Quantization of Adulteration Ratio of Raw Cow Milk by Least Squares Support Vector Machines (LS-SVM) and Visible/Near Infrared Spectroscopy

Ching-Lu Hsieh[1,*], Chao-Yung Hung[2], and Ching-Yun Kuo[3]

[1] Department of Biomechatronics Engineering, National Pingtung University of Science and Technology, Taiwan, R.O.C.
Tel.: 886-8-7703202#7583
chinglu@mail.npust.edu.tw
[2]Department of Biomechatronics Engineering,
National Pingtung University of Science and Technology, Taiwan, R.O.C.
[3] Animal Research Institute, Council of Agriculture, Executive Yuan, Taiwan, R.O.C.

Abstract. Raw cow milk has short supply market in summer and over supply in winter, which causes consumers and dairy industry concern about the quality of raw milk whether is adulated with reconstituted milk (powdered milk). This study prepared 307 raw cow milk samples with various adulteration ratios 0%, 2%, 5%, 10%, 20%, 30%, 50%, 75%, and 100% of powdered milk. Least square support vector machine (LS-SVM) was applied to calibrate the prediction model for adulteration ratio. Grid search approach was used to find the better value of network parameters of γ and σ^2. Results show that R^2 ranges from 0.9662 to 0.9777 for testing data set with plate surface and four concave regions. Scatter plot of testing data showed that adulteration ratio above 10% clearly differs from 0% samples.

Keywords: LS-SVM, Raw milk, Cow milk, Adulteration detection, NIR.

1 Introduction

For thousand years, milk has been an important nutrient source for people. There are more than 50,000 milking cows in Taiwan which produce over 300,000 tons of raw milk annually with a value of about US$ 230 million [1]. In terms of our planet, about 578 million tons of fresh and whole cow's milks are produced with a value of more than 148 billion US dollars per year [2]. Thus, dairy is an important and influential industry.

Weather in Taiwan is hot and humidity during summer, but winter is cooler. This causes dairy industry faces a problem of milk shortages in summer and oversupply in winter. Hu [3] reported that consumption index of cow's milk in summer (Jun. to Oct.) was 101.3-137.4, whereas in winter (Dec. to Mar.), it was 68.1-87.4. This situation makes consumers' worry that whether the raw milk or fresh milk is adulterated with reconstituted (powdered) milk [4]. Milk factories are also concerned

* Corresponding author.

L. Iliadis and C. Jayne (Eds.): EANN/AIAI 2011, Part I, IFIP AICT 363, pp. 130–139, 2011.

that farmers may dilute the raw milk with powdered milk. Therefore, milk factories and consumer agencies need to confirm the quality of raw milk.

Milk may be adulterated either intentionally or accidentally during production and processing. Harding [5] stated that there are many potential adulterants in liquid milk, such as water, neutralizers, salt, sugar, or solid contents. Borin et al. [6] reported adulteration of powdered milk in Brazil and mentioned that the most frequent contaminants were whey, starch, and sucrose that range from 20 to 25%, which does not cause detectable flavor changes. But occasionally contaminant ratio may be as high as 60%. In Taiwan, the media reported that an adulteration ratio of 30% reconstituted milk in fresh milk has been found. Therefore, many researches have been conducted to detect milk contaminants.

For example, Lin et al., [7] used a modified MAD (Metachromatic Agar-Diffusion) technique to detect DNase activity in synthetic raw milk so as to detect reconstituted milk in raw milk. Ding and Chang [4] reported that milk powder is one of the most common forms of adulteration in fresh milk sold in the summer in Taiwan. They used an amino acid analyzer to detect furosine so as to identify the adulteration of milk powder in raw and UHT (ultra-high temperature) treated milk. These studied methods are time consuming and require higher skills.

Water, proteins, lactose and other components (such as somatic or body cell, micro-organisms, antibiotics etc.) are the main composition of milk [8]. Usually, normal milk has an opalescent white to yellowish color because of light dispersion. Conventional testing of compositional quality for milk is a lengthy, labor-intensive, and expensive process, making such methods unsuitable for routine use in quality control [9]. Thus, an infrared (IR) absorption technique that provides faster and easier working conditions, has been adopted in milk analysis.

