JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Article

pubs.acs.org/JAFC

A Near-Infrared Reflectance Spectroscopy Method for Direct Analysis of Several Chemical Components and Properties of Fruit, for Example, Chinese Hawthorn

Wenjiang Dong,[†,‡] Yongnian Ni,[*,†,‡] and Serge Kokot[§]

[†]State Key Laboratory of Food Science and Technology, Nanchang University, Nanchang 330047, China

[‡]Department of Chemistry, Nanchang University, Nanchang 330031, China

[§]School of Chemistry, Physics and Mechanical Engineering, Science and Engineering Faculty, Queensland University of Technology, Brisbane 4001, Australia

Supporting Information

ABSTRACT: Near-infrared spectroscopy (NIRS) calibrations were developed for the discrimination of Chinese hawthorn (*Crataegus pinnatifida* Bge. var. *major*) fruit from three geographical regions as well as for the estimation of the total sugar, total acid, total phenolic content, and total antioxidant activity. Principal component analysis (PCA) was used for the discrimination of the fruit on the basis of their geographical origin. Three pattern recognition methods, linear discriminant analysis, partial least-squares-discriminant analysis, and back-propagation artificial neural networks, were applied to classify and compare these samples. Furthermore, three multivariate calibration models based on the first derivative NIR spectroscopy, partial least-squares regression, back-propagation artificial neural networks, and least-squares-support vector machines, were constructed for quantitative analysis of the four analytes, total sugar, total acid, total phenolic content, and total antioxidant activity, and validated by prediction data sets.

KEYWORDS: *near-infrared spectroscopy, comparative chemometrics, prediction: sugar, acid and phenol content, Chinese hawthorn fruit*

INTRODUCTION

Near-infrared spectroscopy (NIRS) is a well-established analytical technique, which has been applied for the simultaneous determination of various components of food; it affords simple sample preparation and rapid, simultaneous analysis of several analytes in a large number of samples as well as geographical origin classification.[1,2] NIRS combined with chemometrics has been used for food authentication.[2] Also, chemometric methods of data analysis have been applied successfully for the prediction and classification of different analytes with the use of NIRS.[3] Fernandez Pierna et al.[4] applied various chemometrics methods, such as partial least-squares regression (PLS), artificial neural networks (ANN), and least-squares-support vector machines (LS-SVM), for large near-infrared spectroscopic data matrices from feed and related products; the results indicated that ANN and LS-SVM methods are very powerful methods for nonlinear data, and LS-SVM improved the RMSE (root-mean-square error) for independent test sets when compared to results from PLS and ANN models on the same data.

NIRS calibration models have also been used for fruit and related products, particularly for the analysis of sugar, acid, phenol, and antioxidants. Xu et al.[5] utilized several chemometrics methods such as the multiple linear regression (MLR), genetic algorithm-PLS (GA-PLS), interval PLS (iPLS), and the successive projection algorithm-MLR combined with GA (GA-SPA-MLR) for analysis of Vis/NIR spectroscopic data for the determination of sugar content in pears. It was found that the GA-SPA-MLR method was satisfactory for applications in industry. Xie et al.[6] investigated the use of NIR transmittance spectroscopy coupled with PLS regression for the non-destructive, simultaneous measurement of titratable acidity as well as malic and citric acids of bayberry fruit; the results demonstrated that NIR spectroscopic techniques have potential for rapid prediction of titratable acidity and citric acid content in bayberry fruit; however, the accuracy of this method was unsatisfactory. Also, NIRS was applied for the determination of antioxidant activity in food, and this work was supported by principal component regression modeling (PCR).[7] In addition, Zhang et al.[8] reported that total phenols, flavonoid content, and antioxidant capacity of rice grain could be analyzed with the use of the same technique in combination with principal component analysis (PCA), PLS, and modified PLS regression methods.

Crataegus pinnatifida Bge. var. *major* N.E.Br. or *C pinnatifida* BGE. (family Rosaceae), also referred to as "Hawthorn", is associated with about 280 wood plant species; these are distributed in the Northern Hemisphere, mainly in China, Europe, and North America.[9] Hawthorn fruit, which is rich in phenolic compounds, is commonly used in food and for traditional medicinal applications. In China, hawthorn plants are widely grown in Shandong, Hebei, Henan, and Liaoning provinces, and the chemical composition, particularly the

Received: September 2, 2012
Revised: December 23, 2012
Accepted: December 24, 2012
Published: December 24, 2012

bioactive constituents, of the fruit are affected by plant varieties, cultivation methods, soil, climate, and geographical origin. In general, in this context, the "Shanlihong" variety from the Shandong province is often preferred.[10] Consequently, rapid, cost-effective and readily accessible analytical methods for the discrimination of fruit varieties are of commercial and health importance.

Hawthorn fruit is consumed worldwide principally because of their high phenolic content, which has been demonstrated by many studies to have some benefit for the prevention of cardiovascular disease.[11] The analyses of total phenol content (TPC) and total antioxidant activity (TAA) of fruit have been investigated with the use of the Folin−Ciocalteu method (FC), and 1,1-diphenyl-2-picrylhydrazyl (DPPH) free radical scavenging activity, Trolox equivalent antioxidant capacity (TEAC), and ferric reducing antioxidant power (FRAP) procedures.[12] In addition, the flavor and sensory characteristics of the fruit are directly influenced by the content and composition of acids, sugars, and sugar alcohols, and such analytes have been commonly determined by gas chromatography and mass spectrometry.[13] The total acid content is used as an important criterion to assess the quality of the hawthorn fruit in a TCM.[14] However, these methods are generally time-consuming and require many chemicals for analysis, which are costly, especially when many fruit samples are analyzed for quality control.

The aims of this study were to research and develop a rapid, simple and nondestructive NIRS method of analysis to discriminate the Chinese hawthorn fruit, collected from different districts (Shandong, Hebei, and Henan provinces), with the aid of unsupervised PCA, and supervised linear discriminant analysis (LDA), PLS-DA, BP (back-propagation)-ANN modeling, and to build calibration models for the prediction of TS, TA, TPC, and TAA of Chinese hawthorn fruit from NIRS data with the use of linear PLSR as well as the nonlinear BP-ANN and LS-SVM methods. In addition, the analytical performance of the three prediction methods was compared.

