Neurocomputing 174 (2016) 494–499

Contents lists available at ScienceDirect

Neurocomputing

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

Biogeography-based optimization for identifying promising compounds in chemical process

Haiping Ma [a,b,*], Minrui Fei [a], Zhile Yang [c]

[a] Department of Electrical Engineering, Shaoxing University, Shaoxing, Zhejiang, China
[b] Shanghai Key Laboratory of Power Station Automation Technology, School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China
[c] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, UK

ARTICLE INFO

Article history:
Received 24 October 2014
Received in revised form
27 April 2015
Accepted 16 May 2015
Available online 31 August 2015

Keywords:
Promising compound
Chemical process
Biogeography-based optimization
Evolutionary algorithms

ABSTRACT

Identifying promising compounds from a vast collection of potential compounds is an important and yet challenging problem in chemical engineering. An efficient solution to this problem will help to reduce the expenditure at the early states of chemical process. In an attempt to solve this problem, the industry is looking for predictive tools that would be useful in testing optimal properties of a candidate compound earlier. This paper explores the application of biogeography-based optimization (BBO) to achieve such predictive work. BBO is a new evolutionary algorithm that is based on the science of biogeography. BBO is a population-based search method that achieves information sharing by species migration. The performance of BBO is compared with genetic algorithm (GA) and particle swarm optimization (PSO) on a set of test functions and the cases of identifying promising compounds. Simulation results show that BBO is a competitive method in determining an optimal solution to the optimization of promising compounds.

1. Introduction

Finding global optima of complex chemical processes with large search space is one of the primary goals in many scientific investigations. For example, in the realm of compound design, the chemical process approaches involve in solving global optimization problems [1–3]. Scientists can often produce a large number of compounds. However, only a few of them serve as good candidates for a potential compound. To find an effective compound or series of compounds, scientists explore a large number of new potential compounds, and then evaluate their efficacy and safety based on the clinical symptoms and experimental responses. Such chemical process takes many years and it is very costly. Recently, the technology, known as combinatorial chemistry, is now being widely applied in the chemical industry, and is getting interest in several areas of chemical engineering. The work focuses on tweaking the molecules to eliminate potential negative effects and to improve the molecule's ability to interact with the body. An important problem in combinatorial chemistry is how to design the experiments to explore and optimize the high-dimensional solution space while minimizing the number of trials to achieve an optimal solution. For example, sampling conformational space in order to determine molecular formulas of organic compounds involves in finding a global maximum of a fitting function [4]. Several biology-inspired evolutionary computing methods are used to solve such optimization problems because of their intuitive appeal and ability to solve hard optimization problems. The methods used for compound design include genetic programming (GP) [5] and genetic algorithm (GA) [6], which are based on the concept of evolving solutions from less accurate ones. Evolutionary algorithms frequently are used to a specific problem at hand, and they are well-suited for the optimization of promising compounds in chemical process.

Simon [7] initially proposed biogeography-based optimization (BBO) in 2008 which is a new evolutionary algorithm for global optimization. BBO is a population-based search algorithm based on the model of the immigration and emigration of species between habitats. In BBO, individuals, referred to as habitats, are changed through information sharing between candidate solutions. Migrations to solution features within the search space are based on the goodness of the solutions to optimize individuals. One distinctive feature of BBO is that the original population is not discarded after each generation. It is rather modified by migration. Another distinctive feature is that, for each generation, BBO uses the fitness of each solution to determine its immigration and emigration rates. BBO has demonstrated good performance on various single-objective benchmark functions [8–10].

* Corresponding author at: Department of Electrical Engineering, Shaoxing University, Shaoxing, Zhejiang, China.
E-mail address: Mhping1981@126.com (H. Ma).

http://dx.doi.org/10.1016/j.neucom.2015.05.125
0925-2312/

It has also been applied to real-world optimization problems, including sensor selection [7], power system optimization [11], groundwater detection [12] and satellite image classification [13]. In addition, a web-based BBO graphical user interface is used in the reference [14].

The remainder of the paper is organized as follows. Section 2 explains the optimization model of promising compounds in chemical process. Section 3 describes the BBO algorithm and lists the pseudo-code required for its implementation. In Section 4, the results of simulation experiments are given and the performance of the algorithms is compared. The concluding remarks and topics for future investigations are given in Section 5.

2. Combinatorial compound model

Identifying promising compounds from a vast collection of feasible compounds is a challenging problem in chemical process. A feasible compound involves in attaching reagents to each of several locations along a core molecule. In a typical compound, each attachment location may have tens or hundreds of potential reagents, and as the numbers of locations and reagents increase, the number of compounds increases exponentially. Because of the vast number of potential compounds, the entire compound library is rarely created in practice. For example, consider the feasible compound presented in Fig. 1, where three positions require additions. These three positions of the core molecule are denoted by A, B and C. The different reagents can be added to each of those positions. If each position has ten possible reagents denoted by x_i, that is, $x_A = x_B = x_C = 10$, there are 10^3 possible compounds, among which, we are constrained to create only a fraction in the laboratory.

