Identification of maturity of cuiguan pear based on RobustICA and PSO-ELM

Yu Zhang¹, Yongmei Huang¹*, Min Lin¹, Huijun Liu¹ and Xinqi Zhou²

¹College of Metrology and Measurement Engineering, China Jiliang University Hangzhou, China
²Focused Photonics (Hangzhou),Inc.Hangzhou 310052,China
* Corresponding author. Email:ymhuang@cjlu.edu.cn

Abstract. In order to improve the stability of extreme learning machine (ELM), it was proposed to use the combination of robust independent component analysis (RobustICA) and particle swarm optimization ELM algorithm. It was applied to identify the maturity of cuiguan pear at different growing periods. First the two order derivative and the discrete wavelet transform were adopted to remove the background noise and compression spectra. Then RobustICA was used to decompose the spectra to extract independent components and mixed coefficient matrices. Finally, by using particle swarm optimization, the best input layer weight and hidden layer deviation were obtained and adjusted the number of hidden layer nodes to get the best ELM model to differentiate maturity. The method was used to identify the maturity of cuiguan pear at four different stages, and has strong robustness. The results showed that RobustICA-PSO-ELM had better discrimination ability compared with RobustICA - ELM and ELM, and the accuracy increases from 75% to 94.4%.

1. Introduction
Cuiguan pear is one of the most widely distributed fruit trees, which is rich in nutrition and taste well. After harvesting, the photosynthesis of the cuiguan pear basically stops. The nutrients consumed in the process of storage and transportation which lead to the quality drop and lose the commodity value, resulting in the loss rate as high as 30%. Therefore, strengthening the identification of the fruit maturity of the cuiguan pear is meaningful to reduce the loss and increase the economic benefits. Near infrared spectroscopy is a fast and non-destructive method. Compared with the traditional method, it is more convenient which widely used in many aspects of the analysis.

Traditional methods such as PCA and SVD can only decompose uncorrelated components, and they can’t guarantee that these components are independent of each other. Therefore it reduces the typicality of the extracted features. Independent component analysis (ICA) is a kind of blind source separation (BBS), which can separate signals into independent components. Under the premise of the independence of spectral components, ICA can be used to decompose spectra. Fang Limin established a prediction model of moisture, protein and starch content in corn by using ICA method and neural network. [1] Sun Tong use uninformative variable elimination method to eliminate irrelevant variables. Then he used ICA method to establish the soluble solid model of NanFeng Orange. [2] Shao Yongni, who applied the ICA method to obtain different vintages of rice, establish the identification model of rice year by using the characteristic bands as input to artificial neural network and got sensitive bands corresponding to major components of paddy rice.[3]

Extreme learning machine (ELM) is a new neural network algorithm. It can get the unique optimal solution in the training process when the input layer weight and the hidden layer bias are given.
randomly. However, the hidden layer parameters generated by the ELM will lead to poor generalization performance. In order to improve the prediction accuracy, the number of hidden layer nodes needs to be increased which will increase the complexity of the network and lead to overfit. Aiming at the shortage of ELM, a particle swarm optimization algorithm is proposed to optimize the input layer weight and the hidden layer bias, and an optimal model is obtained.

Through the analysis of near infrared spectroscopy of cuiguan pear, RobustICA was used to extract the independent components and corresponding mixing. Particle swarm optimization was used to optimize the input layer weight and hidden layer deviation of ELM and established the cuiguan pear maturity model which can identify maturity successfully.

2. Basic Theory

2.1. Robust Independent Component Analysis

The article uses the RobustICA arithmetic which is based on the kurtosis. The RobustICA algorithm takes kurtosis as the control function, then optimizes the kurtosis by searching the best step length, then get the independent component (basic component, spectral matrix) and mixed coefficient matrix with kurtosis that the kurtosis is not 0. It extracts the main components of the spectrum and discards the unwanted information. The advantage is that the stronger robust and the faster converges.

2.2. ELM Optimization Based on PSO

This paper presents the PSO-ELM algorithm. The concrete steps are as follows:

(a) First, it needs to initial population and set particle number m to generate population. Then it initializes PSO particle velocity v and maximum/ minimum inertia $w_{max}$ / $w_{min}$. Last learning factor $c1$ or $c2$ and maximum iteration time k are set up.