Association of Official Agricultural Chemists [10] stated that analysis of milk by IR is based on absorption of IR energy at specific wavelengths. For instance, CH groups in fatty acid chains of fat molecules absorb at 3.48 μm. The IR spectra usually show the fundamental vibration of molecules where the near infrared (NIR, 750-2500 nm) or visible wavelength (Vis, 400-750 nm) indicate the overtone or combination of molecule vibration and electron transition of constituents [11]. Thus, visible/NIR spectroscopy has been intensively used to evaluate the quality of milk or the adulteration of milk or dairy products. For example, Schmilovitch et al. [12] used NIR to analyze fat and total soluble solids (TSS) content of fresh milk. Chen et al. [13] applied MLR and multiplicative scatter correction method to determine the fat content in raw milk. Kawasaki et al. [14] developed a sensor system for milking robots. The system was equipped with NIR in order to measure fat, protein and lactose, somatic cell count, and milk urea nitrogen of unhomogenized milk.

Visible/NIR spectrum contain high dimensional data usually needs chemometrics to downsize its data dimensions. Spectral calibration tries to find the relationship between spectral absorption of specific wavelength and target composition, which can be applied in further prediction for unknown samples. Neural network has been proven its ability in function fitting that correlates the dependent variable y and independent variables x with y=f(x). Least squares-support vector machine (LS-SVM) that was originally proposed by Suykens and Vandewalle [15] are becoming a popular tool for data classification and function estimation. LS-SVM is modified from support vector machine (SVM) that is a supervised learning methods used for classification

and regression analysis. The LS-SVM is based on the margin-maximization theory performing structural risk minimization. However, it is easier to train than the SVM, as it requires only the solution to a convex linear problem, and not a quadratic problem as in the SVM [15]. Yu et al. [16] compared partial least squares regression (PLSR) with LS-SVM in alcohol content, titratable acidity, and pH prediction and found LS-SVM was slightly better. Siuly and Wen [17] used LS-SVM to cluster EEG signal. Zuao et al. [18] employed on-line LS-SVM for gas prediction. Borin et al. [6] used LS-SVM and NIR spectroscopy for quantification of adulterants (starch, whey, or sucrose) in powdered milk.

In summary, the aims of this study are as follows:

1. take Vis/NIR spectroscopy for raw cow milk samples that were adulterated in various percentages of powdered milk
2. apply LS-SVM approach to quantify the ratio of adulteration
3. use grid search method to test the parameters of LS-SVM so as to find a better model for detecting the percentage of powdered milk in row milk.

Therefore, this study is the first that applies LS-SVM to quantify the adulteration ratio of powdered milk in raw cow milk.

2 Materials and Methods

2.1 Sample Preparation

Raw cow milk samples were collected from a dairy farm near our campus, and they were delivered to our laboratory within 2 hr. in ice box. Sampling process was conducted from Jul. 2010 to Nov. 2010. Powdered milk was made by diluting powdered cow milk with water on the basis of protein composition of raw milk. Protein ratio was chosen as criteria because it is easy changed after processed. After the powdered milk was prepared, it was mixed with raw milk in volume ratio of 0:100 (0%), 2:98 (2%), 5:95 (5%), 10:90(10%), 20:80 (20%), 30:70 (30%), 50:50 (50%), 75:25 (75%) and 100:0 (100%).Number of samples in each ratio and batch were listed in Table 1.

Table 1. Number of prepared samples at each test batch

Batch	0%	2%	5%	10%	20%	30%	50%	75%	100%	Total
1	3	3	3	3	3	3	3	3	3	27
2	9	9	9	9	9	9	9	9	9	81
3	9	9	9	9	9	9	9	9	9	81
4	13	11	11	12	15	11	15	15	15	118
Total	34	32	32	33	36	32	36	36	36	307

Each sample was scanned to record its visible/NIR spectrum and was analyzed to obtain its compositions. Monochrometer (FOSS NIRS 6500, NIRSYSTEM, US) was used to obtain sample spectrum in a range of 400 nm to 2498 nm at 2-nm intervals. A quartz cuvette of light path 0.5 mm was used to record transmittance spectrum of sample at 25℃. Milk composition analyzer (Expert, Scopo Electric, Europe) was employed to observe fat, protein, SNF (solid-not-fat), lactose, and density. From 307

samples, 207 samples were randomly selected as training data set while the remaining samples (100) were used as testing data set.

