Journal of Zhejiang University SCIENCE B
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Use of near-infrared spectroscopy and least-squares support vector machine to determine quality change of tomato juice*

Li-juan XIE, Yi-bin YING†‡

(*College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310029, China*)

†E-mail: ybying@zju.edu.cn

Received Sept. 23, 2008; Revision accepted Jan. 19, 2009; Crosschecked Apr. 28, 2009

Abstract: Near-infrared (NIR) transmittance spectroscopy combined with least-squares support vector machine (LS-SVM) was investigated to study the quality change of tomato juice during the storage. A total of 100 tomato juice samples were used. The spectrum of each tomato juice was collected twice: the first measurement was taken when the tomato juice was fresh and had not undergone any changes, and the second measurement was taken after a month. Principal component analysis (PCA) was used to examine a potential capability of separating juice before and after the storage. The soluble solid content (SSC) and pH of the juice samples were determined. The results show that changes in certain compounds between tomato juice before and after the storage period were obvious. An excellent precision was achieved by LS-SVM model compared with discriminant partial least-squares (DPLS), soft independent modeling of class analogy (SIMCA), and discriminant analysis (DA) models, with 100% of a total accuracy. It can be found that NIR spectroscopy coupled with LS-SVM, DPLS, SIMCA, and DA can be used to control the quality change of tomato juice during the storage.

Key words: Near-infrared (NIR) spectroscopy, Least squares-support vector machine (LS-SVM), Quality change, Tomato juice
doi:10.1631/jzus.B0820299   **Document code:** A   **CLC number:** O65

INTRODUCTION

Determination of food quality is one of the most important issues in food industries. In recent years, consumption of beverage has increased with the development of food processing and storage techniques. Tomato is grown worldwide and is the second most consumed vegetable in the world. It is processed to give various products, such as ketchup and juice. Tomato juice is considered to be one of the most important and common beverages. Fresh tomato juice is rich in organic acids (citric, tartaric, malic acid, etc.), sugars (fructose, sucrose, etc.), vitamins, and some natural pigments (lycopene, β-carotene, etc.). However, vitamin C can easily be oxidized in air and some pigments can be decomposed, which changes the nutrition components in juice. Measuring the change of components with a fast, reliable and adapted analytic method is important to determine the nutritional quality of tomato juice.

Near-infrared (NIR) spectroscopy is a rapid, nondestructive, and accurate technique for the quantitative and qualitative analyses of agricultural products based on overtone and combination bands of specific functional groups, e.g., C–H, N–H, and O–H bands, which are the primary structural components of organic molecules (Cozzolino *et al*., 2004; Liu *et al*., 2006). This opens the possibility of using spectra to determine complex attributes of foods. NIR spectroscopy requires minimal sample processing prior to analysis and can be easily automated (Chen *et al*., 2004; Gestal *et al*., 2004). It has attracted considerable attention and has gained wide acceptance in different fields (Casale *et al*., 2006).

Besides the advantages of NIR spectroscopy, there are drawbacks to it, including its low sensitivity and the characteristic broadness of absorption bands,

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 30825027) and the National Key Technology R&D Program (No. 2006BAD11A12) of China

which often results in overlapped spectra (Thompson *et al*., 1997). For such spectra, multivariate calibration methods, such as principal component analysis (PCA), discriminant analysis (DA), soft independent modeling of class analogy (SIMCA), artificial neural networks (ANN), discriminant partial least-squares (DPLS), and so on, open the possibility to unravel and interpret the optical properties of the sample and allow a classification without the use of chemical information (Liu *et al*., 2006).

