Computers and Electronics in Agriculture 142 (2017) 1–8

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Model updating for the classification of different varieties of maize seeds from different years by hyperspectral imaging coupled with a pre-labeling method

Dongsheng Guo [a], Qibing Zhu [a,*], Min Huang [a], Ya Guo [a], Jianwei Qin [b]

[a] Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, China
[b] USDA/ARS Environmental Microbial and Food Safety Laboratory, Beltsville Agricultural Research Center, Bldg., 303, BARC-East, 10300 Baltimore Ave., MD 20705-2350, USA

ARTICLE INFO

Article history:
Received 9 January 2017
Received in revised form 15 August 2017
Accepted 16 August 2017
Available online 19 August 2017

Keywords:
Classification
Hyperspectral imaging
Maize seeds
Model updating
Pre-labeling

ABSTRACT

The use of hyperspectral imaging technology combined with chemometrics is an effective nondestructive method for sorting seed varieties. However, the performance of the method is susceptible to the influence of time and depends on the training set used in the modeling process. The accuracy of classification models maybe deteriorate when they are used to differentiate the same variety of seeds harvested in different years, due to new variances in the test set are introduced by changes in the cultivation conditions, soil environmental conditions and climatic changes from one year to another. To maintain the accuracy and robustness of model, a model-updating algorithm for differentiating maize seed varieties from different years based on hyperspectral imaging coupled with a pre-labeling method was proposed in this work. The pre-label of each unlabeled sample was obtained using the original classification models developed by the least squares support vector machine classifier. The representative unlabeled samples, which had reliable pre-labels, were selected for updating classification models based on Pearson correlation coefficients. After model updating, the average classification accuracies were improved by 8.9%, 35.8% and 9.6%, compared with those of non-updated models for three test sets, respectively. This shows the effectiveness of the proposed method for classifying maize seeds of different years.

1. Introduction

Maize is one of the major agricultural crops in the world and regarded as a main source of food, feeds, fuel, and industrial raw materials (Tenaillon and Charcosset, 2011). In recent years, the number of maize seed varieties has tremendously increased due to wide applications of seed hybrid technologies. The diversity of maize seed varieties makes seed classification more difficult after harvesting, and ultimately affects crop yields, the plant breeding and crop improvement programs. Hence, developing rapid, accurate, and nondestructive methods for classifying maize seeds is important to maize seed industry.

Hyperspectral imaging technique is a nondestructive and reliable technique, which integrates spectroscopic and imaging techniques in one system for providing both spectral and spatial information simultaneously (Wang et al., 2016). Hyperspectral image technology overcomes the shortcomings that machine vision cannot obtain the chemical composition of samples because of limited spectral information and near infrared spectroscopy technique obtains spectral information of samples by using a single spot and is influenced by the uniformity of sample distribution; and it can be used to rapidly inspect large quantities of samples (Yang et al., 2015). Hyperspectral imaging technique has been successfully used for the classification of maize seeds. Zhu et al. developed a partial least squares discrimination analysis classification model using image entropy extracted from hyperspectral images and obtained 98.90% testing accuracy for 17 varieties of maize seeds (Zhu et al., 2012). Zhang et al. utilized a visible and near-infrared hyperspectral imaging system to differentiate 6 varieties of maize seeds and reported testing accuracy of 98.89% by using least squares support vector machine (LSSVM) coupled with principal component analysis (Zhang et al., 2012).

However, the model performance for maize seed classification depends on the training set used in the modeling process. If the samples in the training set are not representative, the classification results will not be good (Zhu et al., 2015). For maize seed classification, people can only use the seeds from current year or past years to develop the classification model. When developed model

* Corresponding author at: School of Internet of Things, Jiangnan University, 1800 Lihu Avenue, Wuxi, Jiangsu Province 214122, China.
E-mail address: zhuqib@163.com (Q. Zhu).

http://dx.doi.org/10.1016/j.compag.2017.08.015
0168-1699/

is used to classify the seeds from the next year, the model performance will be decreased because new variances were introduced by changes in cultivation conditions, soil environmental conditions, and climatic changes from one year to another among other factors. These new variances of samples were not considered in the original model, thus reducing the accuracy and robustness of the developed model (He et al., 2016). In order to guarantee the classification accuracy of maize seeds harvested in different years, one method is to rebuild a new model just using the new samples (i.e., original training samples are not used in the new model), which will waste the knowledge obtained from the original model, and incur a lot of time and cost consumption because labeling the training samples is time-consuming and extremely laborious. The other method is model updating, which updates the original model by using the original training samples and a small number of representative new samples selected from the new test set. Since only a small number of new samples need to be labeled, the time and cost consumption of model updating is less than that of rebuilding a new model.