To correct the baseline shift and smoothing noise on spectral data, each spectrum was proceed with baseline treatment, Savitzky–Golay smoothing with 4th order 11 data points, and standard normal variate scaling. These data processes were supported by a PLS-Toolbox (Eigenvector co.) compatible to MATLAB.

2.2 LS-SVM

To deal with linear or nonlinear multivariate calibration, the LS-SVM uses a estimation function to map input vectors (training vectors) x_i into a higher dimensional space by the function ϕ. $K(x_i,x_j)= \phi(x_i)^T\phi(x_j)$ is called the kernel function that generally have four basic forms: 1) linear: $K(x_i,x_j)= x_i^Tx_j$; 2) polynomial: $K(x_i,x_j)= (px_i^Tx_j+r)d$, $p>0$; 3) radial basis function (RBF): $K(x_i,x_j)=\exp(-q\| x_i-x_j\|^2)$,$q>0$; 4) sigmoid $K(x_i,x_j)= \tanh(mx_i^Tx_j+r)$, where d, m, p, q, and r are kernel parameters. Therefore, a LS-SVM function estimation is to minimize a cost function (C) that has a regression error, as follows [15]:

$$C = \frac{1}{2}w^T w + \frac{1}{2}\gamma\sum_{i=1}^{N} e_i^2 \tag{1}$$

such that

$$y_i = w^T\phi(x_i) + b + e_i \tag{2}$$

Analyzing Eq. (1) and Eq. (2), we may have a typical problem of convex optimization which can be solved by using the Lagrange multipliers method [16] as follows:

$$L = \frac{1}{2}\|w\|^2 + \gamma\sum_{i=1}^{N} e_i^2 - \sum_{i=1}^{N}\alpha_i\left\{w^T\phi(x_i) + b + e_i - y_i\right\} \tag{3}$$

where $y_i=[y_1, \ldots, y_N]^T$, $e_i=[e_1, \ldots, e_N]^T$, and $i=[\alpha_1,\ldots, \alpha_N]^T$; By conducting $\partial L(w, b, e,\alpha)/\partial w$, $\partial L(w, b, e,\alpha)/\partial b$, $\partial L(w, b, e,\alpha)/\partial e$, and $\partial L(w, b, e,\alpha)/\partial\alpha$, and setting the first derivative to zero, we obtain the optimal solution for training data as equations:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{N}\alpha_i\phi(x_i) = 0 \qquad \text{thus,} \qquad w = \sum_{i=1}^{N}\alpha_i\phi(x_i) \tag{4}$$

$$\frac{\partial L}{\partial e} = \sum_{i=1}^{N}\gamma e - \alpha = 0 \qquad \text{thus,} \qquad \alpha=\gamma e \tag{5}$$

Combining eq.(4) and eq.(5), we have

$$w = \sum_{i=1}^{N}\alpha_i\phi(x_i) = \sum_{i=1}^{N}\gamma e_i\phi(x_i) \tag{6}$$

γ (Gamma) is the regularization constant that balances the model's complexity and the training errors. For nonlinear regression, the kernel function meets Mercer's mapping, then it has formulation [15]:

$$y_i = \sum_{i=1}^{N} \alpha_i K(x_i, x) + b + e_i \tag{7}$$

For a point y_j to be evaluated it is:

$$y_j = \sum_{i=1}^{N} \alpha_i K(x_i, x_j) + b \tag{8}$$

Testing of the kernel function is cumbersome and it depends on each case. For simple and low dimensional problems, kernel function of linear can be adopted, while in high dimensional and nonlinear problems, RBF of Gaussian, $\exp(-\|x_i-x_j\|^2/(2\sigma^2))$, is commonly used. The σ^2 (sigma2) is the width of the Gaussian function. By careful selection on these parameters, a good generalization model could be achieved.

