ISSN 0101-2061 (Print)
ISSN 1678-457X (Online)

DOI: https://doi.org/10.1590/fst.21320

Comparison of multispectral modeling of physiochemical attributes of greengage: Brix and pH values

Xiwei WANG[1], Xiaoyang XING[1], Maocheng ZHAO[1,2]*, Junrong YANG[1]

Abstract

Chemometric modeling concerns both accuracy and computational expense for the prediction of quality-indicating attributes of food materials. Modeling approaches were explored with the hyperspectral images with pH and Brix values of greengages. A two-phase architecture was applied for modeling. Firstly, waveband selection was performed using two approaches, i.e., succession projection algorithm (SPA) and its combination with genetic algorithm (SPA+GA). Secondly, multispectral models based on the two feature sets of wavebands were built via a total of six different modeling methods, i.e., partial least squares regression (PLSR) and extreme learning machine (ELM) in their respective stand-alone versions, their applications combined with genetic algorithm (GA), and their ensemble enhancements with modified Adaboost.RT (MAdaboost.RT). Analysis of accuracy and computational expense showed that supervised feature selection with SPA+GA was superior to unsupervised SPA for better modeling accuracy. MAdaboost.RT-ELM showed high accuracy at low computational expense. ELM models were the better base models than the PLSR ones, for being more randomized and diverse. It indicates that MAdaboost.RT-ELM on SPA is the best choice for a quick test on a newly available dataset, while switching the dimensionality reduction from SPA to SPA+GA may yield more accurate models with added, but well worthy, computational expense.

Keywords: multispectral modeling; supervised feature wavelength selection; modified Adaboost; RT; extreme learning machine; greengage; Brix; pH.

Practical Application: Fresh greengage may be sorted for proper end product-applications according to the physiochemical attributes of Brix and pH values through multispectral imaging via different approaches to chemometric modelling. Supervised feature wavelength selection coupled with MAdaboost. RT-ELM would be the option for accuracy with low computational expense. While MAdaboost. RT-ELM with unsupervised wavelength-selection would be the choice for a blitz modelling.

1 Introduction

Greengage, which currently grows mainly in the coastal areas of Southeast China, is nutritional and categorized as one of the strong alkaline forming foods. It is useful in neutralizing the acidity of blood and keeping the body fluids alkalescent. Besides, it can be used to alleviate the acidification of body fluids and reduce the risk of diseases, such as cardiovascular and cerebrovascular diseases, osteoporosis, kidney stones, arthritis, gout, and cancer (Tian et al., 2018). Greengage fruit has a long history of being used in traditional Chinese foods as pulp served dried or salted, or as an ingredient in fruit wine, as well as for traditional Chinese medicine. But these simple-processed products have low added values. Currently, the surplus supply of greengage due to its increasing growing area in the past decades has already saturated the market of conventional greengage processing industry. Therefore, there is an urgency to upgrade the greengage processing industry with new products of more added values.

Plum essence and greengage cider are such more value-adding non-traditional products. The former is a deep-processed and value-adding food supplement, and the latter is a novel alcoholic beverage from the health-benign fruit material, sometimes referred also as fermented greengage wine. These two end products demand different physiochemical qualities from greengages as raw material: plum essence prefers greengages high in acidity, whereas greengage cider prefers sweetness (Li et al., 2017; Shen et al., 2017a). Sorting harvest of fresh greengage fruit according to physiochemical attributes would be impossible using the conventional method, which involves preprocessing that is both destructive and time-consuming. Spectral imaging technology has gained its favor in the measurement of physiochemical attributes for a variety of food materials because of its rapidity, objectivity, and a synergy of both spectral and spatial dimensions in particular. Previously, we carried out a promising trial of spectral-imaging-based measurement of greengage acidity (Zhao et al., 2017). However, further improvement of the predictive power of spectral imaging approach, via modeling technology, is yet to be studied for multiple physiochemical attributes of greengage with fewer number of spectral components/wavelengths.

Better accuracy and fewer spectral components were defined as the two goals of this study of modeling optimization. Predictive models of partial least squares (PLS) and extreme learning machine (ELM) were first established as the base models to start

Received 19 July, 2020
Accepted 27 Aug., 2020

[1]*Mechanical and Electronic Engineering School, Nanjing Forestry University, Nanjing, Jiangsu, China*
[2]*Taizhou University, Taizhou, Jiangsu, China*
*Corresponding author: samson_njfu@163.com

the optimization procedures. Optimization procedures of SPA and SPA-GA were adopted to serve the purpose of reducing the dimensionality of predictive models as well as modified-Adaboost. RT (MAdaboost.RT) and GA, for accuracy. For repeatability, the same optimization procedures were carried out on the two distinctive physiochemical attributes: pH value for acidity and Brix value for sweetness.

2 Materials and methods

2.1 Samples and data acquisition

A batch of 450 greengages were purchased in May 2016 from Fujian, China. For sample consistency, the greengages that had spots, were excessively over- or under-sized, or were already spoiled were removed from the population, leaving the final set of 435 subjects relatively uniform in size and shape.

The AOTF-based spectral imaging system used in this study was the same one in reference (Wang, 2014) that works in the range of 550-1000 nm with bandwidth approximately 20 nm. Therefore, the spectral sampling interval in the data acquisition was set at 5 nm to take a total of 91 images for each spectral scan, thus resulting in swift data acquisition without too much redundancy while retaining as much information as possible. At each wavelength, the snapshot of the subject was collected with a fixed exposure of 0.08 s at full-frame resolution of 1392 × 1040 pixels. The spectral imaging of each sample only lasted for 7 s; hence, the delicate greengage subject did not lose too much water under strong illumination.

