Stock Portfolio Selection Using Chemical Reaction Optimization

Jin Xu, Graduate Student Member, IEEE, Albert Y.S. Lam, Member, IEEE, and Victor O.K. Li, Fellow, IEEE

Abstract—Stock portfolio selection is a classic problem in finance, and it involves deciding how to allocate an institution’s or an individual’s wealth to a number of stocks, with certain investment objectives (return and risk). In this paper, we adopt the classical Markowitz mean-variance model and consider an additional common realistic constraint, namely, the cardinality constraint. Thus, stock portfolio optimization becomes a mixed-integer quadratic programming problem and it is difficult to be solved by exact optimization algorithms. Chemical Reaction Optimization (CRO), which mimics the molecular interactions in a chemical reaction process, is a population-based metaheuristic method. Two different types of CRO, named canonical CRO and Super Molecule-based CRO (S-CRO), are proposed to solve the stock portfolio selection problem. We test both canonical CRO and S-CRO on a benchmark and compare their performance under two criteria: Markowitz efficient frontier (Pareto frontier) and Sharpe ratio. Computational experiments suggest that S-CRO is promising in handling the stock portfolio optimization problem.

Keywords—Stock portfolio selection, Markowitz model, Chemical Reaction Optimization, Sharpe ratio

I. INTRODUCTION

As a proverb said, “Do not put all your eggs into one basket”, the risk in the stock market can be reduced by holding a variety of stocks rather than owning a few or a single one. The purpose of stock portfolio management is to select an appropriate set of stocks and to compute the portion of the budget allocated to each stock so as to meet the investors’ objectives (return and risk) and economic constraints (liquidity, tax treatment, and unique circumstances).

Modern portfolio theory, established by Nobel laureate Markowitz [1], [2] in 1952, is the core of portfolio management and has been widely used in practice in finance. The Markowitz model takes the variance as the risk and assumes that rational investors are risk averse, which means individuals prefer less risk to more risk. The goal of the Markowitz model is to seek a trade-off between return and risk, i.e., maximizing the expected return for a given level of risk or minimizing the risk for a certain level of expected return. Based on this modern portfolio theory, researchers like Sharpe [3], Sengupta [4], and Stone [5], etc. developed some other schemes for portfolio selection. The Sharpe ratio (also known as reward-to-variability ratio) [3], proposed by Sharpe in 1966, characterizes how well one will be compensated if he bears more risk. It is broadly adopted as a portfolio selection strategy by many financial analysts for its simple and intuitive meaning. In this paper, besides the Markowitz model, we would also employ the Sharpe ratio [3] as our measurement criterion.

Basically, the classical Markowitz model can be reformulated as a quadratic programming problem, and the solutions can be found by the Critical Line Method (CLM) [1]. However, in real operations, there are many other practical constraints, such as cardinality, transaction costs, round-lot, etc. In this paper, we consider the cardinality constraints together with the Markowitz model. Taking into account the administrative costs, we usually limit the total number of stocks in a portfolio. Thus, the Markowitz model with the cardinality constraints will be reduced to a mix-integer quadratic programming problem. This is an NP-hard [6] problem and its optimal solution is computationally intractable when the number of stocks is large. Alternatively, some metaheuristic-based methods which can obtain approximate solutions in a reasonable time have been applied. These metaheuristics include Genetic Algorithm (GA) [7], [8], Simulated Annealing (SA) [9], Particle Swarm Optimization (PSO) [10], Ant Colony Optimization (ACO) [11], etc. However, each of these metaheuristics has its own drawbacks: (1) For GA, since many chromosomes are coded into a similar portfolio or similar chromosomes have very different portfolios, the efficiency is quite low; (2) SA evolves with only one solution and thus will easily get stuck in a local optimum when the search space is large and rugged; (3) For PSO, its application to portfolio selection is still limited, and only employed to find one optimal solution under the criterion of Sharpe ratio, rather than determining the whole Pareto frontier; (4) For ACO, though it can obtain the whole Pareto frontier, points are concentrated on the upper part of the frontier, where both return and risk are high.

Chemical Reaction Optimization (CRO) [12] is an evolutionary metaheuristic approach, motivated by the molecules’ energy exchange in a chemical reaction. Despite being a relatively new evolutionary algorithm, CRO has been shown to enjoy the advantages of both GA and SA and to have a more flexible structure [13]. Moreover, CRO has already demonstrated its excellent performance in handling problems like Quadratic Assignment Problem (QAP), Resource-Constrained Project Scheduling Problem (RCPSP), Channel Assignment Problem (CAP) in wireless mesh networks [12], Grid Scheduling Problem (GSP) [14], and Population Transition Problem in Peer-to-Peer live streaming [15], etc. In this paper, we propose a new CRO with the super molecule scheme (S-CRO), together with the canonical CRO. Both of them are applied to solve the stock portfolio selection problem.