2.3 Performance Evaluation

There are three problems need to be solved when LS-SVM is used, which are kernel function, input feature subset, and kernel parameters. However, no systematic methodology is available for selection the kernel function. In here, a RBF kernel of Gaussian function was used because it was a nonlinear function and could reduce the computational complexity of the training procedure [19]. In order to find proper parameter, a grid search technique and leave one out 10 folds cross-validation were used. Grid search is a two-dimensional minimization procedure based on exhaustive search in a limited range. Leaves one out cross-validation fits a model on the training data points except one and the performance of the model is estimated based on the one point left out. This procedure is repeated for each data point. After the model was set, the testing data were evaluated to the model again. Parameter γ controls the trade-off between structural risk minimization principle and empirical risk minimization. Parameter σ^2 affects the value of function regression error. Small values of σ^2 yield a large number of regressors and eventually it can lead to over fitting. On the contrary, a large value of σ^2 can lead to a less number of regressors, but finally not so accurate. These LS-SVM calculations were conducted using MATLAB 7.0 (The Math Works, USA) and a free LS-SVM toolbox (LS-SVM v 1.5, Suykens, Leuven, Belgium) [20].

3 Results and Discussion

3.1 Composition of Samples

Some important statistics of sample compositions are shown in Table 2. These results indicate that adulteration do modify the composition of samples. Duncan test for these samples also suggests possible groups with 5% significant level. For instance, groups for fat are 3 and 0% to 20% and 50% and 100% are the group a. All samples of adulteration ratio of 0%, 2%, 5%, 10%, and 20% have no significant difference in each composition. This result may suggest too less adulteration will not have statistical difference in composition.

Table 2. Statistics for sample compositions

Adulteration ratio	Fat (%)	SNF (%)	Density (kg/m^3)	Protein (%)	Lactose (%)	Water (%)
0%	3.95 #^{ab}	8.45 ^{a}	1.0284 ^{a}	3.17 ^{a}	4.63 ^{ab}	87.60 ^{c}
2%	3.90 ^{abc}	8.47 ^{a}	1.0285 ^{a}	3.18 ^{a}	4.65 ^{a}	87.63 ^{c}
5%	3.85 ^{abc}	8.46 ^{ab}	1.0284 ^{a}	3.17 ^{a}	4.64 ^{ab}	87.69 ^{bc}
10%	3.82 ^{abc}	8.42 ^{abc}	1.0284 ^{a}	3.16 ^{ab}	4.62 ^{ab}	87.76 ^{bc}
20%	3.85 ^{abc}	8.38 ^{abc}	1.0282 ^{a}	3.14 ^{abc}	4.59 ^{abc}	87.77 ^{bc}
30%	3.81 ^{bc}	8.32 ^{abc}	1.0280 ^{ab}	3.13 ^{abc}	4.57 ^{abc}	87.88 ^{ab}
50%	3.90 ^{abc}	8.27 ^{bc}	1.0278 ^{ab}	3.11 ^{bc}	4.54 ^{bc}	87.84 ^{abc}
75%	3.77 ^{c}	8.23 ^{c}	1.0277 ^{b}	3.10 ^{c}	4.51 ^{c}	88.00 ^{a}
100%	3.96 ^{a}	8.38 ^{abc}	1.0277 ^{b}	3.13 ^{abc}	4.56 ^{abc}	87.66 ^{bc}
Mean±S.D.	3.89±0.25	8.37±0.37	1.0281±0.0014	3.14±0.11	4.59±0.20	87.75±0.42
Maximum	4.17	9.11	1.0301	3.36	5.54	88.81
Minimum	3.08	7.88	1.0226	2.99	4.33	87.04
Duncan Groups	3	3	2	3	3	3

Same letter means same group under Duncan's grouping test.

Boxplot of SNF composition is shown Fig.1. Other similar plots can be obtain to corresponding to other compositions: fat, density, protein, lactose, and water. They are omitted due to space limitation. Figure shows that mean value decreases as adulteration increases, but the 100% adulterated group has higher mean than that of 75% group. This indicates that powered milk contained similar composition of raw milk. But samples of various adulterations will change the composition.