(b) It selects the inertia weight according to $w = w_{max} - \frac{w_{max}-w_{min}}{k} \times a$ (a is the current number of iterations). The fitness value of the particles is calculated according to the fitness function. The fitness values of the particles and the global fitness values are constantly compared until the end condition is satisfied.

(c) It needs to update the position and speed of particles constantly.

(d) When the optimization satisfies the maximum iteration number k or the evaluation value which is less than the specified precision b (b>0), the search is finished and the optimal input weight and the hidden layer bias are obtained.

(e) The best ELM model is obtained by substituting the optimal solution into ELM, and the model is trained by test set.

3. Experiment

3.1. Instrument and Equipment

The near infrared spectra of cuiguan pear samples were collected by using MCS600 array optical fiber spectrometer of Cai Si (Zeiss, Germany) in Germany. The wavelength range was 800-1700nm, and the spectral collection software was Aspect Plus. Matlab2015a was used to analysis Spectral data.

3.2. Sample Preparation and Spectra Acquisition

The 144 samples of cuiguan pear used in this experiment were all from a fruit industry park in Hangzhou, Binjiang. Samples were collected at two weeks before the samples normal maturation, and 36 samples were collected every seven days. The samples were gathered four times, and were divided into I, II, III and IV. After simple cleaning, the spectra were got. The samples were collected several times, and the average spectra were used in the experiment. All samples were divided into calibration set and prediction set by using kennard stone algorithm. The calibration set contained 108 samples for modeling, and the prediction set contained 36 samples for detecting model accuracy.

From Figure 1 can be seen that the spectra of different maturity of cuiguan pear were different and some spectral regions overlapped. Because of the change of sugar and acid of cuiguan pear, it appeared the difference in near infrared spectral absorption. This is the basis for identification of maturity and showed that the identification of cuiguan pear maturity is meaningful.
3.3. Spectral Data Processing and Modeling

Because of the size and surface roughness of cuiguan pear samples were different, and at the same time a large number of useless information in the spectra will reduce the accuracy of the model and spend more time and energy. So this paper uses the two order derivative and discrete wavelet transform to preprocess the original spectra in order to eliminate interference from external factors interference and improve the identification ability of mathematical model. Wavelet base function selected db2 and decomposed to 4 layers. The compressed data is about 7% of the original amount of data.

Mixing coefficient matrix and independent components corresponding are extracted by RobustICA. PSO is used to optimize the input layer weight and the hidden layer deviation to get the optimal ELM model. On the other hand, RobustICA combined with the ELM model, and only through the identification model by pretreatment, the direct use of ELM to establish the comparison. The classification accuracy (CR) was used to evaluate the model. The higher the classification accuracy (CR), the model is better.

\[
CR = \frac{\text{Number of samples which correctly classified}}{\text{total sample}} \times 100\%
\]

3.4. Experimental Process

From figure 2 we can see the experimental process as follows:

(a) The samples of cuiguan pear were collected several times, and the average value was taken as the original spectra. The original spectra were processed by two order derivative and discrete wavelet to eliminate the interference of spectral baseline drift and background light and compressed data to improve modeling efficiency. Wavelets used dbn and n is vanishing moments.

(b) The processed spectra are decomposed by RobustICA and the best independent component number (ICs) is selected to obtain the best independent component and its mixing coefficient matrix.

(c) The input layer weight and hidden layer deviation are as the particle in the particle swarm search space. The relevant parameters were initialized and particle iteration optimize by algorithm until a termination condition is satisfied: when search the maximum number of iterations is reached k or less than a specified value evaluation accuracy of a (a>0), the search is over and get the best input layer and hidden layer deviation.

(d) The optimal input layer weight and hidden layer bias are substituted into the ELM model, and the optimal ELM model is obtained. The mixed coefficient matrix is used as the input, and the actual maturity is taken as the output to get the identification model of cuiguan pear.
4. Results of Experiment and Discussions

4.1. Independent Component Selection of RobustICA

After two order derivative and discrete wavelet pretreatment, the spectrum had basically removed the useless information and the interference of the background light. When RobustICA was used to extract independent components and its corresponding mixed coefficient matrices, the influence of the number of independent components on the model results were investigated. In this paper, the initial independent component number was 5, and 1 is added as step length until get to the 12. The other parameters of the model remain unchanged, which is used to select the best independent components of RobustICA-PSO-ELM and RobustICA-ELM. As shown in figure.3, the RobustICA-PSO-ELM model selects the best independent component number of 7, and the RobustICA-ELM model has the best independent component number of 10.
Figure 3. The accuracy of different choice of ICs in different models

