Research Inventy: International Journal of Engineering And Science
Vol.6, Issue 4 (April 2016), PP -01-05
Issn (e): 2278-4721, Issn (p):2319-6483, www.researchinventy.com

Study of Parametric Kernels for LSSVM Models in NIR Determination of Fishmeal Protein

Jiangbei Wen[1,2], Yuanyuan Liang[1,2], Huazhou Chen[1,*]
[1.] *College of Science, Guilin University of Technology, Guilin, 541004, China*
[2.] *Guangdong Spectrastar Instruments Co. Ltd., Guangzhou, 510663, China*

ABSTRACT - *High quality fishmeal provides suitable protein ratio in the modernizing balance of animal breeding. Near-infrared (NIR) spectroscopy technology was used for the determination of protein concentration in fishmeal samples. In this paper, Least squares support vector machine (LSSVM) method was applied to the NIR analytical procedure to search the optimal calibration model. We focus on the study of three common LSSVM kernels and the tuning of parameter valuing. The parameters of the linear, polynomial and Gaussian radial basis (GRB) kernels were respectively tuned to examine whether they output an optimal or acceptable predictive result, either in the validating or the testing part. These trainings were combined with the LSSVM regulation factor changing at a regular step pace. The results showed that, for polynomial kernel (including the linear kernel), the identified optimal parameters were able to outlet appreciating root mean square error, but the correlation seemed not so good as expected. For GRB kernel, we selected a relatively smaller regularization factor for each fixed kernel width to simplify the complexity in practical applications. In conclusion, the linear, polynomial and GRB kernels are feasible to explicit the NIR features of protein for the LSSVM model in NIR analysis of fishmeal samples.*

Keywords - *Gaussian radial basis, kernels, LSSVM, Near-infrared, polynomial.*

I. INTRODUCTION

Fishmeal is a quality feed containing high protein in the modernizing balance of animal breeding. It plays an irreplaceable role in quantization aquaculture. High quality fishmeal provides suitable protein ratio in nutrition to animals which is essential for their growth. Protein content is an important indicator to evaluate the quality level of fishmeal [1]. With the development of aquaculture feed industry, there is increasing use of fish meal, which requires the endurance of identifying high-protein quality of fishmeal to develop precision aquaculture feedings. Near-infrared (NIR) spectroscopy is a rapid technique to use the material information of the reflectance near-infrared region for quantitative analysis of material components [2]. It owns many merits of fast, easiness and simplicity, reagent-free detection, no pollution and multi-component simultaneous determination. NIR technology has been widely used in the fields of agriculture, food, environment and biomedicine [3-5].

There have been several reports investigating NIR spectroscopy for the detection of fishmeal quality [6-8]. Classic linear regression, principal component regression, multiplicative linear regression and partial least squares regression are widely used in these studies. And, some spectrum pretreatment methods as smoothing, derivative, detrending, and standardization are carried out for these quantitative detections [9-11]. However, the linear models are valid on the basis of the inherent linear correlation between the spectral data and the concentration data. These models are hard to get prospective predictions if the spectrum- concentration relationship is not linear. It is very necessary to study effective chemometric algorithms when using NIR spectroscopy measuring protein content of fishmeal. In response to the detection of specific indicators and spectral information, How to improve the calibration of the model prediction accuracy is the hotspot of investigating the NIR technology for quantitative analysis of target components. There includes many challenging issues.

With the development of nonlinear analytical techniques, near-infrared predictive capability has been improved to a certain extent. Least squares support vector machine (LSSVM) is a popular nonlinear analytical method for classification and regression. The fundamental concept in LSSVM is to map the original data onto a high dimensional space by utilizing kernels, in order to transfer a complicated nonlinear problem into a simple linear problem between the dependent variables and the high dimensional data [12]. The distribution of the features in high dimensional space depends not only on the regularization factor, but also on the selection of the kernel and the corresponding parameters. Thus the training of regularization parameter and the generalization parameters in kernels are very determinate to LSSVM models.

The parameters in kernel functions primarily affect the mapping results from nonlinear to linear space. They play a critical role for the redistribution and comprehensional regulation of the sample data in the linear high-dimensional space [13]. Concerning the choice of the applicable kernel for LSSVM model optimization, a different kernel will lead to different performance of LSSVM models [14-15]. The linear kernel is the most commonly-used and relatively simple kernel for LSSVM mapping, while the polynomial kernel can successfully transfer the NIR information from the nonlinear to the linear space along a smooth polynomial curve. On the other hand, The Gaussian radial basis (GRB) kernel is a high-order kernel for data transformation with applicable effectiveness and faster training process.