Support vector machine (SVM) is a new generation of learning systems, developed by Vapnik (1995) group. It is a binary classification tool that is based on the statistical learning theory and designed to solve the classification problem. SVM has been proven to be a powerful methodology to perform nonlinear classification, multivariate function estimation, and nonlinear regression (van Gestel *et al*., 2004; Pochet *et al*., 2004), and has also led to many other recent developments in kernel-based learning methods in general. In this method kernel maps the data into a higher dimensional input space and constructs an optimal separating hyperplane in this space (Suykens and Vandewalle, 1999). SVM learns in hyperplane with fewer training data, under the control of a selected kernel function (Schölkopf *et al*., 1999). SVM does not need a large number of samples to be trained and is not affected by the presence of outliers (Acevedo *et al*., 2007). One of the typical advantages of SVM, when compared with other methods, is that there are very few parameters to tune or select a priori. Recently, SVM technique has been employed to an extensive application for discrimination. Zhao *et al*. (2006) utilized NIR spectroscopy combined with SVM to identify green, black, and oolong teas. Langeron *et al*.(2007) used SVM to classify NIR spectra of tissue samples. Acevedo *et al*.(2007) discriminated wines according to their denomination of origin by ultraviolet (UV)-visible spectrophotometric techniques combined with SVM. Least squares-support vector machine (LS-SVM) is an alternate formulation of SVM proposed by Suykens *et al*. (2002). Complex calculations as in SVM are avoided in LS-SVM (Li *et al*., 2007). In our study, LS-SVM was used.

The main objective of this study was to investigate the quality change of tomato juice and the differences of NIR spectra during storage period in order to establish that NIR is a good nondestructive method for testing the quality of tomato juice. LS-SVM was used to discriminate the spectra recorded in the tomato juice before and after the storage. Discrimination performances of LS-SVM, DPLS, SIMCA, and DA were compared.

MATERIALS AND METHODS

Samples

Fully ripened tomatoes of good quality, cultivated in a standard greenhouse on the university farm, were harvested in June 2006. One hundred samples were selected and numbered randomly to use in the experiments. Each sample was squeezed and centrifuged. The juice samples were filtered in order to separate the dispersed solid particles, and were placed in the same temperature-controlled room where the spectrometer was located before performing the analysis. Two different datasets were used. Dataset A contained 100 juice spectra measured as soon as the tomatoes were squeezed, centrifuged, and filtered at room temperature (20~24 °C) and the tomato juice had not undergone any oxidation and decomposition processes. Dataset B consisted of 100 spectra of the same juice samples after being stored in airtight glass bottles in a fridge at (4±1) °C for one month. One hundred and twenty samples (60 in dataset A and 60 in dataset B) were used for calibration, and the remaining 80 (40 in dataset A and 40 in dataset B) were for validation. The samples were chosen randomly for calibration and validation.

Spectroscopy measurements and reference methods

Samples were scanned in transmission mode using a commercial spectrometer Nexus FT-NIR (Thermo Nicolet Corporation, Madison, WI, USA), which was equipped with an interferometer, an InGaAs detector, and a wide band light source (Quartz Tungsten Halogen, 50 W). Samples were acquired in a rectangular quartz cuvette of 1-mm pathlength with air as reference at room temperature (20~24 °C). The reference spectrum was subtracted from the sample spectra to remove background noise. The rectangular quartz cuvette was cleaned after each sample was scanned to minimize cross-contamination.

NIR spectra were collected using OMNIC software (Thermo Nicolet Corporation). The instrument gathered spectral data over the range of 800~2400 nm, recording 32 scans, which were averaged to give a spectrum represented as values of log(1/*R*) (*R* means reflection) as a function of the wavelength. The nominal spectral resolution was set to 4 cm^{-1} and mirror velocity was 0.9494 cm/s. Both sample and background spectra were collected in absorbance. No part of the spectrometer was modified in these two experiments. The experiments were controlled to be carried out almost under the same conditions and all the setups were the same.

The soluble solid content (SSC) and pH of the juice samples were determined using a digital refractometer (model: PR-101, Atago Co., Tokyo, Japan) and a pH meter (SJ-4A, Exact Instrumentation Co., Shanghai, China), respectively. All of these measurements were performed immediately after NIR spectroscopy measurements.

Chemometrics analysis

The first discrimination was done with PCA, which is widely used in chemometrics due to the graphical concept and loading plot offers, although it is not a classification method itself. By plotting the principal components (PCs), one can view interrelationships between different variables, and detect and interpret sample patterns, groupings, similarities or differences (Mouazen *et al*., 2006). PCA can provide very important information regarding the potential capability of object separation.

LS-SVM, proposed as a class of kernel machines related to many other well-known techniques, is a new and attractive statistical learning method. It is also related to Gaussian process and regularization networks but uses an optimization approach as in SVM. Therefore, LS-SVM encompasses similar advantages as SVM, but its additional advantage is that it requires only solving a set of linear equations, which is much easier and computationally very simple (Thissen *et al*., 2004). A comprehensive introduction of LS-SVM was presented in literatures (Suykens *et al*., 2002; Thissen *et al*., 2004; Borin *et al*., 2006; Li *et al*., 2007).

