Modeling for mung bean variety classification using visible and near-infrared hyperspectral imaging

Chuanqi Xie, Yong He*

(College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China)

Abstract: This study was carried out to investigate the feasibility of using visible and near infrared hyperspectral imaging for the variety classification of mung beans. Raw hyperspectral images of mung beans were acquired in the wavelengths of 380-1023 nm, and all images were calibrated by the white and dark reference images. The spectral reflectance values were extracted from the region of interest (ROI) of each calibrated hyperspectral image, and then they were treated as the independent variables. The dependent variables of four varieties of mung beans were set as 1, 2, 3 and 4, respectively. The extreme learning machine (ELM) model was established using full spectral wavelengths for classification. Modified gram-schmidt (MGS) method was used to identify effective wavelengths. Based on the selected wavelengths, the ELM and linear discriminant analysis (LDA) models were built. All models performed excellently with the correct classification rates (CCRs) covering 99.17%-99.58% in the training sets and 99.17%-100% in the testing sets. Fifteen wavelengths (432 nm, 455 nm, 468 nm, 560 nm, 705 nm, 736 nm, 760 nm, 841 nm, 861 nm, 921 nm, 930 nm, 937 nm, 938 nm, 959 nm and 965 nm) were recommended by MGS. The results demonstrated that hyperspectral imaging could be used as a non-destructive method to classify mung bean varieties, and MGS was an effective wavelength selection method.

Keywords: visible and near-infrared hyperspectral imaging, mung bean, classification, modeling, wavelength selection

DOI: 10.25165/ijabe.20181101.2655


1 Introduction

Mung bean is welcomed by many people in China, Korea, Burma, India, Japan, Thailand, Pakistan and other Southeast Asian countries due to its high edibility and medicinal value[11]. It contains a lot of nutrients and functional compounds, such as protein, vitamin, lipid, phytochemicals and fiber[2]. However, different varieties of mung beans have distinct agronomical, processing and nutritional characteristics[3]. Thus, identification of the mung bean variety is of great importance. The current standard methods to differentiate varieties of beans are mainly based on artificial sense and biochemical operation[4,5]. However, these methods are inefficient, time-consuming and destructive. Also, professional and qualified technicians are required for such detection methods. Thus, an advanced method (hyperspectral imaging) is needed. This study was carried out to identify the mung bean variety by using the visible and near infrared hyperspectral imaging. Based on the hyperspectral imaging, a multispectral detection system can be designed by the effective wavelengths, which has the potential to be used in industry for online and non-destructive detection.

Hyperspectral imaging combines both spectral and imaging techniques together, and has been widely used in previous researches due to its advantages such as being fast, non-destructive, effective and accurate[6,8]. Using the hyperspectral imaging system, a spatial map can be created when the sample is scanned by the camera. Each pixel of the hyperspectral image has a spectrum covering the full spectral range. The hyperspectral image (hyperspectral cube) is composed of a series of images at the whole wavelengths, and it contains both spectral and spatial information[9]. It can provide a full database with internal and external features of the samples[10], and different varieties of objectives may have various external and internal characteristics, such as color, texture and nutrition content, directly resulting in the spectral signature differently. The spectral signature from a certain pixel of the image is useful for the discrimination and classification of the objectives[11]. This technique has been applied in variety classification in many previous studies. The oat and groat kernels were classified using the near-infrared (NIR) hyperspectral imaging system covering the wavelengths of 1006-1650 nm[12]. The partial least squares-discriminant analysis (PLS-DA) model was built for the classification, and three effective wavelengths were identified (1132 nm, 1195 nm and 1608 nm). Kong et al.[10] identified four cultivars of rice seeds using the NIR hyperspectral imaging in the wavelengths of 1039-1612 nm. PLS-DA, soft independent modeling of class analogy (SIMCA), K-Nearest neighbor (KNN), support vector machine (SVM) and random forest (RF) were established to identify the cultivars. Twelve useful wavelengths (1069 nm, 1079 nm, 1139 nm, 1167 nm, 1183 nm, 1227 nm, 1281 nm, 1304 nm, 1328 nm, 1389 nm, 1467 nm and 1558 nm) were selected by weighted regression coefficients method. Kamruzzaman et al.[13] used the NIR hyperspectral imaging to...
discriminate lamb muscles. Principal component analysis (PCA) was used to compress the dimensionality, and six wavelengths (934 nm, 974 nm, 1074 nm, 1141 nm, 1211 nm and 1308 nm) were determined by PCA loadings. Quality classification of cooked and sliced hams was also studied using NIR hyperspectral imaging[14]. PCA was used for selecting effective wavelengths, and eight wavelengths (980 nm, 1061 nm, 1141 nm, 1174 nm, 1215 nm, 1325 nm, 1436 nm and 1641 nm) were identified to discriminate different turkey ham qualities. Barbin et al.[15] investigated the identification of fresh and frozen-thawed meat by using the NIR hyperspectral imaging. The PLS-DA model achieved the correct classification of 100% based on the effective wavelengths recommended by weighted regression coefficients method. Wei et al.[16] discriminated the ripeness (unripe, mid-ripe, ripe and over-ripe) of persimmon fruit using the visible and near-infrared hyperspectral imaging (400-1000 nm). Three classification models (linear discriminant analysis (LDA), SIMCA and least squares support vector machines (LS-SVM)) were built to classify the different types of persimmons. Three wavelengths (518 nm, 711 nm and 980 nm) were selected by successive projection algorithm (SPA). Based on the spectral and texture feature at the three wavelengths, the LDA model obtained the best correct classification rate of 95.30%. Deng et al.[17] used spectral angle mapping (SAM) model combined with hyperspectral images to distinguish weeds from background and cabbages. All of these studies demonstrated that the hyperspectral imaging has the potential to be used for variety classification.