4.2. PSO-ELM Model

The ELM hidden layer excitation can choose ‘Sigmoidal’ ‘Gauss’ etc. by taking the mixed coefficient matrix which is obtained by RobustICA as the input of the limit learning machine model and using the maturity as output. The article chooses the function named ‘Sigmoidal’. Initialize the parameters of the particle swarm algorithm, set the particle number $m=30$, set all the learning factors as 1.5, the maximum number of iterations $k=30$, set the maximum inertia weight to 0.9 and the minimum inertia weight to 0.4, set the minimum accuracy $b$ to 0.01.

The selection of the hidden layer nodes in ELM is critical. Set the initial number of the hidden layer nodes to 6 then increase to 24 by using the step size of 2, and train the model 15 times. The picture below shows the accuracy of near infrared model maturity in RobustICA-PSO-ELM, RobustICA-ELM and ELM under different hidden layer nodes. It shows that the accuracy decreases dramatically when just using the ELM to build the model with the number of hidden layers increases because of the model has overfitted. After the particle swarm optimization, the situation improves obviously, and the model accuracy also improved.

Figure 4. The accuracy of different number of neurons in different models

From figure 4, the results showed that the RobustICA-PSO-ELM results were optimal when the number of nodes in the hidden layer is 18. The identification maturity model of cuiguan pear which established by RobustICA-PSO-ELM was better than other model methods. Particle swarm optimization was used to find the input layer weight and hidden layer deviation of ELM, and the effect was better, faster. The results of model classification were more accurate. Using the RobustICA
algorithm, the output was non-Gauss which could extract more useful information and increase the accuracy of the model. Particle swarm optimization was used to find the best input layer weight and hidden layer deviation which avoid random selection and obtaining the best ELM model.

4.3. Comparison of Modeling and Prediction Results

| Model name                        | Accuracy  |
|-----------------------------------|-----------|
| RobustICA-PSO-ELM(ICs=7,nods=18)  | 94.44%    |
| RobustICA-ELM(ICs=10,nods=20)     | 88.89%    |
| ELM(nods=6)                       | 75%       |

As shown in the table 1. The accuracy was only 75% when only used the ELM method. After using RobustICA method, the model had some improvement. When RobustICA-PSO-ELM method was employed, the accuracy was best. When the number of independent components is 7 and the number of nodes is 18, the performance of the ELM model was improved and the accuracy rate rose to 94.4%. So it can be seen that the model is reliable and significant effect, improve the accuracy which can be effectively applied to the business and many other things.

5. Conclusion

The independent component was extracted by RobustICA and the mixed coefficient matrix was obtained. The input layer weight and hidden layer deviation of ELM were optimized by PSO, and the maturity of cuiguan pear was identified. The results showed that the accuracy of RobustICA-PSO-ELM is the highest compared with the RobustICA-ELM model and the ELM model alone. When the number of independent components was 7 and the number of nodes in ELM hidden layer was 18, the accuracy of identification was 94.44%. Therefore, RobustICA-PSO-ELM has excellent modeling and identification ability, and it can be used for the accurate identification of cuiguan pear. At the same time, RobustICA-PSO-ELM modeling method can also provide reference for other modeling research methods.

6. Acknowledgment

This work is supported by National major scientific instruments and equipment development projects (2014YQ470377). The authors awfully thanks for anonymous reviewers who made constructive comments.

7. References

[1] Fang Limin, Lin Min. A Method of Near Infrared Multi-component Analysis Based on Independent Component Analysis and Neural Networks Model [J]. Spectroscopy and Spectral Analysis. 2008, 36(6).

[2] Sun Tong, Xu Wenli, Hu Tian, Liu Muhua. Determination of Soluble Solids Content in Nanfeng Mandarin by Vis/NIR Spectroscopy and UVE-ICA-LS-SVM [J]. Spectroscopy and Spectral Analysis. 2013, 12:3235-3239.

[3] Shao Yongni, Cao Fang, He Yong. Discrimination Years of Rough Rice by Using Visible/near Infrared Spectroscopy Based on Independent Component Analysis and BP Neural Network [J]. Journal of infrared and millimeter waves. 2007(6).

[4] Zhao Taoyu, Shao Penghua. Robust Independent Component Algorithm Based on Kurtosis [J]. Computer Development & Applications.2012, 25(8).