Information Sciences 418–419 (2017) 561–574

Contents lists available at ScienceDirect

Information Sciences

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

A return-cost-based binary firefly algorithm for feature selection

Yong Zhang[a], Xian-fang Song[a], Dun-wei Gong[a,b,*]

[a] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China
[b] School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, Shandong, 266061, China

ARTICLE INFO

Article history:
Received 24 October 2016
Revised 9 August 2017
Accepted 11 August 2017
Available online 12 August 2017

Keywords:
Firefly algorithm
Feature selection
Return-cost
Pareto dominance
Binary movement

ABSTRACT

Various real-world applications can be formulated as feature selection problems, which have been known to be NP-hard. In this paper, we propose an effective feature selection method based on firefly algorithm (FFA), called return-cost-based binary FFA (Rc-BBFA). The proposed method has the capability of preventing premature convergence and is particularly efficient attributed to the following three aspects. An indicator based on the return-cost is first defined to measure a firefly's attractiveness from other fireflies. Then, a Pareto dominance-based strategy is presented to seek the attractive one for each firefly. Finally, a binary movement operator based on the return-cost attractiveness and the adaptive jump is developed to update the position of a firefly. The experimental results on a series of public datasets show that the proposed method is competitive in comparison with other feature selection algorithms, including the traditional algorithms, the GA-based algorithm, the PSO-based algorithm, and the FFA-based algorithms.

1. Introduction

Various real-world applications in pattern recognition and machine learning involve more and more attributes (features) as the capabilities in acquiring and storing information increase. Among these features, many may be irrelevant or/and redundant, given the fact that it is generally difficult to determine which one is useful without any prior knowledge [19,29]. On this circumstance, feature selection (FS) becomes very important. Feature selection is a process of selecting a subset formed by the most relevant features from the original feature set, and the selected subset should be necessary and sufficient to describe a target concept, with maintaining a high accuracy in representing the original feature set [40].

There have been a variety of methods to tackle FS problems, which can be generally classified into the following three categories, i.e., the filter, the wrapper, and the hybrid approaches [24,42]. For the first category, the filter approach first ranks all the features according to a series of criteria, and then omits those with high ranks. Given the fact that it is computation-efficient [17], this approach is very popular for feature selection problems with high dimensions. Representative filter methods include minimum-redundancy maximum-relevancy (mRMR) [32], F-score criterion [11], Gini index [30], and correlation coefficient [27] among many others. With respect to the wrapper approach, it generally involves a learning algorithm determined in advance, which is evaluated by the selected feature subset [20]. Compared with the filter approach which is independent of any learning algorithm, the wrapper approach usually has better performances for most cases [42].

* Corresponding author.
E-mail addresses: yongzh401@126.com (Y. Zhang), songxf0614@126.com (X.-f. Song), dwgong@vip.163.com (D.-w. Gong).

http://dx.doi.org/10.1016/j.ins.2017.08.047
0020-0255/

Regarding the last category, the hybrid approach attempts to combine the advantages of both the filter and the wrapper approaches. This study belongs to the wrapper approach.

Along the line of the wrapper approach, there have been a variety of methods to seek an optimal feature subset, and the representatives include the complete search, the greedy search, the heuristic search, and the random search [9,42]. Among these methods, most, however, suffer from stagnation in local optima and/or a high computational cost [36]. Recently, nature-inspired algorithms have received much attention on solving FS problems, due to their capabilities in seeking competitive solutions using strategies with a good performance in exploration [10,15]. These approaches include genetic algorithm (GA) [28,31], memetic algorithm (MA) [23,50], differential evolution [1], ant colony optimization (ACO) [35,38], bee colony optimization (BCO) [43,51], gravitational search algorithm [18], flower pollination algorithm [12], bat algorithm [34], and particle swarm optimization (PSO) [40,41,49]. A detailed review of nature-inspired feature selection approaches can be found in [10], and a survey on evolutionary computation approaches to feature selection can be found in [42].

In recent years, Yang developed a new nature-inspired algorithm, known as firefly algorithm (FFA), to tackle continuous optimization problems [44]. Later, FFA has been extended to solve multimodal optimization problems [45], discrete optimization problems [25,33], dynamic and uncertain optimization problems [2], and multi-objective optimization problems [46]. The applications of FFA in binary optimization problems are, however, very limited. Recently, a binary firefly algorithm (BFFA) [48] was proposed to solve FS problems, in which the transformation from a real value to a binary number is achieved by the probability calculated based on the sigmoidal function. Chandrasekaran et al. proposed another function, called Tanh, to be used in the transformation [4]. However, developing appropriate transformation functions is still a challenging and open research problem. To this end, Vandana and Surekha proposed an integer encoding strategy, where a firefly is encoded with the order number of features [37]. In this method, a firefly is regarded as a point which can move in a D-dimensional space, where D is the total number of features in a dataset. In other words, a firefly is represented as a vector, $(x_1, x_2, \cdots, x_D)$, where $x_i \in \{1, 2, \cdots, D\}$. Taking a dataset with four features as an example (i.e. $D = 4$), a firefly of (1,4,4,1) means that the first and the fourth features are selected. Compared with the binary strategy, the value range of each dimension of a firefly in the integer encoding strategy increases from $\{0, 1\}$ to $\{1, 2, \cdots, D\}$, which results in an increased complexity in seeking an optimal feature subset.

The current FFA has the following limitations when tackling FS problems:

- In the standard FFA algorithm, a firefly moves toward all the other fireflies that have higher brightness when updating its position. When there is more than one brighter firefly located in the same region, the firefly will search the region completely, which results in the waste of computation resource.
- When calculating the firefly's attractiveness value to a brighter one, currently only their distance is considered. The return from the brighter one is, however, not taken in account. Since the attractiveness value is inversely proportional to the distance, without the return-cost considered, a firefly that has excellent brightness but is far from the population will have a low opportunity to reproduction; on the contrary, the one that has inferior brightness but is close to the population will have a high chance to reproduction. As a result, the current mechanism will result in a low efficiency in searching optimal regions.
- Further, there are many control parameters of FFA, such as the attractiveness and the light absorption coefficients, having a great influence on the search behavior of a firefly [46], which needs a lot of efforts to tune them.

Motivated by the above analysis, we propose a return-cost-based binary firefly algorithm (Rc-BBFA) in this paper to improve the capability of FFA in tackling feature selection problems, where the main contributions are summarized below.

- Instead of the distance-based attractiveness, an indicator based on the return-cost is defined to measure the firefly's attractiveness, which is determined automatically by using both the return and the cost from brighter fireflies.
- A Pareto dominance-based strategy is proposed to seek the attractive one for each firefly. Taking the ith firefly as an example, different from the traditional strategy that takes all the brighter ones as its attractive fireflies, the proposed Rc-BBFA selects only one firefly with good values of the cost and the return as its attractive firefly.
- A new binary movement operator based on the return-cost attractiveness and the adaptive jump is developed to update the position of a firefly. Since this operator needs only one control parameter, namely the jump probability, the proposed algorithm is much easier to tune than current FFA.
- Finally, a comprehensive comparison study between Rc-BBFA and the other four algorithms for feature selection problems is conducted, namely the traditional algorithms, the GA-based algorithm, the PSO-based algorithm, and the FFA-based algorithms.

This paper is organized as follows. A preliminary knowledge is provided in Section 2. The proposed algorithm is detailed in Section 3. The efficiency of the proposed algorithm is demonstrated via experimental results in Section 4. Finally, the paper is concluded in Section 5.