MATERIALS AND METHODS

Chemicals and Reagents. Folin−Ciocalteu reagent, 2,2-diphenyl-1-picrylhydrazyl radical (DPPH), gallic acid, 2,2′-azino-bis-(3-ethylbenzothiazoline-6-sulfonate) diammonium salts (ABTS), 2,4,6-tris-(2-pyridyl)-*s*-triazine (TPTZ), and 6-hydroxy-2,5,7,8-tetramethyl-2-carboxylic acid (Trolox) were obtained from Sigma-Aldrich (St. Louis, MO); hydrochloric acid, sulfuric acid, sodium carbonate, phenol and phenolphthalein, ferric chloride ($FeCl_3$), sodium acetate, and potassium persulfate (Xilong Chemical Ind., Co., Ltd., Guangzhou, China) as well as sodium hydroxide and methanol (Damao Chemical Reagent Factory, Tianjin, China) and glucose (Donghong Chemical Reagent Factory, Guangzhou, China) were purchased from different companies. All reagents and solvents were analytical grade or high performance liquid chromatography (HPLC) grade, and freshly doubly distilled water was used throughout for aqueous solutions.

Hawthorn Samples and Sample Preparation. Ninety-six hawthorn samples were collected from three different cultivated regions in China: 36 from Shandong (SD), 31 from Hebei (HB), and 29 from Henan (HN). All samples were harvested during the crop season in September or October 2011; they were sliced and dried in a cool and shady place after harvesting, and the seeds were removed manually. Then, the seedless fruit was ground into powder using a grinder and passed through an 80-mesh sieve. Powdered samples were placed in plastic, sealed bags (85 × 60 mm) and stored in a dry, dark place at 4 °C and 85% relative humidity until analysis. To eliminate moisture interference, the samples were dried at 40 °C for 24 h before chemical analysis and NIRS sampling. The data matrix of pretreated NIR spectra (N = 96 samples) was divided into a calibration set (Nc = 72 samples) and a prediction set (Np = 24 samples) with the use of the Kennard−Stone (K−S) algorithm,[15] to evaluate the performance of the calibration models.

Analysis by Common Methods. Total sugar (TS), total acid (TA), total phenolic content (TPC), and total antioxidant activity (TAA) were determined according to the acid−base titration or UV/vis spectroscopic methods.

Determination of Total Sugar. Total sugar (TS) content was determined according to the phenol sulfuric acid method.[16] An accurately weighed, powdered fruit sample (0.25 g) was transferred to a 100 mL conical flask; water (50 mL) and hydrochloric acid (15 mL) were added; the sample was hydrolyzed in a thermostatic vibrator at 100 °C for 60 min. Subsequently, the sample was cooled to room temperature and filtered through #202 filter paper (Wohua Co., Hangzhou, China). The flask was rinsed twice with 30 mL of water, and all filtrates were pooled and diluted to volume in a 250 mL volumetric flask for sugar determination.

Standard solutions containing 100 mg/L glucose were prepared. Aliquots of 0.2, 0.4, 0.6, 0.8, and 1.0 mL were transferred to five different 10 mL test tubes and diluted to 1.0 mL with water, respectively. Each test tube solution was then mixed with 1.0 mL of 5% phenol and 5.0 mL of sulfuric acid; the absorbance of each solution was then measured at 490 nm with a model 8453 spectrophotometer (Agilent, Waldbronn, Germany) in a 1 cm quartz cuvette after 20 min at 30 °C; a standard glucose absorbance curve was then obtained from these measurements. A 0.2 mL aliquot of each hawthorn fruit sample solution was measured in the same manner as the standard glucose solutions. Results were expressed as grams of glucose equivalent/100 g dry weight (g glucose/100 g DW). Each sample was determined in duplicate, and the mean of two measurements was used for further analysis.

Determination of Total Acid. Reference analysis of the total acid content (TA) was carried out according to the procedure in the Chinese Pharmacopoeia.[14] Hawthorn powder (1.0 g) was extracted in 100 mL of water in a magnetic stirring apparatus for 4 h at 25 ± 0.2 °C; the extract was filtered through filter paper. Filtrate (50 mL) was pipetted into a conical flask, diluted with 50 mL of water and titrated with 0.1 mol/L sodium hydroxide against the phenolphthalein indicator. Since 1.0 mL of 0.1 mol/L NaOH is equivalent to 6.404 mg of citric acid, the results were expressed as grams of citric acid equivalent/100 g dry weight (g citric acid/100 g DW) of the fruit sample. Titrations were performed in duplicate, and the average of two determinations was used for further interpretation.

Extraction of the Phenolic Constituents. Powdered samples (0.5 g) were extracted with 25 mL of 80% methanol in a thermostatic water bath, the sample was refluxed twice (80 °C, 3 h), and the extracts were filtered through filter paper. The two filtrates were combined in a 100 mL volumetric flask, made up to the mark with 80% methanol, and stored at 4 °C.

Determination of The Total Phenolic Content. Total phenolic content (TPC) was analyzed in duplicate.[17] The Folin−Ciocalteu reagent was diluted 10-fold with water. Aliquots (1.0 mL each) of hawthorn extracts or standard solutions of gallic acid (50, 100, 150, 200, and 250 mg/mL) were added to different 25 mL test tubes; the diluted Folin−Ciocalteu reagent (5.0 mL) was added to each tube, which was then shaken. The tubes were allowed to equilibrate at ambient temperature (25 ± 0.2 °C) for 4 min, and then 4 mL of 7.5% sodium carbonate (w/v) solution was added; this solution was then immediately diluted to 25 mL with water and mixed thoroughly. This mixture was kept for 90 min at ambient temperature (25 ± 0.2 °C) in the dark and was then transferred to a 1 cm quartz cuvette for absorbance measurements at 750 nm against a water blank with a model 8453 spectrophotometer. The TPC estimates were expressed as grams of gallic acid equivalent/100 g dry weight (g gallic acid/100 g DW) of sample with the use of a gallic acid calibration plot.

