VOL. 10, NO. 19, OCTOBER 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

ADAPTIVE LINEAR NEURON IN VISIBLE AND NEAR INFRARED SPECTROSCOPIC ANALYSIS: PREDICTIVE MODEL AND VARIABLE SELECTION

Kim Seng Chia
Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Malaysia
E-Mail: kschia@uthm.edu.my

ABSTRACT

Near infrared (NIR) spectroscopic analysis has been widely evaluated in various areas due to its potential to be an alternative of numerous conventional measurement approaches that are time consuming, expensive, or destructive. This study evaluated the feasibility of adaptive linear neuron (Adaline) to be implemented as a variable selection approach to identify effective NIR wavelengths that can be used to predict the soil organic matter (SOM) so that a parsimonious model can be built. Adaline was optimized using its optimal learning rate and training adaptation cycles. After that, the effective wavelengths were identified based on the weight values of the best Adaline. The best predictive accuracy was achieved by the proposed Adaline that used 40 of the total 891 wavelengths with the root mean square error of prediction (RMSEP) and correlation coefficient of prediction (*rp*) of 2.163% and 0.9849, respectively. Findings show that the proposed variable selection approach by means of Adaline is capable of producing a parsimonious model that was able to predict the soil organic matter with better accuracy.

Keywords: adaptive linear neuron, near infrared spectroscopy, variable selection, soil organic matter.

INTRODUCTION

Near infrared (NIR) spectroscopic analysis has been widely evaluated in various areas e.g. agriculture, aquaculture, industry, and medication due to its potential to be an alternative of numerous conventional measurement approaches that are time consuming, expensive, or destructive. Since NIR spectroscopic analysis is an indirect or secondary measurement approach, its performance is highly dependent on the accuracy of primary measurement that provides references for training, and also the ability of a predictive model to establish the relationship between the spectral data and the component of interest. The former can be optimized by following the standard operating procedure of conventional approaches strictly. The latter, on the other hand, is one of interesting research areas that industries and researchers concern about.

Since the number of variables is normally more than the number of available samples in NIR spectroscopic analysis, variable reduction approaches e.g. principal component analysis (PCA) and partial least squares (PLS) has been widely implemented to remove unwanted or redundant variables. In order to further improve the predictive accuracy, variable selection approach has been advocated. This is because variable selection strategies could ultimately produce a parsimonious model in ways that over-fitting problems can be avoided. One popular variable selection approach is using a projection model of PLS to identify optimal spectral intervals using the coefficient values of the model (Xu, Qi *et al.* 2012, Yang, Kuang *et al*. 2012, Álvarez-Sánchez, Priego-Capote *et al*. 2013, Heinze, Vohland *et al*. 2013, Ouyang, Chen *et al*. 2013). However, these proposed algorithms require a combination of few algorithms (e.g. uninformative variable elimination coupled with successive projections algorithm) that makes the algorithm much complex and difficult to be used to interpret the identified effective wavelengths directly.

Adaptive linear neuron (Adaline) is a single layer artificial neuron network. Adaline has been proposed and investigated in many applications e.g. tracking power system harmonics (Zouidi, Fnaiech *et al*. 2008, Sarkar, Choudhury *et al*. 2011) and speed control (Kaminski and Orlowska-Kowalska 2012). Adaline is popular in many applications because Adaline can be optimized using Widrow-Hoff delta rule that does not involve inverse matrix, and has a low computational load. Recently, Adaline has been successfully implemented to model high dimensional NIR spectral data to the boiling point of diesel fuel without any data reduction approach (Chia 2015). Since Adaline contains only one layer, the weight values of each variable may be directly proportional to the influence of the variable. In other words, effective variables may be identified based on the weight values after the Adaline has been optimized.

In this study, a set of public data was used to evaluate the potential of Adaline to model the relationship between soil organic matter (SOM) and near infrared (NIR) spectral data. After that, the weight values of the optimized Adaline were evaluated to identify effective wavelengths.

