A COMPARISON OF LINEAR REGRESSION METHODS FOR THE DETECTION OF APPLE INTERNAL QUALITY BY NEAR INFRARED SPECTROSCOPY

Dazhou Zhu[1], Baoping Ji[1], Chaoying Meng[2], Bolin Shi[3], Zhenhua Tu[1]
Zhaoshen Qing[1,*]

[1] College of Food Science and Nutritional Engineering, China Agricultural University, Beijing, P. R. China 100083

[2] College of Information and Electrical Engineering, China Agricultural University, Beijing, P. R. China 100083

[3] Institute of Food and Agriculture Standardization, China National Institute of Standardization, Beijing, P. R. China 100088

[*] Corresponding author, Address: College of Food Science and Nutritional Engineering, China Agricultural University, Beijing 100083, P. R. China, Tel: +86-010-81906784, Fax: +86-010-62347334, Email: qingzhaoshen@cau.edu.cn

Abstract: Hybrid linear analysis (HLA), partial least-squares (PLS) regression, and the linear least square support vector machine (LSSVM) were used to determinate the soluble solids content (SSC) of apple by Fourier transform near-infrared (FT-NIR) spectroscopy. The performance of these three linear regression methods was compared. Results showed that HLA could be used for the analysis of complex solid samples such as apple. The predictive ability of SSC model constructed by HLA was comparable to that of PLS. HLA was sensitive to outliers, thus the outliers should be eliminated before HLA calibration. Linear LSSVM performed better than PLS and HLA. Direct orthogonal signal correction (DOSC) pretreatment was effective for PLS and linear LSSVM, but not suitable for HLA. The combination of DOSC and linear LSSVM had good generalization ability and was not sensitive to outliers, so it is a promising method for linear multivariate calibration.

Keywords: linear regression, partial least-squares, hybrid linear analysis, least square support vector machine, near infrared spectroscopy, apple

1. INTRODUCTION

The soluble solids content (SSC) of apple is one of the main factors that influence the consumer's decision on purchasing. The nondestructive determination of the internal quality of apple, such as SSC, is of great importance for quality evaluation. As a fast, green, and nondestructive technique and its ability of on-line detection, near infrared (NIR) spectroscopy is widely used for the determination of the internal quality of apple (Ventura et al., 1998; Lammertyn et al., 1998). Until now, various chemometrics algorithms have been introduced to improve the robustness of model. For analyzing the quality of apple, traditionally used linear multivariate calibration methods were multiple linear regression (MLR) (Ventura et al., 1998; Murakami et al., 1994), principle component regression (PCR) (Lu et al., 2000), and the partial least-squares (PLS) regression (Peirs et al., 2000; Zude et al., 2006). In order to develop the relation between SSC and the spectra by taking into account some non-linear variations, artificial neural network (ANN) was used to construct the model (Ying et al., 2005). However, some problems still exist for these methods. The determination of the optimal number of PLS factors may be a difficult problem, particularly when an automatic analysis process is concerned. In some cases, the empirical criteria (Malinowski, 1986) for factor selection may fail to achieve a satisfying result (Xu & Schechter, 1997). In addition, ANN is based on the empirical risk minimization principle, whereby robust calibration modeling requires massive representative samples to be analyzed, and ANN can easily lead to local minimization. Therefore, new excellent methods are needed to construct the calibration model for SSC of apples.

Berger et al. (1998) have introduced a new method, which combines the advantages of different standard methods and is therefore called hybrid linear analysis (HLA). HLA incorporates the spectrum of the desired species into the calibration procedure, and it is based on the theory of net analysis signal (NAS) (Lorber, 1986). Berger et al. have used HLA to analyze the near infrared Raman spectra of aqueous mixtures of glucose, lactic acid, and creatinine, and found that HLA produced lower prediction errors than PLS. However, this original HLA can only be used when the pure spectrum of the desired species can be obtained. Two variant HLA algorithms were introduced which did not require the pure spectrum to be known. The hybrid linear analysis developed by Xu & Schechter was called HLA/XS (Xu & Schechter, 1997), and the hybrid linear analysis developed by Goicoechea & Olivieri was called HLA/GO (Goicoechea & Olivieri, 1999a). HLA was used to determinate the tetracycline in blood serum (Goicoechea & Olivieri, 1999a), the bromhexine in cough-cold syrups (Goicoechea & Olivieri, 1999b), and several components of ophthalmic solutions (Collado et al., 2000). Overall, the performance of HLA seems to be better than that of PLS. Therefore, HLA was used for the determination of SSC of apple.