The total antioxidant activity (TAA) of the samples was determined by three common chemical methods, which are summarized below:

Determination of TAA-FRAP Assay. Antioxidant activity of hawthorn extract was estimated by the FRAP method.[18] The FRAP

reagent was prepared in an acetate buffer (300 mM, pH 3.6), 2,4,6-tripyridyl-*s*-triazine (TPTZ; 10 mM in 40 mM HCl), and $FeCl_3$ (20 mM). The proportions of acetate buffer, TPTZ, and $FeCl_3$ were 10:1:1 (v:v:v), respectively. For the determination of the antioxidant activity, the FRAP reagent (4.00 mL) was mixed with 1 mL of sample extract and a Trolox standard or control (80% CH_3OH); the reaction was kept at 37 °C for 4 min before absorbance was measured at 593 nm. Antioxdant activity was expressed in terms of Trolox (0.15 mg/mL), and 20, 40, ..., 100 μL aliquots in steps of 20 μL of standard solutions were used to establish a calibration curve. TAA was expressed as micromoles of Trolox/g dry weight of sample (μmol Trolox/g DW). All measurements were made in duplicate.

Determination of TAA-DPPH Assay. The DPPH method involved the free radical, 2,2-diphenyl-1-picrylhydrazyl radical (DPPH).[19] A hawthorn extract (0.5 mL) was added to 0.6 mol/L DPPH in 80% methanol solution (0.5 mL); this solution was diluted to 5 mL with 80% methanol. The mixture was shaken vigorously and kept for 30 min at ambient temperature (25 ± 0.2 °C) in the dark. The absorbance was measured in a 1 cm quartz cuvette at 515 nm against 80% methanol as a blank. A calibration curve was constructed using a series of standard Trolox concentrations similarly to the FRAP method. TAA was expressed as micromoles of Trolox/g dry weight of sample (μmol Trolox/g DW). All measurements were made in duplicate.

Determination of TAA-ABTS Assay. Antioxidant activity was measured using the improved ABTS method. [20] $ABTS^+$ was prepared by the reaction of 7 mmol/L ABTS solution with 2.45 mmol/L potassium persulfate; the mixture was kept in the dark at ambient temperature (25 ± 0.2 °C) for 12−16 h before use. $ABTS^+$ was diluted with ethanol to an absorbance of 0.700 ± 0.002 at 734 nm. For the determination of samples, 800 μL of sample extract was reacted with 3.92 mL of $ABTS^+$, and the absorbance at 734 nm was recorded after 6 min. A calibration curve was constructed using a series of Trolox concentrations as standards as for the FRAP method. TAA was expressed as micromoles of Trolox/g dry weight of sample (μmol Trolox/g DW). All measurements were made in duplicate.

NIR Spectroscopic Measurements. NIR spectra were measured in the diffuse reflectance (DR) mode at ambient temperature (25 ± 0.2 °C) on a U-4100 UV/vis/NIR spectrophotometer (Hitachi, Ltd., Tokyo) equipped with a standard integrating sphere and a PbS detector. The raw spectral data were collected with the use of the UV Solutions 2.1 program (Hitachi, Ltd., Tokyo). Samples (0.55 g each) were placed into a glass cell (diameter = 22 mm, depth = 2.5 mm), which was gently compacted with quartz glass. Spectra (32 scans, 2 nm resolution, 800 to 2500 nm, and 851 points/spectrum) were collected in the log(1/*R*) mode (*R* = relative reflectance; $BaSO_4$, the optical reference standard). Each sample was measured in triplicate, and the average spectrum was used for further analysis.

Data Preprocessing and Software. NIR spectroscopic data (96 objects × 851 wavenumbers) were exported in text format, organized in Excel spreadsheets, and then transferred into MATLAB (version 6.5, Mathworks Inc., Natick, MA, United States) for multivariate analysis. It is common to apply data pretreatment methods for comparison of NIR spectroscopic profiles before constructing calibration models, and in this work, several spectroscopic preprocessing methods were applied and compared; they included multivariate scatter correction (MSC),[21] standard normal variate (SNV),[22] detrending,[22] and Savitzky−Golay filter (9 points, second-order polynomial, and first or second derivatives).[23] The first-derivative transformation was finally selected as the optimal pretreatment method on the basis of the best prediction results, which were compared with the use of the correlation coefficient of prediction (R_{pre}) and root-mean-square error of prediction (RMSEP) of the calibration models.

Development and Validation of Chemometrics Models. Chemometrics models were developed with the use of the first-derivative NIR spectroscopy. For the discrimination of hawthorn fruit on the basis of the geographical origin, the data matrix was submitted for interpretation with the use of the unsupervised pattern recognition technique, PCA, and supervised pattern recognition techniques, LDA, PLS-DA, and BP-ANN. The analytical performance for the determination of chemical components and antioxidant activities was compared with the use of the linear regression method, PLSR, and nonlinear regression methods, BP-ANN and LS-SVM.

Unsupervised Pattern Recognition Methods. PCA is a method which transforms the original data matrix into one composed of orthogonal variables called principal components (PCs). Each PC is a linear combination of the original data, and there are as many PCs extracted from the data matrix as there are original variables, i.e., sampled NIR wavenumbers in this work. Each PC accounts for consecutively decreasing the amount of data variance, which results in the compression of significant data into just a few PC variables. Each data object has a score value on each PC, and each original variable is associated with a loadings value on each PC. The new PC data is commonly displayed in two-dimensional score biplots.[24]

Supervised Pattern Recognition Methods. LDA is a common method for data classification. The method optimizes data vectors to achieve maximum separation between objects. Discrimination functions are linear combinations of descriptors that maximize the ratio of between-class variance and minimize the ratio of within-class variance, and the number of linear discrimination functions is equal to the number of sample classes minus one. Commonly, a plot based on the initial linear discrimination functions is used for classification studies.