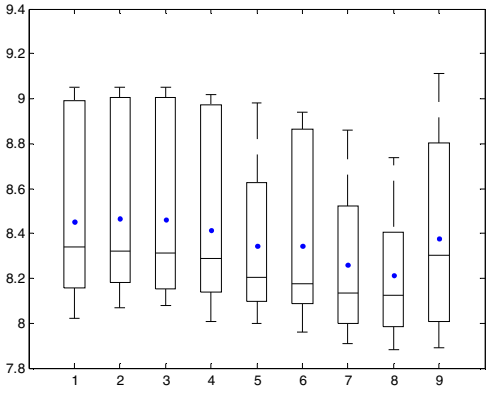

Fig. 1. Boxplot for SNF composition (%) of test samples (xabel 1 to 9 represent adulteration ratio 0% to 100%)

Mean spectra of partial region are shown in Fig. 2 to demonstrate the difference of each adulterated sample. Line on the figure is the average of samples for the adulteration ratio. They have trends of higher ratio samples have lower absorbance, except 2% samples that have the highest value. Reason for that needs more study.

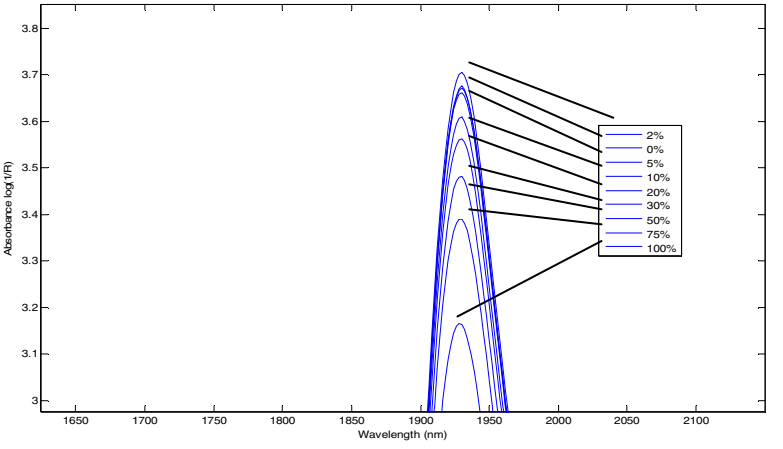

Fig. 2. Average spectra of adulterated samples (No. of samples = 307)

3.2 Grid Search for Training Data and Testing Data

Parameter of γ(gamma) and σ^2(sigma2) of LS-SVM were tested from 1E-15 to 1E+15 in 1E3 step. Coefficient of determination R^2 for training set at each test are shown in mesh plot (Fig.3), which shows a zigzag surface ranging from 0.9588 to 0.9873. From Fig. 3 we found that two regions of near γ(1E-5), σ^2 (1E0) and γ(1E0), σ^2 (1E0) have higher R^2.

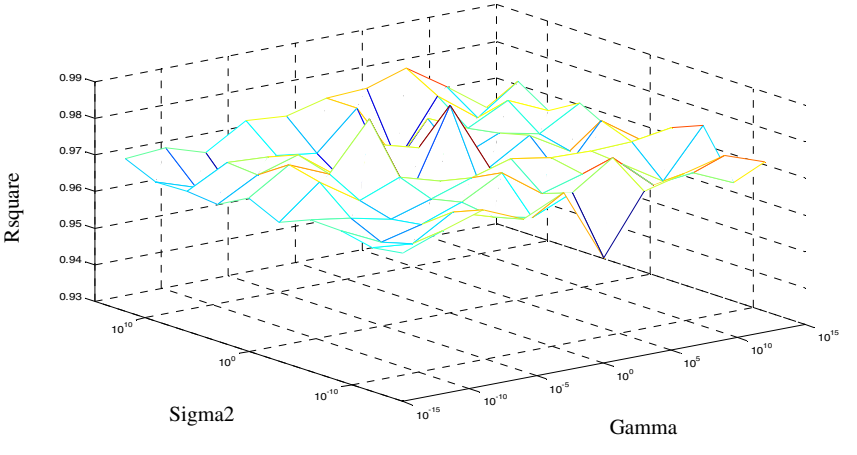

Fig. 3. Mesh plot of R^2 (Rsquare) for training data at combination of γ (Gamma) and σ^2 (Sigma2) for LS-SVM

Contour plot for testing set are shown in Fig.4. Basically it shows a plate surface that R^2 ranges from 0.9662 to 0.9777 with four regions of concave. These four concaves are easily identified in contour plot, which they are near γ(1E-10), σ^2(1E5); γ(1E-5), σ^2(1E0); γ(1E0), σ^2(1E-5); and γ (1E0), σ^2(1E0).

