Computers and Electrical Engineering 47 (2015) 114–125

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

A two-stage feature selection method with its application

Xuehua Zhao [a], Daoliang Li [b,*], Bo Yang [c], Huiling Chen [d], Xinbin Yang [a], Chenglong Yu [a], Shuangyin Liu [e]

[a] School of Digital Media, Shenzhen Institute of Information Technology, Shenzhen 518172, China
[b] College of Information and Electrical Engineering, China Agricultural University, P. O. Box 121, 17 Tsinghua East Road, Beijing 100083, China
[c] College of Computer Science and Technology, Jilin University, Changchun 130012, China
[d] College of Physics and Electronic Information, Wenzhou University, Wenzhou 325035, China
[e] College of Information, Guangdong Ocean University, Zhanjiang 524025, China

ARTICLE INFO

Article history:
Received 22 April 2014
Revised 17 August 2015
Accepted 24 August 2015
Available online 24 September 2015

Keywords:
Foreign fibers
Feature selection
Information gain
Binary particle swarm optimization

ABSTRACT

Foreign fibers in cotton seriously affect the quality of cotton products. Online detection systems of foreign fibers based on machine vision are the efficient tools to minimize the harmful effects of foreign fibers. The optimum feature set with small size and high accuracy can efficiently improve the performance of online detection systems. To find the optimal feature sets, a two-stage feature selection algorithm combining IG (Information Gain) approach and BPSO (Binary Particle Swarm Optimization) is proposed for foreign fiber data. In the first stage, IG approach is used to filter noisy features, and the BPSO uses the classifier accuracy as a fitness function to select the highly discriminating features in the second stage. The proposed algorithm is tested on foreign fiber dataset. The experimental results show that the proposed algorithm can efficiently find the feature subsets with smaller size and higher accuracy than other algorithms.

1. Introduction

Foreign fibers in cotton refer to non-cotton fibers and dyed fibers, such as hairs, binding ropes, plastic films, candy wrappers, and polypropylene twines, etc. The foreign fibers in cotton, especially in lint, will seriously affect the quality of the final cotton textile products, even low content of foreign fibers in cotton [1,2].

To reduce the harm of foreign fiber in cotton textile products, online detection systems based on machine vision have been studied for evaluating the quality of the cotton in recent years [2–5]. In such systems, the classification of foreign fibers in cotton is the basic and key technology, which is closely related to system's performance. To improve the accuracy of classification and efficiency of systems, finding the optimum feature sets with the small size and high accuracy is essential because they not only simplify the design of classifier, but also improve the efficiency of online detection. It is a feature selection (FS) problem in nature.

FS is the technique of selecting a subset of relevant features for building robust learning models, which is commonly used in machine learning. FS aims at simplifying a feature set by reducing its dimensionality and identifying relevant underlying features without sacrificing predictive accuracy [6]. Unfortunately, finding the optimal feature subsets has been proved to be NP-hard [7] so that a number of FS algorithms are proposed to find the near optimal solutions in smaller amount of time. These methods are generally divided into three categories: the filter approach, the wrapper approach and the embedded method. In the first

[*] Corresponding author. Tel.:+86 10 62736764; fax: +86 10 62737741.
E-mail address: dliangl@cau.edu.cn, hdlsyxlq@126.com (D. Li).

http://dx.doi.org/10.1016/j.compeleceng.2015.08.011
0045-7906/© 2015 Elsevier Ltd. All rights reserved.

category, the filter approaches are first utilized to select the subsets of features before the learning algorithms are applied. On the other hand, the wrapper approaches [8] utilize the learning algorithms as a fitness function and search for the best subsets of features in the space of all feature subsets. Besides the filters and wrappers, the embedded methods incorporate variable selection as a part of the training process, and feature relevance is obtained analytically from the objective of the learning model [9].

1.1. Related work

Currently, several FS approaches have been applied to select feature sets of foreign fibers in cotton for online detection.