In this work, each sample was assigned a dummy variable for calibration modeling (SD = 1, HB = 2, and HN = 3), and these values were used for LDA modeling; a cutoff value of ±0.5 was used as criterion of a value of the dummy variable. In general, it has been suggested that for LDA modeling, the number of objects should be at least three times larger than the number of variables.[25] In this work, the number of variables (851 wavenumbers) was much larger than the number of objects (96 samples), so PCA was applied to reduce the number of variables, i.e., the variables were transformed to the smaller number of PCs for LDA.[26] To obtain the optimum number of PC variables (extracted with the use of PCA) for LDA, models from 3 (77.9% variance) to 32 PCs (96.1% variance) were constructed. The optimal model was selected on the basis of prediction performance and contained 27 PCs (criteria: R_{pre} and RMSEP).

PLS-DA is a variant of PLS. In this method, the response variables are replaced by a set of dummy variables describing the origin as a reference value; a dummy variable was assigned to the samples from a particular origin (SD = 1, HB = 2, and HN = 3), and the classification of samples was based on a cutoff value of ±0.5 of dummy variable. The optimum number of factors was chosen by the leave-one-out cross-validation (LOOCV) method.[27]

BP-ANN has a feed forward network structure including input, hidden, and output layers, and this method was used in the present work. The details of this model have been previously described,[28] and model parameters have to be optimized; this involves the number of neurons in the hidden layer, scale functions, learning rate factor, momentum factor, and initial weights. Dummy variables were also assigned to the samples from each geographical origin as with the LDA and PLS-DA modeling.

Linear and Nonlinear Regression Methods. PLSR is a common, linear method for quantitative analysis, and it investigates relationships between spectral and concentration data.[29] In this method, data is compressed into orthogonal factors, which have similar properties to PCs in PCA. In this work, the number of significant factors for prediction modeling was selected by the leave-one-out cross-validation method and the sought model should have the lowest RMSE of cross-validation.[30]

Presence of nonlinearity in NIRS measurements is well-known, and generally, it occurs because of the following: Beer's law is not followed at high analyte concentrations; nonlinearity in the detector response; light source drift; and particle size. Since nonlinear relationships may occur from time to time with the spectroscopic data, it is appropriate to compare the performance of the PLSR prediction models with nonlinear ones such as BP-ANN,[28] and LS-SVM.[31] For this work, the BP-ANN model parameters were estimated.

LS-SVM is a modified algorithm based on the classical SVM,[31] and it is another nonlinear regression method. Conveniently, and importantly, in this method, only linear equations are used for modeling support vectors. Radial basis function (RBF) is commonly used and incorporates a Gaussian kernel function; it involves the optimization of two important parameters: the regularization parameter, gam (γ), and the RBF kernel function parameter, sig^2 (σ^2). These are extracted with the use of the two-step grid search technique and LOOCV.[27]

It is generally suggested to use spectroscopic data compressed into statistically significant factors or latent variables (LV) for modeling with BP-ANN or LS-SVM methods; it is an important step to select appropriate input variables. In this work, the PLSR LVs were used as input variables for BP-ANN and LS-SVM modeling, respectively; the significant LVs were estimated by the LOOCV method.

RESULTS AND DISCUSSION

NIR Spectroscopy Analysis. Original NIR-DR spectra of 96 hawthorn samples from three different provinces (Figure 1A) are quite similar and broad; they are generally known to consist of many overlapping narrow bands of different vibrational modes. The first-derivative preprocessed spectra of all samples (Figure 1B) show peaks at about 1160 and 1425 nm, which are related to the stretch−bend combination mode of water and the OH first overtone band of phenols,[32,33] respectively. The peaks at 1495 and 1610 nm are attributed to the NH_2 group of nitrogen compounds and a glucose overtone band,[34,35] respectively. A band observed at 1695 nm corresponds to the first overtone of the C−H stretch,[36] while two bands located at 1906 and 1960 nm correspond to the second overtone of C═O and O−H vibrations and O−H first overtone of water,[37,38] respectively. The distinct peaks at 2040 and 2160 nm are assigned to the N−H combinations and the aromatic C−H stretch of the polystyrene-like backbone,[39,40] respectively. When the sample spectra from the three different provinces are examined, they do not show any significant differences and, thus, it is very difficult to discriminate them on the basis of geographical origin. Consequently, the NIRS data were submitted to PCA, LDA, PLS-DA, and BP-ANN analysis.

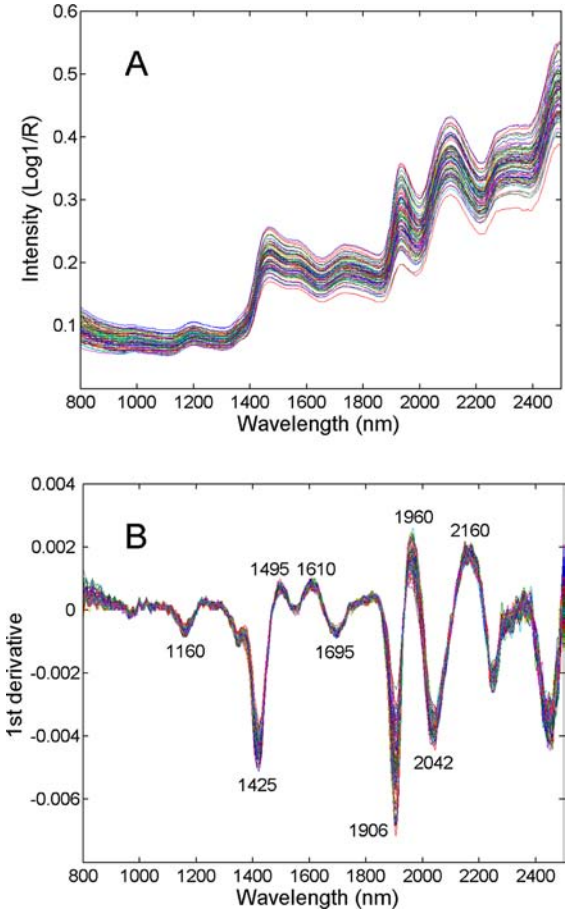

Figure 1. Representative NIR spectra: unprocessed (A) and first-derivative (B) spectra of 96 Chinese hawthorn fruit samples.