For each subject, the sweetness indicating Brix value (unit:°Brix) was measured using the Brix meter of ATAGO PAL-1 (Atago Co.Ltd, Tokyo, Japan), and pH value was measured using a precision pH meter of Raymond PHS-2F (INESA Scientific Instrument Co., Ltd, China) to indicate its acidity. The measurement procedure for the reference data collection is as follows. After spectral images of greengages were obtained, the skin and kernel of each sample were removed. The greengage juice was obtained through extrusion and dropped onto sensors of the two measurement instruments. Readings from the instruments were recorded when stabilized. For each physiochemical attribute, the average of 3 repetitions was used as the reference value for each subject.

2.2 Feature wavelength selection

Spectral imaging technology can be subdivided into hyperspectral imaging technology and multispectral imaging technology according to the number of wavelengths and spectral continuity. Though hyperspectral imaging acquisition systems have the advantage of acquiring large number of images from consecutive spectral bands; however, their high costs often forbid them from being used in production lines. Multispectral imaging is a very cost-effective solution: only a few images are acquired, but well-selected wavelengths carry the information key to the target attribute(s). Multispectral imaging is suitable if required prediction accuracy could be achieved with fewer number of wavelengths, i.e., ideally no more than 10 wavelengths.

Succession projection algorithm (SPA) is a method for the quick selection of effective wavelengths with low redundancy based on the calculation of correlation (Fernandes et al., 2016). This paper used SPA to reduce the dimensionality of spectral images. Only the final 10 characteristic wavelengths of the images that were selected went into the prediction model as spectral input.

Genetic algorithm (GA) is an iterative optimization method mimicking natural evolutionary processes with random selection, crossover and mutation operations. The wavelength selection problem finds its solution when the gene or the set of feature wavelengths eventually converges to the most suitable for the environment, i.e., having the lowest cost according the fitness criterium (Zhang et al., 2017). In this study, the spectral dimension of the original hyperspectral cubes was first reduced from 91 to a smaller quantity through SPA, and then, the remaining feature wavelengths were encoded to form the pool of genomes. The root mean square error of PLSR was used as the cost function of fitness. Moreover, the best 10-genome long gene of the last epoch that survived the iterative processing was exactly the top 10 characteristic wavelengths chosen by GA. This method of reducing the data dimension is called SPA-GA (Zhang et al., 2017). It balances time complexity and prediction accuracy by letting SPA to first reduce the pool of genomes before starting GA. The population of genes in each epoch in this study was 100, the value of mutation probability was 0.01, the value of crossover probability was 0.9, and 2000 epochs of evolution were applied to the GA iterations. Before the GA process began, the spectral dimension was first reduced to 40 with SPA.

2.3 Regression models

Adaboost.RT is an algorithm for regression modeling (Hu et al., 2017; Tian & Mao, 2010). By introducing a fixed threshold as criterion, the regression model is converted to a binary model. However, the classification result of the samples is directly affected by the fixed threshold. The updating of the sampling weight and the sampling method of the next iteration are also affected. Therefore, the selection of appropriate threshold is particularly crucial. When the threshold is too large or too small, it will have an adverse effect on weak learners and eventually lead to instability of the algorithm. To deal with this problem, Modified Adaboost.RT was adopted, in which the threshold e_t is automatically adjusted based on the root mean square error of the predicted values $f(x_i)$ of greengages in each iteration per Equation (1).

$$e_t = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f(x_i) - y_i)^2} \qquad (1)$$

The threshold, φ_{i+1}, decreases when $e_t < e_{t-1}$; Otherwise, the threshold increases, as is governed by Equation (2).

$$\begin{cases} \varphi_{t+1} = \varphi(1-\lambda), & e_t \le e_{t-1} \\ \varphi_{t+1} = \varphi(1+\lambda), & e_t > e_{t-1} \end{cases}, \quad \lambda = \frac{1}{2}\left|\frac{e_t - e_{t-1}}{e_t}\right| \qquad (2)$$

The threshold is automatically adjusted by the rate of the root mean square error in each iteration, and the weight of these training samples that are misclassified is increased, so that the subsequent weak learners pay more attention to these samples that are misclassified. For rapidity and efficiency, this algorithm is widely used (Shrestha & Solomatine, 2006). In this study, this algorithm was used to optimize the base models of respective PLSR and ELM. Because the weighted samples cannot be used to train models when PLSR is used for weak learners, the "resampling" method was used such that subsets of greengage from training set were resampled to create the PLSR base models. The new training set for the next iteration was obtained via resampling based on the distribution of sample weights obtained in previous iteration. Weak predictors are obtained through the new training set. According to the result of reference (Shrestha & Solomatine, 2006), the initial threshold was set between 0 and 0.4 for a stable performance.

GA was also used to improve the base models that predict sweetness and acidity of greengage. The GA-ELM denotes the improvement of ELM with GA (Shen et al., 2017a). The ELM model is a supervised learning algorithm based on a single hidden layer forward feedback network (Huang et al., 2006). The matrix of weight and bias between the input layer and the hidden layer is randomly given and not modifiable. The weight matrix between the hidden layer and the output layer is calculated from the Moor–Penrose matrix of the output matrix of hidden layer (Huang et al., 2012; Iosifidis et al., 2016). L is the number of neurons in the hidden layer, and n is the dimension of a single sample. These n × L + L values of the weight matrix and bias matrix between the input layer and the hidden layer in the ELM are artificially coded into an individual matrix, and we can randomly generate multiple such individuals to form an initial population. After iterations with selections, crossovers, and mutations, one individual with the optimal weight and bias would be finally obtained.