In LS-SVM, the determination of proper kernel function and optimum kernel parameters is the crucial problem. Kernel functions include the linear function, the polynomial function, the radial basis function (RBF), the sigmoid function, etc. By comparison and analysis, LS-SVM with RBF kernel had a higher accuracy. Therefore, RBF kernel was used as the kernel function of LS-SVM in this paper. In addition, proper parameter setting plays a crucial role in building a good LS-SVM classification model with high prediction accuracy and stability. The parameters γ, which determines the tradeoff between minimizing the training error and model complexity, and σ^2, the width, should be optimized. To obtain the optimal parameter values, an intensive grid search technique with leave-one-out cross-validation was used. Grid search is a two-dimensional minimization procedure based on exhaustive search in a limited range. Leave-one-out cross-validation was used to avoid overfitting. It can provide an almost unbiased estimate of the generalization ability of LS-SVM.

All LS-SVM algorithms were performed using MATLAB v7.0 (the Math Works, Natick, Massachusetts, USA). The free LS-SVM toolbox for MATLAB (LS-SVM v1.5, Suykens, Leuven, Belgium) was applied.

Other classification methods, including DPLS, SIMCA, and DA, were performed using the commercial software package, TQ Analyst v6.2.1 (Thermo Nicolet Corporation).

RESULTS AND DISCUSSION

Quality change of tomato juice during storage

Juice in translucent bottles produces physicochemical and organoleptic changes due to the oxidation and decomposition of some of the compounds present in juice during storage. A series of reactions in the juice happen because of the existing of oxygen and light. From the appearance, darkening of color and presence of precipitates can be observed. The loss of aromas can also be found. Table 1 shows the descriptive statistics for the SSC and pH values analyzed in both datasets. The mean values of SSC and pH reduced during the storage. By comparison, SSC had obviously significant difference ($P<0.01$) and pH value had significant difference ($P<0.05$) between the juice in datasets A and B.

Due to the quality changes of juice described above, there were certain modifications between the original and second derivative NIR spectra (Figs.1a and 1b), which are directly related to the physico-chemical changes that took place. Some bands correspond to the compounds that have undergone changes during storage. The 2210~2216 nm region might be due to O–H stretching and C=O stretching of the carboxylic group of the citric acid or other acids affected by oxygen (Casale *et al*., 2006). The differences of loading spectra for SSC and pH could also be found (Fig.2). This figure indicates that the region before 2000 nm cannot provide useful information of spectral difference. The region that has influence on the classification might be after 2000 nm.

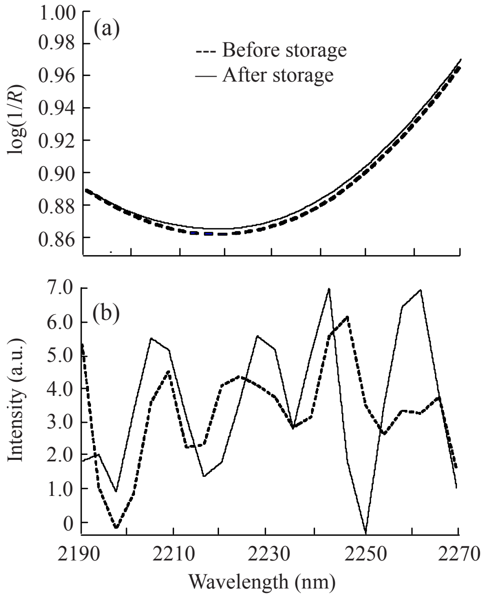

Fig.1 Average (a) raw and (b) second derivative NIR spectra of tomato juice before and after the storage