Classification of Hawthorn Samples by PCA, LDA, PLS-DA, and BP-ANN Methods. PCA was applied to the first-derivative spectroscopic data matrix (96 objects × 851 variables) from the hawthorn samples. The PC1 versus PC2 scores plot (Figure 2A; PC1, 61.0%, PC2, 11.9% data variance described) shows that, when the objects are projected onto PC1, the SD objects have mostly negative scores while the HN ones mostly have positive scores; thus, these two groups are reasonably well separated. However, when the HB objects are projected onto PC1, they comprehensively overlap the other two object groups; this indicates that the three groups cannot be separated on the basis of their origin. On PC2, the HB objects have positive scores and are reasonably separated from the other two groups, which overlap each other with mostly negative scores, i.e., again there is no clear-cut discrimination of the three types of sample on the basis of their origin. However,

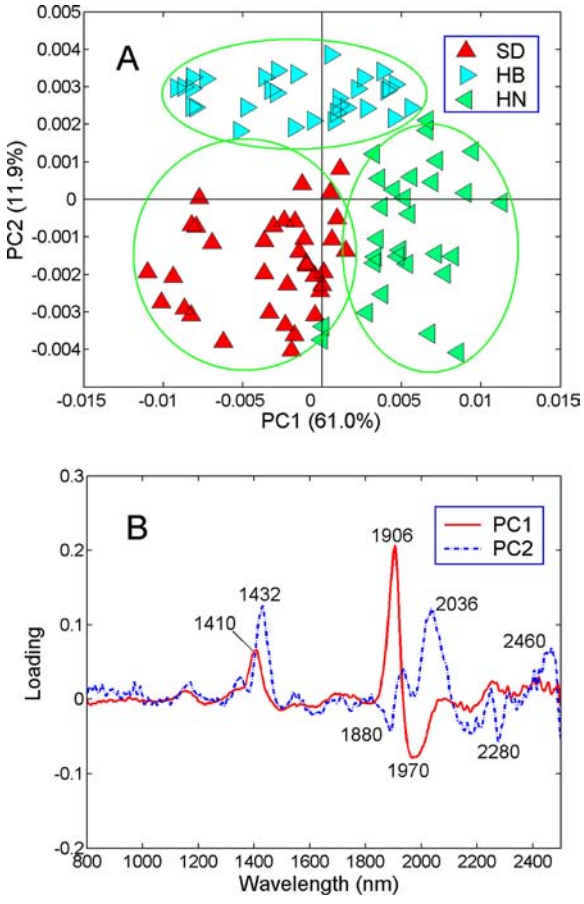

Figure 2. (A) PCA scores plot for 96 samples of Chinese hawthorn fruit: Shandong (SD, red), Hebei (HB, blue), and Henan (HN, green). (B) Loadings profiles for the first two PCs of the NIR spectroscopic data (range: 800−2500 nm).

dx.doi.org/10.1021/jf305272s | J. Agric. Food Chem. 2013, 61, 540−546

in the PC1 versus PC2 plane, the three groups are more or less separated on the basis of their combined properties on the two PCs.

A PC1 and PC2 biplot accounted for most of the NIRS data variance (72.9%) and, hence, was used to demonstrate any object−object clustering (Figure 2B). The loadings profiles indicate the wavenumber variables, which significantly contribute to the discrimination of the objects described in the biplot. The highest loadings with positive values are observed at about 1410 and 1906 nm, and are assigned to the first overtone band of the OH-stretching vibration and the second overtone of C═O as well as O−H vibrations,[37,41] respectively; also, there is one relatively strong negative loading at about 1970 nm, which is assigned to the overtones and combination transitions of the O−H groups.[42] The highest loadings for PC2 with positive values at 1432 and 2036 nm result from the cellulosic OH of a carbohydrate and second overtone carbonyl band,[43,44] respectively. In addition, there is one relatively strong positive loading at about 2460 nm, which is a combination band for aromatic compounds.[45] Furthermore, there are relatively strong, negative loadings values at about 1880 and 2280 nm, which correspond to the O−H and C−O stretch second overtone bands,[40] and C−H combinations.[46]

Thus, PCA indicated qualitatively that sufficient NIR spectroscopic differences exist between the sample types to discriminate on the basis of their origin. The PCA scores were then used in an attempt to classify the three types of samples with the use of LDA. The total variance of the first 27 PCs selected was 95.0%, and these were used as input data for LDA; the reliability of this classification model was studied on the basis of the correct classification rate. The classification results from LDA, PLS-DA, and BP-ANN modeling indicate that satisfactory performance was obtained with respect to the Recognition set for the three types of samples, "SD", "HB", and "HN", with only two samples being misclassified overall; the misclassification occurred with the LDA model. For the Prediction set, the performance was less acceptable with 3 out of 9 samples misclassified. A biplot of the first two LDA discriminant functions displays the scatter of the 96 recognition and prediction data sets (Figure 3) and indicates that the three analyte groups are rather better discriminated on the basis of their geographical origin when compared with the results of the PCA biplot (Figure 2). The PLS-DA model misclassified six samples, giving a 97.2% recognition rate and 83.3% prediction rate, and the BP-ANN method misclassified one sample in the prediction set, giving 95.8% correct classification rate. The prediction results indicated that the BP-ANN method performed better than the other two on this data set. Thus, these observations suggest that NIR spectroscopy combined with the BP-ANN data classification method can be used effectively for the discrimination of hawthorn fruit on the basis of geographical origin.

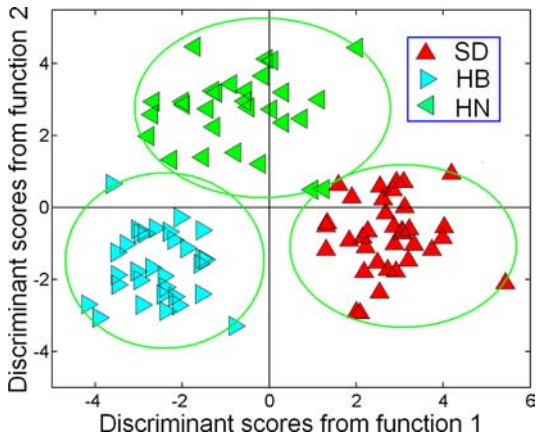

Figure 3. Distribution of 96 samples of the Chinese hawthorn fruit from Shandong (SD, red), Hebei (HB, blue), and Henan (HN, green) on the plane defined by the first two functions of the LDA model.