In this study, 16 weak predictors were used in the MAdboost. RT optimization with the initial threshold of 0.2 for ELM and 200 resampling for PLSR in the model optimization. For the GA optimization, the population of 50 individuals was created, with mutation probability at 0.1, crossover probability at 0.8, and a total of 200 epochs for the optimization process.

The root square error ($RMSE_c$) of training set, the correlation coefficient (R_c) of training set, the root mean squared error ($RMSE_p$) of prediction set, and the correlation coefficient (R_p) of prediction set were used to evaluate the performance of a model (Monteiro et al., 2018; Shen et al., 2017b; Yousefi et al., 2018). The values of R_c and $RMSE_c$ were used to represent the robustness of a model. A robust and accurate model is not indicated when R_c is close to 1 and $RMSE_c$ is small, but when R_p is close to 1 and $RMSE_p$ is small at the same time.

3 Data pre-processing

3.1 Reflectance calibration and characteristic spectra

At the time of collecting hyperspectral images, factors such as uneven illumination, dark current noise of the sensor, uneven distribution of diffraction efficiency in the AOTF space, and different transmittance at different positions of the lens may affect the results. Therefore, the original spectral images in terms of CCD counts need to be calibrated into spectral images in terms of reflectance values (Wang et al., 2013).

These spectral images of greengages were collected along with dark images and a 99% standard reflectance calibration plate. The calibration formula is as follows:

$$I_R = \frac{I_A - I_B}{I_w - I_B} \times 100\% \tag{3}$$

where I_R is the calibrated reflectance spectral images, I_A is the captured spectral images of a greengage to be calibrated, I_W is the spectral images of the 99% reflectance standard plate, and I_B is the dark images that acquired when the lens is completely covered up (Wang et al., 2013). When the calculation encounters a denominator value of 0, the average of the difference between the spectral image of the standard plate and the dark image of the current wavelength was used instead.

A 5-point Savitzky–Golay smooth-filtering (Zhao et al., 2017) was applied to the calibrated reflectance hyperspectral images to eliminate random noise, while retaining as much useful information as possible. The characteristic spectra that were deemed as the average of the spectra extracted from each greengage subject, before and after the spectral preprocessing, are compared in Figure 1.

3.2 Training set and prediction set

The data on 435 greengage subjects were divided into a training set of 350 subjects and a prediction set of 85 subjects using an equal-probability sampling according to the physicochemical values. The range, average value, and standard deviation of physicochemical indices (pH and Brix values) are listed in Table 1.

4 Results and analysis

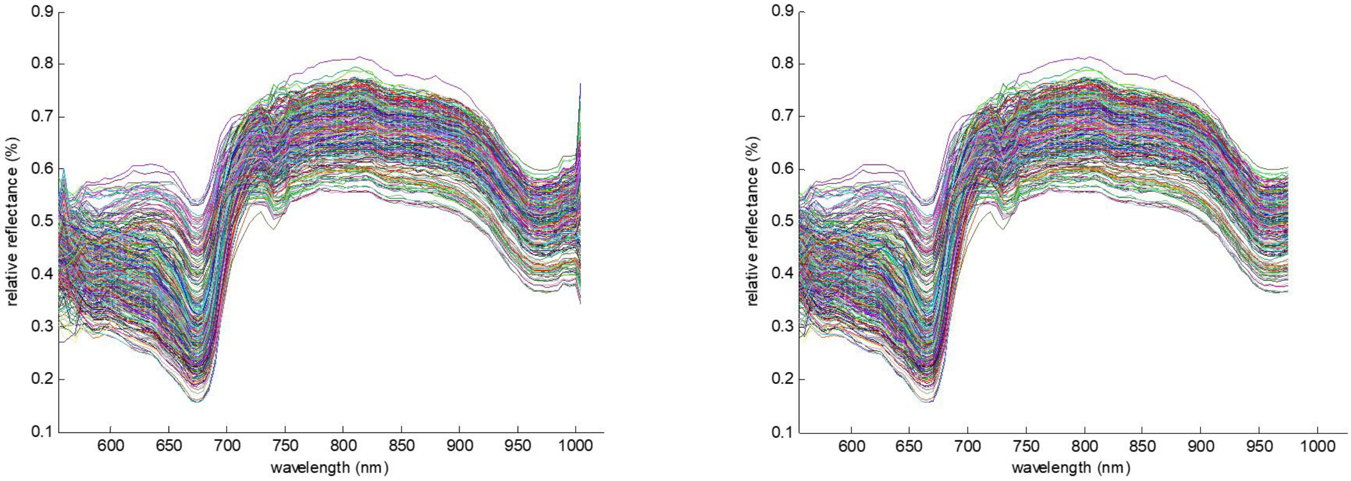
4.1 Chosen sets of wavelengths

Sets of feature wavelengths were chosen from the original hyperspectral range of 550–1000 nm for the multispectral modeling procedures. A total of four feature sets, each containing 10 wavelengths, were selected using SPA and SPA+GA, i.e., 2 for Brix value and the other 2 for pH value.

Using SPA, the set of 10 wavelengths for the modeling of Brix value for sweetness was centered at 740, 640, 550, 965, 725, 685, 720, 665, 805, and 880 nm. Another set for pH value, or acidity, was centered at 885, 640, 550, 725, 685, 720, 960, 665, 800, and 750 nm. Both sets are in descending order of importance. Similarly, using SPA-GA, the top 10 spectral bands for sweetness modeling comprised 550, 590, 685, 760, 830, 860, 865, 890, 930, and 965 nm, while the top 10 acidity wavelengths were 565, 600, 610, 805, 840, 885, 905, 945, 955, and 975 nm.