Support vector machine (SVM) is a relatively new learning algorithm based on the statistical learning theory (Vapnik, 1995). Least square support vector machine (LSSVM) (Suykens et al., 2002) is the reformulation of the principles of SVM. Compared with SVM, LSSVM is much easier due to its solution of only a set of linear equations. In addition, LSSVM has fewer parameters to be tuned. LSSVM was applied to NIR analysis for discrimination (Belousov et al., 2002; Goodacre, 2003) and quantitative predictions (Cogdill & Dardenne, 2004; Chauchard et al., 2004). In all of these cases, LSSVM was used as a non-linear modeling technique, where the RBF kernel function was mostly used. Delightfully, LSSVM performed well when the spectra had non-linear information, and it had some advantages over other non-linear multivariate regression techniques such as ANN. For analyzing apple quality by NIR, our work before have proved that both ν-SVM and LSSVM were powerful for developing the non-linear relation between spectra and chemical values (Zhu et al., 2007; Zhu et al., 2008). Unfortunately, a fact has been ignored in the chemometrics: The biggest difference between SVM and ANN is the use of a regularized parameter which determines the trade-off between minimizing the training error (or empirical risk) and the model complexity in SVM. Similarly, Bayesian regularization that uses a regularization parameter to modify the performance function is a useful method to improve the generalization ability of ANN. Therefore, linear LSSVM may have advantages over PLS and other linear methods due to the use of regularization parameter.

In this study, PLS, HLA, and linear LSSVM were used to construct the linear calibration models between the SSC of apples and the NIR spectra, the performance of these three linear regression methods was compared. In order to improve the prediction error, direct orthogonal signal correction (DOSC) (Westerhuis et al., 2001) was used to preprocess the spectra.

2. EXPERIMENTAL

2.1 Samples

A total of 113 'Fuji' apples coming from the Shandong province of China was used. The weights of all the apples were larger than 200 g. Apples were stored at room temperature (~25^{o}C) for at least 12 hours before NIR measurements. Three outliers were eliminated according to the plots of studentized residuals versus leverage value (Otto, 2003), and the retained

110 samples were divided into a calibration set and a prediction set. Samples were sorted according to SSC. Every third sample was included in the prediction set, while all remaining samples made up the calibration set. The calibration set was used to construct the model, and the prediction set was used to test the model performance for external samples. The statistical characteristics of the SSC for two data sets are summarized in Table 1.

Table 1. Statistic SSC values of calibration and prediction sets of apple[a]

Sample Set	n[b]	Range	Mean	SD[c]
Calibration	73	10.5-18.0	13.18	1.75
Prediction	37	10.5-17.7	13.14	1.74

[a] Unit used, °Brix. [b] n=number of samples. [c] SD=standard deviation.

2.2 Spectra measurement

FT-NIR spectra were recorded on an ANTARIS FT-NIR spectrometer (Thermo Nicolet Corporation, USA) equipped with a NIR fiber-optic probe. It has a spectral range of 3800～12000 ㎝$^{-1}$. The resolution was set to 2 ㎝$^{-1}$, and the number of spectrum scans was set to 64. The reference was measured every 30 minutes. All the reflectance spectra of apples were measured in the lab at room temperature (~25 ^{o}C). For each apple, the spectra were measured on four evenly distributed equatorial positions, and the mean of these four spectra was calculated. Since the lower and higher parts of the spectra had noise, only the spectral range of 4500-9500 ㎝$^{-1}$ was selected for latter analysis. Thus, each spectrum had 2593 data points.

