Partial Least Square-Support Vector Machine for Rapid Detection of Egg Storage Life by Chemometric Processing of Voltammetric Signals

Jingjing Liu, Xiaoting Zhang, Jingyi Li, Ruixia Wen, and Hong Men[*]

School of Automation Engineering, Northeast Dianli University, Changchun Road, Jilin City, Jilin Province, China

The rapid detection of the grade and storage life of eggs is important for consumers and producers. In this study, an electrochemical system based on voltammetry was used for the rapid detection of egg yolk and egg white mixtures for classification and prediction of storage days. Measuring the Haugh units (H.U.) index of Lindian eggs is aimed at identifying the grades of these eggs and the corresponding storage days. Discriminant function analysis was used to identify accurately the grades of eggs on the basic of storage life. The Gaussian kernel function and a suitable penalty factor were selected to establish a support vector machine for predicting the storage days of eggs. By avoiding deviation in the result due to the correlation of data, which was acquired by square-wave voltammetry technology, the partial least squares and support vector machine were combined, and the predictive accuracy of egg storage days was improved by up to 97.14%. A tenfold cross-validation has been used to evaluate the classification performance of the model; the average accuracy of prediction of storage life was 88.57%.

1. Introduction

The high quality of energy provided by nutritious eggs is known all over the world. The quality of eggs being sold relates to the interests of consumers. According to the standard stipulated by the U.S. Department of Agriculture, eggs are divided into four grades: AA, A, B, and C. Eggs being sold must be of B grade or above. Because the surface biofilm of eggs is removed in a week or so, air and microorganisms can enter the egg and cause protein membranes to be oxidized, water to evaporate, and the egg yolk to degenerate so as to finally reduce the consumption and commodity values of the eggs. An accurate and quick method of detecting the storage life and grade of eggs is needed for the convenience of producers.[1–5]

With the rapid development of detection technology, more and more detection methods have been applied to the quality detection of eggs, such as high-performance liquid chromatography, atomic absorption spectroscopy, electronic nose technology, and electronic tongue technology. The evaluation of eggs can be roughly divided into shelf-life detection and internal composition detection.[6–9]

---

[*]Corresponding author: e-mail: menhong_china@hotmail.com

In this study, we focused on detecting the mixture of egg yolk and egg white using an electrochemical system based on a three-electrode structure; the traditional egg Haugh units (H.U.) of measurement is used as a reference. First, using discriminant function analysis (DFA) based on the analysis of variance to identify the grade of eggs, we predict egg storage days to enable people to understand the freshness of eggs better. Thus, the prediction of egg storage days using a support vector machine (SVM) was achieved using data acquired with the electrochemical system. At the same time, in order to avoid the deviation in results based on the correlation of data that was acquired by square-wave voltammetry technology, the partial least squares (PLS) and SVM were combined to improve the forecast accuracy of egg storage days.

2.   Materials and Methods

2.1   Experimental system

In this study, an electrochemical system that consists of a three-electrode structure, an electrochemical workstation, and a computer has been built to detect the grade and storage life of eggs.[10–12] The three-electrode structure was composed of a working electrode, a reference electrode, and an auxiliary electrode. The details of its structure are shown in Fig. 1.

The working electrode, which is also called a sensing electrode, can produce an obvious variation of component concentration, which needs to be tested in the measurement process. In this study, a gold disk electrode was used as the working electrode. The reference electrode consists of a constant-component phase. The electrode potential of the reference electrode did not change during the measurement. The change in the electromotive force of a cell reflects the variation of the electrode potential of the working electrode. In this study, we chose a saturated calomel electrode as the reference electrode. The auxiliary electrode constitutes a battery with the other electrode. We chose a platinum wire electrode as the auxiliary electrode. The reference electrode and the working electrode pole constitute the cell, which forms a current loop. There was a constant voltage system between the working electrode and the reference electrode to maintain the stability of the voltage in the liquid system. Under the control of the voltage, an excitation current was produced in the solution. The loop current, which was produced between the working electrode and the auxiliary electrode, was the output that would be tested as the signal.