As shown in Figure 2a and b, all feature wavelengths are plotted along the first derivative of the characteristic spectrum of greengages or the average of their spectral curves.

The feature wavelengths that were selected by SPA was usually centered at the local peaks of the first derivative spectrum, and almost



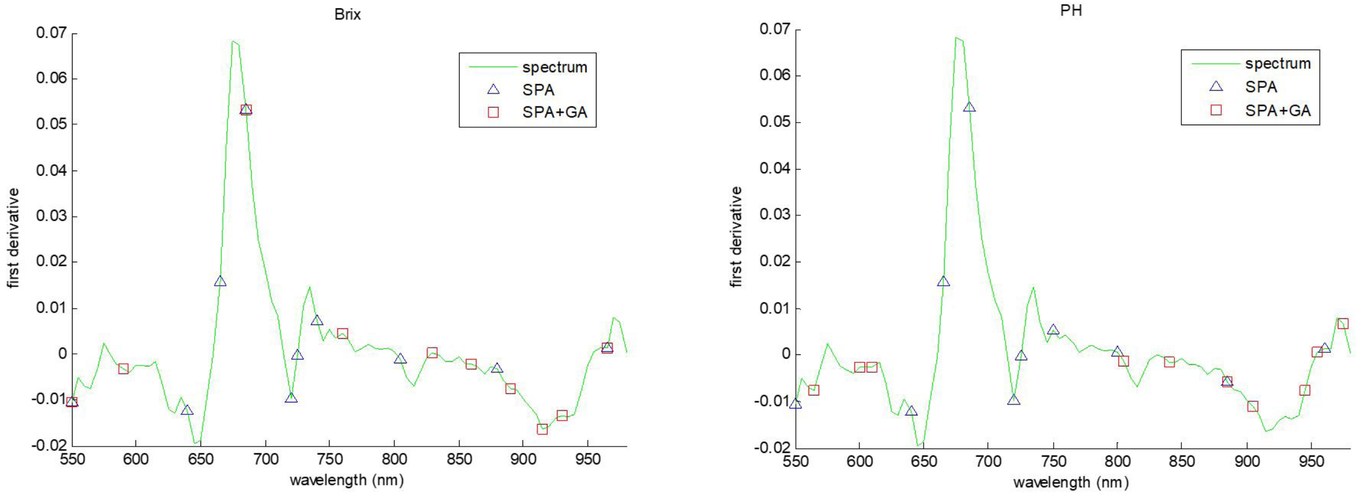
（a）Original spectral data

（b）Filtered spectral data

Figure 1. Spectral data of greengages.

Table 1. Physicochemical distributions of data partition.

Index	Data set	size (/1)	Range (/°Brix, or 1)	Mean (/°Brix, or 1)	Standard Deviation (/°Brix, or 1)
Brix	Calibration	350	5.1-11.8/°Brix	7.4500/°Brix	1.1923/°Brix
	Prediction	85	5.6-10.3/°Brix	7.3871/°Brix	1.0793/°Brix
pH	Calibration	350	2.06-2.62	2.3017	0.1035
	Prediction	85	0.08-2.55	2.3020	0.0953



（a）Characteristic wavelengths of sweetness

（b）Characteristic wavelengths of acidity

Figure 2. Feature wavelengths chosen for sweetness and acidity with two methods for dimensionality reduction.

no wavelengths were selected in flat portions. These phenomena indicate the low redundancy of these wavelengths selected by the SPA. Simultaneously, it was found that these characteristic wavelengths used to predict sweetness and acidity of greengages were remarkably close to each other. The reason the chosen sets by SPA were not identical for both Brix and pH is simply because the divisions of dataset for Brix values were different from those for pH values, in order to keep training sets and test sets from the same distributions for each physiochemical target attribute. Otherwise, if the same data partition were used for both target attributes, the SPA feature wavelengths for both target attributes would be the same, since the SPA algorithm takes only the spectral data, and no input is taken from the measurement of target physiochemical attribute.

In contrast, the sets of wavelengths chosen with SPA-GA were much more different for the two target attributes. And the feature sets for different physiochemical attributes have much few wavelengths in common. It seems that the second stage of the GA process drew information well from the reference data of target attribute and tailored resulting feature sets accordingly. However, it is yet unknown whether the feature sets tailored to the GA criteria of PLS error apply to other modeling algorithms; this needs to be analyzed in Section 4.2 on multispectral regression modeling.

