Feature Selection for Cotton Matter Classification

Xuehua Zhao, Ying Huang, Zhao Li, Shukai Wu, Xiuhong Ma, Hua Chen, Xu Tan

Feature Selection for Cotton Matter Classification

Xuehua Zhao[1], Ying Huang[1], Zhao Li[1], Shukai Wu[1], Xiuhong Ma [1], Hua Chen[1], Xu Tan[1] (✉)

[1]Shenzhen Institute of Information Technology, Shenzhen 518172, China
lcrlc@sina.com,Tanx@sziit.edu.cn

Abstract. Feature selection are highly important to improve the classification accuracy of recognition systems for foreign matter in cotton. To address this problem, this paper presents six filter approaches of feature selection for obtaining the good feature combination with high classification accuracy and small size, and make comparisons using support vector machine and k-nearest neighbor classifier. The result shows that filter approach can efficiently find the good feature sets with high classification accuracy and small size, and the selected feature sets can effectively improve the performane of recognition system for foreign matter in cotton. The selected feature combination has smaller size and higher accuracy than original feature combination. It is important for developing the recognition systems for cotton matter using machine vision technology.

Keywords: Filter approaches, foreign matter, classification.

Introduction

Foreign matter in cotton including hair, plastic films, polypropylene twines and so on, seriously damage the quality of cotton [1, 2]. Currently, recognition systems based on machine vision are an affective approach to detect the foreign matter in cotton [2]. For recognition systems, to determine an good feature combination with high classification accuracy and small size is very important to improve the performance of classifier.

Feature Selection (FS) is finding optimal feature subsets by reducing the useless features without sacrificing predictive accuracy [4]. However, this is difficult because it is NP-hard problem, so that lots of FS methods are used to find the near optimal feature subsets, such as filter approaches [5], branch and bound algorithm [6], sequential forward/backward search [7], metaheuristic-based algorithms [8]. Currently, there are three kinds of FS methods: filter methods, wrapper methods and embedded models. The filter methods does not involve any prediction model. the wapper methods need use prediction model to evaluate the performce of selected feature sets. The methods with embedded model consider feature selection into the training process, and learning model is used to evaluate the relevance between features. The wrapper mothds have high time complexity and the feature sets only are good for the corresponding prediction model. In comparison, the filter methods are computationally efficient.

* Corresponding author; E-mail: Tanx@sziit.edu.cn

adfa, p. 1, 2011.

In this paper, six filter FS methods are considered to determine the good feature combination of foreign matter in cotton for improving the classifiction accuracy of recognition systems and the feature sets selected by six methods are validated in dataset of cottom foreign matter using support vector machine (SVM for short) and k-nearest neighbor classifier (kNN for short). This study aims at determining the good feature combination to improve the classification accuracy of foreign matter in cotton. To determine the good feature combination, 75 features are extracted in cotton images to build the original feature vector and comparisons are made to determine the good feature combination. The results illustrates filter FS methods can efficiently determine the good feature combination with high classifiction and small size, and improve the classification accuracy of cotton foreign matter.

The remainder is organized as follows. Section 2 describes six filter FS appoaches. Section 3 presents the experiments. Section 4 describes the conclusions.

Methods

We describe the six filter FS methods [9], which include fisher FS method (FisherFS), reliefF FS (ReliefFS), Chi-square FS method (ChiFS), Gini index FS method (GiniFS), information gain FS method (IGFS) and Kruskal Wallis FS method (KruskalFS).

FisherFS. The FisherFS evaluates each feature according to the following criterion which is formulated as:

$$FisherFS(f_i) = \frac{\sum_{j=1}^{c} n_j (\mu_{ij} - \mu_i)^2}{\sum_{j=1}^{c} n_j \sigma_{ij}^{\ 2}} \tag{1}$$

where f_i is the feature, μ and σ respectively denotes the mean and variance, n denotes the number of samples, j is the class.

ReliefFS. The ReliefFS evaluates each feature according to the following criterion which is formulated as:

$$ReliefFS(f_i) = \frac{1}{p}\sum_{t=1}^{p}\left\{-\frac{\sum_{x_j \in NH(x_t)} d(f_{ti} - f_{ji})}{m_{x_t}} + \sum_{y \neq y_{x_t}} \frac{1}{m_{x_t y}} \frac{P(y)}{1-P(y)} \sum_{xj \in NM(x_t y)} d(f_{ti} - f_{ji})\right\} \tag{2}$$

where p is the number of samples, $d(\cdot)$ denotes the distance, f_{ti} is the value of feature f_i in sample x_t, yx_t denotes the class of the sample x_t, $P(y)$ denotes the probability of sample. $NH(x)$ is a set of nearest sample to $\mathbf{x}$, m_{xt} denotes the sizes of $NH(x_t)$, $m_{xt,y}$ denotes the size of $NM(x_t, y)$.

