34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Determination of Glucose Concentration from Near-Infrared Spectra Using Locally Weighted Partial Least Square Regression

Bilal Malik and Mohammed Benaissa, *Senior Member, IEEE*

Abstract— **This paper proposes the use of locally weighted partial least square regression (LW-PLSR) as an alternative multivariate calibration method for the prediction of glucose concentration from NIR spectra. The efficiency of the proposed model is validated in experiments carried out in a non-controlled environment or sample conditions using mixtures composed of glucose, urea and triacetin. The collected data spans the spectral region from 2100nm to 2400nm with spectra resolution of 1nm. The results show that the standard error of prediction (SEP) decreases to 23.85 mg/dL when using LW-PLSR in comparison to the SEP values of 49.40 mg/dL, and 27.56 mg/dL using Principal Component Regression (PCR) and Partial Least Square (PLS) regression respectively.**

I. Introduction

Research over the years has identified near infra red (NIR) spectroscopy coupled with chemometrics and signal processing methods as one of the most promising techniques for non-invasive blood glucose measurement [1-3] to help with the management of diabetes. NIR spectroscopy is fast, easy and accurate in comparison to other techniques. However, the challenge is to extract the glucose dependent information in the presence of dominating signals from spectra of their mixtures, associated spectral variations, and the underlying spectral noise. Sophisticated multivariate data-analysis algorithms such as principle component analysis (PCA), principle component regression (PCR) and partial least square regression (PLSR), coupled with advanced signal processing techniques have been proposed to build robust regression models for calibration and prediction of glucose concentrations [4-8]. Linear models are sometimes limited due to the chemical properties of a measuring object, which have an intricate effect on NIR spectra. Another issue in building robust models is how to manage the variations in the process characteristics, which is vital in the chemical industry. Thus, maintenance of the models is an important issue to consider in soft-sensors[9]. Traditional linear models generally discard the data after the training phase losing potentially valuable extra information during the prediction phase, which is why a class of the so called "memory-based" methods such as locally weighted regression (LWR) that retain the training data and use it for each prediction has been advocated for dynamic processes. LWR is a technique for non-parametric regression, which performs regression around a point of interest, using only training data that are "local" to that point[10, 11].

In LWR, the goal is to fit θ to minimize $\sum_i w^i(y^i - \theta^T x^i)^2$ and output $\theta^T x$; Where, x is the input data matrix, y is output vector, and the w^i's are non-negative valued weights. For a particular value of i, if w^i is large then θ is chosen in such a way as to make $(y^i - \theta^T x^i)^2$ small. However, the $(y^i - \theta^T x^i)^2$ error term will be ignored in the fit if w^i is small.

$$w^i = \exp(-\frac{((x^i - x)^2)}{2\tau^2})$$

The weights depend on the particular point x. Moreover, if $|x^i - x\ |$ is large, then w^i is small and if $|x^i - x\ |$ is small, then w^i is close to 1. Thus, the value of θ is selected in such a way as to give a higher "weight" to the errors on training examples close to the query point x. τ is called the bandwidth parameter and it controls how quickly the weight of a training example falls off with the distance of its x^i from the query point x.

A fair standard choice for weights for a vector x and an appropriate choice of τ could be generalized by the following equation.

$$w^i = \exp(-\frac{((x^i - x)^T(x^i - x))}{2\tau^2})$$

LWR has been used in agriculture and the food industry [12]. Kim et al have used LWR for estimation of active pharmaceutical ingredients[13]. However, to our knowledge, no attempt has been made to date to use LWR for predicting glucose concentrations from NIR spectra.

In this work, a local linear regression model using LW-PLSR is developed for the quantitative analysis of glucose using NIR spectra, we believe for the first time. It is also shown using practical data in a non-controlled environment that the proposed LW-PLSR technique performs better than the conventional linear regression techniques.