4.2 Regression modeling

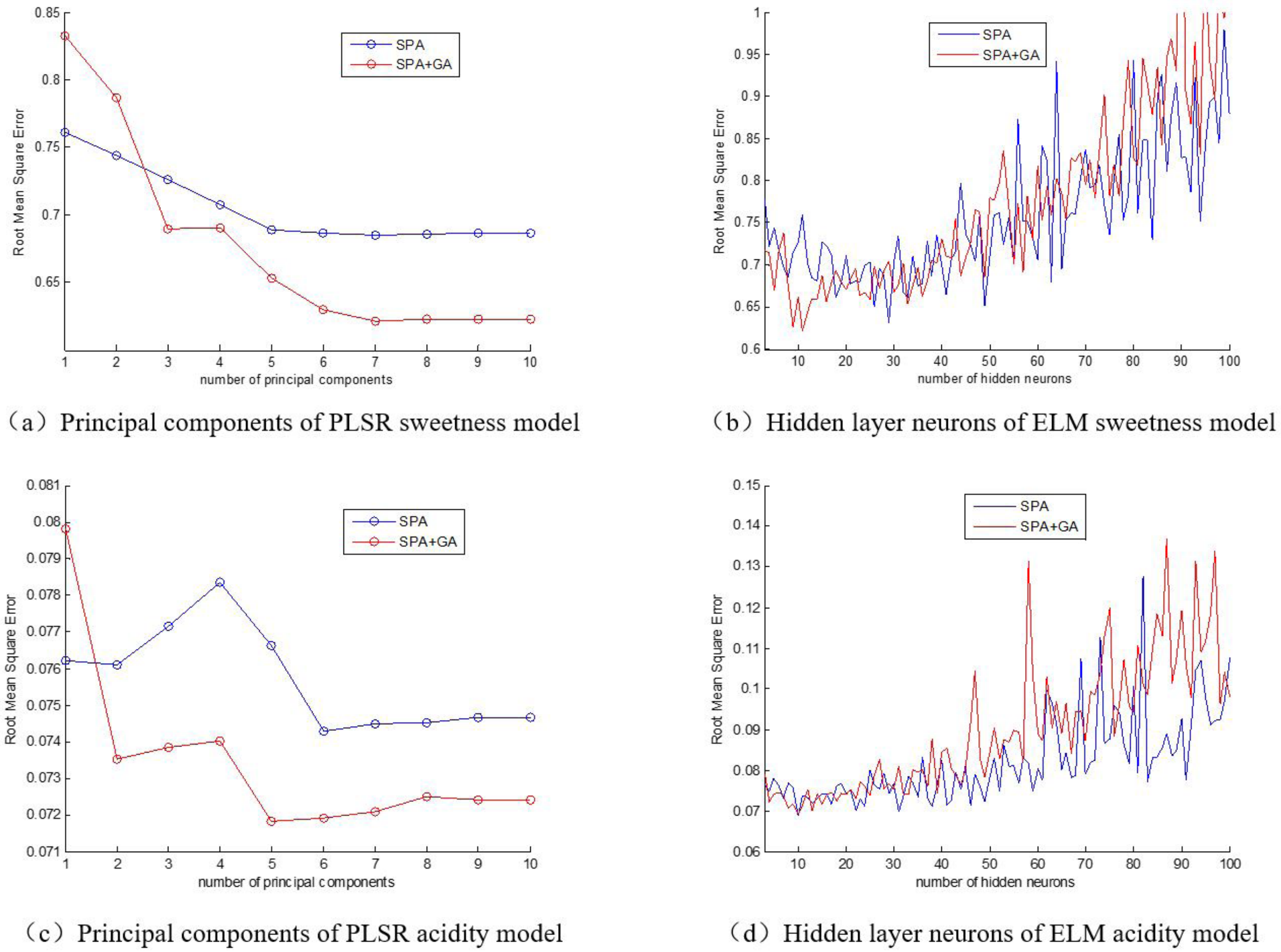
Configuration

After data dimensionality reduction, the four sets of feature wavelengths were fed to the multispectral modeling, which followed two different branches of base modeling algorithms, i.e., PLSR and ELM. Each branch of the multispectral modeling consists of three different models, starting from the base model in its original sense, then another standalone model as modified form of this base model, and ending in the optimized ensemble of such base models combined under the framework of MAdaboost.RT.

To furnish the framework of MAdaboost.RT, a group of base models need to be created. For the best use of the base modeling algorithm with given target attributes, parameters of the number of principal components of the PLSR and the number of hidden-layer neurons of the ELM were fine-tuned with a 10-fold cross validation on the training dataset. Figure 3 shows the cross-validation results in root-mean-square errors of the physiochemical indices, where (a) and (b) show the Brix values for sweetness, and (c) and (d) show pH values for acidity.

As shown by the results of PLSR modeling in Figure 3a and c, the RMSE$_c$ of both sweetness and acidity gradually decreased with the increase in the number of principal components until passing a certain point; then, the error values stopped dropping or even rose again as an indication of overfitting. Considering the compatibility with the set chosen with SPA as well as that with SPA+GA, eight principal components were chosen for the PLSR base models to be used in the further MAdaboost. RT optimization of Brix values (or sweetness), whereas six PLS principal components were chosen for pH values (or acidity).

As shown by the results of ELM modeling in Figure 3b and d, the initial dropping of R.M.S. error with the increase in number of hidden-layer neurons of both Brix values for sweetness or pH values for acidity was soon replaced with a rising trend as a result of overfitting. For similar consideration of the compatibility with both the set chosen with SPA and that



（a）Principal components of PLSR sweetness model

（b）Hidden layer neurons of ELM sweetness model

（c）Principal components of PLSR acidity model

（d）Hidden layer neurons of ELM acidity model

Figure 3. Selection of base model parameters.

chosen with SPA+GA, the number of hidden-layer neurons of ELMs in sequent MAdaboost.RT optimization was 15 for Brix values (or sweetness), whereas the number was 13 for pH values (or acidity).

In addition, the set of wavelengths chosen with SPA+GA exhibited better performance over that with SPA for both Brix values and pH values while working with PLSR, as shown in Figure 3a and c, whereas no such difference between these two sets of feature wavelengths were exhibited while working with ELM, as shown in Figure 3b and d. The apparent superiority of SPA-GA with PLSR in $RMSE_c$ of the training data was a bias, thus to be ignored, because the same data and performance indices had already been used in the wavelength selection process. Therefore, no comparison was done at this stage between the SPA-GA and the SPA, in terms of accuracy; this comparison was left to the following Section Accuracy that compares all modeling methods with test datasets.