In this paper, we investigate the functional effects of the three different kernels in LSSVM model for the NIR analysis of protein in fishmeal samples. By comparing the validating results as well as the testing results, the optimal kernel along with its effective parameters can be determined for further applications.

II. SAMPLES AND EXPERIMENT

The rapid detection on fishmeal protein by using NIR spectrometry requires the collection of the spectral data of each fishmeal and the laboratory measured data of protein. The spectral data was collected by FOSS NIR Systems 5000 grating spectrometer. The calibration model is established based on the prepared samples, and is applied for other unmeasured samples.

Ninety-four portions of fishmeal powder were collected and each portion is reserved as nondestructive and treated as an original sample. The protein concentration was detected by using the Kjeldahl method. All of the 94 samples were prepared for the spectrum collection. The spectral detection was launched in a conditioning laboratory at a temperature of 26ºC and a relative humidity of 52%. The spectrometer automatically scanned 64 times for each sample in about 2 minutes and output the average spectrum as the final measured data. The full-scanning range was set as 1100-2500nm. The resolution was set as 2 nm, thus we have 700 points of wavelengths in the full range.

For NIR analysis, all 94 samples were divided into 3 sample sets for different uses in the modeling and testing processes. A randomly selected bundle including 24 samples were used for model test after establishing calibration models. The remaining samples were automatically divided into the calibrating set (46 samples) and validating set (24 samples) by Kennard-Stone method [16]. The calibration samples were used for modeling and the validation samples for optimization.

III. LSSVM MODEL AND EVALUATION METHOD

The idea of LSSVM regression is to use the kernel function to map the target data into a high-dimensional space to form a linear relationship between the spectra and the concentrations [17], so that the regression models can be defined using the following equation,

$$\hat{c}_j = \sum_{i=1}^{m} \alpha_i K(x_j, x_i) ,$$

where $\hat{c}_j$ the predictive concentration of the *j*-th sample, $K(x_j, x_i)$ is the kernel function depending on the spectral data and α_i is the Lagrange multiplier which is defined as

$$\alpha_i = \left((A_i)^{\mathrm{T}} A_i + \frac{1}{2\gamma} \right)^{-1} ,$$

where γ is called the regularization factor [12] and A_i is a linear combination of all the calibration spectra (the NIR spectra with *m* wavenumbers), weighted by the concentration values.

The effect of LSSVM regression depends on the selection of the kernel and corresponding parameters. Moreover, when the kernel function is selected, the corresponding kernel parameters for tuning are identified, and will be trained in combination with γ to accomplish LSSVM model optimization.

In this study, to search for the moderate robustness and stability to enable nonlinear modeling for the acquired NIR dataset, we investigated three different kernels for LSSVM,

Linear kernel	$K(x_j, x_i) = x_j \cdot x_i + b$,
Polynomial kernel	$K(x_j, x_i) = (x_j \cdot x_i + b)^d$,
Gaussian radial basis kernel	$K(x, x_i) = \exp(\|x_j - x_i\|^2 / 2\sigma^2)$,

where *d* represents the degree of polynomial and σ^2 represents the GRB kernel width, used to adjust the degree of generalization. It is obvious that the polynomial kernel will degenerate to a linear kernel when *d* equals to 1. Thus we only need to operate the polynomial and the GRB kernel in LSSVM modeling.

The evaluation for LSSVM model mainly depends on the predictive bias and the correlation, so the modeling indicators include root mean squared error (RMSE) and the correlation coefficient (R). The modeling indicators in the validating process were denoted as RMSEv and Rv, while the indicators in testing were denoted as RMSEt and Rt.

IV. RESULTS AND DISCUSSIONS

LSSVM models were established for the calibration of fishmeal samples targeting the concentration of protein. As the training of polynomial kernel includes the case of linear kernel, we only discussed the procedures for the polynomial kernel and the GRB kernel. With the increase of the regularization factor, the error of experience is over-punished and the model would be overlearning. When the regularization factor exceeds a certain value, the empirical risk and learning feature would not change any more. On this basis, it is worth noting that the regularization factor (γ) is necessary to be trained in combination with the degree of polynomial (d) in the polynomial kernel, and with the kernel width (σ^2) in GRB kernel.