MATERIAL AND METHODS

Spectral data

The visible and near infrared (VIS-NIR) spectral data of the 108 soil samples that measured in Abisko, Northern Sweden (681210N, 181490E) were used in this study. This data set was one of public data sets that was available on the webpage of

http://www.models.kvl.dk/datasets/. 72 spectral data were acquired from soil samples with a depth of 0 to 5cm; while the remainders were acquired from soil samples with a depth of 5 to 10 cm (Rinnan and Rinnan 2007). These VIS-NIR spectral data ranged from 400nm to 2498nm, with an interval of 2nm. After removing visible spectral data, the range of NIR spectral data that used in this study was from 700nm to 2498nm. With an interval of 2nm, there were 900 wavelengths per spectrum.

Component of Interest

Soil organic matter (SOM) that used in this study was measured using loss on ignition test at 550 OC. Half of the data were randomly separated for training, while the remainders were used for testing. The mean and standard deviation of training data were 85.27% and 11.15%, respectively. The mean and standard deviation of testing data were 85.58% and 10.59%, respectively. The ranges of training and testing data sets were from 42.91 to 95.85% and from 44.11 to 95.52%, respectively. Extrapolation prediction was avoided in this study because the range of the testing data set was within the range of the training data set.

Data processing

MATLAB (version R2009b, win64) was used to process and analyse data in this study. Firstly, Savitzky-Golay (SG) second order derivative with filter length of 34nm was used to remove high frequency noises, baseline shift and slope effects that existed in the acquired spectral data. After that, the spectral data of the training set were normalized into a range of -1 and 1. After that, the parameter of normalization was retained and used to pre-process the spectral data of the testing set.

Besides, the soil organic matter (SOM) values of training data were normalized with minimum and maximum values of -1 and 1, respectively. The return parameter was used to post-process all predicted values into their normal scale so that the accuracy of Adaline can be analyzed in the original scale.

Adaptive linear neuron

Adaptive linear neuron (Adaline) that coupled Widrow-Hoff delta rule (also known as least mean square (LMS) algorithm) was used to predict the soil organic matter (SOM) based on near infrared (NIR) spectral data. Two parameters affect the performance of Adaline, i.e. learning rate and adaptation cycle. Learning rate, μ controls the magnitude of the change in each training process. Adaptation cycles, on the other hand, must be sufficient so that Adaline can achieve its optimal performance with a given learning rate.

In order to identify the best performance of Adaline with a given learning rate, the maximum number of adaptation cycles was arbitrarily set to a high value of 5000. The effects of six different learning rate, μ of 0.01, 0.001, 0.0005, 0.0001, 0.00005, and 0.00001 on the predictive accuracy of Adaline were evaluated. The training of Adaline was continued until either the maximum number of adaptation cycles was reached or the prediction accuracy cannot be further improved using a learning curve.

Effective wavelength selection

The effective wavelengths will be selected based on the coefficient value (also known as the weight value) of each wavelength from the best Adaline model. This idea is analogue to the variable selection approach that is based on the coefficient values of multiple linear regression. Since multiple linear regression is incapable of modeling high dimensional data, such as near infrared spectral data, an alternative e.g. Adaline is worthy to be investigated.

Effective wavelengths should have higher weight values than the wavelengths that are less influential. By removing these less influential wavelengths that have lower weight values, the number of input variables can be reduced significantly. Ultimately, the objective of variable selection could be achieved, that is, to produce a parsimonious model that can achieve comparable or even better predictive accuracy.

Predictive accuracy

The predictive accuracy of Adaline was estimated using the root mean square error (RMSE) as that stated in Eq. 1.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \quad (1)$$

Where, y_i and $\hat{y}_i$ are the actual and the predicted SOMs for *i*-th sample, respectively; and n is the total number of samples. When the prediction data set is used, the computation of RMSE is defined as root mean square error of prediction (RMSEP). When the training data set is used, on the other hand, the computation of RMSE is defined as the root mean square error of calibration (RMSEC). Since the RMSE is calculated based on the derivation between the actual and predicted SOMs, the best predictive model should achieve the smallest RMSEP value.