2.3 SSC Measurement

The SSC of apples was measured by a refractometer (WAY, Shanghai Precision & Scientific Instrument Co., Ltd, China). About 100 g eatable flesh of each apple was cut and the juice was obtained by a juice extractor (AF2000, Dondxing desheng food mechanicl factory, Shantou, China). Then the juice was centrifuged for ten minutes, and a little pellucid juice was dropped onto a refractometer to record the ^{o}Brix.

2.4 Data processing

DOSC was used to preprocess the spectra, the optimal number of DOSC components removed from the spectra and the tolerance factor was selected by leave-one-out (LOO) cross-validation. The cross-validation process was applied on the calibration set. The calibration models of SSC were constructed by PLS, HLA/XS, HLA/GO, and linear LSSVM, respectively. The optimal number of PLS factors and HLA factors were selected by LOO cross-validation and the criterion proposed by Haaland and Thomas (Haaland & Thomas, 1988). In this study, the value of F corresponding to a probability smaller than 0.75 yielded the optimum number of factors for PLS and HLA. For linear LSSVM, the regularization parameter C was selected by two-step LOO cross-validation. First, C increased with an exponential sequence and a cursory range of C was obtained; and then, the optimal C was selected with an equal increasing sequence within the cursory range.

The quality of the calibration model was evaluated by the determination coefficient (R^2) for the model of calibration set, the standard error of cross-validation (SECV), and the relative standard deviation (RSD) for the prediction set. All the calculations were performed in the software system MATLAB 2006a (The Math Works, Inc., Natick, MA, USA).

3. RESULTS AND DISCUSSION

3.1 Influence of parameters on the calibration model

The number of factors for PLS and HLA, as well as the parameter C of LSSVM, had significant influence on the quality of models. SECV decreased sharply with the increase of PLS factor numbers and then reached a local minimum (SECV=0.7102). When the factor number was very large (>20), SECV kept constant (SECV=0.6946) (Fig. 1a). It should be noted that the model was over-fitted with a large factor number (R^2=0.9996 with a factor number of 20). In this case, the traditional cross-validation with minimum PRESS criterion was invalidated. Therefore, the criterion proposed by Haaland and Thomas (1988) was used to select the optimal number of PLS factors. Similar result was obtained for HLA. For both HLA/XS and HLA/GO, SECV decreased with the increase of HLA factors first and then kept constant (Fig. 1a). This was consistent with the result reported by Xu and Schechter (1997), where RMSEP reached a constant value. Xu suggested that all the *m* (the sample number of calibration set) factors can be used. In fact, R^2 was very close to one when the number of HLA factors was larger than 30 (R^2=1.0000 with a factor number of 73), suggesting that the model was over-fitted. Therefore, we selected the optimal HLA factors by cross-validation for both HLA/XS and HLA/GO. For the linear LSSVM model, SECV decreased with the increase of C when C was small (Fig. 1b), suggesting that the increase of model complexity gave better result. When C was very large, the model was too complicated, and the model was over-fitted. Therefore, SECV increased subsequently.

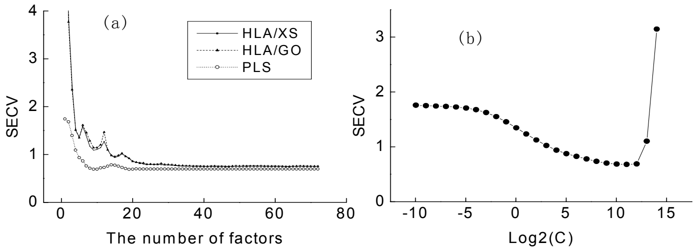

Fig. 1: (a) SECV values versus the numbers of factors for PLS, HLA/XS and HLA/GO models; (b) SECV values versus the parameter C for LSSVM models with linear kernel.