The remainder of this paper is organized as follows. We introduce both the Markowitz model and the Sharpe ratio in Section II, and both of them are adjusted to fit the
application of CRO. Section III firstly describes the basic idea and framework of the canonical CRO, and then gives a detailed illustration of our proposed S-CRO. Computational experiments and results are shown in Section IV. Finally, the last section contains a summary of our work and topics for future investigation.

II. THE OPTIMIZATION MODEL

A. The Markowitz mean-variance model

According to modern portfolio theory, the investors are rational which means they are risk-opposing. Moreover, depending on their own economic conditions, people have different levels of tolerance to risk. In the Markowitz model, the return on a portfolio is calculated by the expected value of the portfolio return, and the corresponding risk is quantified by the variance of the portfolio return. Markowitz assumes that the aim of the investors is to determine a set of portfolio which can minimize the risk while fulfilling a predetermined expected return. Mathematically, the standard Markowitz mean-variance model can be formulated as follows [1]:

\[
\text{min. } \sigma_p^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} \omega_i \omega_j \text{Cov}_{i,j} \tag{1}
\]

subject to: \(E_p = \sum_{i=1}^{N} \omega_i R_i \geq E_{pre}\), \(\sum_{i=1}^{N} \omega_i = 1\), \(\omega_i \geq 0\) for \(i = 1, 2, ..., N\), where \(N\) is the number of available stocks and \(\omega_i\) is the proportion of capital assigned to Stock \(i\). For each \(i\), \(R_i\) is its expected return in a given time period. \(\text{Cov}_{i,j}\) represents the covariance between stocks \(i\) and \(j\), and when \(i\) equals \(j\) in (1), \(\text{Cov}_{i,j}\) becomes the variance of \(i\). \(\sigma_p^2\) and \(E_p\) stand for the variance and expected return of the portfolio, respectively. Constraint (2) guarantees the expected return of the portfolio will not be less than a predefined value \(E_{pre}\). The weights sum to one as shown in Constraint (3), while Constraint (4) implies shot selling is not allowed.

Strictly speaking, with enough computer power, we can solve the optimization model by computer. However, the degree of an investor’s risk aversion is difficult to be quantified since it also relies on many factors, such as investor’s age, family situation, current cash reserves, insurance coverage, etc. Thus, in most situations, we also need to calculate the Sharpe ratio as a reference for investors. The Sharpe ratio is used to measure the risk-adjusted performance of a portfolio:

\[
\text{max } SR = \frac{E_p - R_f}{\sigma_p}, \tag{9}
\]

where \(R_f\) is the risk-free return (like treasury bond rate) and it can be regarded as a constant, and \(\sigma_p\) is the standard deviation of the portfolio return. In order to convert it to a minimization problem, we reconstruct it as:

\[
\text{min } S = 1 - \frac{E_p - R_f}{\sigma_p}. \tag{10}
\]
In the following sections, the proposed CRO approaches will be employed to determine the generalized Markowitz efficient frontier as well as compute the greatest Sharpe ratio.

III. THE PROPOSED ALGORITHMS

A. The canonical CRO

Detailed discussions of Chemical Reaction Optimization (CRO) can be found in [12]. Here we only give a brief description of this method.

Chemical Reaction Optimization is a new population-based strategy used to approximate optimal solutions to optimization and search problems. The underlying idea of this approach arises from an analogy with the chemical reaction process. By following the phenomenon that products are always more stable than the reactants, molecules are inclined to stay at the most stable energy state through a sequence of intermediate changes. Similarly, solutions in CRO tend to reach the global minimum by performing predefined elementary reactions.

Each molecule (solution) is characterized by attributes, such as potential energy \( (PE) \), kinetic energy \( (KE) \), number of hits, minimum structure. Among these attributes, \( PE \) and \( KE \) correspond to the objective function value and the ability to accept worse solution, respectively, and the others are used in the selection of elementary reactions. Moreover, the chemical reaction is assumed to take place in a closed container, and there are four kinds of elementary reactions including on-wall ineffective collision, decomposition, inter-molecular ineffective collision, and synthesis. The former two involve only one molecule which collides with the wall of container, while the latter two involve more than one molecule (usually two) that interact with each other. In addition, the number of molecule(s) remains constant in the two ineffective collisions (i.e. on-wall ineffective collision and inter-molecular ineffective collision) and only the neighborhoods of original solution are searched. For the other two elementary reactions, one molecule is divided into several in the decomposition, while synthesis combines many molecules into one. These two reactions generate new solutions very different from the original ones and they help the algorithm jump out of the local optimums.