Besides, the correlation coefficient of calibration, r_c and the correlation coefficient of prediction, r_p were used to estimate the strength of the relationship between the predicted SOM and actual SOM in training and prediction data sets, respectively.

RESULTS AND DISCUSSION

Preprocessed spectral

Figure-1 illustrates that the acquired spectral data contain baseline shift effect because each spectral data have different baselines. This unwanted effect is hardly to be eliminated during data acquisition because any deviation in terms of light source intensity, surface of samples, temperature, or angle of spectral acquisition could contribute this kind of unwanted effects.

Nonetheless, the baseline shift effect can be reduced significantly using Savitzky-Goley second order preprocessing approach, as that illustrated in Figure-2. One drawback of the second order pre-processing approach is it alters the shape of the original spectrum. However, this should not have any negative impact to the predictive accuracy of a predictive model.

Learning rate and adaptation cycles

Figure-3 illustrates the predictive performance of Adaline that used different learning rate of 0.001, 0.0005, 0.0001, 0.00005, and 0.00001 in the first 600 adaptation cycles. Adaline that used learning rate of 0.01 was excluded from the discussion because this model could not reach its convergence in this study.

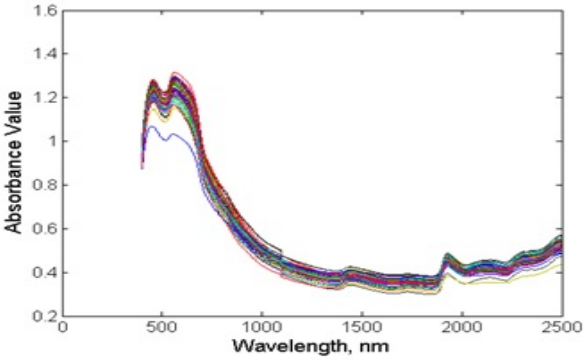

Figure-1. Spectral data without any pre-processing.

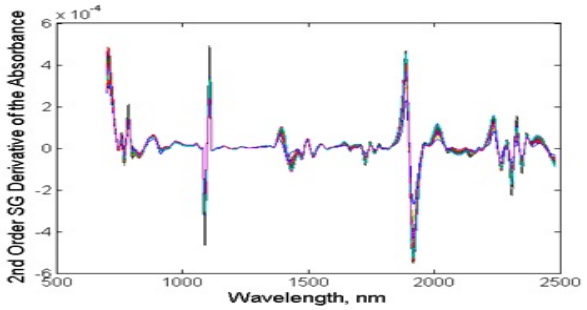

Figure-2. Spectral data after preprocessing via second order Savitzky-Golay derivative.

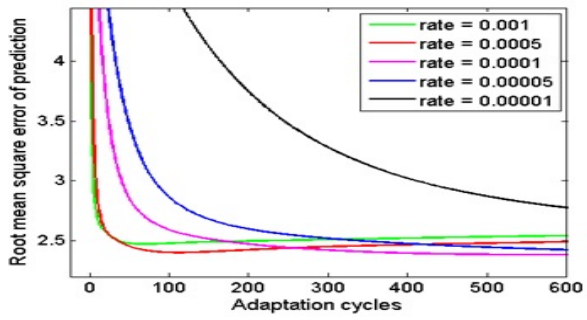

Figure-3. Learning Curve - the predictive performance of Adaline that used different learning rate in the first 600 adaptation cycles.

The performance of Adaline in Figure-3 indicates that Adaline that used the highest value of the learning rate (i.e. 0.001) reached its optimal performance with only 65 adaptation cycles. However, the RMSEP of the Adaline was increased if the training continued. In other words, the Adaline could achieve a worse predictive accuracy if the training was not stopped with its optimal adaptation cycles. Besides, the optimal accuracy of the Adaline was not the best among the other Adaline models that used lower learning rates. This suggests that the Adaline that used a learning rate of 0.001 was under-fitted.