Indeed, the optimization methods could be applied to this problem, and the goal is to obtain sets of reagents that maximize the target efficacy of a compound, which is measured by its pre-specified chemical property. From an experimenter's viewpoint, the chemical property of interest of a compound can be viewed as the response of the process, which is to be minimized. That is, the response of interest $y(x)$ is the desired chemical property of a compound with reagents specified by $x = (x_1, x_2, \ldots, x_d)$, where d is the number of independent locations. The response surface is likely to have various shapes. Some parts of the surface are smooth, other parts are filled with local optima, and others have unexpected extreme peaks of activity. Hence, some random search methods (e.g., genetic algorithms (GA), particle swarm optimization (PSO)) are best to find a global optimum on such response in which there are many local optima. Although we only consider scalar response y in this paper, multi-objective response (e.g., optimization of more than one property) can also be accommodated by modeling the desired function of the compounds. On the other hand, in practice the feasible compounds that satisfy the essential constraints are more important to the scientists. These constraints include some chemical properties [15]: chemical reactivity, occurrence of toxicological features, molecular weight, number of rotated bonds, and so on. So a more general viewpoint is to formulate the problem as a constrained optimization function, and the response of process is selected as the objective function and the other chemical properties are treated as constraints.

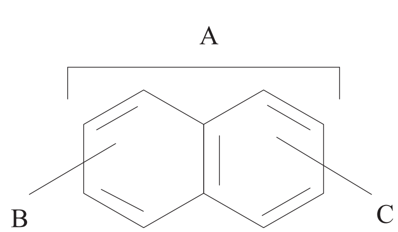

Fig. 1. The core molecule of a compound with three reagents locations.

3. Biogeography-based optimization

BBO is a new population-based global optimization algorithm [7]. As its name implies, BBO is based on the study of the distribution of species over time and space. This study, which is a subset of biology, is called biogeography [16]. Suppose that we have a global optimization problem and a population of candidate solutions (individuals). Each individual is considered to be analogous to a habitat and is characterized by a habitat suitability index (HSI). The value of HSI, which is the same as fitness in other population-based optimization algorithms, and which measures the goodness of the solution, depends on many features of the habitat. A habitat with a high-HSI is a good solution, and a habitat with a low-HSI is a poor solution. High-HSI solutions tend to share their features with low-HSI solutions by emigrating solution features to other habitats. Low-HSI solutions accept a lot of new features from high-HSI solutions by immigration from other habitats. Immigration and emigration tend to improve the solutions and thus evolve a solution to the optimization problem. Namely, BBO views the value of HSI as the objective function, and the evolution procedure of BBO is to determine the solutions which maximize the HSI by immigrating and emigrating features of the habitats. In BBO, there are two main operators: migration (which includes both emigration and immigration) and mutation.

Migration is a probabilistic operator that adjusts a habitat H. We use the migration rates of each habitat to probabilistically share features between habitats. The probability that H_i is modified is proportional to its immigration rate λ_i, and the probability that the source of the modification comes from H_j is proportional to the emigration rate μ_j. Migration is defined by

$$H_i(\mathrm{SIV}) \leftarrow H_j(\mathrm{SIV}) \tag{1}$$

In biogeography, an SIV is a suitability index variable which characterizes the habitability of a region [7], that is, the HSI is determined by many SIVs. In BBO, an SIV is a solution feature, equivalent to a gene in GA. In other words, an SIV is a search variable and the set of all possible SIVs is the search space from which an optimal solution will be determined.

In BBO, each solution H_i has its own immigration rate λ_i and emigration rate μ_i. They can be calculated as

$$\lambda_i = I\left(1 - \frac{k(i)}{n}\right)$$
$$\mu_i = E\left(\frac{k(i)}{n}\right) \tag{2}$$

where I is the maximum possible immigration rate, E is the maximum possible emigration rate, $k(i)$ is the fitness rank of the ith individual (1 is worst and n is best), and n is the number of solutions in the population. I and E are often set to 1, or slightly less than 1. Eq. (2) indicates that a good solution has relatively high μ and low λ, while the converse is true for a poor solution. So the immigration rate and the emigration rate are functions of the fitness of the solution.

Because migration relies entirely on the quality of existing solutions as well as migration topology, BBO will converge to the global optimum if and only if the population already contains features of the optimal solution. If one feature of the optimal solution is missing from the population, it is impossible for BBO to