More precisely, the steps of canonical CRO are implemented as follows for the stock portfolio selection problem:

1) Randomly generate a population of initial molecules (solutions), calculate each solution’s objective function value as its \( PE \), and initialize each molecule’s other attributes.

2) Until a stopping criterion is met, do:
   i) According to the parameter \( MoleColl \in [0, 1] \), randomly choose one molecule or two molecules from the population.
   ii) Based on the decomposition criterion \( \alpha \) or synthesis criterion \( \beta \), select one of the four elementary reactions.
   iii) Generate the new molecule(s) according to the corresponding reaction scheme.

   iv) If there is enough energy\(^1\) for the new molecule(s) to be generated, replace the original molecule(s) with the new one(s), and update the relevant \( KE \).
   v) Else maintain the original molecule(s)

3) Output the global minimum solution and its corresponding values.

B. The super molecule-based CRO

One advantage of CRO over other metaheuristic methods is its flexible structure which can be easily adjusted to fit the problem. We can reconstruct the CRO process by choosing different combination of elementary reactions. In our proposed super molecule-based CRO (S-CRO) as shown in Fig. 2, the main body of the algorithm can be divided into three stages:

1) The S-CRO evolves with only two elementary reactions, i.e. on-wall ineffective collision and inter-molecular ineffective collision. This ensures the number of molecules remains the same, and the goal is to make the molecules explore as much as possible the solution space in their initial solutions’ neighborhoods.

2) Analyze the characteristics of all the molecules resulted from the previous stage, and then produce a super molecule based on that.

3) The super molecule is added to the container, and together with the molecules from the first stage, performs canonical CRO. The only difference is that the super molecule will not participate in decomposition and synthesis reactions. The main purpose is to prevent the super molecule from changing dramatically, which may destroy its good quality inherited from Stage 2.

For stock portfolio selection, which is a mixed-integer quadratic programming problem, we need not only to select a mixture from a huge number of stocks, but also determine the proportion for each chosen stock. Thus, in our algorithms, two vectors are used to represent the solution. The stock vector is employed to denote the selected stocks, while the proportion vector depicts the corresponding percentage of invested capital. Moreover, the schemes for the four elementary reactions are listed for reference (Suppose there are 20 stocks to choose from, and the cardinality is set to 5).

- On-wall ineffective collision: one-weight change in the proportion vector. One element (bold) in the vector will be selected randomly. Then, a random real number generated in the range \([t, 1] \) (t is the step size) is added to it. Finally, we normalize the vector to make all the elements sum to 1.

\[
0.110, 0.250, 0.170, 0.330, 0.140 
\rightarrow 0.110, 0.280, 0.170, 0.330, 0.140 
\rightarrow 0.107, 0.272, 0.165, 0.320, 0.136
\]

- Decomposition: half-random in the stock vector [14].

Strings before decomposition
\[
\omega: [1, 6, 8, 15, 20]
\]

\(^1\)For example, on-wall ineffective collision happens when \( PE_\omega + KE_\omega \geq PE_{\omega'}, \) where \( \omega \) and \( \omega' \) represent the original and the new molecule’s structures, respectively. Readers can refer to [12] for the energy requirements of other reactions.
String after decomposition

\[ \omega'_1: [1, 4, 8, 10, 20] \]
\[ \omega'_2: [2, 6, 9, 15, 17] \]

- Inter-molecular ineffective collision: one-stock change in the stock vector \([14]\).
  \[ [1, 6, 8, 15, 20] \rightarrow [1, 4, 8, 15, 20] \]
- Synthesis: keep the same stocks of the two molecules in the new stock vector, and randomly generate the remaining.

Strings before synthesis

\[ \omega_1: [1, 6, 8, 15, 20] \]
\[ \omega_2: [3, 6, 8, 11, 20] \]

String after synthesis

\[ \omega': [2, 6, 8, 12, 20] \]

In our problem, the super molecule is produced via three steps: 1) calculate the frequency of each stock existing among the molecules left in the previous stage; 2) choose the most popular stocks for the stock vector of a super molecule; (3) compute the corresponding normalized proportion vector of the super molecule. To a certain extent, the super molecule is similar to the “elite” in genetic algorithms. However, the “elite” in GA is usually generated from two chromosomes, while the super molecule is based on all other molecules. Furthermore, for both canonical CRO and S-CRO, when drawing the Markowitz efficient frontier, we will choose 100 different values of the risk tolerance coefficient \(\gamma\) spaced evenly in the range \([0, 1]\).