The electrode array connected to the electrochemical workstation was inserted into the sample solution when we performed the experiment. On one hand, the electrochemical workstation provided a series of electrical pulses; on the other hand, it collected the response and stored it in the
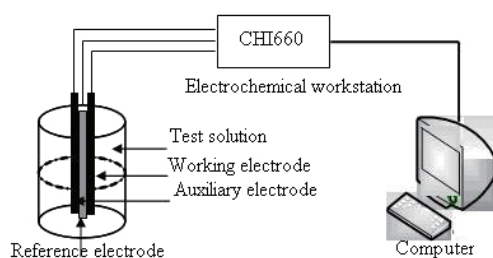

Fig. 1.   Electrochemical system.

computer. Then, feature extraction was carried out on the stored signal. Finally, pattern recognition algorithms were used to identify the signals in order to distinguish and identify the different substances to obtain taste information on the different substances.

2.2   Experimental methods

2.2.1   H.U. detection

The traditional methods of measuring the freshness of eggs, called H.U. detection, were used in this study to determine the storage life and grade of eggs. Measuring the H.U. index of the same batches of Lindian eggs was aimed at identifying the grades of these eggs and the corresponding storage days.

H.U. is a character index based on regression relationships between egg quality and egg white height. H.U. values are calculated using Eq. (1). H.U. values change with storage time. When the H.U. value is greater than 72, the grade of eggs is AA. The A grade denotes that the H.U. value is between 72 and 60. The B grade denotes that the H.U. value is between 60 and 30. The C grade denotes that the H.U. value is less than 30. The H.U. value of fresh eggs is usually between 75 and 82, with some being as high as 90.

$$HU = 100 \cdot \log(H + 7.57 - 1.7w^{0.37}) \tag{1}$$

In the equation, *H* denotes the height of the egg white (mm), and *w* denotes the weight of the egg (g).

The height of the egg white can be measured using an albumen altimeter. First, the level of a glass plate is corrected. After the egg is opened on the glass plate, three equidistant points are taken, which are evenly distributed in the middle of the edge between the egg yolk and the protein (avoiding the protein lacing); the height of the egg white is measured using the albumen altimeter, and the average is calculated.

The H.U. tester was used to measure 70 Lindian eggs that were in the same batch with the electrochemical system experiment based on voltammetry. As soon as the eggs were obtained, 10 eggs were randomly selected for testing. The H.U. value could be determined according to the egg weight and egg albumen height measured in the experiment. The remaining 60 eggs were divided into 20 groups and detected every two days with a handy instrument to determine the H.U. value.

2.2.2   Experiment with eggs based on the electrochemical system

The purpose of detecting the mixed liquor of egg yolks and egg whites with the electrochemical system based on the three-electrode structure was to obtain a more accurate storage life and grade of raw eggs. Seventy Lindian eggs from the same batch were selected to be used in the experiment. The eggs were about 5 cm long and 60 to 70 g in weight. There are five steps in each experiment. First, the beakers used in the examination were cleaned with deionized water. Second, three electrodes were soaked in alcohol for 6 min. Then, they were cleaned with deionized water and dried with filter paper. They were fixed in the electrode rack and connected to the electrochemical workstation later. Third, one egg was chosen for the measurement and the egg's yolk and white were mixed uniformly to obtain the mixed liquor. The mixed liquor was divided into 5 portions after being diluted fivefold with deionized water. Then, we put the mixtures in numbered beakers.

The experiment on the mixed liquor should begin in accordance with the purchase date. Fourth, the electrochemical workstation was turned on and the corresponding technology parameter for the experiment was set. Then, the electrode array was put into the mixed liquor to begin the measurement. Fifth, the measurement of the 5 portions of mixed liquor from the same egg was carried out in order. When the experiment on one egg was finished, the electrode was cleaned with deionized water and dried with filter paper before the measurement of the next egg. Ten eggs were chosen daily for the measurement. The experiment was carried out every 6 d by the same method until a total of 50×7 sets of data were obtained in 36 d.

The experiment was based on square-wave voltammetry technology. The corresponding parameters were as follows: 1.1 V of initial voltage, −0.6 V of terminal voltage, potential increments of 0.005 V, amplitude of 0.04 V, frequency of 15 Hz, and 15 s of standing time.

