

Identification of Wine According to Grape Variety Using Near-Infrared Spectroscopy Based on Radial Basis Function Neural Networks and Least-Squares Support Vector Machines

Jing Yu¹  · Jicheng Zhan¹ · Weidong Huang¹

Received: 9 February 2017 / Accepted: 24 March 2017 / Published online: 20 April 2017
© Springer Science+Business Media New York 2017

Abstract This paper describes how near-infrared (NIR) spectroscopy combined with radial basis function neural networks (RBFNN) and least-squares support vector machines (LS-SVMs) based on principal component analysis (PCA) can be used to classify wines from grape varieties. The effects of different preprocessing methods (standard normal variate (SNV) and multiplicative scattering correction (MSC)) on classification results were also compared. The results show that the use of NIR preprocessing spectral data with optimum RBFNN parameters produced a very high level of correct classification rate, 90.16–98.36%. For RBF LS-SVM, identification rates were from 91.80 to 98.36%. The results demonstrate that, combined with chemometrics with appropriate spectral data pretreatment, NIR spectroscopy has potential to rapidly and non-destructively differentiate wine according to grape variety. The results of this study are helpful to develop a more rapid and nondestructive detection method of wine.

Keywords Near-infrared spectroscopy · Wine · Discrimination · Radial basis function neural networks · Least-squares support vector machines

Introduction

Wine is a fermented alcoholic beverage containing various compounds of different types (Azcarate et al. 2015). Many

factors, including grape variety, yeast strain, wine making technology, and human practice, contribute to the chemical composition in wine and thus affect wine quality (Son et al. 2009). Although these factors are essential to the wine quality, the grape variety is the most basic and important factor for improving the quality of wine (Xiao et al. 2015; Son et al. 2009). Also, authenticity and commercial value of wines are often linked to its geographical origin, and certain countries or regions are known for producing excellent wines of high commercial value (Azcarate et al. 2015; Šelih et al. 2014). Due to its composition and cost, the classification of grape variety is critical for avoiding the adulteration or fraud in wines (de Villiers et al. 2003; Wang et al. 2003).

Some analytical methods have been developed to evaluate the quality of wine worldwide. Gas or liquid chromatography coupled to mass spectrometry (GC-MS or LC-MS) is the most common instrumentation (de Villiers et al. 2012). Recently, near-infrared (NIR) spectroscopy, allied to chemometrics techniques, such as soft independent modeling of class analogy (SIMCA) (Sáiz-Abajo et al. 2004; Chen et al. 2005), principal component analysis (PCA) (Cozzolino et al. 2003), partial least-squares discriminant analysis (PLSDA) (Andre 2003), linear discriminant analysis (LDA) (Casale et al. 2006), locally weighted regression (LWR) (Roussel et al. 2001), radial basis function neural networks (RBFNN) (Fidêncio et al. 2001), support vector machines (SVMs) (Rossi and Villa 2006), and so on, has gained wide acceptance in different fields, like agriculture, food, chemistry, and so on (Blanco and Villarroya 2002). The SVM is a new generation of learning systems, developed by Vapnik and his coworker, and designed to solve the classification problem. It maps input data to high dimensional feature space to transform a linearly separable problem and learns in hyperplane with fewer training data, under the control of a selected kernel function (Vapnik 1995; Schölkopf et al. 1999). Recently, the SVM

✉ Weidong Huang
huanggd@263.net

¹ College of Food Science and Nutritional Engineering, Beijing Key Laboratory of Viticulture and Enology, China Agricultural University, Beijing 100083, People's Republic of China

technique has been employed to an extensive application for nonlinear discrimination and quantitative prediction (Belousov et al. 2002; Goodacre 2003). It has been proved to be a powerful methodology for solving problems in nonlinear classification, function estimation, and density estimation (Pochet et al. 2004), and has also led to many other recent developments in kernel-based learning methods in general. There are several programs for SVM calculation including ν -SVM, LS-SVM, weighted SVM, direct SVM, etc. (Schölkopf et al. 2000; Suykens and Vandewalle 1999; Suykens et al. 2002; Roobaert 2002).