IV. COMPUTATIONAL EXPERIMENTS

We test our algorithms on a public OR library maintained by Beasley \([16]\). The benchmark for the stock portfolio selection problem includes five sets of data, which are derived from Hang Seng Index in Hong Kong with 31 stocks, DAX 100 in Germany with 85 stocks, FTSE 100 in UK with 89 stocks, S&P 100 in USA with 98 stocks and Nikkei 225 in Japan with 225 stocks. These data record the weekly prices from March 1992 to September 1997, and the mean return and the covariance between stocks are publicly available at \([16]\).

The parameters for canonical CRO and S-CRO are shown in Table I, and they are coded in C++ and the simulations are implemented on a PC with Intel Core 2 Duo-E677@2.66Hz CPU and 2GB RAM. In addition, we will compare the canonical CRO and S-CRO with the unconstrained efficient frontiers, which are also provided by the benchmark.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Parameter</th>
<th>Assigned value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canonical CRO</td>
<td>(KE) loss rate</td>
<td>0.8</td>
</tr>
<tr>
<td>CRO</td>
<td>MoleColl</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>Initial KE</td>
<td>initial minimal fitness</td>
</tr>
<tr>
<td></td>
<td>Initial Energy</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Iteration Number</td>
<td>100000</td>
</tr>
<tr>
<td></td>
<td>Cardinality (M)</td>
<td>10</td>
</tr>
</tbody>
</table>

For CRO

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>First stage iterations</th>
<th>Assigned value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canonical CRO</td>
<td>500000</td>
<td></td>
</tr>
<tr>
<td>S-CRO</td>
<td>500000</td>
<td></td>
</tr>
</tbody>
</table>

The graphical results of the Markowitz efficient frontier for the canonical CRO and S-CRO are shown in Figs. 3, 4, 5, 6, and 7. It is clear that S-CRO is much better than canonical CRO in terms of closeness to the true Pareto frontier without constraint. In fact, S-CRO almost coincides with the true Pareto frontier. This also confirms the rule of thumb.
that putting a limited number of carefully chosen stocks in a portfolio can probably achieve better performance than working with all the stocks when taking into account the administrative costs. However, we should also note that the distribution of S-CRO points along the true Pareto frontier is not uniform. This is due to the choice of $\gamma$, and if we evaluate enough different values of $\gamma$, accurate Markowitz efficient frontier can be obtained. Thus, the S-CRO algorithm can be used as a tool to generate the efficient frontier for the investors.

In order to reduce the random impact from the algorithm itself, we repeat 50 times to calculate the Sharpe ratio and compute the average and standard deviation. The results of canonical CRO and S-CRO are listed in Table II. For all five sets of data, the performance of S-CRO is superior to that of canonical CRO in terms of Sharpe ratio, expected return and variance. Specifically, when the number of stocks is huge like the example of Nikkei 225, the advantage of S-CRO becomes more significant. Moreover, another merit of S-CRO is its smaller standard deviation. This is quite useful in practice since we do not need to repeat so many times to get a reliable good solution, which can save us a lot of time. From Table II, we can also observe that for each run, S-CRO consumes more CPU time than canonical CRO. This is mainly caused by the super molecule generation in S-CRO. However, the difference is not substantial, and it is worthwhile to spend a little more time to achieve much better solutions. Thus, it is a good idea for us to apply S-CRO to optimize the Sharpe ratio.

V. CONCLUSION

Stock portfolio selection is one of the most challenging problems in finance. We formulate it as a mix-integer quadratic programming problem with the cardinality constraint. It is NP-hard and the optimal solution is computationally intractable. CRO is a new metaheuristic inspired by the molecular evolution in a chemical reaction. This paper proposes a new CRO scheme (named S-CRO) by adding a super molecule to the evolutionary process. Accordingly, the structure of canonical CRO was tailored to fit S-CRO. Then, both S-CRO and canonical CRO have been tested and compared under five different scenarios in terms of the Markowitz efficient frontier and the Sharpe ratio. Simulation results show that our proposed S-CRO performs much better than the canonical CRO and achieves the Pareto frontier, demonstrating its power in solving the stock portfolio selection problem.