Calibration Models. The K−S algorithm is a common method for extracting an object subset in multidimensional space, which includes all of the most diverse samples in, for example, a calibration set; this enables the selection of a subset of representative samples. Thus, in the context of NIR spectroscopy, it has been demonstrated that a K−S built training set has a better prediction performance than a randomly built set or one constructed from some other well-known data selection methods such as Kohonen self-organized mapping and D-optimal designs.

Calibration models were constructed with the use of three regression methods, i.e., the PLS model, the BP-ANN model, and the LS-SVM procedure. When compared statistically on the basis of the TS, TA, TPC, and TAA values in hawthorn fruit samples, the calibration and prediction sets are apparently well balanced and analyte values are similarly and evenly spread over the calibration range. Statistical parameters used to evaluate the performance of the models included the following: the correlation coefficient of calibration (R_{cal}) and prediction (R_{pre}) and the root-mean-square error of calibration (RMSEC), cross-validation (RMSECV), and prediction (RMSEP). In addition, the residual predictive deviation (RPD) was calculated for each model as the ratios between the standard deviation (SD) of the TS, TA, TPC, and TAA values of the prediction samples and the RMSEP values. A well behaving model should have large R_{cal}, R_{pre}, and RPD values and low RMSEC and RMSEP; generally, RPD values larger than 3 indicate a well performing prediction model.[47]

PLS Model. A PLS calibration model was constructed from the relevant first-derivative NIRS matrix and each of the four analyte variables (TA, TS, TPC, and TAA). The significant factors were factors in each of the four cases and were selected with the aid of the well-known LOOCV method using the first lowest RMSECV for all models, i.e., 4, 5, 4, and 3 LVs were chosen accordingly. In the calibration step, the model was optimized with the use of the LOOCV method, and generally, a relatively large R_{cal} is expected; in the prediction step, another quite independent data set was taken for prediction, and with the use of the above-mentioned calibration model, a somewhat lower R_{pre} value was obtained in comparison with that from the previous calibration set. The results of calibration and prediction sets for the above-mentioned six variables indicate that only the calibrations for TA and TPC have $R_{cal} > 0.90$ and relatively low RMSEC values, and for prediction, only the TA variable produced satisfactory results, with the R_{pre} and RMSEP values of 0.935 and 0.135. Consequently, it would appear that linear modeling was insufficient to account for the NIRS data.

BP-ANN Model. A one-hidden-layer feed-forward network was used for prediction of TS, TA, TPC, and TAA; the first-derivative NIRS matrix with the four analyte variables (TA, TS, TPC, and TAA) was processed by this method. The sigmoid function was chosen for the hidden layer because it facilitates the processing of a large quantity of nonlinear data and purelin linear transfer function was selected for the output layer due to

its stability.[25] The learning rate, goal of the least mean square difference, and the biggest training epochs were set to 0.1, 0.001, and 800, respectively. The BP-ANN modeling involved the LVs derived from the PLS work as input variables, i.e., input layer neurons. The number of neurons within the hidden layers varied from 1 to 10 as their number was sequentially increased one at a time so as to investigate the model's prediction performance; this was compared with the use of the squared error loss function. The results for the each of the four analytes generally indicated a somewhat improved prediction performance in comparison with the PLS model. This observation indicated that the BP-ANN model performed somewhat better than the PLS one is in general agreement with previous analytical work.[48]

LS-SVM Model. In general, previous studies using LS-SVM for multivariate calibration models suggested that this method performed well in prediction modeling of nonlinear data. In this work, to compare the performance of the LS-SVM with the BP-ANN method, the same number of LVs used in BP-ANN modeling were applied as inputs to the LS-SVM models. Also, the radial basis function (RBF) was included in the LS-SVM algorithm. Two parameters related to the performance of the LS-SVM model were optimized: the regularization parameter gam (γ), which minimizes the training error and minimizes the model complexity, and the bandwidth sig^2 (σ^2). The optimal values of (γ, σ^2) were obtained with the use of the two grid search technique and LOOCV. The calibration and prediction results for the four analytes, TS, TA, TPC, and TAA, indicate that the performance of this method is considerably better than that of the BP-ANN models. This finding is in agreement with a building body of evidence from previous studies.[49,50] These observations are supported by a comparison of the calibration and prediction set values obtained from the traditional wet methods, the NIRS measurements and associated LS-SVM chemometrics models. The correlation equations for prediction of TS, TA, TPC, TAA-FRAP, TAA-DPPH, and TAA-ABTS are as follows: $Y_{TS} = 0.981X_{TS} + 0.703$, $Y_{TA} = 0.994X_{TA} + 0.020$, $Y_{TPC} = 0.984X_{TPC} + 0.056$, $Y_{TAA\text{-}FRAP} = 0.970X_{TAA\text{-}FRAP} + 5.437$, $Y_{TAA\text{-}DPPH} = 0.919X_{TAA\text{-}DPPH} + 10.603$, and $Y_{TAA\text{-}ABTS} = 0.990X_{TAA\text{-}ABTS} + 1.638$, respectively. Regarding the nature and the quality of the NIR spectra, when working with complex substances such as the hawthorn fruit samples, it is common to observe lower sensitivity and some nonlinearity; consequently, relatively lower RPD values for some analytes or activity properties may arise, and the present study is no exception in this regard. However, in this work, when the prediction results were compared statistically with the calibration data, they were quite satisfying. Thus, when the data scatter is examined at 95% confidence level, i.e., $Y \pm 2s$, (s = standard deviation), very few sample points are outside these limits, and those that are, lie quite closely to the lines representing these limits. These observations suggest that the NIRS measurements and their subsequent analytical interpretation are quite acceptable for the prediction of the six components and property analytes in the hawthorn fruit. In addition, further support for this conclusion is provided by the slope values of the six fitted lines (slopes = 0.919−0.994), which are very close to the ideal correlation lines (slope = 1.000).