ChiFS. The ChiFS evaluates each feature according to the following criterion which is formulated as:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{C}\frac{(n_{ij}-\mu_{ij})^2}{\mu_{ij}} \tag{3}$$

$$\mu_{ij} = \frac{n_{*j}n_{i*}}{n} \tag{4}$$

where r denotes the different values of one feature, C denotes the classes, n_{ij}, n_{i*}, n_{*j} and n respectively denote the number for instances with the i-th value, the i-th value, in the *j-th* class and instances.

GiniFS. GiniFS evaluates each feature according to the following criterion which is formulated as:

$$GiniFS(f_j) = 1 - \sum_{i=1}^{C}[p(i \mid f)]^2 \tag{5}$$

where C denotes the class set. Smaller the value of the GiniFS for one feature is, more relevant the feature is.

IGFS. IGFS evaluates each feature according to the following criterion which is formulated as:

$$IGFS(X,Y) = -\sum_i P(x_i)\log_2(P(x_i)) + \sum_j P(y_j)\sum_i P(x_i \mid y_j)\log_2(P(x_i \mid y_j)) \tag{6}$$

where X denotes the feature and Y denotes the class labels, $H(\cdot)$ denotes the *entropy*, $H(X|Y)$ denotes the *entropy* of X after observing Y.

KWFS. KWFS evaluates each feature according to the following criterion which is formulated as:

$$K = N - 1\frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i} n_i(r_{ij} - \bar{r})^2} \tag{7}$$

where n_i denotes the number of samples included in class i, r_{ij} denotes the rank of sample j in the class I, N denotes the number of samples included in all classes,

Experiments

Data Preparation

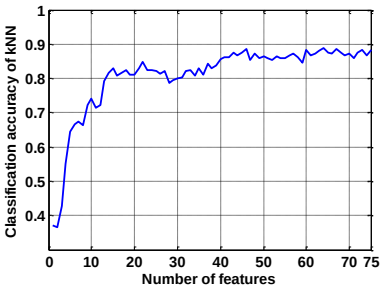
The data is obtained according the following steps: (1) 1800 images containing foreign fibers first are collected. (2) These images are divided into six classes: black plastic film, polypropylene, hemp rope, cloth, hair and feather, every class includes 300 images. (3) These images are segmented and the foreign matter objects are generated, the number of hair, polypropylene, film, rope, cloth and black plastic objects is 210, 720, 422541,, 504, 395, respectively. (4) The features are extracted in these objects. In this study, the number of the extracted features is 75, the 75-dimensional feature vector is built to denote the samples of cotton foreign matter.

Experimental Set

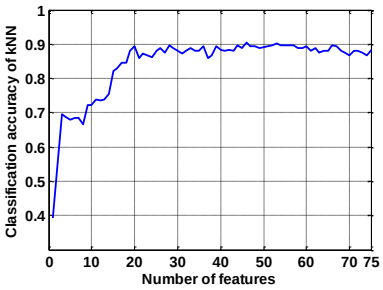
In our experiments, Lenovo personal computer with Windows 7, Intel i5, 8.0GB main memory is adopted, six algorithms are coded by the Matlab 2010b, kNN [10] and SVM [11] are selected as classifiers, the cross validation is used to evaluate of the algorithms.

Results and Discussion

Fig.1 shows the performance with kNN of the selected feature combination from six methods. For six methods, the accuracy is improved when the size of feature combination increases, the curve start smooth until the size of feature combination arrives a specific value. For example, the number of features is more than 20 for ReliefFS, 30 for GiniFS and 30 for KruskalFS. This indicates the original feature set includes the redundant and irrelevant features and the good feature combination can found by six methods.