NIR spectra are affected by both the concentration of the chemical constituents and the physical properties of the analyzed product. The physical properties, like shape, size, and so on, account for the majority of the variance among spectra while the variance due to chemical composition is considered small (Lafargue et al. 2003). It is necessary to perform mathematical pretreatments to reduce the effects of scatter (Yan 2005) and enhance the contribution of the chemical composition. The advantages of NIR spectroscopy versus abovementioned techniques lie in its characteristics of rapidity, simplicity, and nondestructive measurement (Chen et al. 2004; Gestal et al. 2004).

The aim of this paper is to demonstrate the feasibility of the combination of NIR spectroscopy and multivariate statistical analysis in the analysis of wine quality. In this paper, NIR transmittance spectroscopy was collected from wines with different grape varieties. RBFNN and least-squares support vector machines (LS-SVMs) based on PCA were applied to classify samples. The effects of different preprocessing methods on classification results were also compared.

Materials and Methods

Samples

One hundred and ninety-one commercial red wine samples for two wine-producing provinces of China were included: 101 from Xinjiang (abbreviation XW) and 90 from Shandong (abbreviation SW). Wine samples were selected from the 2013–2015 vintages. The alcoholic content ranged from 12.2 to 14.8% vol/vol ethanol. All the samples were purchased from a local supermarket.

Apparatus and Software

NIR transmittance spectra were collected with a commercial spectrometer Nexus FT-NIR (Thermo Electron Corp., Madison, WI, USA) which was equipped with a bifurcated optic fiber cable, an InGaAs detector (800–2500 nm), and a wide band light source (Quartz Tungsten Halogen, 50 W).

Both light source beams and receptor beams were enclosed in the fiber probe. They were distributed uniformly.

NIR spectra were collected using specific software OMNIC 6.1a (Thermo Electron Corp., Madison, WI, USA), which can be used to modify spectrometer set-up and store acquired spectra. The mirror velocity was 0.9494 cm s^{-1} , and the resolution was 8 cm^{-1} in this work. Each spectrum consisted of an average of 64 successive scans. Three replicates of each sample were taken, and their mean value was calculated using OMNIC 6.1a.

Samples were analyzed directly using a 1-mm quartz cell, and the transmittance spectra were stored as *T*. Before sample spectra acquisition, a reference spectrum was collected from the empty quartz cell.

Data Analysis

In this study, the pretreatments used were standard normal variate (SNV) and multiplicative scattering correction (MSC).

PCA was performed before RBFNN and LS-SVM classification models were developed to extract the main information on the NIR spectra recorded on wine samples. PCA can reduce spectral data and construct a few new uncorrelated variables, known as principal components (PCs) (Martens and Naes 1989). That is to say, the second PC is orthogonal to the first PC covered as much of information of the variation in the data. The second PC covers as much of the remaining variation as possible, and so on. Each spectrum will have its own unique set of scores; therefore, a spectrum can be represented by its PCA scores in the factor space instead of intensities in the wavelength space (Park et al. 2003). By plotting the PCs, one can view interrelationships between different variables, and detect and interpret sample patterns, groupings, similarities, or differences (Mouazen et al. 2006).

PCA on either the pure spectral data or the pretreated data can provide very important information regarding the potential capability of separation of objects. PCA was carried out using the commercial software package, TQ Analyst v6.2.1 (Thermo Nicolet Corporation, Madison, WI, USA).

The RBFNN is used to solve modeling and classification problems (Pulido et al. 1999). It has superiorities in function approximation and learning speed (Qu et al. 2007). The RBFNN can be considered as three-layer feed-forward neural network with a simple architecture. In RBFNN, the first layer does not process the information; it only distributes the input variables to the hidden layer. In our case, the inputs were scores from the PCA of different preprocessed spectra data. Each neuron of the hidden layer represents a radial function, and the number of radial functions depends on the problem to be solved (Fidêncio et al. 2001). The RBFNN discussed in this study have two neurons in the output layer for two classes to be determined.

