Applied Soft Computing xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Applied Soft Computing

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

Feature selection with modified lion's algorithms and support vector machine for high-dimensional data

Kuan-Cheng Lin [a], Jason C. Hung [b,*], Jhen-ting Wei [a]

[a] Department of Management Information Systems, National Chung Hsing University, 250 Kuo Kuang Rd., Taichung 402, Taiwan
[b] Department of Information Technology, Overseas Chinese University, Taichung 40721, Taiwan

ARTICLE INFO

Article history:
Received 26 January 2017
Received in revised form 7 January 2018
Accepted 9 January 2018
Available online xxx

Keywords:
Feature selection
High-dimensional data
Lion's algorithm
Support vector machine

ABSTRACT

The lion's algorithm is a novel evolutionary algorithm designed to imitate the behavior observed in a pride of lions. This study solves the classification problem by employing the lion's algorithm for the selection of feature subsets in high dimensional data. The proposed feature selection process identifies and removes irrelevant/redundant features to reduce data dimensionality and thereby improve the efficiency and accuracy of classification. We devised three versions of the lion's algorithm in which greedy search was applied to the territorial defense strategy and/or territorial takeover strategy. Experiments using datasets in the UCI machine learning database demonstrate the superiority of the modified versions over the original algorithm. Ultimately, the approach involving the application of greedy search to territorial defense proved the most effective.

1. Introduction

Information technology is being used in a wide variety of surprising applications. The use of computer algorithms to identify diseases [1] is an example of data classification via machine learning. Briefly, a training dataset (previously classified data) is used to establish a classifier used to identify the class of unknown data. Numerous classification techniques have been developed, such as artificial neural networks [2], Bayes classifiers [3], and support vector machines (SVM).

Classification algorithms often have difficulty dealing with data sets that include a large number of features, which greatly increase the temporal and spatial complexity. Many of the features in an input data set are irrelevant or redundant and should therefore be eliminated. Feature selection is the process of identifying the subset of features that would allow the classifier to perform most effectively. Numerous previous studies [4–6] have coupled classification algorithms with feature selection methods based on global search methods, such as evolutionary algorithms.

Evolutionary computation [7] is a universal meta-heuristic algorithm used to resolve optimization problems. Many of these systems are modeled on biological mechanisms, such as genetic algorithms [8] and particle swarm algorithms [9]. Swarm intelligence uses population-based searches to solve optimization problems. The lion's algorithm was developed in 2012 as an evolutionary algorithm based on swarm intelligence. This algorithm has been shown to outperform genetic algorithms in all benchmark functions [10]. In a previous study [11], the lion's algorithm outperformed the genetic algorithm when implemented as a wrapper feature selection method and support vector machine classifier for problems of classification. In this study, we illustrate modified versions of the lion's algorithm that enable it to outperform the original.

This paper is organized as follows: Section 2 presents a short introduction to feature selection, support vector machines, and the lion's algorithm. Section 3 describes the modified lion's algorithm. Section 4 outlines the experiment used to evaluate the proposed algorithm and our results. Conclusions are drawn in Section 5.

2. Background

In this section, we review previous research efforts into feature selection, support vector machines, and the lion's algorithm.

2.1. Feature selection

High dimensional data can greatly undermine the efficiency of machine learning. Feature selection makes it possible to reduce the

* Corresponding author.
E-mail addresses: kclin@nchu.edu.tw (K.-C. Lin), jhung@ocu.edu.tw (J.C. Hung), p4585424@gmail.com (J.-t. Wei).