Accuracy

The performance of all the multispectral models on the test dataset, in terms of the respective sets of feature wavelengths chosen with SPA and SPA-GA, is shown in Table 2. Both R_p and $RMSE_p$ were used as gaging indices.

With R_p as the gaging index, the performance data of all the modeling methods using the selected sets of wavelengths for SPA or SPA-GA from Table 2 are plotted in Figure 4.

The superiority of SPA+GA over SPA for feature-wavelength selection was evident in both physiochemical attributes of the Brix value in Figure 4a and the pH value in (b); the columns from the SPA+GA group were taller than their counterparts from the SPA group. This pattern of increased modelling accuracy was repeated over all the six versions of the ELM models; this strongly indicates that although the wavelengths were chosen based on the PLS error in the GA process, the chosen sets were specific to the target attribute and did not depend on the gaging index used for the GA criteria. The specificity to the target attribute makes it possible to delineate the data dimensionality reduction and the modelling process, because it allows the wavelength set chosen with SPA+GA to operate effectively on the modelling methods based on algorithms other than its internal GA criteria.

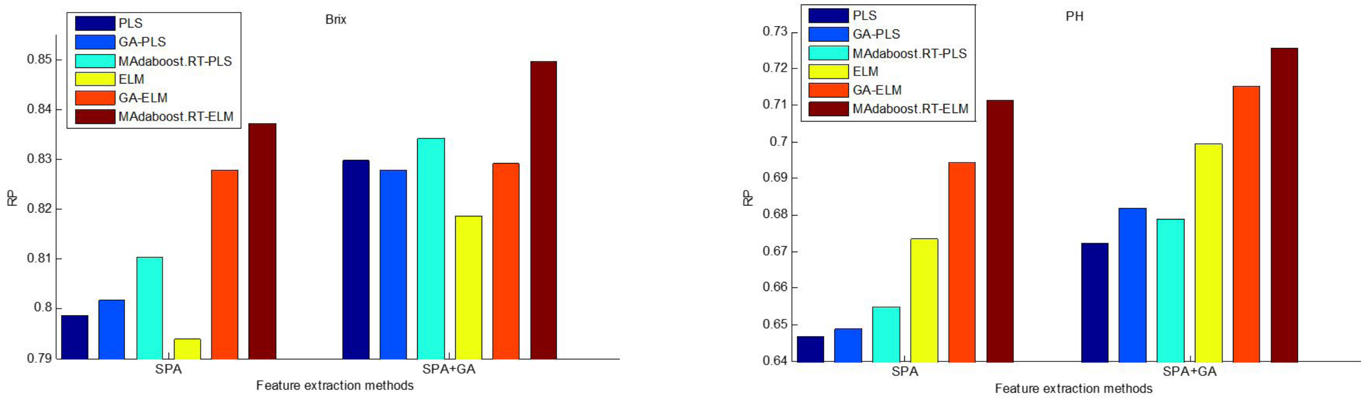
The superiority of MAdaboost.RT as an effective modelling optimization framework was apparent as its application with ELM, i.e. MAdaboost.RT-ELM, topped the prediction of both target attributes regardless of the feature wavelength selection method; moreover, better accuracy was achieved with MAdaboost.RT-PLSR in comparison with the base PLSR models, although MAdaboost.RT performed better with ELM than PLSR did. The difference in the accuracy improvement over the corresponding base models, ELM or PLSR, may be due to the different methods used for generating the base models. The PLSR base models in MAdaboost.RT-PLSR were created by resampling the training set. Thus, the models had a stronger correlation and were less different from each other, leaving less room for improvement for MAdaboost.RT by optimizing the thresholds of these weak predictors. In contrast, the ELM-based weak models in the MAdaboost.RT framework were produced with random matrices and were thus considerably more dissimilar. By adjusting the thresholds according to the performance of weak predictors, focusing specifically on the misclassified samples in the previous round, the learning iterations significantly improved the more diverse weak predictors by maximizing their advantages with the weak predictors being optimally weighted.

GA-ELM modeling also showed a considerable improvement in prediction accuracy over the base ELM model on both target attributes, with feature wavelength sets chosen using both methods. GA appeared to improve the seeking of the weights and biases between the input-layer and hidden-layer neurons in the ELM. Through the internal crossovers and mutations of GA, the guided evolution yielded better performance than that achieved by generating matrices out to pure randomness.

Table 2. Prediction accuracy of different models.

Wavelength selection	Modeling	Brix		pH	
		R_p	$RMSE_p$/°Brix	R_p	$RMSE_p$
SPA	PLSR	0.7938	0.6864	0.6467	0.0743
	ELM	0.7986	0.6832	0.6734	0.0719
	GA-PLSR	0.8017	0.6802	0.6489	0.0732
	GA-ELM	0.8277	0.6247	0.6944	0.0699
	MAdaboost.RT -PLSR	0.8103	0.6654	0.6549	0.0739
	MAdaboost.RT -ELM	0.8372	0.5986	0.7113	0.0678
SPA+GA	PLSR	0.8297	0.6228	0.6724	0.0721
	ELM	0.8187	0.6541	0.6994	0.0695
	GA-PLSR	0.8278	0.6294	0.6820	0.0704
	GA-ELM	0.8291	0.6373	0.7151	0.0704
	MAdaboost.RT -PLSR	0.8342	0.6029	0.6790	0.0710
	MAdaboost.RT -ELM	0.8498	0.5892	0.7254	0.0645

R_p: gaging index.



