Applied Soft Computing 31 (2015) 293–307

Contents lists available at ScienceDirect

Applied Soft Computing

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

A novel hybrid system for feature selection based on an improved gravitational search algorithm and *k*-NN method

Jie Xiang [b], XiaoHong Han [a,*], Fu Duan [b], Yan Qiang [b], XiaoYan Xiong [a], Yuan Lan [a], Haishui Chai [b]

[a] *Key Laboratory of Advanced Transducers and Intelligent Control Systems, Ministry of Education China, Taiyuan University of Technology, Taiyuan, Shanxi, People's Republic of China*

[b] *College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan, Shanxi, People's Republic of China*

ARTICLE INFO

Article history:
Received 15 March 2013
Received in revised form 12 January 2015
Accepted 23 January 2015
Available online 10 February 2015

Keywords:
Feature selection
Binary classification
Gravitational search algorithm
k-nearest neighbor
Leave-one-out cross-validation
Function optimization

ABSTRACT

Feature selection is an important pre-processing step for solving classification problems. This problem is often solved by applying evolutionary algorithms in order to decrease the dimensional number of features involved. In this paper, we propose a novel hybrid system to improve classification accuracy with an appropriate feature subset in binary problems based on an improved gravitational search algorithm. This algorithm makes the best of ergodicity of piecewise linear chaotic map to explore the global search and utilizes the sequential quadratic programming to accelerate the local search. We evaluate the proposed hybrid system on several UCI machine learning benchmark examples, comparing our approaches with feature selection techniques and obtained better predictions with consistently fewer relevant features. Furthermore, the improved gravitational search algorithm is tested on 23 nonlinear benchmark functions and compared with 5 other heuristic algorithms. The obtained results confirm the high performance of the improved gravitational search algorithm in solving function optimization problems.

1. Introduction

In many fields such as pattern recognition [31], data mining [7], gene selection from microarray data [12], text categorization [35] and multimedia information retrieval [22,26], datasets containing huge numbers of features are often involved. In such cases, feature selection will be necessary. Due to the abundance of noisy, irrelevant or misleading features, the ability to handle imprecise and inconsistent information in real world problems has become one of the most important requirements for feature selection [29].

Feature selection is a process of choosing a subset of features from the original set of features forming patterns in a given dataset. The subset should be necessary and sufficient to describe target concepts, retaining a suitably high accuracy in representing the original features. The importance of feature selection is to reduce the problem size and search space for learning algorithms. In the design of pattern classifiers it can improve the quality and speed of classification. In other words, the objective of feature subset selection is to reduce the number of features used to characterize a dataset so as to improve a learning algorithm's performance on a given task. The maximization of the classification accuracy in a specific task for a certain learning algorithm is the ultimate objective.

Most existing feature selection algorithms consist of four basic issues that determine the nature of the search process: a starting point in the search space, an organization of the search, an evaluation strategy of the feature subsets, and a criterion for halting the search. The search method is in charge of identifying promising candidate subsets, while the evaluator must assess the goodness of a given subset. Evaluators can be of two types: filter or wrapper [6,25]. Filter techniques evaluate the goodness of an attribute or set of attributes by using only intrinsic properties of the data. On the other hand, wrapper algorithms use a classifiers to assess the quality of the attribute subset proposed as candidate by the search algorithm. Wrappers usually perform better than filters but with the disadvantage of being more time-consuming [21].

Basically, the problem of feature selection could be treated as a problem of optimization in a search space. Feature selection method based on stochastic search algorithms has attracted great attention. Several methods have been proposed to perform feature selection using evolutionary techniques. Raymer and Punch [40] suggested using Genetic Algorithms (GA) to tackle the problem. Authors in [4,45,48] proposed to use binary Particle Swarm Optimization (PSO) for feature selection. Zhang and Sun applied tabu search in this problem [53].

* Corresponding author. Tel.: +86 13485325973.
E-mail address: jmqchs@sohu.com (X. Han).

http://dx.doi.org/10.1016/j.asoc.2015.01.043
1568-4946/

Recently, a new population-based heuristic algorithm, namely gravitational search algorithm (GSA), based on the metaphor of gravitational interaction between masses was introduced mainly designed for problems in electrical engineering [38]. In this algorithm, the searcher agents are a collection of masses which interact with each other based on the Newtonian gravity and the laws of motion. A comprehensive comparison between GSA and some well-known heuristic algorithms such as GA and PSO was presented by Rashedi et al. Their results indicated that in optimization field, the GSA approach has some merit over other methods [38,39].