The radial functions which are most used are Gaussian function (Derks et al. 1995).

$$K = \exp\left(\frac{-\|x-c_j\|}{2\sigma^2}\right) \quad j = 1, 2, \dots, m \quad (1)$$

where K represents the radial basis function; $\|x-c_j\|$ is the Euclidean distance between x input vector and c_j , the centroids; and σ is the width. The outputs from the radial functions are fully connected to the neurons of the output layer by the strength of weight coefficients w_{jk} . The expression of output is

$$y_k = \sum_{j=1}^m w_{jk}K + b_j \quad k = 1, 2, \dots, p \quad (2)$$

where m is the total number of hidden layer neurons, j represents the j th node in the hidden layer, and b_j is the bias. Finally, the response of each output neuron is calculated by a linear least-squares regression of its inputs, that is, the output of the hidden layer. For a fixed σ value, the learning procedure means finding the optimum c_j and w_{jk} in order to obtain minimum sum-squared-error (SSE) values (Fidêncio et al. 2001). When the error of network output reaches the pre-set error goal value in RBFNN, the procedure of adding hidden neurons will stop (Qu et al. 2007).

LS-SVM considers constraint conditions of loss function, which constructs the Lagrangian by solving the linear Karush-Kuhn-Tucker (KKT) system for the classification problem in least-squares sense. In this study, LS-SVM was used. Resulting from this procedure, solutions can derive directly from solving a set of linear equations (Suykens and Vandewalle 1999)

$$\begin{bmatrix} 0 & I_n^T \\ I_n & K + \gamma^{-1}I \end{bmatrix} \begin{bmatrix} b_0 \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \quad (3)$$

where K denotes the kernel matrix, I_n is an $[n \times 1]$ identity matrix, T means transpose of a matrix or vector, γ is a weight vector, b is regression vector, and b_0 is the model offset. The solution of Eq. (3) can be found by using the most standard methods of solving sets of linear equations. Therefore, LS-SVM avoids complex calculations.

The attainment of the kernel function is cumbersome, and it will depend on each case. Here, a simple Gaussian function, RBF, which is more used nowadays, was chosen. Therefore, only two parameters (γ , σ) are needed for LS-SVM.

All the samples were divided randomly into two parts. One part that contained 130 wine samples (70 for XW samples and 60 for SW ones) was used as calibration set, and the remaining 61 (31 for XW samples and 30 for SW ones) were used as prediction

samples. All calculations were made with a program implemented in the Matlab environment (The Math Works, Natick, USA).

Results and Discussion

The Transmittance Spectra of Wines

Figure 1 shows the mean NIR transmittance spectra of 191 wines with two grape varieties. The spectra were overlapped each other. Figure 1 illustrates the lowest molecular absorptivities are in the short wavelength region (833–1408 nm) with higher values in the first overtone region (1490–1852 nm) and still higher absorbance levels in the combination region (2083–2380 nm). The absorptions around 1923 nm were saturated and with high noise signals. Thus, one segment of the spectrum was removed: from 1894 to 1990 nm due to the saturation of the spectrum caused by the strong combination band of -OH from water. Apart from this region, other regions from 1000 to 1400 nm correspond to with C-H second and third overtones of ethylene and O-H second overtone of water (Gierlinger et al. 2004). The region 1450 and 1850 nm is related to water absorbance. C-H first overtone stretch vibration modes in CH_3 and CH_2 groups occur around 1660–1760 nm region. The 2225 nm band corresponds to C-H combinations of ethylene. The 2340–2400 nm region represents the combination of bending and stretching vibrations associated with C-H and C=C bands.

Unique spectral differences between wines with different grape varieties were expected to present and provide qualitative information.