3.2 Calibration models obtained by three regression methods

The optimal models constructed by the three kinds of regression methods were obtained. The results of both HLA/XS and HLA/GO were comparable to PLS (Table 2). Compared with PLS, HLA had higher SECV while their RSD showed no obvious difference. In fact, HLA is a NAS based calibration method, which decomposes the total signal into analyte contribution and contributions from interferences. However, the decomposition may be too radical, thus could not exactly extract useful spectra information. In addition, the optimal number of factors for HLA was relatively large. This was different from the result of Collado et al. (2000); Collado reported that HLA used fewer factors than PLS for the determination of Tetracycline in Blood Serum. Maybe the components of serum were simple or the content of the main components of serum were relatively high, therefore, only few factors could define the projection matrix for HLA. However, the component of apple is very complicated, thus more factors were needed. Among all the three kinds of regression methods, linear LSSVM gave the best results (with R^2 of 0.9304, SECV of 0.6627, and RSD of 4.41%). Due to the foundation of structural risk minimization, linear LSSVM has excellent generalization ability.

Table 2. The calibration models of SSC constructed by three kinds of linear regression methods with DOSC pretreatments.(^{N} The number of factors for PLS and HLA.)

Method	Parameters	R^2	SECV (^{o}Brix)	RSD (%)
PLS	N=8	0.8930	0.7102	5.13
HLA/XS	N=23	0.9635	0.7978	5.17
HLA/GO	N=24	0.9739	0.7954	5.02
LSSVM	C=4800	0.9304	0.6627	4.41
DOSC-PLS	N=1	0.8989	0.5731	4.05
DOSC-HLA/XS	N=18	0.9409	0.8584	5.54
DOSC-HLA/GO	N=19	0.9488	0.8718	5.53
DOSC-LSSVM	C=0.05	0.8988	0.5732	4.01

3.3 Influence of DOSC pretreatment on the three regression methods

DOSC was used to preprocess the spectra. For DOSC-PLS, the optimal number of DOSC components removed from the spectra and the tolerance factor selected by leave-one-out (LOO) cross-validation were 2 and 0.01, respectively. The parameters of DOSC in other regression methods were set the same as that of PLS for latter comparison. After DOSC preprocessing, the model quality of PLS and LSSVM improved, both SECV and RSD decreased (Table 2). The factors of PLS model decreased from eight to one, making the model easier to interpret. Similar results were obtained for HLA, where the factors decreased from 23, 24 to 18, and 19, respectively (Table 2). However, the combination of DOSC and HLA gave even worse results then single HLA. For both HLA/XS and HLA/GO, SECV and RSD increased. Since DOSC had removed the information that was linearly unrelated (orthogonal) to the SSC of apples, the regression methods that directly built the relation between spectra and SSC, such as PLS and LSSVM, would obtain better results. As for HLA, all other components were used to define the projection matrix of HLA and to calculate the NAS. The use of DOSC would remove part of the information of these other components, thus influence the accuracy of NAS. Moreover, the principle of DOSC and HLA was similar, that is, dividing the spectra into two parts: useful signal and interference. However, DOSC is a “mild” one, and HLA is a rather “stringent” one. So the mildness of DOSC is lost on combining it with HLA. This indicated that it may not be appropriate to apply DOSC and HLA simultaneously. Among all the calibration models of SSC, DOSC-LSSVM gave the best results.