The aim of this study was to develop an effective method to classify mung bean varieties using the hyperspectral imaging. The specific objectives were to: (1) classify different varieties of mung beans based on spectral reflectance information; (2) identify effective wavelengths that play the most significant roles for the variety classification using modified gram-schmidt (MGS); and (3) compare the performance of different classification models.

2 Materials and methods

2.1 Work flow

The process of this study can be seen in Figure 1. In the first step, raw hyperspectral images of four varieties of mung beans were acquired by the visible and near infrared hyperspectral imaging camera. All raw hyperspectral images were then corrected by the white and dark reference images. Spectral reflectance values were extracted from the corrected hyperspectral images and treated as the dependent variables. All samples were divided into two sets (training and testing) at a ratio of 2:1. One classification model (extreme learning machine, ELM) was established based on the full spectral wavelengths. Then, MGS method was used to select effective wavelengths. Based on these selected wavelengths, the ELM and LDA models were built, respectively. The optimal model was finally determined by the values of correct classification rates (CCRs).

2.2 Samples

Four different brands of mung beans (Jinlongyu, Ouyi, Saiwengfu and Xiangmangyu) were used in this study. A volume of 60 cm³ of mung beans were evenly distributed in a glass dish (d=90 mm). For each brand, there were 90 dishes. Each dish represented one sample, and was then scanned individually by the hyperspectral imaging system. Finally, a total of 360 samples for the four varieties were used in this study.

2.3 Experiment device

A hyperspectral imaging system which covers the spectral wavelengths of 380-1023 nm was used in this study (Figure 2). This system consisted of a CCD camera (C8484-05, Hamamatsu City, Japan), a lens (OLE-23, USA), an imaging spectrograph (V10E, Specim, Oulu, Finland), two light sources (Oriel Instruments, USA) provided by two 150 W quartz tungsten halogen lamps, a computer, and a conveyor belt operated by a stepper motor (IRCP0076, Isuzu Optics Corp., Taiwan, China). The area CCD array detector has 672×512 (spatial × spectral) pixels, and the spectral resolution is 2.8 nm. Samples were scanned by the hyperspectral imaging system line by line, and the reflected light was dispersed by the spectograph and captured by the area CCD array detector in spatial-spectral axes. A hyperspectral image can be generated when the sample is scanned by the camera. Each hyperspectral image contains hundreds of gray images, and each pixel contains both spectral and spatial features (Figure 3). The ENVI 4.7 software (Research System Inc., Boulder, Co., USA) was used for obtaining spectral reflectance information from hyperspectral images. MATLAB R2009a (The Math Works Inc., Natick, MA, USA) was used to identify effective wavelengths and establish classification models.
back-propagation (BP) algorithm\textsuperscript{[22]}]. LDA is a supervised recognition algorithm, and can be used to find a linear combination of features that can classify different varieties of objectives\textsuperscript{[16]}. The LDA algorithm produces a number of orthogonal linear discriminant functions which makes the samples to be classified in different categories\textsuperscript{[23]}]. Both classification models were built using MATLAB R2009a software.