PCA

PCA was calculated on SNV preprocessed spectra of the samples to verify the separation based on wine types although PCA itself cannot be used as a classification tool. All of the

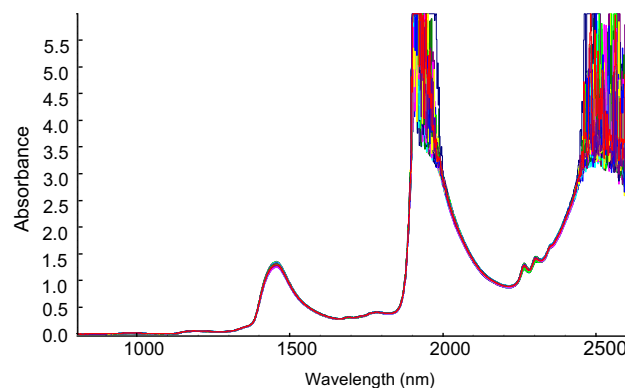


Fig. 1 Mean spectra for the XW and SW wine samples obtained from raw spectral data

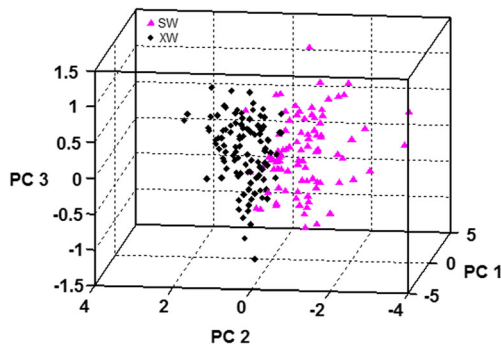


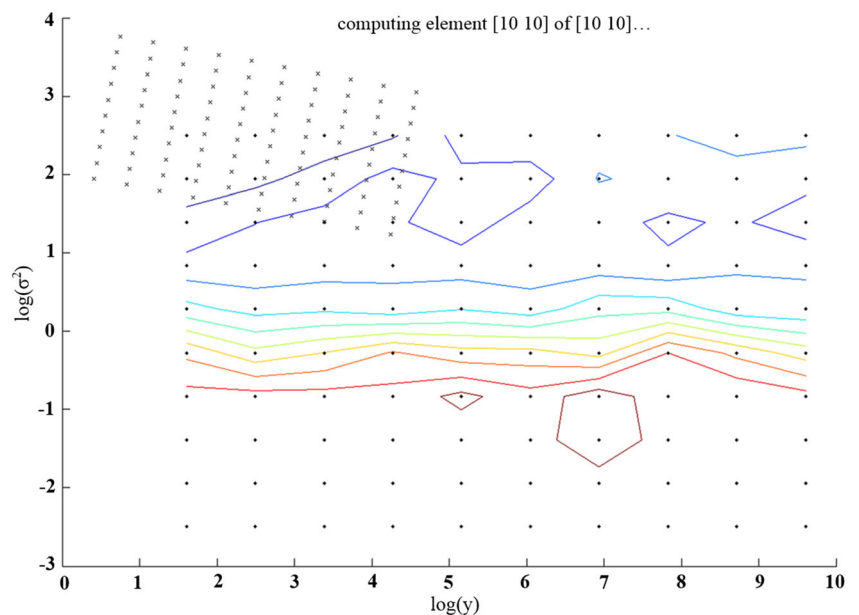
Fig. 2 The first three PC score plots of wines using SNV NIR spectra. *Triangular marks* represent SW samples and *rhombus marks* represent XW samples

spectra of 101 XW wine samples and 90 SW wine ones were used for PCA.

To visualize the data trend and the discriminating efficiency, the three-dimensional (3D) scatter plot of data using the first three PCs, PC1, PC2, and PC3, was obtained as shown in Fig. 3. The initial three factors account for the most spectral variations 90.860% (69.685, 16.236, 4.939% for PC1, PC2, PC3, respectively) related to chemical quality and indicated as positive or negative. So, PCA was performed based on mainly chemical spectral information. As shown in Fig. 2, there was a clear cluster trend in the 3D PC plot. Only a partial overlap is observed between the samples. It can be assumed that the samples can be differentiated by using PCA in terms of the different genes. Figure 2 also shows that most XW samples have positive scores in the second component; however, SW samples have negative scores.