2.3 Identification methods

2.3.1 DFA

DFA is a multivariate statistical method based on observational variables to determine the classification of samples.[13–16] It is used to classify type-unknown data by modeling type-known data. Common types of discriminant function analysis include the Bayes discriminant method, the distance discriminant method, and the Fisher discriminant method. In this study, the Fisher discriminant method based on analysis of variance is used to analyze the data. The one-dimensional linear function can replace the n-dimensional space point in order to transform the n-dimensional space observation points into m-dimensional points, and then classify the observation set in the m-dimensional space. The differences in the data, which have reduced dimension, should be extended in order to obtain a high identifiable efficiency. The optimal linear function is selected on the basis of the principle of the maximum ratio of the mean square error in the group and the mean square error between groups.

2.3.2 SVM

The basic idea of SVM is to build a separating hyperplane as the decision surface to obtain the maximum isolated edge between the counterexample and the positive example.[17–19] SVM could be the vector that is closest to the hyperplane .

An sample is the n-dimensional vectors; the L sample sets can be expressed as

$$(x_1, y_1), ...,(x_l, y_l) \in R^n \times \{+1,-1\}. \tag{2}$$

The constructed hyperplane can be expressed as

$$\omega \bullet \Psi(x) + b = 0, \tag{3}$$

where ω is the weighting vector of the hyperplane, $\Psi(x)$ is the nonlinear mapping function, and b is the value for the category field.

$$y_i(\omega \bullet \Psi(x) + b) \geq 1 \tag{4}$$

By normalization and introducing the relaxation $\xi_i \geq 0$,

$$y_i(\omega \bullet \Psi(x) + b) > 1 - \xi_i, \tag{5}$$

in which $i = 1, 2, \ldots, L, \ldots, n$.

The objective function is

$$\Psi(x) = \frac{1}{2}\omega \bullet \omega + C \sum_{i=1}^{n} \xi_i. \tag{6}$$

That means that the misclassified sample is the minimum while assuring the maximum class intervals so that the optimal classified surface can be obtained. $C > 0$, known as the penalty factor, is used to control the degree of penalty for the misclassified sample. The optimization problems are translated into Langrange dual problems for solving.

$$\max\left[\sum_{l}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \left(\Psi(x_i) \bullet \Psi(x_j)\right)\right] \tag{7}$$

Constraints: $0 \leq \alpha_i, \sum_{i=1}^{n} y_i \alpha_i = 0, i = 1, 2, \ldots, L, n$

There are only a small number of coefficients that are nonzero in the solutions of optimization problems. The vector corresponding to the coefficients is the support vector to determine the optimal generalized classified surface.

Discriminated functions can be expressed as

$$f(x) = \text{sgn}\left[\sum y_i \alpha_i (\Psi(x_i) \bullet \Psi(x)) + b\right], \tag{8}$$

and $\text{sgn}(x) = \begin{cases} +1 & x > 0, \\ -1 & x < 0. \end{cases}$

For nonlinear classification, the algorithm of point multiplication can be difficult and time-consuming. The Kernel functions $K(x_i, x_j) = \Psi(x_i) \bullet \Psi(x_j)$ that meet the Mercer Theorem in the original input space could be used instead of dot product operations in the high-dimensional feature space. The method is called experienced kernel mapping or kernel trick. There are four types of common kernel function: linear, polynomial, Gaussian, and sigmoidal.

2.3.3 PLS

PLS is an algorithm that is used to fit the relationships of variables to two or more independent variables.[20–22] Prediction data of PLS were divided into several regions described by the feature vector. In the same regions, vectors are orthogonal to each other; in the different regions, vectors are associated with the model based on the previous channel.

For vectors of n-dimensional samples, the PLS individually extracts 't' and 'u' ingredients from the independent and dependent variable data. The values 't' and 'u' must meet the conditions that carry as far as possible the variations of information of each data table and reach the maximum level of correlation between them. The regression of independent variables to 't' and dependent

variables to 'u' is achieved. If there is no satisfactory regression, then the next round of extraction will be carried out until a satisfactory precision is obtained.

2.3.4   PLS-SVM

The PLS-SVM is the combination of the PLS and the SVM.[23–25]   First, PLS is used for data analysis and processing; then, the training samples are disposed of with the SVM to obtain the optimal hyperplane and classify them.