### 2.7 Wavelength selection

The spectral feature, which covers the wavelengths from 380 nm to 1023 nm, usually contains highly redundant information among different wavelengths\textsuperscript{[24]}. Thus, the selection of effective wavelengths is a crucial step in spectral analysis\textsuperscript{[25]}. The final aim of selecting effective wavelengths is to establish a subset of spectral wavelengths to replace the whole ones. The selected wavelengths can be equal to or even more effective than the full wavelengths\textsuperscript{[26]}. Also, these wavelengths can not only simplify the model but also be used to develop a multispectral imaging classification system. MGS was used to select useful wavelengths in this study.

The principle of MGS is to build a new orthogonal basis based on the original one by projection. An eigenvector, which is selected from the feature matrix $X=[x_1, x_2, \ldots, x_n]$, can be treated as the first orthogonal basis $x_1$. Then the projection of other vector quantities onto this orthogonal basis can be calculated

$$x_{proj} = \frac{X_n - x_1 x_{1}^{T} x_{ns}}{\| x_{ns} - x_1 x_{1}^{T} x_{ns} \|}$$

obtained from other vector quantities and this projection and $x_1$ orthogonal, resulting in $n-1$ eigenvectors of the first orthogonal basis. Similarly, the projection of the class label vector $Y$ onto the first orthogonal basis ($Y_{proj} = (Y, x_1)$) can be calculated. Make the difference value obtained from $Y$ and projection and $x_1$ orthogonal. Finally, the eigenvector can be identified. The main procedures can be seen in Table \textsuperscript{[27]}.

Step 1: Each eigenvector of characteristic matrix $X=[x_1, x_2, \ldots, x_n]$ is normalized, $x_n = \frac{x_n}{\| x_n \|}$ ($i=1,2,\ldots,n$);

Step 2: The inner product of class label vector $Y$ and each normalized eigenvector is calculated $y_{proj,s} = (Y, x_s)$ ($s=1,2,\ldots,n$), the eigenvector $x_s$ ($s=1,2,\ldots,n$) that maximizes $y_{proj,s}$ is selected as normalized orthogonal basis;

Step 3: The eigenvector of feature matrix $X=[x_1, x_2, \ldots, x_n]$ is projected onto normalized orthogonal basis $x_{ns}$, $x_{ns} = (x_s - x_1 x_{1s}^{T} x_{ns}) / \| x_{ns} - x_1 x_{1s}^{T} x_{ns} \|$, the difference value ($x_s = x_s - x_{proj,s}$) of eigenvector and projection is calculated;

Step 4: The class label vector $Y$ is projected onto normalized orthogonal basis $x_{ns}$, $y_{proj} = (Y, x_{ns})$, the difference value ($Y = Y - y_{proj} = Y - Y_{proj}$) of $Y$ and projection is calculated;

Step 5: $X=[x_1, \ldots, x_{ns}], x_{ns+1}, \ldots, x_n$, $Y=Y, n=n-1$;

Step 6: Calculation will stop when the required eigenvectors are obtained, or it will return to step 2.

This method belongs to the backward feature selection, which executes the selection by removing one or more feature from the whole features. The backward feature selection method can keep...
the informative features as many as possible. On the other hand, MGS considers the relationship between the feature structures and label information, which can improve the performance of the subsequent classification or prediction model. As a whole, the merits of MGS can be summarized as follows: (1) backward feature selection, and (2) the feature selection combines the feature structure and label information.

Table 1  Main procedures of modified gram-schmidt (MGS) algorithm

<table>
<thead>
<tr>
<th>Input</th>
<th>Feature matrix X=[x1, x2, …, xM], class label vector Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output</td>
<td>feature matrix Xs selected by MGS</td>
</tr>
<tr>
<td>1</td>
<td>for k=1:featnum %Main loop</td>
</tr>
<tr>
<td>2</td>
<td>xk = x1/(x1') %Normalized feature matrix</td>
</tr>
<tr>
<td>3</td>
<td>xn = argmax(Y, xk) . xk(+) = xk %Finding the direction of maximum inner product</td>
</tr>
<tr>
<td>4</td>
<td>x' = x - (x . x') / x'.x %Projecting X onto xk, computing the residual X</td>
</tr>
<tr>
<td>5</td>
<td>Y' = Y - (Y . x') / x'.x %Projecting Y onto xk, computing the residual Y</td>
</tr>
<tr>
<td>6</td>
<td>X=[x'1, X2, …, Xk, xk+1, …, xM], Y=Y', m=1 %Changing the variables to iterate</td>
</tr>
<tr>
<td>7</td>
<td>end</td>
</tr>
<tr>
<td>8</td>
<td>return x0</td>
</tr>
</tbody>
</table>

3 Results and discussion

3.1 Spectral feature analysis

The mean spectral reflectance curves of the four varieties of mung beans were shown in Figure 4. It can be seen that the general trends of the spectral reflectance curves of different varieties were very similar with some noise at the beginning of the wavelength, and some peaks as well as valleys covering the whole wavelengths. Different peaks and valleys are corresponding to various spectral bands, for example, the wavelength at 970 nm indicates there is a water absorption band here. Also, obvious difference can be seen in the spectral region of 700-1023 nm, which means most of the effective wavelengths for the classification should be located in this region. In order to minimize the influence of the noise, the wavelengths at the beginning were rejected, resulting in spectral wavelengths of 395-1023 nm (band 14-band 512) being used in this study.