The PCA provides very important information on the potential capability of separation of objects (Andre 2003). The result suggests that the discrimination between wines is

Fig. 3 Contour plot of the optimization the parameters γ and σ^2 for classification of wine with different grape varieties using raw NIR spectra



possible, and the spectral differences associated with characteristics of the sample provided important information for further qualitative analysis.

RBFNN Classification

As has been shown in Table 1, parameters of optimum RBFNN, the number of neurons in the hidden layer, and the value of σ were determined corresponding to different preprocess methods. Table 2 shows the classification results in the validation set. The use of SNV of NIR spectral data with number of neurons = 23 and $\sigma = 1.7$ produced a very high level of classification rate. Overall, a classification accuracy of 98.36% was reached, which demonstrated the good discriminatory power to differentiate wine samples. For SW wines, 100% of the samples were correctly assigned. For XW ones, one sample (number 19) was assigned to SW wine sample and correct classification rate was 96.77%. The worst model that produced the lowest correct classification rate on validation sample set involved 22 neurons, and σ of 3 was obtained using raw spectral data. Three XW samples (numbers 12, 13, and 19) were incorrectly classified (90.32% accurate classification), and the same number of SW samples (numbers 1, 10, and 24) was misclassified (90% accurate classification). Overall identification rate was 90.16%. That is to say, RBFNN method produced a perfect classification for wine with different grape varieties.

LS-SVM Classification

In this study, wines would be classified through LS-SVM. Therefore, it is a problem of two-level classification. To obtain a good performance, the input data and some parameters, γ ,

Table 1 Parameters of optimum RBFNN using raw, SNV, and MSC spectral data

Parameter	Raw spectra	SNV spectra	MSC spectra
σ	3	1.7	1
Number of neurons	22	23	25

which determine the trade-off between minimizing the training error and minimizing model complexity, and σ in LS-SVM have to be chosen carefully. In this study, the top 25 PCs were extracted by PCA and input to the classifiers as latent variables when training. The top 25 principal components almost contained 99.801% of total variance, so they could almost express the total spectral information. To find the optimal model parameters, a “grid search mechanism” was performed on basis of 10-fold cross-validation on the training set. Figure 3 shows the contour plot of the optimization the parameters using raw NIR spectra.

Table 3 shows the identification results of the LS-SVM with different model parameters and preprocess methods. The RBF kernel with $\gamma = 4.03$, $\sigma^2 = 17.55$ and 10 neurons using MSC spectra resulted in the best identification rate for wines with different grape varieties, whose accuracies of identification were 96.67 and 100%, respectively. The use of SNV of NIR spectral data produced a little lower level of classification rate. Overall, a classification accuracy of 96.72% was reached. For XW wines, 100% of the samples were correctly assigned also. For SW ones, two samples (numbers 10 and 24) were assigned to XW wines. The same with RBFNN models, the worst model involved using raw spectral data. One XW sample (numbers 19) was incorrectly classified (96.77% accuracy rate) and four SW samples (numbers 10, 15, 24, and 26) were misclassified (86.67% accuracy rate). Overall identification rate was 91.80%.

In summary, the LS-SVM models using RBF classifiers are helpful to determine the quality of wine based on NIR spectroscopy.

Table 2 Overall prediction results for the validation set objects

Pretreatment	Real class	Predicted class		Identification rate (%)
		XW	SW	
Raw spectra	XW	28	3	90.16
	SW	3	27	
SNV spectra	XW	30	1	98.36
	SW	–	30	
MSC spectra	XW	29	2	93.44
	SW	2	28	

