Applied Soft Computing 18 (2014) 261–276

Contents lists available at ScienceDirect

Applied Soft Computing

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

Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms

Bing Xue*, Mengjie Zhang, Will N. Browne

School of Engineering and Computer Science, Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand

ARTICLE INFO

Article history:
Received 31 December 2012
Received in revised form 30 July 2013
Accepted 25 September 2013
Available online 27 October 2013

Keywords:
Particle swarm optimisation
Feature selection
Classification

ABSTRACT

In classification, feature selection is an important data pre-processing technique, but it is a difficult problem due mainly to the large search space. Particle swarm optimisation (PSO) is an efficient evolutionary computation technique. However, the traditional personal best and global best updating mechanism in PSO limits its performance for feature selection and the potential of PSO for feature selection has not been fully investigated. This paper proposes three new initialisation strategies and three new personal best and global best updating mechanisms in PSO to develop novel feature selection approaches with the goals of maximising the classification performance, minimising the number of features and reducing the computational time. The proposed initialisation strategies and updating mechanisms are compared with the traditional initialisation and the traditional updating mechanism. Meanwhile, the most promising initialisation strategy and updating mechanism are combined to form a new approach (PSO(4-2)) to address feature selection problems and it is compared with two traditional feature selection methods and two PSO based methods. Experiments on twenty benchmark datasets show that PSO with the new initialisation strategies and/or the new updating mechanisms can automatically evolve a feature subset with a smaller number of features and higher classification performance than using all features. PSO(4-2) outperforms the two traditional methods and two PSO based algorithm in terms of the computational time, the number of features and the classification performance. The superior performance of this algorithm is due mainly to both the proposed initialisation strategy, which aims to take the advantages of both the forward selection and backward selection to decrease the number of features and the computational time, and the new updating mechanism, which can overcome the limitations of traditional updating mechanisms by taking the number of features into account, which reduces the number of features and the computational time.

1. Introduction

In classification problems, a dataset usually involves a large number of features, often including relevant, irrelevant and redundant features. However, irrelevant and redundant features are not useful for classification and they may even reduce the classification performance due to the large search space, which is termed “the curse of dimensionality” [1,2]. Feature selection is proposed to select a subset of relevant features from a large number of available features to achieve similar or even better classification performance than using all features [2]. By eliminating/reducing irrelevant and redundant features, feature selection could reduce the number of features, shorten the training time, simplify the learned classifiers, and/or improve the classification performance [2]. Existing feature selection algorithms can be broadly classified into two categories [3,4]: filter approaches and wrapper approaches. Filter approaches are independent of a learning algorithm and they are argued to be computationally cheaper and more general than wrappers. Wrapper approaches include a learning algorithm as part of the evaluation function. Therefore, wrappers can often achieve better results than filter approaches.

Feature selection is a difficult combinatorial problem. The best feature subset is usually a group of features with the presence of feature complementarity because of the feature interaction problem. There could be two-way or multi-way interactions among features [1,5]. As a result, an individually relevant feature may become redundant when working together with other features so that eliminating some such features will remove or reduce unnecessary complexity. On the other hand, an individually redundant or weakly relevant feature may become highly relevant when working with others. Therefore, an optimal feature subset should be a group of complementary features. The feature selection task is challenging due mainly to the large search space. The size of the search space increases exponentially with respect to the number

* Corresponding author. Tel.: +64 220327481.
E-mail address: Bing.Xue@ecs.vuw.ac.nz (B. Xue).

1568-4946/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2013.09.018

of available features in the dataset [1]. Therefore, an exhaustive search is practically impossible in most situations. Although many different search techniques have been applied to feature selection, most of these algorithms still suffer from the problems of stagnation in local optima or being computationally expensive [2,1,4]. In order to better address feature selection problems, an efficient global search technique is needed.

Evolutionary computation (EC) techniques are well-known for their global search ability. Particle swarm optimisation (PSO) [6,7] is a relatively recent EC technique, which is computationally less expensive than some other EC algorithms. Therefore, PSO has been used as an effective technique in feature selection [8,4]. However, there are a number of limitations about current PSO for feature selection. Firstly, PSO has not been tuned to the feature selection task. Gutierrez et al. [9] show that initialisation strategies in PSO perform differently in different problems with high dimensional search spaces. However, no existing initialisation strategies are specifically proposed for feature selection problems except our previous work [10]. Secondly, the traditional personal and global best updating mechanism may miss some feature subsets with high classification performance, but a small number of features (detailed discussions in Section 3.2). Therefore, the potential of PSO for feature selection has not been fully investigated and we will continue our previous work [10] to further study the initialisation and the updating mechanism in PSO for feature selection.