An NIR spectroscopic analytical method was developed for the discrimination of Chinese hawthorn fruit from different regions and also for the quantitative determination of the chemical components. However, NIRS of these samples are complex and overlapped, so generally suitable chemometrics methods, including linear and nonlinear multivariable calibration methods, should be used to resolve the NIRS and extract the effective information. A comparison of performance of the NIR method and the six standard methods suggests a general agreement between the two approaches, but with the NIRS analysis the potential of high sample throughput and low costs, as well as a significant reduction in solvents and other chemicals, encourages the application of this method with hawthorn fruit discrimination and development of similar methods with other similar agricultural products.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Additional figures and tables. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +86 791 83969500. Fax: +86 791 83969500. E-mail: ynni@ncu.edu.cn.

Funding

This research was supported by grants from the Chinese National Natural Science Foundation (NSFC-21065007) and the State Key Laboratory of Food Science and Technology of Nanchang University (SKLF-MB-201002 and SKLF-TS-200919).

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Williams, P. C. Implementation of near-infrared technology. In *Near-Infrared Technology in the Agricultural and Food Industries*, 2nd ed.; Williams, P. C., Norris, K. H., Eds.; American Association of Cereal Chemists: St. Paul, MN, 2001; pp 145−169.

(2) Oliveri, P.; Egidio, V. D.; Woodcock, T.; Downey, G. Application of class-modelling techniques to near infrared data for food authentication purposes. *Food Chem.* **2011**, *125*, 1450−1456.

(3) Ni, Y. N.; Mei, M. H.; Kokot, S. Analysis of complex, processed substances with the use of NIR spectroscopy and chemometrics: Classification and prediction of properties—the potato crisps example. *Chemom. Intell. Lab. Syst.* **2011**, *105*, 147−156.

(4) Fernandez Pierna, J. A.; Lecler, B.; Conzen, J. P.; Niemoeller, A.; Baeten, V.; Dardenne, P. Comparison of various chemometric approaches for large near infrared spectropic data of feed and feed products. *Anal. Chim. Acta* **2011**, *705*, 30−34.

(5) Xu, H. R.; Qi, B.; Sun, T.; Fu, X. P.; Ying, Y. B. Variable selection in visible and near-infrared spectra: Application to on-line determination of sugar content in pears. *J. Food Eng.* **2012**, *109*, 142−147.

(6) Xie, L. J.; Ye, X. Q.; Liu, D. H.; Ying, Y. B. Prediction of titratable acidity, malic acid, and citric acid in bayberry fruit by near-infrared spectroscopy. *Food Res. Int.* **2011**, *44*, 2198−2204.

(7) Zhang, M. H.; Luypaert, J.; Fernandez Pierna, J. A.; Xu, Q. S.; Massart, D. L. Determination of total antioxidant capacity in green tea by near-infrared spectroscopy and multivariate calibration. *Talanta* **2004**, *62*, 25−35.

(8) Zhang, C. Y.; Shen, Y.; Chen, J.; Xiao, P.; Bao, J. S. Nondestructive prediction of total phenolics, flavonoid contents, and antioxidant capacity of rice grain using Near-infrared spectroscopy. *J. Agric. Food Chem.* **2008**, *56*, 8268−8272.

(9) Zhang, Z. S.; Ho, W. K. K.; Huang, Y.; Chen, Z. Y. Hypocholesterolemic activity of hawthorn fruit is mediated by regulation of cholesterol-7α-hydroxylase and acyl CoA: cholesterol acyltransferase. *Food Res. Int.* **2002**, *35*, 885−891.