Previous studies found that model updating was an effective method to improve the developed model by updating the training set (Xie and Ying, 2012; Farrell et al., 2012). Several methods were used to select representative samples from the test set to update models. Zhang updated a classification model by selecting samples with the lowest certainty values to label variety information by chemical methods and adding the samples to the training set to retrain the classifier (Zhang, 2014). Huang et al. proposed an online updating model based on the improved incremental learning capability of the support vector data description (ISVDD). Several test samples that were rejected by the ISVDD models were given correct labels by experts or chemical methods and used to update LSSVM and ISVDD models (Huang et al., 2016). However, all of these methods need true label information of representative samples selected from the test set for model updating, which still require significant amount of time and cost consumption. A pre-labeling approach is one of the semi-supervised classification methods. This method obtains the pre-labels of the unlabeled samples using the original model, and retrains the model using the original training set and partial pre-labeled samples with the highest degree of confidence (Han et al., 2016; Ma et al., 2016). Through the pre-labeling approach, some unlabeled samples with reliable pre-labels as representative samples were extended to the original training set for increasing the representativeness of the training set, thus improving the model performance. The pre-labeling approach could reduce time and cost consumption because the labels of selected samples for model updating are obtained by the model itself.

This study investigated a model updating method for the classification of maize seeds harvested in different years using hyperspectral imaging coupled with a pre-labeling method. The specific objectives were: (1) developing an LSSVM model for maize seed classification using the spectral features extracted from hyperspectral images, and (2) updating the LSSVM model using a pre-labeling algorithm, to improve the model performance for maize seed classification harvested in different years.

2. Materials and methods

2.1. Experimental samples

A total of 3600 samples of four varieties (JIDAN7, JUNDAN18, JUNDAN20, and LUDAN818) were used in this study, and each variety contained 900 maize seeds from three years (300 maize seeds from each year). The 1000 samples (250 samples for each variety) were randomly selected from JIDAN7 (2012), JUNDAN18 (2007), JUNDAN20 (2009), and LUDAN818 (2012)) as the training samples to develop the original classification model. The 800 samples (200 samples for each variety) were random selected form JIDAN7 (2013), JUNDAN18 (2011), JUNDAN20 (2010), and LUDAN818 (2013) as the Unlabeled set 1. The samples in Unlabeled set 2 were random selected form JIDAN7 (2014), JUNDAN18 (2014), JUNDAN20 (2011), and LUDAN818 (2014). Unlabeled set 3 is the pool samples of Unlabeled set 1 and Unlabeled set 2). The three unlabeled sets (Unlabeled set 1, Unlabeled set 2, and Unlabeled set 3) were used to update the original models using a pre-labeling strategy. Four test sets (Test set 1, Test set 2, Test set 3, and Test set 4) were used to test the model performance. The 250 samples (50 samples for each variety) in JIDAN7 (2012), JUNDAN18 (2007), JUNDAN20 (2009), and LUDAN818 (2012)) were used as Test set 1 to confirm the ability of the original model to classify maize seeds from the same year. Test set 2 (JIDAN7 (2013), JUNDAN18 (2011), JUNDAN20 (2010), and LUDAN818 (2013), 100 samples for each variety), Test set 3 (JIDAN7 (2014), JUNDAN18 (2014), JUNDAN20 (2011), and LUDAN818 (2014), 100 samples for each variety), and Test set 4 (the pool samples of Test set 2 and Test set 3) were used to evaluate the performance of the updated model. The detailed sample arrangement is shown in Table 1. It should be noted that, in order to ensure the consistency of seed variety, all samples were provided by the same seed production company (Fengle Seed Inc., Hefei, China). Therefore, it is difficult to ensure that each seed variety used in this study has the same production year because of varied production plan and other affecting factors. Although the sample production year for each variety is not exactly the same, it does not affect testing the influences of different year samples on the performance of the models and the effectiveness of the proposed method in this study. The seeds were stored in a closed vessel in a refrigerator at 4 °C before experiment. Fig. 1 shows the true appearance of the four varieties of maize seeds.

2.2. Hyperspectral imaging acquisition

The hyperspectral images of the maize seeds in the reflectance mode were attained by the line-scan hyperspectral imaging system in the visible/NIR range (400–1000 nm). The details of the hyperspectral image acquisition system are discussed in the paper by Huang et al. (2016).

In the scanning system, the hyperspectral image of each seed was attained by a constant speed scanning line at a 250 ms exposure time. A total of 750 scanning lines for each hyperspectral image were obtained at 60 mm scan length and 80 μm step size, and the hyperspectral reflectance images had 6.38 nm/pixel spectral resolution intervals. Finally, the hyperspectral reflectance images of the 94 wavelengths of each maize seed were obtained. The entire process was completed in a closed black box to avoid interference from external light sources. After the hyperspectral images of the seeds were collected, the dark and reflectance images of the white Teflon were acquired for image correction based on the following equation to compensate light source variation effect:

$$R_J = \frac{R_I - R_B}{R_W - R_B} \tag{1}$$

where R_J andR_I are the corrected and raw spectral images of the samples, respectively; R_B is the dark reference image obtained by completely blocking the lens; and R_W is the white reference image obtained from the diffuse reflectance white standard.

To obtain the mean spectral features of the maize seeds, the segmentation of the seeds from the hyperspectral image background is essential. The reflectance image at 782.59 nm was selected to identify the seeds because of the outline of maize seeds in the image was the clearest (Fig. 2 step a). Before image