Firstly, we investigated the LSSVM modelling results for the polynomial kernel. Generally, we have γ consequently valued as $\{2^{-11}, 2^{-9}… 2^{-1}, 2^{1}…2^{33}, 2^{35}\}$, and, d valued as {0, 1, 2, 3}. The model optimization was done for the validating samples and the predictive results were obtained according to every parametric model (showed in Fig. 1). It can be seen from Fig. 1 that the predictive results were awful when d equals to 0, which means that using a constant as the kernel would not lead to a prospective modeling result. The linear (d=1), quadratic (d=2) and the cubic (d=3) polynomials were able to acquire the optimal effect when γ was properly valued. Thus we selected the most optimal models (the 10 solid circles in Fig. 1) to perform the predictions for the testing samples, and the testing results as well as the validating results were listed in Table 1. The data in Table 1 showed that the LSSVM models with polynomial kernels worked quite well in the validating part, just because the optimal parameters were outlet based on the validating samples, while the performance went down for the testing samples that are excluded in the optimizational step. The correlation results (Rt) were generally lower than 0.9.

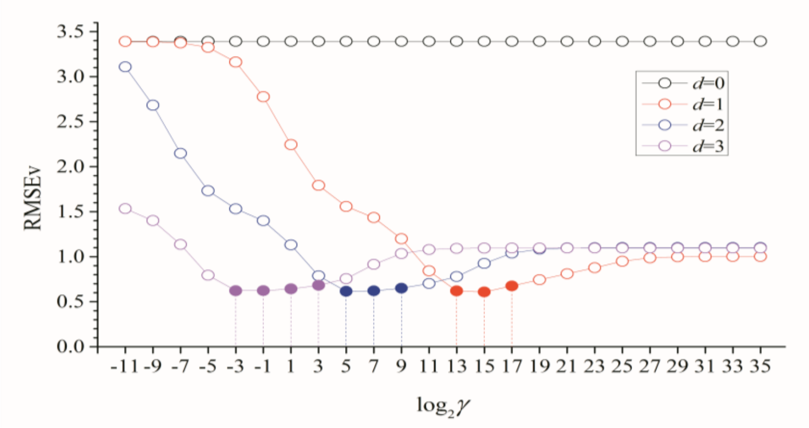

Figure 1 The validating results corresponding to the tuning of γ based on different degrees of polynomial kernel (d represents the degree of polynomial)

Table 1 The predictive results for the optimal polynomial kernels in LSSVM modeling

	γ	RMSEv	Rv	RMSEt	Rt
d=1					
	2^{13}	0.6199	0.9896	2.1157	0.8571
	2^{15}	0.6088	0.9855	2.2115	0.8395
	2^{17}	0.6749	0.9804	2.0838	0.8283
d=2					
	2^{5}	0.6158	0.9881	2.0722	0.8642
	2^{7}	0.6206	0.9847	2.0084	0.8593
	2^{9}	0.6501	0.9822	1.9229	0.8421
d=3					
	2^{-3}	0.6226	0.9874	2.0222	0.8694
	2^{-1}	0.6232	0.9846	1.9357	0.8661
	2^{1}	0.6428	0.9826	1.8787	0.8471
	2^{3}	0.6824	0.9796	2.1069	0.7910

Note: d represents the degree of polynomial kernel; γ represents the regularization parameter in LSSVM modeling

Secondly, we investigated the LSSVM modeling results GRB kernel. The regulation factor γ was tuned within the same range as for the polynomial kernel, i.e. γ consequently equals to 2^{-11}, 2^{-9}… 2^{-1}, 2^{1}… 2^{33}, 2^{35}, while the GRB kernel width σ^2 was tuned valuing to 2^{-7}, 2^{-5}… 2^{-1}, 2^{1}… 2^{15}, 2^{17}. The model optimization was done for the validating samples and the predictive results were obtained according to every parametric model (showed in Fig. 2). It can be seen from Fig. 2 that all the curves corresponding to σ^2 dropped down to the stable minimal line except σ^2 equaling to 2^{-7}, 2^{15} and 2^{17}, which were the border values in this tuning strategy. With this resultant phenomenon we believed that the optimal values of σ^2 were included in this research. Additionally, the dropping curves reached the minimal level at different values of γ. The first minimum-reaching point for each curve was sketched as a solid circle in Fig. 2. It is interesting that with the increasing value of σ^2, the first minimum-reaching point identified a relatively large value of γ. Meanwhile, these points gave out the optimal validating results at the same predictive level.