II. Data preparation

The samples were prepared by dissolving glucose, urea and triacetin in a phosphate buffer solution. Thirty samples were prepared with different concentrations of the analytes to span their physiological range in blood. The concentration of glucose was between 20 mg/dL to 500 mg/dL, triacetin concentration between 10 to 190 mg/dL and urea between 0-50 mg/dL in the prepared samples. A Fourier transform spectrometer (FTIR Cary 5000 version 1.09) was used to collect the spectrum from these prepared samples. From

Bilal Malik is a Ph.d. student at Sheffield University, Sheffield, UK in the department of Electrical and Electronics Engineering. (phone: 0044-144-5188; fax: 0044-144-5143; e-mail: elp09bm@ Sheffield.ac.uk).
Mohammed_Benaissa. is with the Electrical and Electronics Engineering Department,SheffieldUniversity,Sheffield,UK,
e-mail:m.benaissa@sheffield.ac.uk

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

each sample, three spectra were collected, and a total of 90 NIR spectra were collected from the spectrometer in this manner. The wavelength region chosen for collecting the spectra was 2100nm to 2400nm, with a spectral resolution of 1 nm. The absorbance spectra of the buffer solution were used as reference. These experiments were conducted in a non-controlled environment in order to test the ability of the proposed model to determine the concentration of glucose.

III. Model development

The quantitative evaluations were carried out using matlab version R2010a. The collected data was pre-processed before calibration using Savitzky-Golay filter [14], which reduces the effect of noise. The window size and the polynomial order in the filter were 117 and 5, respectively.

The spectra were divided randomly into two sets for the calibration and validation of the model. The calibration model was built using the first set containing the three replicate spectra of 20 samples. The calibrated model was tested using the second set containing the triplicate spectra of 10 samples. This basic procedure was kept the same in development of the three models PCR, PLSR and LW-PLSR. The cross validation was employed to determine the number of principle components (PCs) in case of PCR and latent variables (LVs) in case of PLSR and LW-PLSR to build the best model. The minimal value of root mean square error of cross validation (RMSECV) indicated the best number of latent variables in each case. RMSECV is much better indication of model fit as compared to root mean square error of calibration (RMSEC), as the latter does not point out when the model is over-fitted to the data. The predictive accuracy of models was tested by estimating the root mean square error of prediction (RMSEP) on the independent test data set.

The PCR model was developed with different values of principle components (PCs) and the best results were obtained using 6 PCs as suggested by cross validation. In case of PLSR and LW-PLSR 6LVs were suggested by cross validation which resulted in the best models.

The variance in the training data is described by latent variables. The latent variables not only capture variance but also the correlation with the analytical data. The process is controlled by the cross validation in order to make sure that the correlation is authentic rather than some selective fitting of noise. The cross validation was performed with venetian blinds w/ 7 splits utilizing PLS-Toolbox version 3.5 from Eigenvector Research Inc [15]. The implementation of venetian blinds is simple and easy [16]. The square of regression coefficients **(R^2),** standard error of calibration (SEC), standard error of cross validation (SECV) and standard error of prediction (SEP) were used to evaluate the capacity of the calibration models to predict the glucose concentration from the testing spectra.

IV. Results

The motivation for using the locally weighted regression was to compare the ability of PLS-LWR to predict the concentration of glucose in the mixture solution of glucose, triacetin and urea in comparison to PCR and PLS.

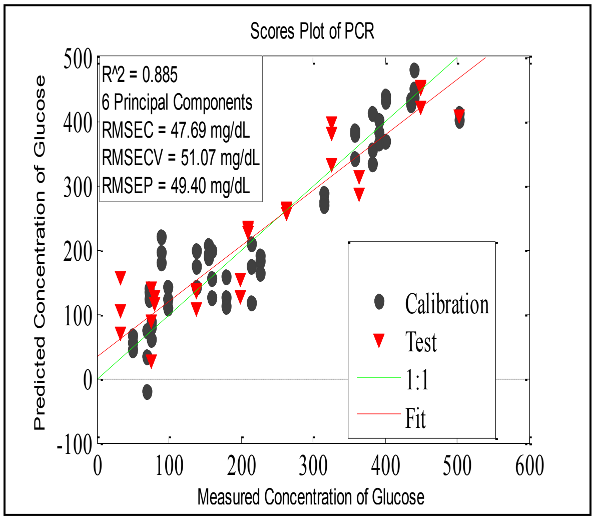

Figure 1. Glucose Prediction Performance using the PCR

Figure 1 above shows the scores plot generated using the PCR. As shown in the plot, the correlation coefficient R^2 is 0.88, RMSEC is 47.69 mg/dL, RMSECV is 49.40 mg/dL and RMSEP is 49.4 mg/dL. From these results it is clear that the PCR is not the best option for predicting the glucose from the dataset. We therefore tried to build the model on this data using the PLS. The scores plot generated using PLS is shown in figure 2 where R^2 , RMSEC, RMSECV and RMSEP were improved to 0.97, 22.54 mg/dL-, 31.59 mg/dL and 27.56 mg/dL respectively. However, as can be seen the RMSEP is still very high.