3.4 Sensitivity to outliers for the three kinds of regression methods

Outliers are data that have a rather large influence on the regression solution and the occurrence of such data points can lead to considerable deviations from normality (Philips & Eyring, 1983). To investigate the sensitivity to outliers for the three kinds of regression methods, three outliers that had been eliminated were again put into the calibration set, while the prediction set kept constant. The models constructed with and without DOSC pretreatment are shown in Table 3. Both SECV and RSD increased compared with the models without outliers (compare Table 2 and Table 3), suggesting that the existence of outliers had affected the model robustness. Among all of the four models, linear LSSVM had the smallest increase extent for RSD. In addition, the C value of linear LSSVM model decreased from 4800 to 750, this indicated that linear LSSVM gave a smaller weight to the prediction error (see equation 10), thus minimized the influence of outliers. As a result, linear LSSVM performed better than PLS, HLA/XS, and HLA/GO, indicating the latter three methods were more sensitive to outliers than linear LSSVM. Specially, the results of HLA models were very bad, suggesting that outliers must be eliminated when HLA was used to construct a regression model. After DOSC pretreatment, the RSD of PLS and linear LSSVM decreased, while the RSD of HLA/XS and HLA/GO increased. This again indicated that DOSC preprocessing was not suitable for HLA. In addition, PLS and linear LSSVM obtained almost the same result, with RSD of 5.33% and 5.28%, respectively. The same phenomenon has appeared in Table 2, where the RSD of DOSC-PLS and DOSC-LSSVM were 4.05% and 4.01%, respectively. It seems that linear LSSVM performed better than PLS for the original spectra; while the use of DOSC pretreatment helped PLS to achieve a result as good as that of linear LSSVM.

Table 3. The calibration models constructed by three kinds of linear regression methods with three outliers in the calibration set. (^{N} The number of factors for PLS and HLA.)

Method	Parameters	R^2	SECV (°Brix)	RSD (%)
PLS	N=7	0.7449	1.1216	6.87
HLA/XS	N=27	0.9585	1.4785	7.93
HLA/GO	N=27	0.9697	1.4497	8.02
LSSVM	C=750	0.7817	1.0504	5.66
DOSC-PLS	N=1	0.7715	0.8695	5.33
DOSC-HLA/XS	N=24	0.8950	1.7085	8.91
DOSC-HLA/GO	N=22	0.8795	1.8652	9.21
DOSC-LSSVM	C=0.03	0.7714	0.8695	5.28

4. CONCLUSIONS

For the determination of SSC in apples, linear regression methods of PLS, HLA/XS, HLA/GO, and linear LSSVM all obtained satisfying results. HLA was sensitive to outliers, thus the application of HLA requires the elimination of outliers first. The predictive ability of SSC model obtained by HLA was comparable to that of PLS. Among the three kinds of methods, linear LSSVM performed better than PLS and HLA, since it uses a regularized parameter to determine the trade-off between minimizing the training error and the model complexity. DOSC pretreatment was effective for PLS and linear LSSVM, but seems to be not suitable for both HLA/XS and HLA/GO. The combination of DOSC and linear LSSVM had good generalization ability and was not sensitive to outliers, thus gave the best calibration model for the SSC of apples.

REFERENCES

A. Belousov, S. Verzakov, J. von Frese, Applicational aspects of support vector machines, J. Chemom. 2002, 16:482-489

A. J. Berger, T. W. Koo, I. Itzkan, M. S. Feld, An enhanced algorithm for linear multivariate calibration, Anal. Chem. 1998, 70: 623-627

A. Lorber, Error propagation and figures of merit for quantification by solving matrix equations, Anal. Chem. 1986, 58: 1167-1172

A. Peirs, K. Ooms, J. Lammertyn, B. Nicolaï, Prediction of the optimal picking date of different apple cultivars by means of VIS/NIR spectroscopy, Postharvest Biol. Tec. 2000, 21:189–199

D. M. Haaland, E. V. Thomas, Partial Least- Squares Methods for Spectral Analyses, Anal. Chem. 1988, 60: 1193-1208

D. Zhu, B. Ji, C. Meng, B. Shi, Z. Tu, Z. Qing, The application of direct orthogonal signal correction for linear and non-linear multivariate calibration , Chemom. Intell. Lab. Syst. 2008, 90:108-115

D. Zhu, B. Ji, C. Meng, B. Shi, Z. Tu, Z. Qing, The performance of ν -support vector regression on determination of soluble solids content of apple by acousto-optic tunable filter near-infrared spectroscopy, Anal. Chim. Acta. 2007, 598:227–234

E. R. Malinowski, Factor Analysis in Chemistry, Wiley, New York (1986)

F. Chauchard, R. Cogdill, S. Roussel, J. M. Roger, V. Bellon-Maurel, Application of LS-SVM to non-linear phenomena in NIR spectroscopy: development of a robust and portable sensor for acidity prediction in grapes, Chemom. Intell. Lab. Syst. 2004, 71: 141-150