The m training samples with k categories (x_1, y_1), (x_2, y_2), …, (x_m, y_m) are set, in which $x_i = (x_{i1}, x_{i2}, …, x_{in})$, $y_i = \{1, 2, …, k\}$, $i = 1, 2, …, m$, and the data is normalized.

(1) Integrated components with PLS $t_{i1}, t_{i2}, …, t_{is}$, $s < n$ are extracted to eliminate the correlation between classes. The integrated components $t_{i1}, t_{i2}, …, t_{is}$ have the maximum correlation between classes. Simple samples are marked as (t_1, y_1), (t_2, y_2), …, (t_m, y_m), in which $t_i = (t_{i1}, t_{i2}, …, t_{is})$, $i = 1, 2, …, m$.

(2) Training samples are simplified with SVM to obtain the support vector and optimal hyperplane, and then the forecast is completed.

3.   Results and Discussion

3.1   H.U.

During the experiments on the quality of eggs, the first experiment determined whether the purchased eggs were fresh, namely, whether the grade of eggs was AA. The next experiment was for determining the eggs' grade for different storage lives and the eggs' corresponding storage days. The results showed that the grade of the same batch of Lindian eggs was AA. The H.U. value was 92.55 ($\pm$ 8.5). Table 1 shows the result for different storage days by the H.U. method.

As can be seen from Table 1, the H.U. values of eggs were all above 85. Under the conditions of room temperature and 70 to 80% relative humidity, the qualities of eggs would be reduced to A grade when stored for 6 d. The qualities of eggs would become B grade when stored more than 30 d. Eggs stored for more than 36 d would basically have gone bad and become C grade. These eggs were not available for sale any more.

3.2   Feature data extraction

After the experiments on mixed liquors of egg yolks and egg whites with the electrochemical system based on the three-electrode structure, the corresponding current–voltage curve was obtained. To compare these curves easily, current–voltage curves under different storage lives were superimposed. Figure 2 shows the current–voltage curve of the mixed liquors of egg yolks and egg whites tested under the conditions of 0, 6, 12, 18, 24, 30, and 36 d.

Because a current–voltage curve of the mixed liquors of egg yolks and egg whites tested using the electrochemical system contained a huge amount of data, it was important to extract the characteristic value. The current values at five special points were selected on the voltammetry curves (the first peak current, the second minimum current, the second peak current, the third peak current, the first minimum current) as characteristic values for each sample.

Table 1
H.U. values of eggs under different storage days.

Storage days(d)	Minimum value	Maximum value	Mean value
1	85.69	95.78	92.55
3	84.34	89.86	85.98
5	70.98	77.22	73.67
7	66.12	73.81	71.56
9	65.79	71.18	70.29
11	65.67	70.45	69.76
13	64.62	69.82	68.46
15	63.01	68.50	67.98
17	62.45	68.31	67.45
19	61.31	65.85	64.87
21	60.94	65.23	64.45
23	60.32	64.01	63.23
25	60.10	63.79	62.71
27	58.89	62.35	61.50
29	57.51	61.98	60.05
31	53.19	60.05	57.85
33	40.98	45.47	43.34
35	29.75	32.95	31.39
37	28.97	30.06	29.78

Fig. 2.   (Color online) Current–voltage curves of the egg mixed liquors for different storage times.

The relationships between the H.U. values of eggs for different storage times and the characteristic value could be established. As shown in Fig. 3, the responses to changes in the first peak current, the second minimum current, the second peak current, and the third peak current varied widely, except for the response to change in the first minimum current. Thus, for the results of the egg mixed liquors, it was feasible to select the first peak current, the second minimum current, the second peak current, and the third peak current as characteristic values.
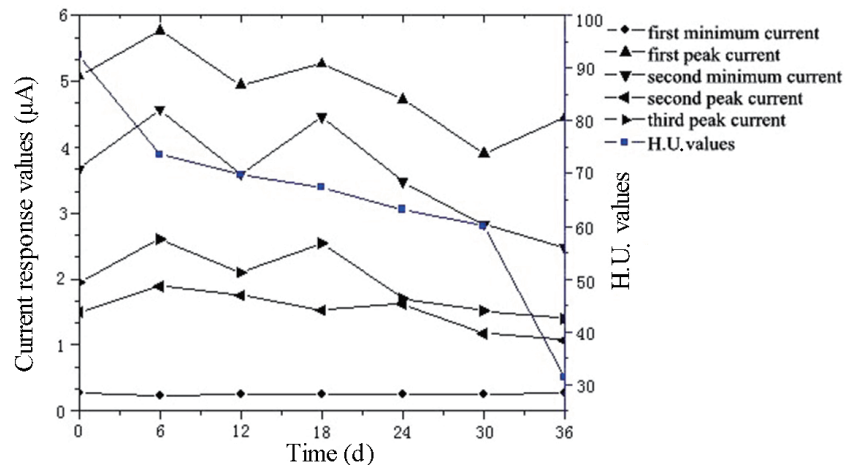
Fig. 3.   Corresponding relationships between H.U. values for different storage times and current response values of the egg mixed liquors.