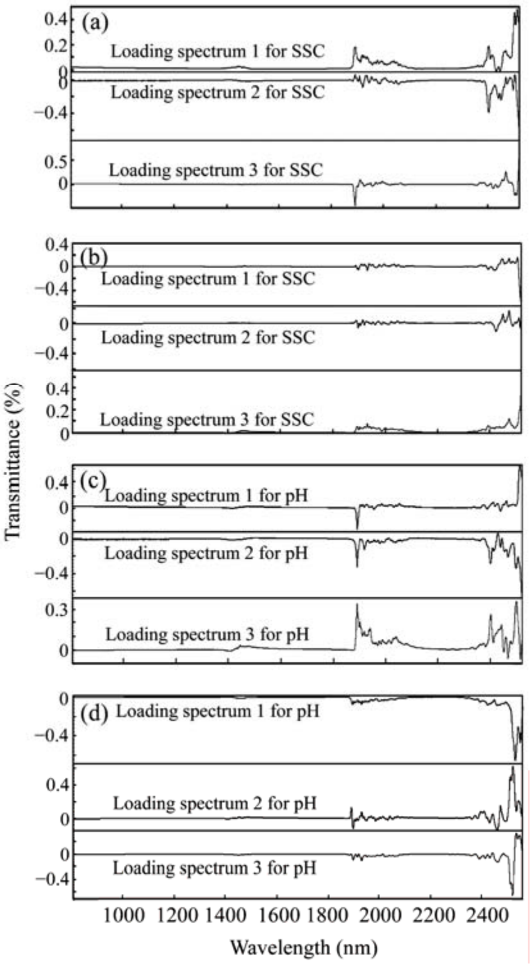

Fig.2 The first three loading spectra for SSC and pH before and after the storage

(a) The first three loading spectra for SSC before storage; (b) The first three loading spectra for SSC after storage; (c) The first three loading spectra for pH before storage; (d) The first three loading spectra for pH after storage

Table 1 Statistic values of SSC and pH in the calibration and validation sets

Parameter	Set	Dataset	Number of samples	Range	Mean	*SD*	*P*
SSC (°Brix)	Calibration	A	60	3.80~4.40	4.08	0.074	<0.01
		B	60	3.90~4.20	4.06	0.110	
	Validation	A	40	4.00~4.30	4.13	0.099	<0.01
		B	40	3.90~4.10	4.03	0.054	
	All	A	100	3.80~4.40	4.10	0.109	<0.01
		B	100	3.90~4.20	4.05	0.070	
pH	Calibration	A	60	4.19~4.48	4.25	0.048	<0.05
		B	60	4.12~4.37	4.23	0.048	
	Validation	A	40	4.19~4.29	4.21	0.021	<0.05
		B	40	4.19~4.24	4.21	0.013	
	All	A	100	4.19~4.48	4.24	0.045	<0.05
		B	100	4.12~4.37	4.22	0.043	

To estimate the changes in juice during the storage, the separation between the samples of the two datasets was quantified using the classification methods.

PCA

PCA was performed to reveal the presence of several groups as well as a number of outlier samples using raw spectra.

Fig.3 shows the two dimensional (2D) principal component score plot using the first two score vectors, PC1 and PC2, derived from raw spectra of the samples. The initial two factors, which account for the most spectral variation 99.882% (99.554% and 0.328% for the first two principal components, PC1 and PC2, respectively) related to chemical quality and indicated as positive or negative, are used to make differentiation clearer. There was significant separation between the samples in the two datasets. Fig.3 also shows that spectra in dataset A mainly have negative scores in the first component; however, those in dataset B mainly have positive scores in the first component.

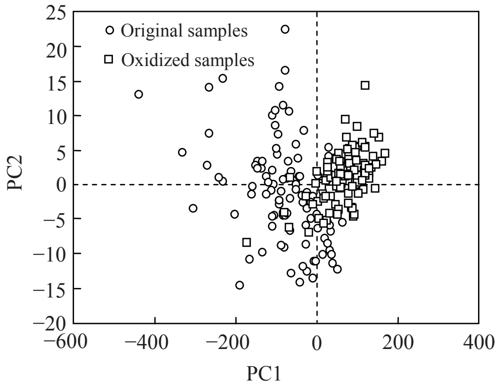

Fig.3 Score plot of PC1 and PC2 based on raw spectra

PCA displayed substantial discriminating information between spectra recorded before and after the storage. The result suggests that the discrimination between the two datasets is possible and that different spectral attributes of samples are associated with characteristics of the sample.

As PCA only indicates the visualizing dimension spaces, different classification methods were utilized for an improved separation.