We located these first minimum-reaching points and extracted the models, perform the predictions for the testing samples, and the testing results as well as the validating results were listed in Table 2. Table 2 showed that the LSSVM models with GRB kernels worked awfully well in the validating part, just as the performance with polynomial kernel, while the predictive effects went down for the testing samples. But obviously, the correlation results (Rt) were generally above 0.9. In comparison with the polynomial kernel, LSSVM models will be evaluated performing slightly better when using the GRB kernel.

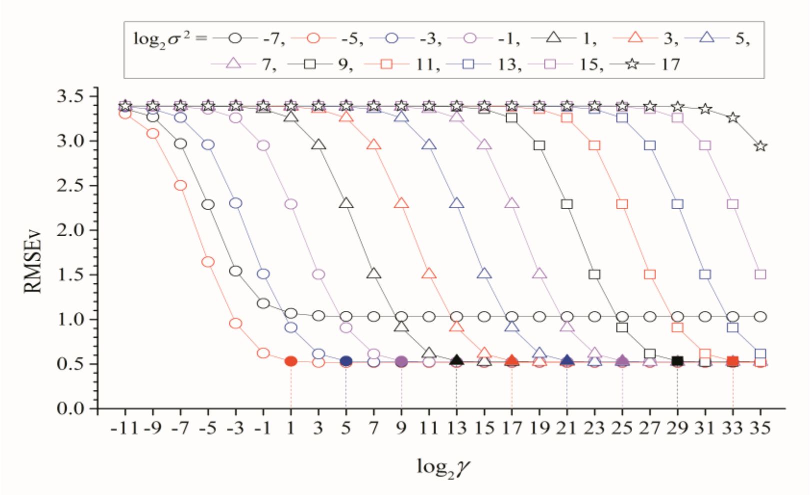

Figure 2 The validating results corresponding to the tuning of γ based on the comparison of σ^2 valuing in Gaussian radial basis kernel

Table 2 The predictive results for the optimal GRB kernels in LSSVM modeling

σ^2	γ	RMSEv	Rv	RMSEt	Rt
2^{-5}	2^{1}	0.5311	0.9921	1.5112	0.9078
2^{-3}	2^{5}	0.5335	0.9920	1.3955	0.9203
2^{-1}	2^{9}	0.5338	0.9919	1.3873	0.9212
2^{1}	2^{13}	0.5338	0.9919	1.3868	0.9213
2^{3}	2^{17}	0.5338	0.9919	1.3868	0.9213
2^{5}	2^{21}	0.5338	0.9919	1.3868	0.9213
2^{7}	2^{25}	0.5338	0.9919	1.3868	0.9213
2^{9}	2^{29}	0.5338	0.9919	1.3868	0.9213
2^{11}	2^{33}	0.5330	0.9919	1.3855	0.9213

Note: σ^2 represents degree of generalization in Gaussian radial basis kernel and γ represents the regularization parameter in LSSVM modeling

To view the trend of predictive effects in testing part, we chose two specific models from Table 2 to sketch the RMSEt curves versus γ (see Fig. 3). One is the first minimum-reaching point with a minimum value of γ ($\sigma^2=2^{-5}$), the other is the one with an obvious drop in RMSEt ($\sigma^2=2^{-1}$). The testing results also gave out the similar changing trend to the validating results. The RMSEt curve dropped down at an exact γ point and then kept stable at the minimal line. These results will conduct us to select a right value of the regulation factor.

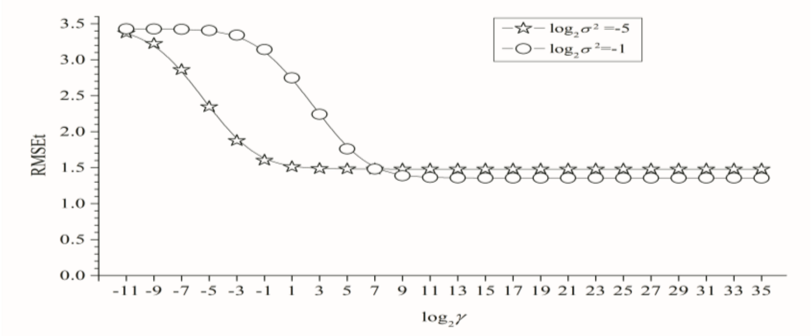

Figure 3 The testing results for two specific values of σ^2 in Gaussian radial basis kernel

