991	600809	600577
SR_p	0.9403	3.479	4.2771

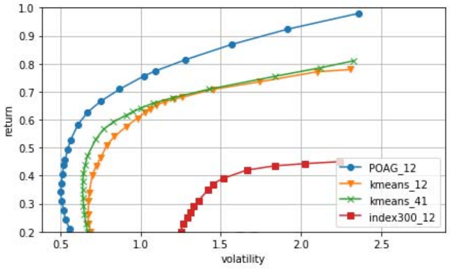

Fig.2. Effective Frontier Comparison of Different Portfolio Optimization Algorithms

VII. REFERENCES

[1] Dongju Li. "A Study on the Relationship between Stock Market Development and Economic Growth-An Explanation from Econometrics" J. Journal of Financial Research, vol. 9, Sep. 2006, pp. 78-83.

[2] Honghai Liu. "Chinese monetary policy under the Fed rate hike" J. China Finance, vol. 6, Jun.2016, pp. 86-87.

[3] Kolm P N, Tütüncü R, Fabozzi F J. "60 Years of portfolio optimization: Practical challenges and current trends" J. European Journal of Operational Research, vol. 234, April 2014, pp. 356-371, doi: 10.1016/j.ejor.2013.10.060

[4] Karceski J, Lakonishok J. "On Portfolio Optimization: Forecasting Covariances and Choosing the Risk Model" J. Review of Financial Studies, vol.12, October 1999, pp.937-974, doi: 10.1093/rfs/12.5.937

[5] Markowitz H. "Portfolio Selection" J. The Journal of Finance, vol. 7, Jul. 1952, pp. 77-91.

[6] Markowitz H M. "Foundations of Portfolio Theory" J. The Journal of Finance, vol. 46, Jun. 1991, pp. 469-477, doi: 10.1111/j.1540-6261.1991.tb02669.x

[7] Tola V, Lillo F, Gallegati M, et al. "Cluster analysis for portfolio optimization" J. Journal of Economic Dynamics & Control, vol. 32, Jan. 2008, pp. 235-258, doi: 10.1016/j.jedc.2007.01.034.

[8] Tao Zhong, Qin-ke Peng. "Portfolio selection method based on social network analysis" J. System Engineering Theory and Practice | System Eng Theor Prac,vol. 35, Dec. 2015, pp. 3017-3024, doi: 10.12011/1000-6788(2015)12-3017.

[9] Gallegati M, Tola V, Lillo F, et al. "Cluster analysis for portfolio optimization" J. Journal of Economic Dynamics & Control, vol. 32, Jan. 2005, pp. 235-258, doi: 10.1016/j.jedc.2007.01.034.

[10] Wendi Wu, Xijun Cheng, Feng Liu. "CVaR portfolio model based on K-means clustering with the constraint of generalized entropy" J. Journal of University of Chinese Academy of Sciences, vol. 33, Mar. 2016, pp. 31-36, doi: 10.7523/j.issn.2095-6134.2016.01.005.

[11] XiaoXue Qin, Jiangtao Qin. "The Application of Improved k - means Clustering in Stock Price Fluctuation" J. Science Technology and Industry | Sci Technol Ind, vol. 16, May. 2016, pp. 144-148, doi: 10.3969/j.issn.1671-1807.2016.01.028.

[12] Lemieux V, Rahmdel P S, Walker R, et al. "Clustering Techniques And their Effect on Portfolio Formation and Risk Analysis" Proc. DSMM, Jun. 2014, pp. 1-6, doi: 10.1145/2630729.2630749

[13] Huiping Wang, Hua Han, Yan Li. "A Comparative Study on Portfolio Based on Financial Network and Key" J. Journal of Wuhan University of Technology | J Wuhan Univ Technol, vol. 4, Sep. 2014, pp. 585-588, doi: 10.3963/j.issn.2095-3852.2014.04.032.

[14] Ming Yuan. "Construction of Stock Market Network Based on Multi-scale Curve and Its Community Mining" J. Computer Engineering, vol. 4, Apr. 2015, pp. 60-64, doi: 10.3969/j.issn.1000-3428.2015.04.011.

[15] Sharpe W F. "The sharpe ratio" J. Journal of Portfolio Management, vol.21, Jan. 1994, pp. 49-58, doi: 10.3905/jpm.1994.409501.

[16] Frey B J, Dueck D. "Clustering by passing messages between data points" J. Science, vol. 315, Feb. 2007, pp. 972-976, doi: 10.1126/science.1136800.

[17] Treynor J L, Black F. How to Use Security Analysis to Improve Portfolio Selection, vol. 46, Chicago: The University of Chicago Press, 1973, pp. 66-86.

[18] Holland J H. Adaptation in natural and artificial systems, vol. 6, Massachusetts: MIT Press, 1992, pp. 126–137.

[19] He-hua Liu, Chao Cui, Jing Chen. "AnImproved Genetic Algorithm for Solving Travel Salesman Problem" J. Transaction of Beijing Institute of Technology, vol. 33, Jul. 2013, pp. 390-393, doi: 10.3969/j.issn.1001-0645.2013.04.013.