LS-SVM classification

To obtain the optimal models, the input data and some parameters, γ and σ^2 in LS-SVM, have to be chosen carefully. In this study, the top 10 PCs were extracted by PCA and input to the classifiers as latent variables when training. The top 10 PCs almost contained 100% of total variance, so they could express the total spectral information. The optimal combination of γ and σ^2 within the region of 10^{-2}~10^4 was set based on experience. The contour plot of the optimization of the parameters γ and σ^2 for the classification of juice is shown in Fig.4. The grids' "●" in the first step was 10×10, and the searching step was a crude search with a large step size. The optimal search area was determined by the error contour line. The grids' "×" in the second step was also 10×10, and the searching step was the specified search with a small step size. The optimal combination of γ and σ^2 for juice discrimination was achieved with γ=4.2305 and σ^2=14.733.

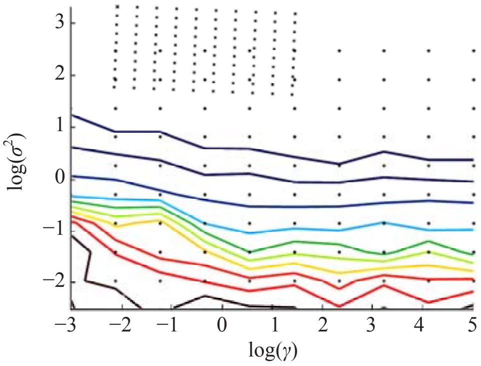

Fig.4 Contour plot of the optimization of the parameters γ and σ^2 for discrimination of juice

Different lines indicate lines of the same error. The grids' "●" in the first step (a crude search with a large step size) and "×" in the second step (specified search with a small step size) are both 10×10

Calibration and validation models for juice discrimination were developed by LS-SVM, DPLS, SIMCA, and DA (Table 2). LS-SVM method with reasonable parameter settings performed quite well and was significantly better than the other three methods, for the classification accuracy is the primary criterion for estimating the performance of the four classification methods. The accuracy of the calibration and validation sets for LS-SVM was 100%. The comparison of the average accuracy shows that SIMCA performed worst, with the lowest correct classification rate, involving 13 samples incorrectly classified in the calibration set and 7 ones in the

Table 2 Performance comparison results for discrimination models developed by LS-SVM, DPLS, SIMCA, and DA

Method	Total accuracy (%)	Number of sample misclassified				Accuracy (%)			
		Dataset A		Dataset B		Dataset A		Dataset B	
		Calibration set (*n*=60)	Validation set (*n*=40)	Calibration set (*n*=60)	Validation set (*n*=40)	Calibration set (*n*=60)	Validation set (*n*=40)	Calibration set (*n*=60)	Validation set (*n*=40)
LS-SVM	100	0	0	0	0	100	100	100	100
DPLS	97.00	1	0	2	3	98.33	100	96.67	92.50
SIMCA	90.00	7	3	6	4	88.33	92.50	90.00	90.00
DA	95.50	3	1	3	2	95.00	97.50	95.00	95.00

validation set. DA performed a bit better, with 95.5% of a total accuracy. It is worth mentioning that DPLS, SIMCA, and DA, although extensively used in chemometrics, did not provide results as well as LS-SVM, at least for this particular problem.

Considering above classification results, it is logical to suppose that juice may be differentiated by NIR spectroscopy on the basis of the change of the substances formed during the storage.

CONCLUSION

In this work, LS-SVM, DPLS, SIMCA, and DA coupled with NIR spectroscopy were used to control the quality change of tomato juice during the storage. The differences of spectra recorded in the juice before and after the storage period might be attributed to the changes of certain specific compounds of tomato juice. The results show that this investigation is feasible. An excellent precision and accuracy were achieved by LS-SVM model compared with DPLS, SIMCA, and DA models. More researches can be done to improve the proposed method and obtain simpler classification methods.

