Available at www.sciencedirect.com

INFORMATION PROCESSING IN AGRICULTURE XXX (2018) XXX–XXX

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

Efficient detection method for foreign fibers in cotton

Xuehua Zhao [a], *Xiangyun Guo* [b,*], *Jie Luo* [a], *Xu Tan* [c]

[a] School of Digital Media, Shenzhen Institute of Information Technology, Shenzhen 518172, China
[b] School of Information Management, Beijing Information Science & Techonology University, Beijing 100192, China
[c] School of Software Engineering, Shenzhen Institute of Information Technology, Shenzhen 518172, China

A R T I C L E I N F O

Article history:
Received 28 September 2017
Received in revised form
10 April 2018
Accepted 13 April 2018
Available online xxxx

Keywords:
Classification
Feature selection
Foreign fibers in cotton
Detection system

A B S T R A C T

Since foreign fibers in cotton seriously affect the quality of the final cotton textile products, machine-vision-based detection systems for foreign fibers in cotton are receiving extensive attention in industrial equipment. As one of the key components in detection systems, the suitable and good classifier is significantly important for machine-vision-based on detection systems for foreign fibers in cotton due to it improving the system's performance. In the study, we test five classifiers in the dataset of foreign fibers in cotton, and for finding the best feature set corresponding to the classifiers, we use the four filter feature selection approaches to find the best feature sets of foreign fibers in cotton corresponding to specific classifiers. The experimental results show that the extreme learning machine and kernel support vector machines have the excellent performance for foreign fiber detection and the classification accuracy are respectively 93.61% and 93.17% using the selected corresponding feature set with 42 and 52 features.

1. Introduction

Foreign fibers in cotton refer to non-cotton fibers and dyed fibers, such as hair, binding ropes, plastic films, candy wrappers, and polypropylene twines, etc. which are mixed with cotton during picking, storing, drying, transporting, purchasing, and processing. Foreign fibers in cotton could seriously affect the quality of the final cotton textile products [1]. Until now, the machine-vision-based detection systems have been studied to evaluate the quality of the cotton [2–4]. In such systems, the classifiers are the basic and key component which is closely related to system's performance. Currently, many classifiers have been proposed to apply in all kinds of applications, the representative classifiers are as follows: (1) k-nearest neighbor classifier (kNN) [5], the kNN is the simple and rather trivial classifiers, but it is lazy learners, that is, building the model is cheap, but classifying unknown objects is relatively expensive. (2) Support vector machines (SVM) [6], the SVM offers one of the most robust and accurate methods among all well-known algorithms. It has a sound theoretical foundation, requires only a dozen examples for training, and is insensitive to the number of dimensions. But the SVM only classifies the linearly separable data, for the unlinearly separable data, the kernel SVM (KSVM) [7] is more suitable. (3) Naive Bayesian Classifier (NBC) [8], the NBC is very easy to construct, not needing any complicated iterative parameter estimation schemes. This means it may be readily applied to huge data sets. The extreme learning machine

* Corresponding author.
E-mail address: guo-xiangyun@163.com (X. Guo).
Peer review under responsibility of China Agricultural University.
https://doi.org/10.1016/j.inpa.2018.04.002

(ELM) is a novel machine learning algorithm for single hidden layer feedforward neural networks (SLFNs) [9]. Currently, due to its learning parameters determined randomly, fast training speed and excellent generalization performance, the ELM has drawn increased attention and has been applied in many fields like image recognition [10], fault diagnosis [11], text classification [12], and so on. However, the ELM still has not been applied to foreign fiber detection.

For the classifiers, it is important to find the feature set with the great distinguishing ability because it can efficiently improve the performance of classification. Feature Selection (FS) is the technique of selecting subsets of relevant features which aims at simplifying a feature set by reducing its dimensionality and identifying relevant features without sacrificing predictive accuracy [13]. Unfortunately, finding the optimal feature subset has been proved to be a NP-hard problem, so a number of FS algorithms are proposed to look for the near optimal solutions [14,15]. FS algorithms generally fall into three categories: filter, wrapper and embedded models [16]. The filter approaches are computationally efficient and are preferable for high-dimensional databases because they do not involve a learning machine.

In this paper, five classifiers are presented for foreign fiber detection and comparisons are made to find the suitable classifier for detection systems of foreign fibers in cotton. In addition, to improve the performance of detection systems, four filter FS methods are presented to find the feature set with the great distinguishing ability. The main objective of this study is to find the suitable classifier and corresponding feature set for improving the performance of detection systems of foreign fiber in cotton.