XW wine from Xinjiang Province, SW wine from Shandong Province

Table 3 Overall prediction results of the LS-SVM for the validation set objects

Pretreatment	Parameter			Real class	Predicted class		Identification rate (%)
	γ	σ^2	Neurons		XW ^a	SW ^b	
Raw spectra	25.17	17.93	15	XW	30	1	91.80
				SW	4	26	
SNV spectra	2.34	16.94	12	XW	31	–	96.72
				SW	2	28	
MSC spectra	4.03	17.55	10	XW	31	–	98.36
				SW	1	29	

XW wine from Xinjiang Province, SW wine from Shandong Province

Conclusion

This paper introduced new technique, NIR spectroscopy, to nondestructively differentiate wines with different grape varieties. This study showed that differences between wines of different grape varieties do exist, and NIR spectroscopy combined with RBFNN or LS-SVM with appropriate spectral data pretreatment has potential to judge the relative pattern of the objects that have very similar properties. The results of this study show that the use of SNV NIR spectral data with optimum RBFNN parameters produced a very high level of correct classification rate, 98.36%. The LS-SVM with RBF kernel using MSC spectra resulted in the best identification rate for both wine types, whose accuracies of identification were 96.67 and 100%, respectively. The results demonstrated that NIR spectroscopy combined with chemometrics has the good discriminatory power to differentiate wines from different varieties in a nondestructive way.

Compliance with Ethical Standards

Funding This study was funded by the National “Twelfth Five-Year” Plan for Science and Technology Support (2016YFD0400504).

Conflict of Interest Jing Yu declares that she has no conflict of interest. Weidong Huang declares that he has no conflict of interest. Jicheng Zhan declares that he has no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

- Andre M (2003) Multivariate analysis and classification of the chemical quality of 7-aminocephalosporanic acid using near-infrared reflectance spectroscopy. *Anal Chem* 75:3460–3467