Adaline that used low learning rate values of 0.0001, 0.00005 and 0.00001 cannot reach their convergence in the first 600 adaptation cycles. The prediction performance, however, was improved when the number of adaptation cycles was further increased. In other words, more adaptation cycles were needed to obtain the optimal performance of these Adaline models.

Table-1 summarizes the optimal performance of Adaline that used different learning rates when sufficient adaptation cycles were provided. The best root mean square error of prediction (RMSEP) was achieved by Adaline that used a learning rate of 0.00001 with 561 adaptation cycles. When the learning rate was reduced to 0.00005 or 0.00001, similar RMSEP was achieved when sufficient adaptation cycles were provided. However, it is worth to highlight that the Adaline that used these low learning rate (i.e. 0.00005 or 0.00001) needed double or even more adaptation cycles in order to achieve the similar performance. Adaline that used the best learning rate should compromises the predictive accuracy and the number of adaptation cycles. Thus, the best model is the Adaline that used a learning rate of 0.0001 with 561 adaptation cycles.

Table-1. The optimal performance of Adaline that used different learning rates.

Learning rate	Optimal adaptation cycles	The performance of Adaline			
		Training		Prediction	
		RMSE %	r_c	RMSE %	r_p
0.001	65	0.8445	0.9971	2.4690	0.9856
0.0005	102	0.9130	0.9966	2.3990	0.9856
0.0001	**561**	**0.8859**	**0.9968**	**2.3800**	**0.9860**
0.00005	1135	0.8835	0.9968	2.3800	0.9861
0.00001	5000 (max)	0.9297	0.9965	2.3840	0.9858

Effective wavelengths

The weight values of the best Adaline that used learning rate of 0.0001 and 561 adaptation cycles were used to identify effective wavelengths. A total 891 weight values were available from the second order derivative spectral data. These weight values ranged from -0.0215 to 0.0191, with mean of 0.0005 and standard deviation of 0.0043. Intuitively, the wavelengths that have absolute values less than the double of the standard deviation (i.e. 0.0086) could be considered as insignificant and thus should removed. Consequently, only 40 wavelengths were

selected to be used as the inputs of Adaline to produce a parsimonious model.

Adaline with the selected effective wavelengths

A Adaline was trained using the 40 selected effective wavelengths with a learning rate of 0.001. Figure-4 illustrates the performance of the Adaline with maximum 5000 adaptation cycles. Results indicate that the Adaline that used the proposed variable selection approach was capable of achieving RMSEP of 2.38% with fewer adaptation cycles of 439, compared to the previous best Adaline that used the whole spectrum as its inputs with 561 adaptation cycles. Besides, the RMSEC of the proposed model was 1.694% compared to 0.8859% that was achieved by the Adaline that did not use the proposed variable selection approach. This suggests that the RMSEC of the latter was over-optimistic. In other words, misleading performance estimation could be occurred if RMSEC was used for interpretation when high dimensional and complex inputs e.g. NIR spectral data were involved.

Next, it is noticed that the predictive performance of the Adaline that used the proposed variable selection approach could be further improved when the training was continued. The best RMSEP of 2.163% was achieved when 5000 adaptation cycles were used for the proposed model. This performance was around 9.1% better than the Adaline that did not use the proposed variable selection approach.

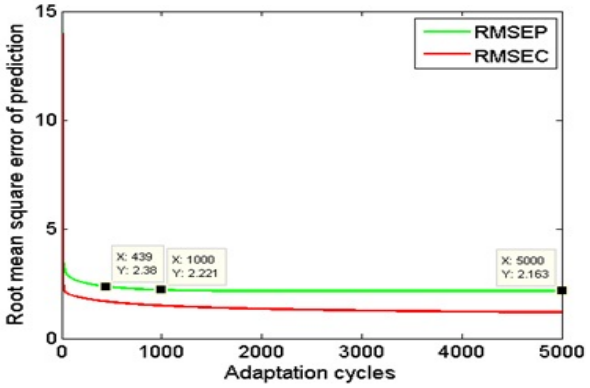

Figure-4. The learning curve of the Adaline that used the proposed variable selection approach.