V. CONCLUSIONS

In this paper, parametric tuning was studied for the LSSVM models in NIR analysis of fishmeal protein concentration. We respectively exert the linear, polynomial and GRB kernels in the nonlinear-to-linear transformation and examine whether they would output an optimal or an acceptable predictive result, either in the validating process or in the testing part. These training combined the LSSVM regulation factor with the tuning of degree of polynomial or with the GRB kernel width, and testing the grid search results for a parameter step-screening procedure. We can find out some optimal parameters for the polynomial kernel that outlet the prospective predictive root mean square error, but the correlation effect was not as good as expected. On contrary, for the GRB kernel, we cannot tell which would be the most optimal model as the RMSEv curve quickly dropped to a minimum value and stay stable, but we chose the first minimum-reaching points for each σ^2 as the aim because we prefer the γ equaling to a relatively small value so that the forward applicable model can be of less complexity.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Scientific Foundation of China (61505037), the Natural Scientific Foundation of Guangxi (2015GXNSFBA139259), and the University Scientific Research Project of Guangxi Education Office (KY2015ZL095).

REFERENCES

[1] R. L. Olsen, M. R. Hasan, A limited supply of fishmeal: Impact on future increases in global aquaculture production, *Trends in Food Science & Technology, 27,* 2012, 120-128.

[2] D. A. Burns, E. W. Ciurczak, *Handbook of near-infrared analysis (3rd ed.)*(Boca Raton, CSC Press LLC, 2006).

[3] N. A. Ngo Thi and D. Naumann, *Analytical and Bioanalytical Chemistry, 387,* 2007, 1769-1777.

[4] J. Sarembaud, G. Platero, M. Feinberg, Stability studies of agrifood reference materials under different conditions of storage by near-infrared spectroscopy, *Food Analytical Methods, 1,* 2008, 227-235.

[5] R. W. Bondi, B. Igne, J. K. Drennen, Effect of experimental design on the prediction performance of calibration models based on near-infrared spectroscopy for pharmaceutical applications, *Applied Spectroscopy, 66,* 2012, 1442-1453.

[6] Z. Y. Niu, L. J. Han, X. O. Su, Z. H. Yang, Quality Predict ion of Fish Meal by Near Infrared Reflectance Spectroscopy , *Transactions of the Chinese Society for Agricultural Machinery, 36,* 2005, 68-71.

[7] Z. L. Yang, L. J. Han, X. Liu, Q. F. Li, Rapidly qualitative discrimination of meat and bone meal in fishmeal by visible and near infrared reflectance spectroscopy, *Transactions of the CSAE, 25,* 2009, 308-311.

[8] H. Z. Chen, F. Chen, K. Shi, Q. X. Feng, Near-Infrared Analysis of Fishmeal Protein based on Random Forest, *Transactions of the Chinese Society for Agricultural Machinery, 46(5),* 2015, 233-238.

[9] M. Clupek, P. Matejka, K. Volka, Noise reduction in Raman spectra: Finite impulse response filtration versus Savitzky-Golay smoothing, *Raman Spectroscopy, 38,* 2007, 1174-1179.

[10] D. Syvilay, N. Wilkie-Chancellier, B. Trichereau, Evaluation of the standard normal variate method for Laser-Induced Breakdown Spectroscopy data treatment applied to the discrimination of painting layers, *Spectrochimica Acta Part B:Atomic Spectroscopy, 114,* 2015, 38-45.

[11] D. Sabatier, P. Dardenne, L. Thuries, Near infrared reflectance calibration optimization to predict lignocellulosic compounds in sugarcane samples with coarse particle size, *Journal of Near Infrared Spectroscopy, 19,* 2011, 199-209.

[12] Barman, N. C. Dingari, G. P. Singh, J. S. Soares, *Analytical Chemistry, 84,* 2012, 8149-8156.

[13] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewaller, Least Squares Support Vector Machines, *World Scientific Publishing,* 2002.

[14] X. An, S. Xu, L. D. Zhang, Multiple Dependent Variables LS-SVM Regression Algorithm and Its Application in NIR Spectral Quantitative Analysis, *Spectroscopy and Spectral Analysis, 29(1),* 2009, 127-130.

[15] H. Z. Chen, A. Wu, Q. X. Feng, G. Q. Tang, FT-MIR Modelling Enhancement for the Quantitative Determination of Haemoglobin in Human Blood by Combined Optimization of Grid-Search LSSVR Algorithm with Different Pre-Processing Modes, *Analytical Methods, 7,* 2015, 2869-2876.

[16] R. W. Kennard, L. A. Stone, Computer Aided Design of Experiments, *Technometrics, 11,* 1969, 137-148.

[17] N. Cristianini, J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods* (New York: Cambridge University Press, 2000).