https://doi.org/10.1016/j.asoc.2018.01.011
1568-4946/

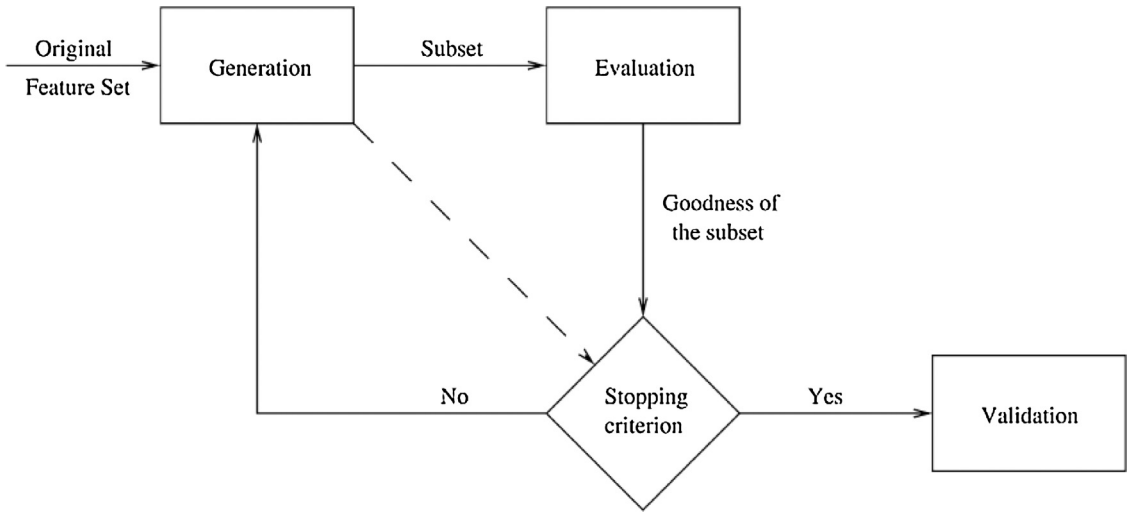

Fig. 1. Feature selection followed by validation.
(Cited from Manoranjan Dash, Huan Liu, 1997).

number of unnecessary features in order to improve classification accuracy, reduce error rates, and improve efficiency (Fig. 1).

Feature selection is the process of identifying the optimal subset of features according to a given evaluation criteria; however, the computational overhead tends to be very high [12]. Other approaches based on heuristics or random search methods have been developed to reduce computational complexity by compromising optimality. A typical feature selection method includes four basic components:

(1) Generation of candidate subsets
(2) Evaluation of candidate subsets
(3) Stopping criterion
(4) Validating the validity to identified subset

Despite the various methods used in feature selection, the basic steps of generation, evaluation, and stopping are typical of nearly all methods.

2.2. Support vector machines

In 1995, Vapnik [13] proposed the support vector machine (SVM), a supervised learning model based on structural risk minimization. SVM has been widely applied in the fields of classification and regression. SVM is based on statistical learning theory in which two types of data are divided up on a hyper-plane with the maximize possible separation. For data located in n-dimensional space, we use the following formula for the hyper-plane:

$$w \cdot x + b = 0 \quad (1)$$

where w is the vector of the data dimension, b is a constant representing the degree of bias, the purpose of which is to ensure that the hyper-plane falls within the correct position space after a horizontal movement. SVM classification involves using the hyper-plane as a decision function, as follows:

$$f(x) = \mathrm{sgn}(w \cdot x + b) \quad (2)$$

When each data x is substituted into the decision function (2), the result is only {+1, −1}. If the result is +1, the data x is belongs to a class that is positive, −1 represents a class that is negative. Hence, all data can be divided into "positive" and "negative" categories.

A suitable hyper-plane situates the two categories at a maximal distance. Based on the separation of categories, hyper-plane in Fig. 2(b) is superior to that in Fig. 2(a).

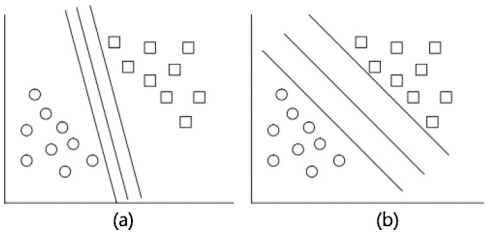

Fig. 2. (a) Non-optimal hyper-plane; (b) optimal hyper-plane.

Table 1
Elements used in the lion's algorithm.

Element	Role
pride	Group of lions
$\mathcal{L}^{male}$	Lion
$\mathcal{L}^{female}$	Lioness
$\mathcal{L}^{cub}$	Cub generated by lion and lioness
$\mathcal{L}^{nomad}$	Nomadic lion

Previous studies have demonstrated that SVM classifiers combined with a bio-inspired feature selection mechanism, such as a cat swarm optimization [4,5] or school of fish [6] outperform those without feature selection. In this study, we did not make changes to the SVM mechanism, but rather used SVM as a black box to generate a classification model for feature selection.

2.3. Lion's algorithm

The lion's algorithm was developed by Rajakumar in 2012 [14], and adopted for standard- and large-scale bilinear systems in 2014 [15]. Lions live in a grouping called a pride with one male leader. Male cubs live in their birth pride until they reach early adulthood, whereupon they leave the pride to wander alone as nomadic lions. If during their wandering, a nomadic male encounters a new pride, it may challenge the leader for dominance. In the event that the nomadic male wins this encounter, it becomes the new leader of the pride. In the lion's algorithm, each lion represents a solution. As shown in Fig. 3 and Table 1, the algorithm proceeds through four basic steps: pride generation, mating, territorial defense and territorial takeover.

Please cite this article in press as: K.-C. Lin, et al., Feature selection with modified lion's algorithms and support vector machine for high-dimensional data, Appl. Soft Comput. J. (2018), https://doi.org/10.1016/j.asoc.2018.01.011