2. Methods

For determining the suitable classifier and best feature set, we test five classifiers and four filter FS approaches in the data set of foreign fiber in cotton based on the framework in Fig. 1. In this framework, the FS approaches are used to find the good feature subset, the classifiers are used to evaluate the quality of feature subset. Finally, we can find the most suitable classifier and corresponding feature set by our framework.

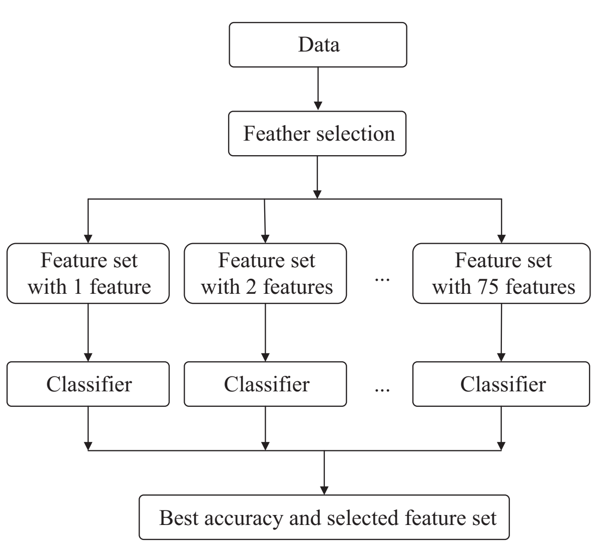

Fig. 1 – Flow graph of methods.

For classifiers, we select five classifiers which are the ELM, kNN, SVM, KSVM and NBC, respectively. The reasons that five classifiers are selected in our experiments are as follows: (1) The ELM is a learning algorithm for single hidden layer feed-forward neural networks which has the high efficiency and unification of classification and regression. (2) The kNN is the simple and efficient classifier which is used in various classify tasks. (3) The SVM is a powerful classification algorithm that has shown the excellent performance in a variety of classification tasks. (4) The KSVM has the excellent performance in the un-linear data by tracing maximum margin hyperplanes in the Kernel space where samples are mapped. (5) The NBC is a probabilistic classifier based on the assumption of conditional independence among the predictive attributes given the class. For filter FS approaches [16], we select fisher for feature selection (FisherFS), reliefF for feature selection (ReliefFS), Chi-square for feature selection (ChiFS), Gini index for feature selection (GiniFS). Next, we briefly introduce these classifiers and FS approaches.

2.1. Clssifiers

2.1.1. ELM

The ELM is first designed for single hidden layer feed-forward neural networks and then extended to generalized single hidden layer feed-forward networks (SLFNs) which did not necessarily resemble neurons. Different from traditional neural SLFN learning algorithms, ELM aims to minimize both training error and the norm of output weights. Due to its high efficiency and unification of classification and regression, ELM has been an active research topic over the past a few years and has been successfully applied to many applications [10–12].

Given a training set $D = \{x_i, t_i | i = 1, 2, \ldots, N\}$, where x_i is the $n \times 1$ input feature vector and t_i is a $m \times 1$ target vector. The standard SLFNs which has an activation function $g(x)$, and the number of hidden neurons $\tilde{N}$ can be mathematically modeled as follows:

$$\sum_{i=1}^{\tilde{N}} \beta_i g(w_i \cdot x_i + b_i) = o_j, j = 1, 2, \ldots, N \quad (1)$$

where β_i is the weight vector connecting the i-th hidden neuron and the output neurons, w_i is the weight vector connecting the i-th neuron and the input neurons, b_i is the bias of the i-th hidden neuron, and o_j is the output of the j-th input data. If SLFNs can approximate these N samples with zero error, we will have $\sum_{j=1}^{N} \|o_j - t_j\| = 0$, i.e., there exist β_i, w_i, b_j such that $\sum_{j=1}^{\tilde{N}} \beta_i g(w_i \cdot x_j + b_j) = t_j, j = 1, 2, \ldots, N$. The above equation can be written compactly as:

$$H\beta = T \quad (2)$$

where H is the hidden layer output matrix:

$$H(w_1, \ldots, w_{\tilde{N}}, b_1, \ldots, b_{\tilde{N}}, x_1, \ldots, x_N) = \begin{pmatrix} g(w_1 \cdot x_1 + b_1) & \cdots & g(w_{\tilde{N}} \cdot x_1 + b_{\tilde{N}}) \\ \vdots & \ddots & \vdots \\ g(w_1 \cdot x_N + b_1) & \cdots & g(w_{\tilde{N}} \cdot x_N + b_{\tilde{N}}) \end{pmatrix}_{N \times \tilde{N}} \quad (3)$$

Please cite this article in press as: Zhao X et al. Efficient detection method for foreign fibers in cotton. Info Proc Agri (2018), https://doi.org/10.1016/j.inpa.2018.04.002