Yang et al. [10] proposed an improved genetic algorithm by which the optimal feature subset can be selected effectively and efficiently from a multi-character feature set. The algorithm adopted the segmented chromosome management scheme to implement local management of chromosome, and can obtain strong searching ability at the beginning of the evolution and achieved accelerated convergence along the evolution. Zhao et al. [11] proposed an FS algorithm that used ant colony optimization to select the feature subsets of foreign fibers in cotton. To improve the efficiency of ant colony optimization, Zhao et al. [12] proposed an improved ant colony optimization for feature selection, whose objective is to find the (near) optimal subsets in multi-character feature sets. In the algorithm, group constraint is adopted to limit subset constructing process and probability transition for reducing the effect of invalid subsets and improve the convergence efficiency. Li et al [13] used PSO (particle swarm optimization) to select the optimal feature sets of foreign fibers in cotton.

However, these algorithms belong to wrapper approaches, their advantages include the interaction between feature subset search and model selection, and the ability to take into account feature dependencies. A common drawback of these algorithms is that they have a higher risk of overfitting than filter techniques and are very computationally intensive.

1.2. Motivation and contribution

Filter approaches assess the relevance of features by looking only at the intrinsic properties of the data. In most cases a feature relevance score is calculated, and low-scoring features are removed. Afterwards, this subset of features is presented as input to the classification algorithm. Advantages of filter approaches are that they easily scale to very high-dimensional datasets due to their simplicity and fastness, and are independent of the classification algorithm. As a result, feature selection only need to be performed only once, and then different classifiers can be evaluated. A common disadvantage of filter approaches is that they ignore the interaction with the classifier, and that the most approaches are univariate. This means that each feature is considered separately, thereby ignoring feature dependencies, which may lead to worse classification performance when compared to other types of feature selection approaches.

Whereas filter approaches are independent of the model hypothesis, wrapper methods embed the model hypothesis search within the feature subset search. In this setup, a search procedure in the space of possible feature subsets is defined, and various subsets of features are generated and evaluated. The evaluation of a specific subset of features is obtained by training and testing a specific classification model, rendering this approach tailored to a specific classification algorithm. To search the space of all feature subsets, a search algorithm is then 'wrapped' around the classification model. However, as the space of feature subsets grows exponentially with the number of features, heuristic search methods are used to guide the search for an optimal subset. Advantages of wrapper approaches include the interaction between feature subset search and model selection, and the ability to take into account feature dependencies. A common drawback of these techniques is that they have a higher risk of overfitting than filter techniques and are very computationally intensive.

In this paper, we proposed a two-stage feature selection algorithm by combining the IG (Information Gain) filter approach and BPSO (Binary Particle Swarm Optimization) wrapper approach for foreign fiber data. IG method selects the features with higher scores. BPSO, when combined with a learning algorithm, is a successful wrapper but computationally expensive. The integration of IG and BPSO thus leads to an effective feature selection scheme. In the first stage of our algorithm, the IG is applied to find a candidate feature set of foreign fibers. This filters out many unimportant features and reduces the computational load for BPSO. In the second stage, BPSO wrapper is applied to directly select the highest discriminative feature subset from the candidate set. We perform the comprehensive experiments to validate the efficiency of the algorithms on foreign fiber datasets. The experimental results show that the proposed algorithm is very effective for the data of foreign fibers in cotton.

The rest of the paper is organized as follows: Section 2 introduces some basic concepts of classification and FS. Section 3 describes the proposed algorithm. Experimental results are presented in Section 4. The last section draws a general conclusion.

2. Preliminaries

In this section, we introduce the basic notions of classification and FS.

2.1. Classification

Machine learning is usually divided into two main types: supervised learning and unsupervised learning. In the supervised learning approach, the goal is to learn a mapping from inputs x to outputs y, given a labeled set of input–output pairs $D = \{x_i, y_i\}_{i=1}^{n}$. Here D is called the training set, and N is the number of training examples.