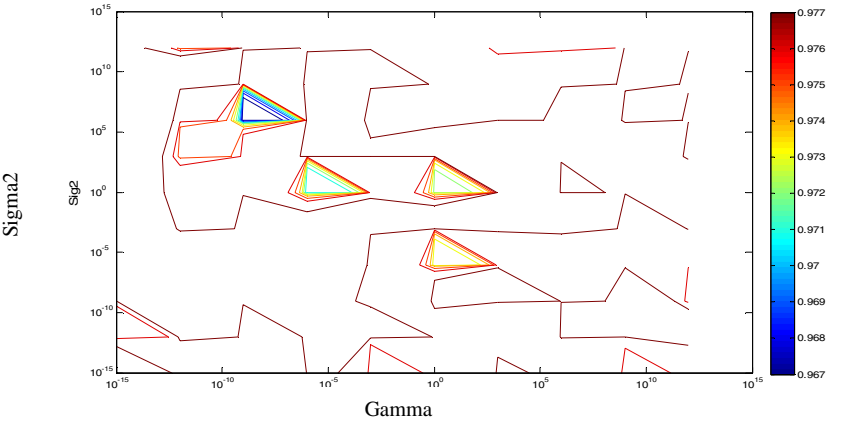

Fig. 4. Contour plot of R^2 for testing data at combination of γ (Gamma) and σ^2 (Sigma2) for LS-SVM

To evaluate the accuracy of prediction, parameters of LS_SVM were set to γ (1E-5), σ^2 (1E-5) that is not a concave region. The R^2 for testing set is 0.9775 and its scatter plot for 100 samples is in Fig. 5. From the plot, we learned that 100% have the most identical prediction, and 0%, 2%, 5%, and 10% are predicted widely. When the adulteration is higher than 10%, most samples have prediction value differs from 0%, which suggested that the sensibility for the model is about 10%. For benefit, 20% to 30% adulteration ratio is more attractive in market. Therefore, 10% sensibility has its application potential in market.

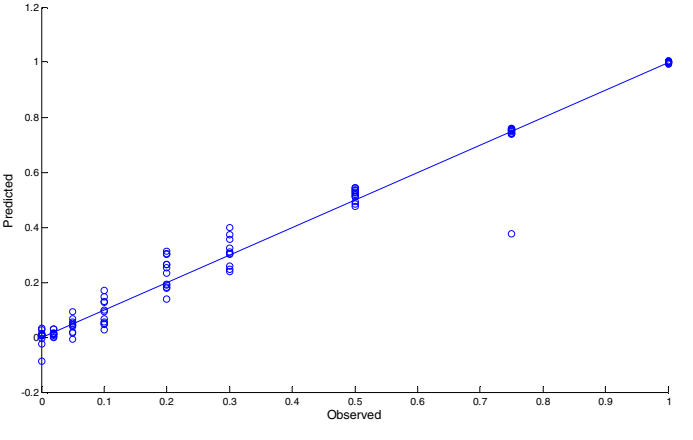

Fig. 5. Scatter plot of testing set for observed value and predicted valued. (Noted value has been normalized. No. of samples = 100).

4 Conclusion

This study applied LS-SVM and visible near infrared spectroscopy to quantize adulteration ratio of powdered milk in raw cow milk. Network parameter of γ and σ^2 were examined with grid search approach both from 1E-15 to 1E+15. Results show that R^2 for training data set has a zigzag surface with two obvious peaks of near γ(1E-5), σ^2(1E0) and γ(1E0), σ^2(1E0). Results also showed that based on 100 testing samples, most regions have high R^2 except four concaves regions. Parameters γ and σ^2 were chosen to 1E-5 that is not concave region and had R^2 0.9775. Its scatter plot indicated that most samples can be correctly separated from 0% when it is adulteration ratio is equal or higher than 10%.