3.3   Classification of egg grade based on DFA

Figure 4 shows the results of the mixed liquors of egg yolks and egg whites analyzed by discriminant function. DFA was used to determine the different grades of the mixed liquors. The sum of the contribution rates of DF1 and DF2 was 96.7%. Different grades of eggs can be distinguished on the basis of the contribution rate, and the individual sample in the same grade of egg was relatively concentrated.

3.4   Prediction of storage life of eggs

In this study, there are some corresponding relationships between the grade and the storage life of eggs. In fact, the prediction of the storage life of eggs is more important than the detection of egg grade because we can know the freshness degree of eggs better. The prediction of the storage life of eggs was the main focus in subsequent research.

3.4.1   SVM

The Gaussian radial basis function (RBF) kernel function was used in the SVM model to predict the storage days of eggs. Training data of 80% data points and 20% for testing purpose were randomly chosen. The training data were used to find the model parameters, and the testing data were used to evaluate the classification performance of the learned model. When the training model was established, we began the prediction for the classification of the testing set. The results are as follows: the accuracy of the prediction of storage life is 94.29%, the penalty factor C is 100, the mean square error (MSE) is 0.1189, and the training time is 1.89 s. Figure 5 shows the result of the classification of the testing set.
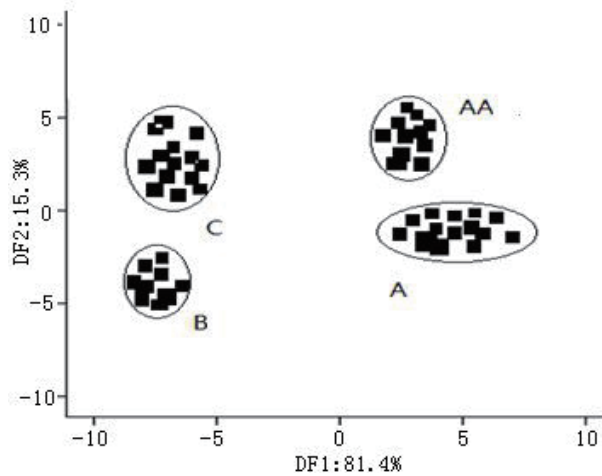
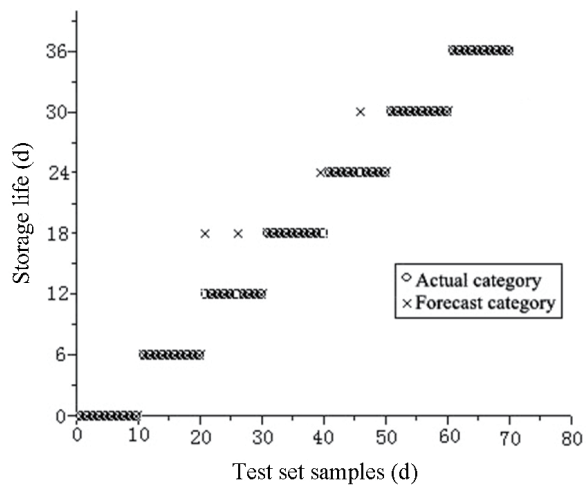Fig. 4. Results of egg grade by discriminant function.

Fig. 5. Analysis results of egg storage days based on SVM.