1.1. Goals

The overall goal of this paper is to propose a new PSO based feature selection approach to selecting a smaller number of features and achieving similar or even better classification performance than using all features and traditional/existing feature selection methods. In order to achieve this goal, we propose *three new initialisation strategies*, which are motivated by forward selection and backward selection, and *three new personal and global best updating mechanisms*, which consider both the number of feature and the classification performance to overcome the limitation of the traditional updating mechanism. Specifically, we will:

- propose new initialisation strategies in PSO to reduce the number of features without decreasing the classification performance of the feature subsets achieved by using traditional initialisation strategy,
- develop new updating mechanisms in PSO to guide the algorithm to search for the feature subsets with high classification performance and a smaller number of features, and to outperform the traditional updating mechanism,
- develop a new PSO based feature selection algorithm using one of the proposed initialisation strategies and one of the proposed updating mechanisms, and
- investigate whether the proposed feature selection algorithm can obtain a feature subset with a smaller number of features and better classification performance than using all features, and outperform two traditional feature selection methods, the standard PSO based algorithm, and a PSO based algorithm using a single fitness function combining both the number of features and the classification performance.

1.2. Organisation

The remainder of the paper is organised as follows. Section 2 provides background information. Section 3 describes the proposed new initialisation strategies, the proposed personal best and global best updating mechanisms, and the pseudo-code of the proposed algorithm. Section 4 describes experimental design and Section 5 presents experimental results with discussions. Section 6 provides conclusions and future work.

2. Background

2.1. Particle swarm optimisation (PSO)

PSO is an EC technique proposed by Kennedy and Eberhart in 1995 [6,7]. PSO simulates the social behaviour such as birds flocking and fish schooling. In PSO, a population, also called a *swarm*, of candidate solutions are encoded as particles in the search space. PSO starts with the random initialisation of a population of particles. Particles move in the search space to search for the optimal solution by updating the position of each particle based on the experience of its own and its neighbouring particles. During the movement, the current position of particle i is represented by a vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$, where D is the dimensionality of the search space. The velocity of particle i is represented as $v_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$, which is limited by a predefined maximum velocity, v_{max} and $v_{id}^t \in [-v_{max}, v_{max}]$. The best previous position of a particle is recorded as the personal best called *pbest* and the best position obtained by the swarm so far is the global best called *gbest*. PSO searches for the optimal solution by updating the position and the velocity of each particle according to the following equations:

$$x_{id}^{t+1} = x_{id}^t + v_{id}^{t+1} \tag{1}$$

$$v_{id}^{t+1} = w \times v_{id}^t + c_1 \times r_{1i} \times (p_{id} - x_{id}^t) + c_2 \times r_{2i} \times (p_{gd} - x_{id}^t) \tag{2}$$

where t denotes the tth iteration in the evolutionary process. $d \in D$ denotes the dth dimension in the search space. w is inertia weight. c_1 and c_2 are acceleration constants. r_{1i} and r_{2i} are random values uniformly distributed in [0, 1]. p_{id} and p_{gd} represent the elements of *pbest* and *gbest* in the dth dimension.

2.2. Related work on feature selection

(1)*TraditionalFeatureSelectionApproaches*

Relief [11] is a filter feature selection algorithm, which assigns a weight to each feature to denote the relevance of the feature to the target concept. However, Relief does not deal with redundant features because it attempts to find all relevant features regardless of the redundancy between them. FOCUS [12], also a filter algorithm, exhaustively examines all possible feature subsets, then selects the smallest feature subset. However, the FOCUS algorithm is computationally inefficient because of the exhaustive search.

Greedy search based sequential forward selection (SFS) [13] and sequential backward selection (SBS) [14] are two typical wrapper methods. SFS (SBS) starts with no features (all features), then candidate features are sequentially added to (removed from) the subset until the further addition (removal) does not increase the classification performance. However, these two methods suffer from the problem of so-called nesting effect, which means once a feature is selected (eliminated) it cannot be eliminated (selected) later. This problem can be solved by combining both SFS and SBS into one algorithm. Therefore, Stearns [15] proposes a "plus-l-take away-r" method, which performs l times forward selection followed by r times backward elimination. However, it is difficult to determine the optimal values of (l, r).

(2)*EvolutionaryComputationTechniquesforFeatureSelection*

EC techniques have been applied to feature selection problems, such as genetic algorithms (GAs), genetic programming (GP), ant colony optimisation (ACO) and PSO.

Zhu et al. [16] propose a feature selection method using a memetic algorithm that is a combination of local search and GA. In this algorithm, individual features are firstly ranked according to a filter measure. GA employs the classification accuracy as the fitness