However, there are three things we need to note. Firstly, the Markowitz model has its own limitations. It assumes the stock return follows the normal distribution while many researchers argue that the returns in real stock market are asymmetrically distributed and other statistics (like skewness, kurtosis) should be considered. The second problem is that besides the cardinality constraint, there are also many other
TABLE II

<table>
<thead>
<tr>
<th>Index</th>
<th>SHARPE RATIO RESULTS</th>
<th>canonical CRO</th>
<th>S-CRO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hang Seng</td>
<td>Sharpe ratio (Avg)</td>
<td>0.1908629</td>
<td>0.2104100</td>
</tr>
<tr>
<td></td>
<td>Sharpe ratio (SD)</td>
<td>0.0126028</td>
<td>0.0000010</td>
</tr>
<tr>
<td></td>
<td>Expected return (Avg)</td>
<td>0.0070479</td>
<td>0.011402</td>
</tr>
<tr>
<td></td>
<td>Time(s) (Avg)</td>
<td>0.121</td>
<td>0.125</td>
</tr>
<tr>
<td>DAX 100</td>
<td>Sharpe ratio (Avg)</td>
<td>0.2561531</td>
<td>0.3615554</td>
</tr>
<tr>
<td></td>
<td>Sharpe ratio (SD)</td>
<td>0.0283834</td>
<td>0.0012485</td>
</tr>
<tr>
<td></td>
<td>Expected return (Avg)</td>
<td>0.0057178</td>
<td>0.0067509</td>
</tr>
<tr>
<td></td>
<td>Variance (Avg)</td>
<td>0.0005006</td>
<td>0.0003488</td>
</tr>
<tr>
<td></td>
<td>Time(s) (Avg)</td>
<td>0.124</td>
<td>0.133</td>
</tr>
<tr>
<td>FTSE 100</td>
<td>Sharpe ratio (Avg)</td>
<td>0.2411870</td>
<td>0.2936723</td>
</tr>
<tr>
<td></td>
<td>Sharpe ratio (SD)</td>
<td>0.0140641</td>
<td>0.0011946</td>
</tr>
<tr>
<td></td>
<td>Expected return (Avg)</td>
<td>0.0053986</td>
<td>0.0056605</td>
</tr>
<tr>
<td></td>
<td>Variance (Avg)</td>
<td>0.0005113</td>
<td>0.0003716</td>
</tr>
<tr>
<td></td>
<td>Time(s) (Avg)</td>
<td>0.122</td>
<td>0.139</td>
</tr>
<tr>
<td>S&amp;P 100</td>
<td>Sharpe ratio (Avg)</td>
<td>0.2547994</td>
<td>0.3101401</td>
</tr>
<tr>
<td></td>
<td>Sharpe ratio (SD)</td>
<td>0.0116441</td>
<td>0.0022049</td>
</tr>
<tr>
<td></td>
<td>Expected return (Avg)</td>
<td>0.0051833</td>
<td>0.0056382</td>
</tr>
<tr>
<td></td>
<td>Variance (Avg)</td>
<td>0.0004170</td>
<td>0.0003311</td>
</tr>
<tr>
<td></td>
<td>Time(s) (Avg)</td>
<td>0.123</td>
<td>0.134</td>
</tr>
<tr>
<td>Nikkei 225</td>
<td>Sharpe ratio (Avg)</td>
<td>0.0760307</td>
<td>0.1391573</td>
</tr>
<tr>
<td></td>
<td>Sharpe ratio (SD)</td>
<td>0.0309694</td>
<td>0.0012423</td>
</tr>
<tr>
<td></td>
<td>Expected return (Avg)</td>
<td>0.0025319</td>
<td>0.0034318</td>
</tr>
<tr>
<td></td>
<td>Variance (Avg)</td>
<td>0.0011628</td>
<td>0.0006083</td>
</tr>
<tr>
<td></td>
<td>Time(s) (Avg)</td>
<td>0.121</td>
<td>0.142</td>
</tr>
</tbody>
</table>

practical constraints, such as transaction costs, round lot, tax concerns, etc. They can be quite different in different countries. Finally, the analysis of historical price for a stock does not mean we can predict its future trend precisely because the stock market is very complicated. Therefore, our future work will refine the stock portfolio selection model by considering additional market indicators.

ACKNOWLEDGMENT

This research is supported in part by the University of Hong Kong Strategic Research Theme of Information Technology. A. Y. S. Lam is also supported in part by the Croucher Foundation Research Fellowship.

REFERENCES