References

Acevedo, F.J., Jiménez, J., Maldonado, S., Domínguez, E., Narváez, A., 2007. Classification of wines produced in specific regions by UV-visible spectroscopy combined with support vector machines. *J. Agric. Food Chem.*, **55**(17):6842-6849. [doi:10.1021/jf070634q]

Borin, A., Ferrão, M.F., Mello, C., Maretto, D.A., Poppi, R.J., 2006. Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk. *Anal. Chim. Acta*, **579**(1): 25-32. [doi:10.1016/j.aca.2006.07.008]

Casale, M., Sáiz Abajo, M.J., González Sáiz, J.M., Pizarro, C., Forina, M., 2006. Study of the aging and oxidation processes of vinegar samples from different origins during storage by near-infrared spectroscopy. *Anal. Chim. Acta*, **557**(1-2):360-366. [doi:10.1016/j.aca.2005.10.063]

Chen, J., Arnold, M.A., Small, G.W., 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. [doi:10.1021/ac0498056]

Cozzolino, D., Kwiatkowski, M.J., Parker, M., Cynkar, W.U., Dambergs, R.G., Gishen, M., Herderich, M.J., 2004. Prediction of phenolic compounds in red wine fermentations by visible and near infrared spectroscopy. *Anal. Chim. Acta*, **513**(1):73-80. [doi:10.1016/j.aca.2003.08.066]

Gestal, M., Gómez-Carracedo, M.P., Andrade, J.M., Dorado, J., Fernández, E., Prada, D., Pazos, A., 2004. Classification of apple beverages using artificial neural networks with previous variable selection. *Anal. Chim. Acta*, **524**(1-2):225-234. [doi:10.1016/j.aca.2004.02.030]

Langeron, Y., Doussot, M., Hewson, D.J., Duchêne, J., 2007. Classifying NIR spectra of textile products with kernel methods. *Eng. Appl. Artif. Intell.*, **20**(3):415-427. [doi:10.1016/j.engappai.2006.07.001]

Li, Y., Shao, X., Cai, W., 2007. A consensus least squares support vector regression (LS-SVR) for analysis of near-infrared spectra of plant samples. *Talanta*, **72**(1): 217-222. [doi:10.1016/j.talanta.2006.10.022]

Liu, L., Cozzolino, D., Cynkar, W.U., Gishen, M., Colby, C.B., 2006. Geographic classification of Spanish and Australian Tempranillo red wines by visible and near-infrared spectroscopy combined with multivariate analysis. *J. Agric. Food Chem.*, **54**(18):6754-6759. [doi:10.1021/jf061528b]

Mouazen, A.M., Karoui, R., de Baerdemaeker, J., Ramon, H., 2006. Classification of Soils into Different Moisture Content Levels Based on VIS-NIR Spectra. 2006 ASABE Annual International Meeting Sponsored by ASABE, Oregon Convention Center, Portland, Oregon.

Pochet, N., de Smet, F., Suykens, J.A.K., de Moor, B., 2004. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics*, **20**(17):3185-3195. [doi:10.1093/bioinformatics/bth383]

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.

MIT Press, Cambridge, Massachusets, p.25-43.

Suykens, J.A.K., Vandewalle, J., 1999. Least squares support vector machine classifiers. *Neural Processing Letters*, **9**(3):293-300. [doi:10.1023/A:1018628609742]

Suykens, J.A.K., van Gestel, T., de Brabanter, J., de Moor, B., Vandewalle, J., 2002. Least Squares Support Vector Machines. World Scientific Pub. Co., Singapore.

Thissen, U., Ustun, B., Melssen, W.J., Buydens, L.M.C., 2004. Multivariate calibration with least-squares support vector machines. *Anal. Chem.*, **76**(11):3099-3105. [doi:10.1021/ac035522m]

Thompson, C.J., Danielson, J.D.S., Callis, J.B., 1997. Quantification of hydrofluoric acid species by chemical-modeling regression of near-infrared spectra. *Anal. Chem.*, **69**(1):25-35. [doi:10.1021/ac9604550]

van Gestel, T., Suykens, J.A.K., Baesens, B., Viaene, S., Vanthienen, J., Dedene, G., de Moor, B., Vandewalle, J., 2004. Benchmarking least squares support vector machine classifiers. *Machine Learning*, **54**(1):5-32. [doi:10.1023/B:MACH.0000008082.80494.e0]

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York, USA.

Zhao, J.W., Chen, Q.S., Huang, X.Y., Fang, C.H., 2006. Qualitative identification of tea categories by near infrared spectroscopy and support vector machine. *J. Pharm. Biomed. Anal.*, **41**(4):1198-1204. [doi:10.1016/j.jpba.2006.02.053]