- Azcarate SM, de Gomes A, Alcaraz MR et al (2015) Modeling excitation-emission fluorescence matrices with pattern recognition algorithms for classification of Argentine white wines according grape variety. *Food Chem* 184:214–219
- Belousov AI, Verzakov SA, von Frese J (2002) A flexible classification approach with optimal generalization performance: support vector machines. *Chemom Intell Lab Syst* 64:15–25
- Blanco M, Villarroya I (2002) NIR spectroscopy: a rapid-response analytical tool. *Trends Anal Chem* 21(4):240–250
- Casale M, Sáiz Abajo MJ, González Sáiz JM et al (2006) Study of the aging and oxidation processes of vinegar samples from different origins during storage by near-infrared spectroscopy. *Anal Chim Acta* 557:360–366
- Chen J, Arnold MA, Small GW (2004) Comparison of combination and first overtone spectral regions of near-infrared calibration models for glucose and other biomolecules in aqueous solutions. *Anal Chem* 76(18):5405–5413
- Chen Q, Zhao J, Zhang H et al (2005) Qualitative identification of tea by near infrared spectroscopy based on soft independent modeling of class analogy pattern recognition. *J Near Infrared Spectrosc* 13:327–332
- Cozzolino D, Smyth HE, Gishen M (2003) Feasibility study on the use of visible and near-infrared spectroscopy together with chemometrics to discriminate between commercial white wines of different varietal origins. *J Agric Food Chem* 51:7703–7708
- Derks EPPA, Sánchez MS, Buydens LMC (1995) Robustness analysis of radial base function and multi-layered feed-forward neural network models. *Chemom Intell Lab Syst* 28:49–60
- de Villiers A, Alberts F, Lynen F et al (2003) Evaluation of liquid chromatography and capillary electrophoresis for the elucidation of the artificial colorants brilliant blue and azorubine in red wines. *Chromatographia* 57:393–397
- de Villiers A, Alberts P, Tredoux AGJ et al (2012) Analytical techniques for wine analysis: an African perspective: a review. *Anal Chim Acta* 730:2–23
- Fidêncio PH, Ruisánchez I, Poppi RJ (2001) Application of artificial neural networks to the classification of soils from São Paulo state using near-infrared spectroscopy. *Analyst* 126:2194–2200
- Gestal M, Gómez-Carracedo MP, Andrade JM et al (2004) Classification of apple beverages using artificial neural networks with previous variable selection. *Anal Chim Acta* 524:225–234
- Gierlinger N, Schwanninger M, Wammer R (2004) Characteristics and classification of Fourier-transform near infrared spectra of the heartwood of different larch species (*Larix* sp.). *J Near Infrared Spectrosc* 12:113–119
- Goodacre R (2003) Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules. *Vib Spectrosc* 32:33–45
- Lafargue ME, Feinberg M, Daudin JJ et al (2003) Detection of heterogeneous wheat samples using near infrared spectroscopy. *J Near Infrared Spectrosc* 11:109–121
- Martens H, Naes T (1989) *Multivariate calibration*, 2nd edn. Wiley, Chichester
- Mouazen AM, Karoui R, De Baerdemaeker J, Ramon H (2006) Classification of soils into different moisture content levels based on VIS-NIR spectra, Written for presentation at the 2006 ASABE Annual International Meeting Sponsored by ASABE. Oregon Convention Center, Portland, Oregon
- Park B, Abbott JA, Lee KJ et al (2003) Near-infrared diffuse reflectance for quantitative and qualitative measurement of soluble solids and firmness of delicious and gala apples. *T ASAE* 46:1721–1731
- Pochet N, De Smet F, Suykens JAK et al (2004) Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. *Bioinformatics* 20:3185–3195
- Pulido A, Ruisánchez I, Rius FX (1999) Radial basis functions applied to the classification of UV-visible spectra. *Anal Chim Acta* 388:273–281
- Qu N, Li X, Dou Y et al (2007) Nondestructive quantitative analysis of erythromycin ethylsuccinate powder drug via short-wave near-infrared spectroscopy combined with radial basis function neural networks. *Eur J Pharm Sci* 31:156–164
- Roobaert D (2002) DirectSVM: a simple support vector machine perceptron. *J VLSI Signal Process* 32:147–156
- Rossi F, Villa N (2006) Support vector machine for data classification. *Neurocomputing* 69:730–742
- Roussel SA, Hardy CL, Hurburgh CR et al (2001) Detection of roundup ready™ soybeans by near-infrared spectroscopy. *Appl Spectrosc* 55:1425–1430
- Sáiz-Abajo MJ, González-Sáiz JM, Pizarro C (2004) Near infrared spectroscopy and pattern recognition methods applied to the classification of vinegar according to raw material and elaboration process. *J Near Infrared Spectrosc* 12:207–219
- Schölkopf B, Burges C, Smola A (1999) Three remarks on the support vector method of function estimation in advanced in kernel methods: support vector learning. the MIT Press, Cambridge
- Schölkopf B, Smola AJ, Williamson RC, Bartlett PL (2000) New support vector algorithms. *Neur Comp* 12:1207–1245
- Šelih VS, Šala M, Drgan V (2014) Multi-element analysis of wines by ICP-MS and ICP-OES and their classification according to geographical origin in Slovenia. *Food Chem* 153:414–423
- Son H-S, Hwang G-S, Ahn H-J et al (2009) Characterization of wines from grape varieties through multivariate statistical analysis of 1H NMR spectroscopic data. *Food Res Int* 42:1483–1491
- Suykens JAK, Vandewalle J (1999) Least squares support vector machine classifiers. *Neur Proc Let* 9:293–300
- Suykens JAK, De Brabanter J, Lukas L et al (2002) Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing* 48:85–105
- Vapnik VN (1995) *The nature of statistical learning theory*. Springer, New York
- Wang JF, Geil PH, Kolling DRJ et al (2003) Analysis of zein by matrix-assisted desorption/ionization mass spectrometry. *J Agric Food Chem* 51:5849–5854
- Xiao Z, Fang L, Niu Y, Yu H (2015) Effect of cultivar and variety on phenolic compounds and antioxidant activity of cherry wine. *Food Chem* 186:69–73
- Yan Y (2005) *Foundation of NIR spectral analysis and its application*. China Light Industry Press, Beijing