CONCLUSIONS

This study proposed a variable selection approach by means of Adaline to predict the soil organic matter (SOM) using near infrared (NIR) spectral data. The proposed variable selection approach successfully reduced the input variable number from 891 to 40 wavelengths. With the 40 selected wavelengths as the inputs of Adaline, the Adaline achieved the best RMSEP of 2.163%, with r_p of 0.9849, RMSEC of 1.193%, and r_c of 0.9941. In other words, the objective of variable selection has been achieved by the proposed approach, i.e. to produce a parsimonious model that is capable of achieving better predictive accuracy.

ACKNOWLEDGEMENTS

The author would like to acknowledge Universiti Tun Hussein Onn Malaysia (UTHM) for providing financial support (Vot U351), and Advanced Mechatronic Research Group (AdMiRe), FKEE, UTHM for providing facilities for this study; and Riikka Rinnana and Åsmund Rinnan, University of Copenhagen, Denmark for the data sets used in this study.

REFERENCES

[1] Álvarez-Sánchez B., F. Priego-Capote J. García-Olmo M. C. Ortiz-Fernández L. A. Sarabia-Peinador and M. D. Luque de Castro 2013. "Near-infrared spectroscopy and partial least squares-class modeling (PLS-CM) for metabolomics fingerprinting discrimination of intervention breakfasts ingested by obese individuals." Journal of Chemometrics, Vol. 27, No. 9, pp. 221-232.

[2] Chia K. S. 2015. Predicting the Boiling Point of Diesel Fuel using Adaptive Linear Neuron and Near Infrared Spectrum. 10th Asian Control Conference. Kota Kinabalu, pp. 2438 - 2490.

[3] Heinze S., M. Vohland Rainer G. Joergensen and B. Ludwig 2013. "Usefulness of near-infrared spectroscopy for the prediction of chemical and biological soil properties in different long-term experiments." Journal of Plant Nutrition and Soil Science Vol. 176, No. 4, pp. 520-528.

[4] Kaminski M. and T. Orlowska-Kowalska 2012. Adaline-based speed controller of the drive system with elastic joint. Optimization of Electrical and Electronic Equipment (OPTIM), 13th International Conference on.

[5] Ouyang Q., Q. Chen J. Zhao and H. Lin 2013. "Determination of Amino Acid Nitrogen in Soy Sauce Using Near Infrared Spectroscopy Combined with Characteristic Variables Selection and Extreme Learning Machine." Food and Bioprocess Technology Vol. 6, No. 9, pp. 2486-2493.

[6] Rinnan R. and Å. Rinnan 2007. "Application of near infrared reflectance (NIR) and fluorescence spectroscopy to analysis of microbiological and chemical properties of arctic soil." Soil Biology and Biochemistry, Vol. 39, No. 7, pp. 1664-1673.

[7] Sarkar A., S. R. Choudhury and S. Sengupta 2011. "A self-synchronized Adaline network for on-line tracking of power system harmonics." Measurement Vol. 11, pp. 784-790.

[8] Xu H., B. Qi T. Sun X. Fu and Y. Ying 2012. "Variable selection in visible and near-infrared spectra: Application to on-line determination of sugar content in pears." Journal of Food Engineering Vol. 109, No. 1, pp. 142-147.

[9] Yang H., B. Kuang and A. M. Mouazen. 2012. "Quantitative analysis of soil nitrogen and carbon at a farm scale using visible and near infrared spectroscopy coupled with wavelength reduction." European Journal of Soil Science, pp. 1-11.

[10] Zouidi A., F. Fnaiech K. Al-Haddad and S. Rahmani 2008. Adaptive linear combiners a robust neural network technique for on-line harmonic tracking. Industrial Electronics, IECON. 34th Annual Conference of IEEE.