However, there are two major issues associated with the search performance of GSA: one is premature convergence happening in existing GSA due to rapid reduction of diversity; the other is that GSA converges rapidly in the beginning of the search process while slows down quickly when the global best solution is near the optimum of the local search space. Consequently, to get a more accurate estimation of the local optima, more ineffective iterations are needed; in addition, it is hard to have a good balance between exploration and exploitation. In order to overcome these drawbacks, researchers have made much work to improve the performance of GSA. Sarafrazi et al. [41] introduced an operator called "Disruption" originating from astrophysics to improve the exploration and exploitation abilities of the original GSA. Li and Zhou [23] strengthened the algorithm by combination of the search strategy of particle swarm optimization and applied it to the solution of optimization problems in parameters identification of hydraulic turbine governing system. Binod Shaw et al. [43] proposed an opposition-based GSA (OGSA) which improves the convergence rate of the GSA by utilizing opposition-based learning for population initialization and also for generation jumping. Yin et al. [52] integrated two strategies, the maximum and minimum of the dth dimension and a new $G(t)$, into the original GSA to add the objects' diversity and make them explore more solution spaces. In this paper, we propose an improved gravitational search algorithm (IGSA) by making the best of ergodicity of PWL to help GSA explore the global search while employing the SQP to accelerate the local search.

Since the IGSA is operated in continuous space and many optimization problems are set in binary space, we introduce a binary IGSA (BIGSA) referring to the original BGSA introduced by Rashedi et al. [39]. Based on the BIGSA, a new method for feature selection wrapped by k-nearest neighbor (k-NN) method is presented, in which the k-NN method based on Euclidean distance calculations serves as a classifier for evaluating classification accuracies [33]. We apply the proposed method to several UCI machine learning benchmark examples.

The remainder of this paper is organized as follows: In Section 2 we provide a brief summary of GSA. Section 3 explains the IGSA. Section 4 presents the proposed hybrid system for feature selection. The results of the work are presented and discussed in Section 5. Finally, we provide conclusions in Section 6.

2. Gravitational search algorithm (GSA)

GSA is a newly developed stochastic search algorithm based on the law of gravity and mass interactions [36–38]. This approach provides an iterative method that simulates mass interactions, and moves through a multi-dimensional search space under the influence of gravitation. In GSA, agents are considered as objects and their performances are measured by their masses; these objects attract each other by gravitational force which causes a global movement of all objects toward objects with heavier masses [36–38].

Assumed there are k objects (masses), the position of the ith object is defined as Eq. (1):

$$X_i = (x_i^1, \ldots, x_i^d, \ldots, x_i^n), i = 1, 2, \ldots, k, \tag{1}$$

where x_i^d denotes the position of ith object in the dth direction. The force exerting on the object i from the object j is defined as Eq. (2):

$$F_{ij}^d(t) = G(t)\frac{M_i(t) \times M_j(t)}{R_{ij}(t) + \varepsilon}(x_j^d(t) - x_i^d(t)) \tag{2}$$

where M_j is the mass related to object j, M_i is the mass related to object i, $G(t)$ is gravitational constant at time t, ε is a small constant, and $R_{ij}(t)$ is the Euclidian distance between two objects i and j. The total force $F_i^d(t)$ that exerts on object i in the dth direction is a randomly weighted sum of dth components of the forces from other agents:

$$F_i^d(t) = \sum_{j=1, j \neq i}^{k} rand_j F_{ij}^d(t), \tag{3}$$

where $rand_j$ is a uniform random variable in the interval [0,1].

The acceleration of the object i, $a_i^d(t)$, at time t and in the dth direction, is given as Eq. (4):

$$a_i^d(t) = \frac{F_i^d(t)}{M_{ii}(t)} \tag{4}$$

where M_{ii} is the inertial mass of the object i. Its next velocity $v_i^d(t+1)$ and its next position $x_i^d(t+1)$ are calculated as Eqs. (5) and (6):

$$v_i^d(t+1) = rand_i \times v_i^d(t) + a_i^d(t) \tag{5}$$

$$x_i^d(t+1) = x_i^d(t) + v_i^d(t+1) \tag{6}$$

where $rand_i$ is a uniform random variable in the interval [0,1]. This random number is applied to give a randomized characteristic to the search, $v_i^d(t)$ and $x_i^d(t)$ are its current velocity and position, respectively.

The masses of objects are evaluated by the fitness function. Assuming the equality of the gravitational and inertia mass, the mass $M_i(t)$ is updated by Eqs. (8)–(10) and (11):

$$M_i = M_{ii}, i = 1, 2, \ldots, k, \tag{7}$$

$$m_i(t) = \frac{fit_i(t) - worst(t)}{best(t) - worst(t)}, \tag{8}$$

$$M_i(t) = \frac{m_i(t)}{\sum_{j=1}^{k} m_j(t)}, \tag{9}$$

$$best(t) = \min_{j \in \{1,\ldots,k\}} fit_j(t), \tag{10}$$

$$worst(t) = \max_{j \in \{1,\ldots,k\}} fit_j(t), \tag{11}$$

where $fit_i(t)$ represents the fitness value of the object i at time t. The flow chart of GSA is shown in Fig. 1.

3. Improved gravitational search algorithm (IGSA)

3.1. PWL with dynamic search range

The piecewise linear chaotic map (PWL) has been attached increasing attention in being applied to optimization areas due to its efficiency in implementation, simplicity in representation and