（a）Brix value, or sweetness

（b）pH value, or acidity

Figure 4. Comparison of multispectral modeling methods grouped in feature-wavelength selection methods.

Nevertheless, the final model out of the GA-ELM process would still be a single ELM model, and according to the results, its performance was not comparable to that of the combination of the ELM base models of MAdaboost.RT-ELM.

Unlike MAdaboost.RT or GA-ELM, GA-PLSR modeling did not guarantee an improvement in prediction accuracy. It did not outperform the base PLSR model in all cases, yielding inferior PLSR Brix values on the SPA+GA set of wavelengths. Nevertheless, it did outperform MAdaboost.RT-PLSR in some cases, as in the pH prediction on the SPA+GA set of feature wavelengths. This uncertainty in the performance may due to the mechanism of GA-PLSR modeling. It further dropped wavelengths from the top 10s provided by the feature selection process, and the final GA-PLSR model utilized only a subset from the top 10 wavelengths. In essence, eliminating redundancy may improve the prediction accuracy in some cases; however, modeling on fewer wavelengths may also incur a greater risk of missing key information and may sometimes result in worse accuracy.

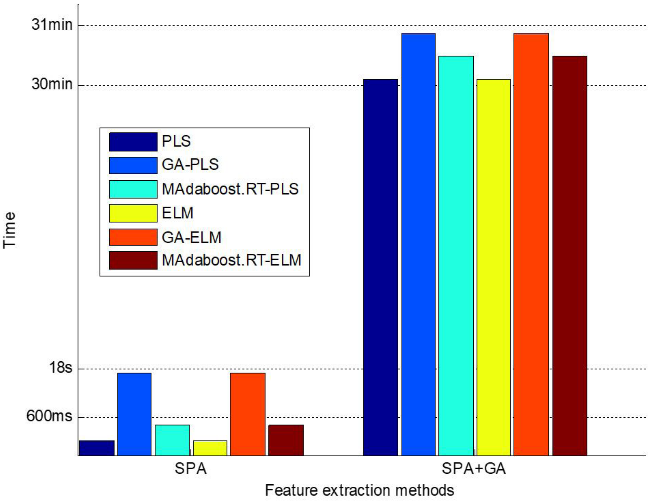


Figure 5. Time complexity of different models.

4.3 Computational expense

Figure 5 shows the time required to build prediction models using different optimization methods, based on the greengage data. The dimensionality reduction method based on SPA+GA required a considerably longer time than based on only the SPA, because a large portion of time was required for the GA iterations for selecting feature wavelengths. Considering the significant improvement in the modeling accuracy, as mentioned in Section 4.2, the additional computational resources required by the GA process for feature wavelength selection were justified.

The computational expense differs negligibly, by less than 1 min, between modelling methods of PLSR, ELM, or their enhancement using MAdaboost.RT or GA. GA-PLSR and GA-ELM required the longest time for the evolving iterations. Considering the uncertainty in the prediction accuracy improvement, as mentioned in Section 3.2, the time invested in the GA process for the modelling did not yield favorable results in all cases. On the contrary, with only a small additional computational overhead for MAdaboost.RT over its base models, PLSR or ELM, a consistent improvement in prediction accuracy could be achieved, as discussed in Section 4.2. Thus MAdaboost.RT was the best option for multispectral modeling for the prediction of greengage physiochemical attributes.

4.4 Best practices

(1) The supervised feature wavelength selection of SPA+GA for data dimensionality reduction is recommended when moderate computational resources are available. The additional time required for the second stage of the GA selection of wavelengths always yielded an improvement in accuracy of the Brix and pH values in every type of multispectral modelling. This is because, after the initial stage of redundancy removal in SPA, the evolutional process guides the selection of wavelengths toward the target attribute and yields tailored final sets of feature wavelengths with a moderate time overhead;

(2) MAdaboost.RT-ELM is the preferred type of multispectral modelling. It consistently provided a reliably high accuracy at a low computational expense, on both the Brix and pH values, and for

feature wavelength sets chosen with both supervised SPA+GA and unsupervised SPA. Unlike the PLSR base models in MAdaboost. RT-PLSR, the ELM base models are more diverse and provided greater accuracy improvements by appropriately configuring the thresholds and weights. Compared with the individual models of PLSR, ELM, GA-ELM, or GA-PLSR, the fine-tuned combination of base models in MAdaboost.RT-ELM showed an overall superiority;

(3) MAdaboost.RT-ELM with feature wavelengths selected using SPA was the best quick model for multispectral modelling as it consistently provided the highest accuracy among the quick modelling methods, in the prediction of both Brix and pH values, with a negligible computational overhead.

5 Conclusion

In this study, exemplified by two target physiochemical attributes—Brix and pH values—multispectral modeling procedures were repeatedly performed on sets of feature wavelengths selected with supervised or unsupervised algorithms, and the following conclusions were drawn:

(1) Supervised feature wavelength selection is preferred over unsupervised selection for better accuracy. SPA+GA improves the accuracy of all types of multispectral modelling when its GA part guides the selection toward the target attribute, with a moderate computational overhead. The wavelengths chosen with SPA+GA are specific to the target attribute and are compatible with different modelling algorithms;

(2) MAdaboost.RT-ELM is the preferred type of multispectral modelling with a consistently high accuracy at a low computational expense, in terms of both the Brix and pH values and feature wavelength sets chosen with both supervised SPA+GA and unsupervised SPA. Overall, MAdaboost.RT is superior to the individual PLSR, ELM, GA-ELM, and GA-PLSR models. Since the ELM base models generated from random matrices in MAdaboost.RT-ELM are more diverse than the PLSR base models from resampled training data in MAdaboost.RT-PLSR, the former allow for greater improvement by appropriately configuring their thresholds and weights;

(3) MAdaboost.RT-ELM with SPA-based selection of feature wavelengths represents the most accurate quick type of multispectral modelling, at a negligible additional computational cost over the quickest approaches of PLSR or ELM on SPA, with reliable top accuracy among the quick modelling methods.