(a) FisherFS



(b) ReliefFS

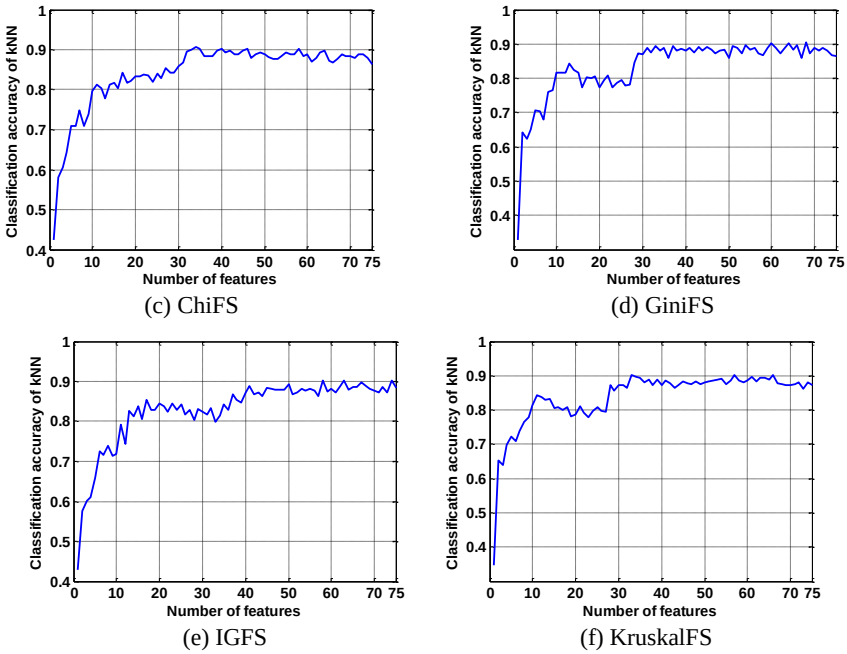


Fig. 1. Accuracy with kNN of feature subset of six methods.

Table 1. The selected best feature sets by six methods with kNN

Methods	Number of features	Accuracy
Original set	75	0.8704
FisherFS	63	0.8969
ReliefFS	46	0.9054
ChiFS	**34**	**0.9062**
GiniFS	68	0.9056
IGFS	74	0.9022
KruskalFS	57	0.9024

Table 1 lists the best selected feature combination by six methods. As we can see in Table 1, for all six methods, the size of features combination is less than the size of the combination built by original features, the accuracy is higher than that of original feature combination. Especially, the feature combination from ChiFS method has the best result with the 34 features and 0.9062 classification accuracy. This indicates the selected feature sets can affectively improve the classification accuracy.

The performance with SVM of feature sets is shown in Fig.2 shows. As we can see, the accuracy rate increases with the size of feature combination, the curve becomes smooth until the size of feature sets arrives a certain specific value. For example, the

number of features is more than 18 for FisherFS, 17 for ReliefFS, 19 for ChiFS, 18 for GiniFS, 20 for IGFS and 18 for KruskalFS.

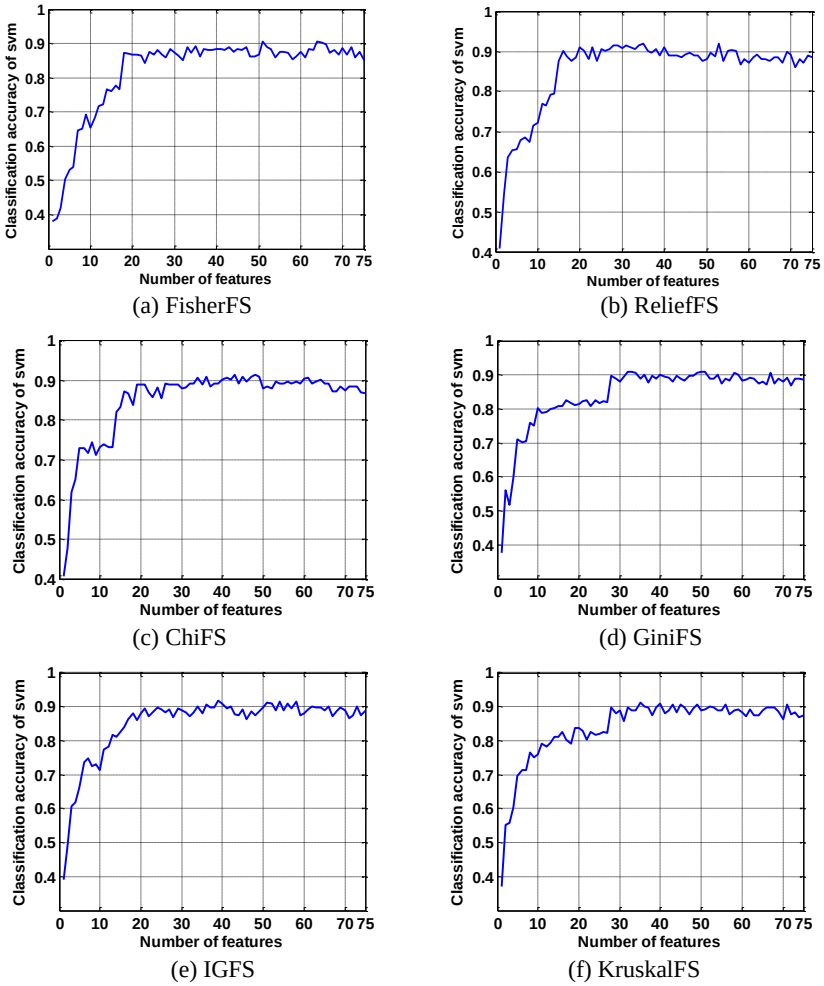


Fig. 2. Accuracy of feature subset of six methods with SVM.