G. R. Philips, E. M. Eyring, Comparison of conventional and robust regression in analysis of chemical data, Anal. Chem. 1983, 55:1134-1138

H. C. Goicoechea, A. C. Olivieri, Determination of bromhexine in cough–cold syrups by absorption spectrophotometry and multivariate calibration using partial least-squares and hybrid linear analyses. Application of a novel method of wavelength selection, Talanta. 1999b, 49: 793–800

H. C. Goicoechea, A. C. Olivieri, Enhanced Synchronous pectrofluorometric Determination of Tetracycline in Blood Serum by Chemometric Analysis. Comparison of Partial Least-Squares and Hybrid Linear Analysis Calibrations, Anal. Chem. 1999a, 71: 4361-4368

J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle, Least Squares Support Vector Machines. World Scientific Publishing, Singapore, 2002

J. A. Westerhuis, S. de Jong, A. K. Smilde, Direct orthogonal signal correction, Chemom. Intell. Lab. Syst. 2001, 56:13-25

J. Lammertyn, B. Nicolaï, K. Ooms, V. D. Semedt, J. D. Baerdemaeker, Non-destructive measurement of acidity, soluble solids, and firmness of Jonagold apples using NIR-Spectroscopy, Trans. of the ASAE. 1998, 41: 1089–1094

L. Xu, I. Schechter, A calibration method free of optimum factor number selection for automated multivariate analysis. Experimental and theoretical study, Anal. Chem. 1997, 69: 3722-3730

M. Murakami, J. Himoto, K. Itoh, Analysis of apple quality by near infrared reflectance spectroscopy, J. Fac. Agr. Hokkaido Univ. 1994, 66: 51–61

M. Otto wrote; X. G. Shao, W. S. Cai, X. J. Xu translated, Chemometrics: Statistics and Computer Application in Analytical Chemistry. Scientific Publishing Company, Beijing, China, 2003

M. S. Collado, V. E. Mantovani, H. C. Goicoechea, A. C. Olivieri, Simultaneous spectrophotometric-multivariate calibration determination of several components of ophthalmic solutions: phenylephrine, chloramphenicol, antipyrine, methylparaben and thimerosal, Talanta. 2000, 52: 909–920

M. Ventura, A. de Jager, H. de Putter, F. P. M. M. Roelofs, Non-destructive determination of soluble solids in apple fruit by near infrared spectroscopy (NIRS), Postharvest Biol. Tec. 1998, 14: 21–27

M. Zude, B. Herold, J.-M. Roger, V. Bellon-Maurel, S. Landahl, Non-destructive tests on the prediction of apple fruit flesh firmness and soluble solids content on tree and in shelf life, J. Food Eng. 2006, 77: 254-260

R. F. Lu, D. E. Guyer, R. M. Beaudry, Determination of firmness and sugar content of apples using near-infrared diffuse reflectance, J. Texture Stud. 2000, 31: 615–630

R. Goodacre, Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules, Vib. Spectrosc. 2003, 32:33-45

R. P. Cogdill, P. Dardenne, Least-squares support vector machines for chemometrics: an introduction and evaluation, J. Near Infrared Spectrosc. 2004, 12:93-100

U. Thissen, B. Üstün, W. J. Melssen, L. M. C. Buydens, Multivariate Calibration with Least-Squares Support Vector Machines, Anal. Chem. 2004, 76:3099-3105

V. Vapnik, The nature of statistical learning theory, New York, Springer, 1995

Y. Ying, Y. Liu, X. Fu, H. Lu, Application of principal component regression and artificial neural network in FT-NIR soluble solids content determination of intact pear fruit, in Optical Sensors and Sensing Systems for Natural Resources and Food Safety and Quality, Ed by Y.-R. Chen, G.E. Meye, S.-I. Tu, Proc. of SPIE, 2005, Vol. 5996, p. 292