References

Fernandes, D. D. S., Almeida, V. E., Pinto, L., Véras, G., Harrop Galvão, R. K., Gomes, A. A., & Ugulino Araújo, M. C. (2016). The successive projections algorithm for interval selection in partial least squares discriminant analysis. *Analytical Methods*, 8(41), 7522-7530. http://dx.doi.org/10.1039/C6AY01840H.

Hu, M., Hu, Z., Zhang, M., & Fu, C. (2017). Research on wind power forecasting method based on improved AdaBoost. RT and KELM algorithm. *Dianwang Jishu/Power System Technology*, 41, 536-542. http://dx.doi.org/10.13335/j.1000-3673.pst.2016.0831.

Huang, G. B., Zhu, Q., & Siew, C. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3), 489-501. http://dx.doi.org/10.1016/j.neucom.2005.12.126.

Huang, G.-B., Zhou, H., Ding, X., & Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, 42(2), 513-529. http://dx.doi.org/10.1109/TSMCB.2011.2168604. PMid:21984515.

Iosifidis, A., Tefas, A., & Pitas, I. (2016). Graph embedded extreme learning machine. *IEEE Transactions on Cybernetics*, 46(1), 311-324. http://dx.doi.org/10.1109/TCYB.2015.2401973. PMid:25751883.

Li, Y., Yu, X., Zheng, L., Yuan, H., Jiang, X., Hu, Y., Shen, C., & Pan, M. (2017). Study on the optimization of greengage wine fermentation process by two strains of yeast. Journal of Sichuan University of Science & Engineering. *Journal of Sichuan University of Science & Engineering*, 30, 10-16.

Monteiro, M., Fonseca, A. C., Freitas, A. T., Pinho e Melo, T., Francisco, A. P., Ferro, J. M., & Oliveira, A. L. (2018). Using machine learning to improve the prediction of functional outcome in ischemic stroke patients. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(6), 1953-1959. http://dx.doi.org/10.1109/TCBB.2018.2811471. PMid:29994736.

Shen, J., Li, H., Wu, S., & Lin, Y. (2017a). Current status and development suggestions of greengage industry in Zhao' an county. *Fujian Agricultural Science & Technology*, 5, 65-68.

Shen, L., Chen, J., Ge, Z., Jin, J., Yang, J., & Zhang, H. (2017b). Improved multiple kernel extreme learning machine based on AdaBoost.RT. In *Proceedings of the 29th Chinese Control and Decision Conference* (pp. 4736-4441). Chongqing, China: Institute of Electrical and Electronics Engineers Inc. http://dx.doi.org/10.1109/CCDC.2017.7979333.

Shrestha, D. L., & Solomatine, D. P. (2006). Experiments with AdaBoost. RT, an improved boosting scheme for regression. *Neural Computation*, 18(7), 1678-1710. http://dx.doi.org/10.1162/neco.2006.18.7.1678. PMid:16764518.

Tian, H.-X., & Mao, Z.-Z. (2010). An ensemble ELM based on modified AdaBoost.RT algorithm for predicting the temperature of molten steel in ladle furnace. *IEEE Transactions on Automation Science and Engineering*, 7(1), 73-80. http://dx.doi.org/10.1109/TASE.2008.2005640.

Tian, T., Yang, H., Yang, F., Li, B., Sun, J., Wu, D., & Lu, J. (2018). Optimization of fermentation conditions and comparison of flavor compounds for three fermented greengage wines. *Lebensmittel-Wissenschaft + Technologie*, 89, 542-550. http://dx.doi.org/10.1016/j.lwt.2017.11.006.

Wang, X. (2014). *New approach to detect freshness of pork using spectral imaging* (Ph.D. thesis). Nanjing Forestry University, China.

Wang, X., Zhao, M., Ju, R., Song, Q., Hua, D., Wang, C., & Chen, T. (2013). Visualizing quantitatively the freshness of intact fresh pork using acousto-optical tunable filter-based visible/near-infrared spectral imagery. *Computers and Electronics in Agriculture*, 99, 41-53. http://dx.doi.org/10.1016/j.compag.2013.08.025.

Yousefi, M., Yousefi, M., Ferreira, R. P. M., Kim, J. H., & Fogliatto, F. S. (2018). Chaotic genetic algorithm and Adaboost ensemble metamodeling approach for optimum resource planning in emergency departments. *Artificial Intelligence in Medicine*, 84, 23-33. http://dx.doi.org/10.1016/j.artmed.2017.10.002. PMid:29054572.

Zhang, S., Wang, Z., Zou, X., Qian, Y., & Yu, L. (2017). Recognition of tea disease spot based on hyperspectral image and genetic optimization neural network. *Nongye Gongcheng Xuebao (Beijing)*, 33(22), 200-207.

Zhao, M., Yang, J., Lu, D., Cao, J., & Chen, Y. (2017). Detection methods of greengage acidity based on hyperspectral imaging. *Transactions of the Chinese Society for Agricultural Machinery/Nongye Jixie Xuebao*, 48(9), 318-323. http://dx.doi.org/10.6041/j.issn.1000-1298.2017.09.040.