3.4.2 PLS

We constructed a PLS model for the storage life prediction of eggs. PLS creates a linear relationship from principal components of the input data. The correlation coefficient obtained from the PLS model was 0.9382 ($r^2 = 0.9382$), which indicated that the model was valid. In the PLS model, the MSE of the model was 0.5146, and the training time was 1.5 s. Figure 6 shows the relevant curve of the actual and predicted storage lives of eggs based on the PLS model.

3.4.3 PLS-SVM

The principal component obtained from the analysis of PLS was selected as a training data set, and then the SVM was used to establish the regression model; finally, predictive analysis was achieved. The kernel function in SVM was the Gaussian RBF. The accurate recognition rate of prediction for classification was 97.14%. The penalty factor was 100 ($C = 100$). The mean square error was 0.0567 and the training time was 1.3 s. Figure 7 shows the prediction for storage days based on PLS-SVM.

From Fig. 7, PLS-SVM shows good performance for the prediction of egg storage days. To avoid the contingency caused by the random choice of training set and testing set, the tenfold cross-validations were used to prove the performance of the PLS-SVM model. Table 2 shows the predicted results of egg storage days based on the tenfold cross-validations. Finally, the results are as follows: the accuracy of prediction of storage life is between 74.29 and 100%, the average accuracy of prediction of storage life is 88.57%, and the MSE is 0.2869.
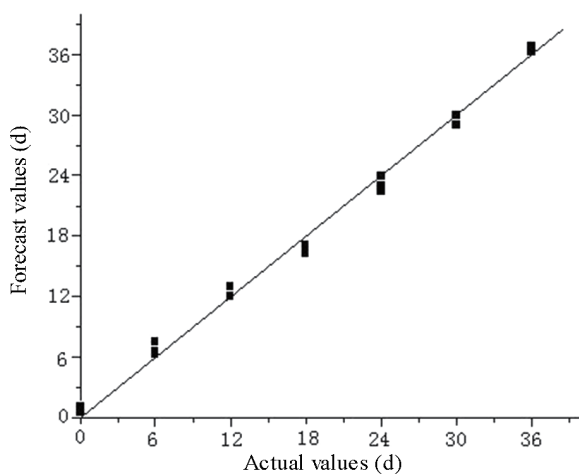
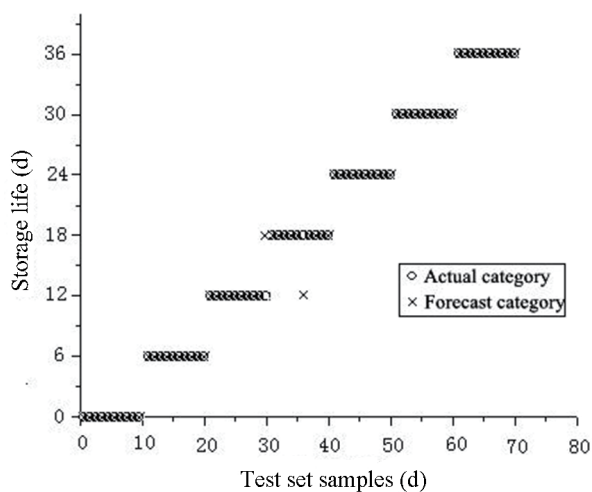Fig. 6.   PLS actual and predicted values of the model curves for egg mixed liquors.

Fig. 7.   Analysis results of egg storage days based on PLS-SVM.

Table 2
Predicted results of egg storage days based on tenfold cross-validation of PLS-SVM.

Data set	Accuracy of training set (%)	Accuracy of testing set (%)	MSE
1	99.37	77.14	0.4781
2	99.68	94.29	0.2390
3	95.56	100	0
4	98.73	91.43	0.2928
5	97.78	88.57	0.3381
6	99.37	74.29	0.5071
7	99.68	100	0
8	99.68	97.14	0.1690
9	100	88.57	0.3381
10	98.10	74.29	0.5071

4.   Conclusions

According to the physicochemical properties of a mixture of egg yolk and egg white, an electrochemical system based on a three-electrode structure and square-wave voltammetry technology were used in this study to identify the grade of eggs and the corresponding storage days.