Acknowledgments. Authors appreciate Council of Agriculture, the Executive Yuan, Taiwan, R.O.C. for financial support.

References

1. COA, Agricultural Statistics Yearbook. Council of Agriculture, Executive Yuan, Taipei, Taiwan, p. 126 (2009) (in Chinese)
2. FAOSTAT (2008), http://faostat.fao.org/default.aspx
3. Hu, Y.W.: Analysis of consumer's behavior for cow milk and goat milk in Taiwan. Grains and Livestock 241, 18–27 (1993) (in Chinese)
4. Ding, H.C., Chang, T.C.: Detection of reconstituted milk in fresh milk. Journal of the Chinese Agricultural Chemical Society 23(4), 406–411 (1986) (in Chinese)
5. Harding, F.: Adulteration of milk. In: Harding, F. (ed.) Milk Quality, ch. 5, pp. 60–74. Chapman & Hall, Wiltshire (1995a)
6. Borin, A., Ferrao, M., Mello, C., Maretto, D., Poppi, R.: Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk. Analytica Chimica Acta 579, 25–32 (2006)
7. Lin, C.W., Chen, S.H., Su, H.P., Ju, C.C.: Detection of reconstituted milk in raw milk through determination of milk Dnase activity. Journal of the Chinese Agricultural Chemical Society 25(3), 332–340 (1987) (in Chinese)
8. Berg, J.: Diary Technology in the Tropics and Subtropics, Pudoc Wageningen, the Netherlands, pp. 223–242, 1–45 (1988)
9. Harding, F.: Composition quality. In: Harding, F. (ed.) Milk Quality, ch. 6, pp. 75–96. Chapman & Hall, Wiltshire (1995b)
10. AOAC official methods of analysis. 972.16 Fat, lactose, protein, and solids in milk. Mid-infrared spectroscopic method. Association of Official Analytical Chemists, Arlington, VA, pp. 816–818 (1990)
11. Norris, K.H.: Making Light Work: Advances in Near Infrared Spectroscopy. In: Murray, I., Cowe, I.A. (eds.), pp. 596–602. VCH Press, New York (1991)
12. Schmilovitch, Z., Maltz, E., Austerweill, M.: Fresh raw milk composition analysis by NIR spectroscopy. In: Ipema, A.H., Lippus, A.C., Metz, J.H.M., Rossing, W. (eds.) Proceedings of the International Symposium on Prospects for Automatic Milking, Wageningen, Netherlands, EAAP Publication No.65 (1992)

13. Chen, J.Y., Iyo, C., Terada, F., Kawano, S.: Effect of multiplicative scatter correction on wavelength selection for near infrared calibration to determine fat content in raw milk. Journal of Near Infrared Spectroscopy 10(4), 301–307 (2002)
14. Kawasaki, M., Kawamura, S., Tsukahara, M., Morita, S., Komiya, M., Natsuga, M.: Near-infrared spectroscopic sensing system for on-line milk quality assessment in a milking robot. Computers and Electronics in Agriculture 63, 22–27 (2008)
15. Suykens, J.A.K., Vandewalle, J.: Least squares support vector machine classifiers. Neural Processing Letters 9, 293–300 (1999)
16. Yu, H.Y., Niu, X.Y., Lin, H.J., Ying, Y.B., Li, B.B., Pan, X.X.: A feasibility study on on-line determination of rice wine composition by Vis–NIR spectroscopy and least-squares support vector machines. Food Chemistry 113, 291–296 (2009)
17. Siuly, Y., Wen, P.: Clustering technique-based least square support vector machine for EEG signal classification. Computer Methods and Programs in Biomedicine (2010), doi:10.1016/j.cmpb.2010.11.014
18. Zuao, X., Wang, G., Zhao, K., Tan, D.: On-line least squares support vector machine algorithm in gas prediction. Mining Science and Technology 19, 194–198 (2009)
19. Wu, D., He, Y., Feng, S., Sun, D.W.: Study on infrared spectroscopy technique for fast measurement of protein content in milk powder based on LS-SVM. Journal of Food Engineering 84, 124–131 (2008)
20. LS-SVM v 1.5, `http://www.esat.kuleuven.be/sista/lssvmlab/home.html`