Table 2. The best feature sets selected by six methods with SVM.

Methods	Number of features	Accuracy
Original set	75	0.8704
FisherFS	66	0.9061
ReliefFS	**35**	**0.9186**
ChiFS	43	0.9148
GiniFS	51	0.9106
IGFS	39	0.9186

KruskalFS	35	0.9110

Table 2 lists the selected best feature combination by six methods. As we can see in Table 2, comparisons with the original feature set, the accuracy rate of the selected feature combination of six methods is higher and the size of feature combination is less. The selected feature combination by ReliefFS is better than that by other five methods, which contain 35 features with 0.9186 accuracy rate. This indicates the selected feature combination can affectively improve the classification accuracy of recognition system.

Conclusions

One of key problems for foreign matter recognition in cotton is to determining the good feature combination representing foreign matter. In our work, six filter methods for feature selection, which are respectively FisherFS, ReliefFS, ChiFS, GiniFS, IGFS and KruskalFS, are used to determine the good feature combination with high classification accuracy and small size. The found feature sets is tested in KNN and SVM, respectively. The results shows the filter methods can get the good feature sets with high accuracy and small size to efficiently improve the classification accuracy of recognition system. For kNN, the ChiFS can obtain the feature subset with 34 features and 0.9062 accuracy rate. For SVM, the ReliefFS can find the optimal feature set with 35 features and 0.9186 accuracy rate. It is significant for improving the classification accuracy of recognition system based on machine vision.

Acknowledgments: This research is supported by MOE (Ministry of Education in China) Youth Fund Project of Humanities and Social Sciences (17YJCZH261), Guangdong Natural Science Foundation (2018A030313339), Guangdong Universities Characteristic Innovation Projects (2017GKTSCX063), Science Research Cultivation Project of Shenzhen Institute of Information Technology (ZY201718), Guangdong College Students Cultivation of Scientific, Technological Innovation Special Funds (pdjh2018b0861) and Shenzhen 13th Five-Year Plan Project of Philosophy and Social Sciences (SZ2018D017).

References

Zhao X, Guo X, Luo J, Tan X. Efficient detection method for foreign fibers in cotton. Information Processing in Agriculture, 5(3):320-328 (2018).

Kuzy J, Li C. A pulsed thermographic imaging system for detection and identification of cotton foreign matter. Sensors, 17(3): 518 (2017).

Zhang M, Li C, Yang F. Classification of foreign matter embedded inside cotton lint using short wave infrared (swir) hyperspectral transmittance imaging. Computers and Electronics in Agriculture, 139: 75-90 (2017).

Li J, Cheng K, Wang S, Wang S., Morstatter F., Trevino R. P., Tang J., Liu H. Feature selection: a data perspective. ACM Computing Surveys (CSUR), 50(6): 94 (2017).

Gu S, Cheng R, Jin Y. Feature selection for high-dimensional classification using a competitive swarm optimizer. Soft Computing, 22(3): 811-822 (2018).

Atashpaz-Gargari E, Reis M S, Braga-Neto U M, Barrera J., Dougherty E. R. A fast Branch-and-Bound algorithm for U-curve feature selection. Pattern Recognition, 73: 172-188 (2018)

Kuo R. J., Huang S. L., Zulvia F. E., Liao T. W. Artificial bee colony-based support vector machines with feature selection and parameter optimization for rule extraction. Knowledge and Information Systems, 55(1): 253-274 (2018)

Alijla B. O., Lim C. P., Wong L. P., Khader A. T., Al-Betar M. A. An ensemble of intelligent water drop algorithm for feature selection optimization problem. Applied Soft Computing, 65, 531-541(2018).

Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., Liu, H.: Advancing feature selection research. ASU feature selection repository, 1-28 (2010)

Gallego A. J., Calvo-Zaragoza J., Valero-Mas J. J., Rico-Juan J. R. Clustering-based k-nearest neighbor classification for large-scale data with neural codes representation. Pattern Recognition, 74, 531-543 (2018).

Tahir M., Jan B., Hayat M., Shah S. U., Amin M. Efficient computational model for classification of protein localization images using Extended Threshold Adjacency Statistics and Support Vector Machines. Computer methods and programs in biomedicine, 157, 205-215 (2018)