The traditional methods of measuring the freshness of eggs, named H.U. detection as a reference, were used in this study to characterize the egg storage quality. In this study, the detection of the mixture of egg yolk and egg white by square-wave voltammetry technology and DFA was used to identify the different grades of eggs successfully.

SVM was established on the basis of the Gaussian RBF kernel function and the penalty factor C ($C = 100$). The accuracy of the prediction of storage life is 94.29%. The partial least squares and support vector machine were combined to improve the forecast accuracy of egg storage days by up to 97.14%. A tenfold cross-validation was used to evaluate the classification performance of the PLS-SVM model and the average accuracy of the prediction of storage life is 88.57%.

Acknowledgements

References

1   M. Omid, M. Soltani, M. H. Dehrouyeh, S. S. Mohtasebi, and H. Ahmadi: J. Food Eng. **118** (2013) 70.
2   W. O. Song and J. M. Kerver: J. Am. Coll. Nutr. **19** (2000) 556.
3   B. Ramos, O. Pinho, and I. M. P. L. V. O. Ferreira: Food Chem. **116** (2009) 340.
4   M. Liu, L. Yao, T. Wang, J. Li, and C. Yu: J. Food Eng. **124** (2014) 117.
5   A. Hidalgo, M. Rossi, and C. Pompei: Food Chem. **94** (2006) 608.
6   Y. Santaladchaiyakit, S. Srijaranai, and R. Burakham: Food Anal. Method. **7** (2014) 1973.
7   T. Zuliani, B. Kralj, V. Stibilj, and R. Milacic: Ital. J. Food Sci. **17** (2005) 155.
8   Q. L. Wang, G. F. Jin, Y. G. Jin, M. H. Ma, N. Wang, C. Y. Liu, and L. Y. He: Eur. J. Lipid Sci. Tech. **116** (2014) 1044.
9   J. J. Liu, H. H. Gao, J. Y. Li, R. X. Wen, and H. Men: Transducer Microsyst. Technol. **33** (2014) 56.
10 V. B. Santos, E. L. Fava, D. M. C. Ns, R. C. Faria, and O. Fatibello-Filho: Talanta **126** (2014) 82.
11 X. X. Zhang, F. Ran, H. L. Fan, L. B. Kong, and L. Kang: Acta Phys. Chim. Sin. **30** (2014) 881.
12 B. Khemchandani, A. K. Jaiswal, E. Sayanna, B. Hinton, and M. Forsyth: Corros. Eng. Sci. Techn. **50** (2015).
13 E. Hernandez-Martinez, J. X. Velasco-Hernandez, and T. Perez-Muñoz: Physica A **392** (2013) 6015.
14 P. García, D. López, and M. V. D. Parga: Theor. Comput. Sci. **448** (2012) 41.
15 Y. C. Zhou, Y. Tang, and X. J. Wang: Proc. Eng. **29** (2012) 3755.
16 M. Banaee and S. H. Tavassoli: Polym. Test. **31** (2012) 759.
17 S. Kang and S. Cho: Expert Syst. Appl. **41** (2014) 4989.
18 C. X. Xue, R. S. Zhang, H. X. Liu, M. C. Liu, Z. D. Hu, and B. T. Fan: J. Chem. Inf. Comput. Sci. **44** (2004) 1267.
19 C. Xie, C. Shao, and D. Zhao: J. Comput. **6** (2011) 1935.
20 Q. Z. Xue and Z. L. Guo: Pattern Recogn. **47** (2014) 3726.
21 B. N. Zhong, X. T. Yuan, R. R. Ji, Y. Yan, Z. Cui, X. P. Hong, Y. Chen, T. Wang, D. S. Chen, and J. X. Yu: Neurocomputing **133** (2014) 317.
22 P. Shan, S. L. Peng, Y. M. Bi, L. Tang, C. X. Yang, Q. Xie, and C. W. Li: Chemom. Intell. Lab. Syst. **138** (2014) 72.
23 P. Samui: Engineering **3** (2011) 431.
24 Y. Wang, M. Yang, G. Wei, R. Hu, Z. Luo, and G. Li: Sens. Actuators, B **193** (2014) 723.
25 F. Song, G. Cheng, F. R. Sun, and G. Q. Zhu: Adv. Mater. Res. **859** (2014) 19.