(10) Zhao, H. C.; Feng, B. T. *Chinese Fruit-Plant Monograph, Hawthorn Flora*; China Forestry Publishing House: Beijing, China, 1996; pp 14−65.
(11) Liu, P. Z.; Yang, B. R.; Kallio, H. Characterization of phenolic compounds in Chinese hawthorn (*Crataegus pinnatifida* Bge. var. *major*) fruit by high performance liquid chromatography-electrospray ionization mass spectrometry. *Food Chem.* **2010**, *121*, 1188−1197.
(12) Stratil, P.; Klejdus, B.; Kuban, V. Determination of phenolic compounds and their antioxidant activity in fruits and cereals. *Talanta* **2007**, *71*, 1741−1751.
(13) Liu, P. Z.; Kallio, H.; Lu, D. G.; Zhou, C. S.; Ou, S. Y.; Yang, B. R. Acids, sugars, and sugar alcohols in Chinese hawthorn (*Crataegus* spp.) fruits. *J. Agric. Food Chem.* **2010**, *58*, 1012−1019.
(14) China Pharmacopoeia Committee. *Chinese Pharmacopoea (I)*; Chinese Medical Science Press: Beijing, China, 2010; p 29.
(15) Kennard, R. W.; Stone, L. S. Computer aided design of experiments. *Technometrics* **1969**, *11*, 137−148.
(16) Dubois, M.; Gilles, K. A.; Hamilton, J. K.; Rebers, P. A.; Smith, F. Colorimetric method for determination of sugars and related substances. *Anal. Chem.* **1956**, *28* (3), 350−356.
(17) Singleton, V. L.; Rossi, J. A., Jr. Colorimetry of total phenolics with phosphomolybdic-phosphotungstic acid reagents. *Am. J. Enol. Vitic.* **1965**, *16*, 144−158.
(18) Frankel, E. N.; Meyer, A. S. The problems of using one-dimensional methods to evaluate multifunctional food and biological antioxidants. *J. Sci. Food Agric.* **2000**, *80*, 1925−1941.
(19) Sun, T.; Tang, J. M.; Powers, J. R. Effect of pectolytic enzyme preparations on the phenolic composition and antioxidant activity of Asparagus juice. *J. Agric. Food Chem.* **2005**, *53*, 42−48.
(20) Re, R.; Pelligrini, N.; Proteggente, A.; Pannala, A.; Yang, M.; Rice-Evans, C. A. Antioxidant activity applying an improved ABTS radical cation decolorization assay. *Free Radical Biol. Med.* **1999**, *26*, 1231−1237.
(21) Geladi, P.; Macdougall, D.; Martens, H. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Appl. Spectrosc.* **1985**, *39*, 491−500.
(22) Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.* **1989**, *43*, 772−777.
(23) Gorry, P. A. General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method. *Anal. Chem.* **1990**, *62*, 570−573.
(24) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*, 2nd ed.; Wiley: New York; 2001.
(25) Jombart, T.; Devillard, S.; Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **2010**, *11* (94), 1−15.
(26) Casale, M.; Sinelli, N.; Oliveri, P.; Egidio, V. D.; Lanteri, S. Chemometrical strategies for feature selection and data compression applied to NIR and MIR spectra of extra virgin olive oils for cultivar identification. *Talanta* **2010**, *80*, 1832−1837.
(27) Baumann, K. Cross-validation as the objective function for variable-selection techniques. *TrAC, Trends Anal. Chem.* **2003**, *22*, 395−406.
(28) Despagne, F.; Massart, D. L. Neural networks in multivariate calibration. *Analyst* **1998**, *123*, 157R−178R.
(29) Geladi, P.; Kowalski, B. R. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.
(30) Martens, H.; Naes, T. *Multivariate calibration*; Wiley: New York, 1989.
(31) Suykens, J. A. K.; Van Gestel, T.; De Brabanter, J.; De Moor, B.; Vandewalle, J. *Least Squares Support Vector Machines*; World Scientific: Singapore, 2002.
(32) Wesley, I. J.; Blakeney, A. B. Investigation of starch-protein-water mixtures using dynamic near infrared spectroscopy. *J. Near Infrared Spectrosc.* **2001**, *9*, 211−220.
(33) Goddu, R. F. Near infrared spectrophotometry. In *Advances in Analytical Chemistry and Instrumentation*; Interscience: New York; 1960, pp 347−417.
(34) Yu, H. Y.; Ying, Y. B.; Fu, X. P.; Lu, H. S. Classification of Chinese rice wine with different marked ages based on near infrared spectroscopy. *J. Food Qual.* **2006**, *29*, 339−352.
(35) Chen, Y. Q.; Bai, G. Xiao, J.; Wang, L.; Luo, Q. M. Noninvasive blood glucose measurement system based on three wavelengths in near-infrared region. *Fifth international conference on photonics and imaging in biology and medicine; Saratov, Russia*; 2007; p 38.
(36) Li, H.; Kikuchi, R.; Kumagai, M.; Amano, T.; Tang, H. N.; Lin, J. M.; Fujiwara, K.; Ogawa, N. Nondestructive estimation of strength deterioration in photovoltaic backsheets using a portable near infrared spectrometer. *Sol. Energy Mater. Sol. Cells* **2012**, *101*, 166−169.
(37) Peiris, K. H. S.; Pumphrey, M. O.; Dowell, F. E. NIR absorbance characteristics of deoxynivalenol and of sound and *Fusarium*-damaged wheat kernels. *J. Near Infrared Spectrosc.* **2009**, *17*, 213−221.
(38) Sato, T.; Uezono, I.; Morishita, T.; Tetsuka, T. Nondestructive estimation of fatty acid composition in seeds of *Brassica napus* L. by near-infrared spectroscopy. *J. Am. Oil Chem. Soc.* **1998**, *75*, 1877−1881.
(39) Moron, A.; Cozzolino, D. Application of near infrared reflectance spectroscopy for the analysis of organic C, total N, and pH in soils of Uruguay. *J. Near Infrared Spectrosc.* **2002**, *10*, 215−221.
(40) Alexander, T.; Tran, C. D. Near-infrared spectrometric determination of di- and tripeptides synthesized by a combinatorial solid-phase method. *Anal. Chem.* **2001**, *73*, 1062−1067.
(41) Iwahashi, M.; Hayashi, Y.; Hachiya, N.; Matsuzawa, H.; Kobayashi, H. Self-association of octan-1-ol in the pure liquid state and in decane solutions as observed by viscosity, self-diffusion, nuclear magnetic resonance and near-infrared spectroscopy measurements. *J. Chem. Soc., Faraday Trans.* **1993**, *89*, 707−712.
(42) Tran, C. D.; De Paoli Lacerda, S. H. Determination of binding constants of cyclodextrins in room-temperature ionic liquids by near-infrared spectrometry. *Anal. Chem.* **2002**, *74*, 5337−5341.
(43) Michell, A. J.; Schimleck, L. R. NIR spectroscopy of woods from Eucalyptus globules. *Appita J.* **1996**, *49*, 23−26.
(44) Miller, C. E.; Edelman, P. G.; Ratner, B. D.; Eichinger, B. E. Near-infrared spectroscopic analyses of poly(ether urethane urea) block copolymers. Part II: phase separation. *Appl. Spectrosc.* **1990**, *44*, 581−586.
(45) Blanco, M.; Maspoch, S.; Villarroya, I.; Peralta, X.; Gonzalez, J. M.; Torres, J. Determination of the penetration value of bitumens by near infrared spectroscopy. *Analyst* **2000**, *125*, 1823−1828.
(46) Golic, M.; Walsh, K.; Lawson, P. Short-wavelength near-infrared spectra of sucrose, glucose, and fructose with respect to sugar concentration and temperature. *Appl. Spectrosc.* **2003**, *57*, 139−145.
(47) Cozzolino, D.; Kwiatkowski, M. J.; Parker, M.; Cynkar, W. U.; Dambergs, R. G.; Gishen, M.; Herderich, M. J. Prediction of phenolic compounds in red wine fermentations by visible and near infrared spectroscopy. *Anal. Chim. Acta* **2004**, *513*, 73−80.
(48) Xu, L.; Deng, D. H.; Cai, C. B. Predicting the age and type of tuocha tea by Fourier transform infrared spectroscopy and chemometric data analysis. *J. Agric. Food Chem.* **2011**, *59*, 10461−10469.
(49) Liu, F.; He, Y.; Sun, G. M. Determination of protein content of *Auricularia auricula* using near infrared spectroscopy combined with linear and nonlinear calibrations. *J. Agric. Food Chem.* **2009**, *57*, 4520−4527.
(50) Shao, Y. N.; Cen, Y. L.; He, Y.; Liu, F. Infrared spectroscopy and chemometrics for the starch and protein prediction of irradiated rice. *Food. Chem.* **2011**, *126*, 1856−1861.