![Figure 4 Average spectral reflectance curves of the four varieties of mung beans](image)

3.2 Classification based on whole wavelengths

In this study, the ELM classification model was firstly established based on full spectral wavelengths. The reflectance values were treated as X variables, and the wavelengths were treated as Y variables (Jinlongyu-1, Ouyi-2, Saiwenfugu-3, Xiangmanyuan-4). The ELM model obtained a satisfying result with CCRs of 99.58% in the training set and 100% in the testing set. No samples were incorrectly identified in the testing set. However, the large number of input variables will increase the calculation time and cannot be used for designing a multispectral classification system. Therefore, effective wavelengths should be identified for establishing simplified classification models.

3.3 Effective wavelengths

In order to improve the performance of classification models and simplify the calculation, MGS was carried out to identify the key wavelengths. Based on the selected wavelengths, the spectral data set was then reduced into a matrix with a dimension of mnx (m was the number of samples, and x was the number of selected wavelengths). As a result, a total of fifteen wavelengths (432 nm, 455 nm, 468 nm, 560 nm, 705 nm, 736 nm, 760 nm, 841 nm, 861 nm, 921 nm, 930 nm, 937 nm, 938 nm, 959 nm and 965 nm) were obtained. Compared with full spectral wavelengths, the number of selected wavelengths was greatly reduced. It only accounted for 3.01% of the number of whole wavebands (band 14-band 512). These selected wavelengths were then used as the new input variables for establishing classification models.

3.4 Classification based on selected wavelengths

The ELM and LDA models were then built using the selected wavelengths suggested by MGS (Table 2). MGS-ELM model performed excellently with the CCRs of 99.17% in both training and testing sets. MGS-LDA model acquired the CCRs of 99.17% in the training set and 98.33% in the testing set. Compared with the full wavelengths-based ELM model, there was a little decrease of CCR in the testing set of MGS-ELM model. However, the result was still excellent (99.17%). Also, the number of input variables were largely reduced, which made the classification model simple. It indicated that these wavelengths selected by MGS were very useful for classifying different varieties of mung beans. As mentioned above, the spectral reflectance curves were obviously different in the wavelengths of 700-1023 nm. Most of the selected wavelengths were located in this region. This might be why the selected wavelengths-based models could also obtain good classification results.

Table 2  Classification results by using different models

<table>
<thead>
<tr>
<th>Classification model</th>
<th>Number of wavelengths</th>
<th>Training set</th>
<th>Testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. Missed</td>
<td>CCR%</td>
<td>No. Missed</td>
</tr>
<tr>
<td>MGS-ELM</td>
<td>15</td>
<td>240</td>
<td>99.17</td>
</tr>
<tr>
<td>MGS-LDA</td>
<td>15</td>
<td>240</td>
<td>99.17</td>
</tr>
</tbody>
</table>

Note: CCR: correct classification rate.

4 Conclusions

This study demonstrated that the visible and NIR hyperspectral imaging could be used to classify the four different varieties of mung beans. MGS was efficient for selecting useful wavelengths. The numbers of selected wavelengths suggested by MGS only accounted for 3.01% of that of the full wavelengths. The full wavelengths-based ELM model obtained an excellent result with a very high CCR value (100%) in the testing set. Those models established based on selected wavelengths obtained prominent results with CCRs ranging from 98.33% to 99.17% in the testing sets. Although the results acquired by selected wavelengths were a little lower than that obtained by full wavelengths, they are still excellent. Moreover, a large amount of the full wavelengths were not used and the number of input variables were largely reduced, which made the model simple and efficient.
Reduced into a few ones, which have the potential to be used for designing an online detection system. However, in order to get more robust results, more samples and varieties should be considered in further studies.

Acknowledgements

This work was supported by the National Key Scientific Instrument and Equipment Development Projects (2014YQ470377), the Scientific Research Foundation for Returned Overseas Students and the Fundamental Research Funds for the Central Universities of China (2012FZA6005, 2013QNA6011).

References