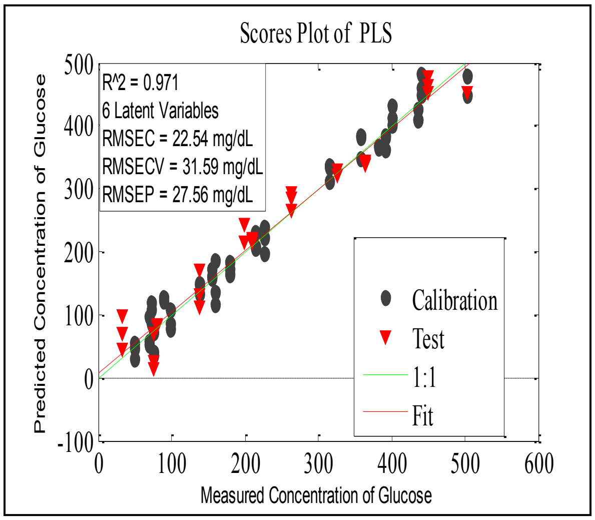

Figure 2. Glucose Prediction Performance using the PLS

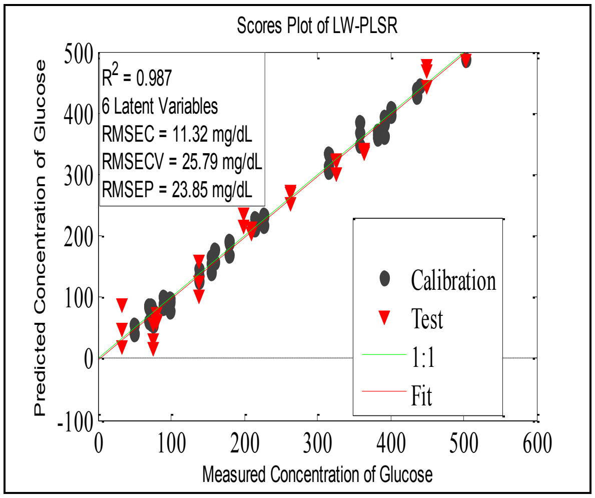

Figure 3. Glucose Prediction Performance using the LW- PLS

Finally, we developed the model using LW-PLS. As shown in figure 3. R^2 , RMSEC, RMSECV and RMSEP have improved to 0.98, 11.32 mg/dL, 25.79 mg/dL and 23.85 mg/dL respectively.

Table 1 below summarises the comparative results of the three calibration models developed. The data was pre-processed using the first derivative in all these models.. As is evident from the table, LWR performed best.

Table 1 comparison between PCR, PLS, LW-PLS

	R^2	**RMSEC (mg/dL)**	**RMSECV (mg/dL)**	**RMSEP (mg/dL)**	**Pre-processing**
PCR	0.88	47.69	51.07	49.40	First Derivative
PLS	0.97	22.54	31.59	27.56	First Derivative
LW-PLS	0.98	11.32	25.79	23.85	First Derivative

V. Conclusion

The LW-PLSR model has been applied to predict the glucose concentration from a mixture composed of triacetin, urea and glucose. The results of the LWR model are compared with the models developed with the PCR and PLS on the same data under the same pre-processing conditions. The LW-PLSR shows better results in terms of R^2 , RMSEC, RMSECV and RMSEP.

The improvement in prediction is encouraging and may lead to the possibility of more people in the area of chemometrics using the LW-PLSR-based approach for regression. In future, performance of the LW-PLSR would be evaluated for extraction of glucose from blood plasma.

References

[1] M. J. Wabomba, G. W. Small, and M. A. Arnold, "Evaluation of selectivity and robustness of near-infrared glucose measurements based on short-scan Fourier transform infrared interferograms," *Analytica Chimica Acta,* vol. 490, pp. 325-340, 2003.

[2] A. A. Al-Mbaideen, T. Rahman, and M. Benaissa, "Determination of glucose concentration from near-infrared spectra using principle component regression coupled with digital bandpass filter," 2010, pp. 243-248.

[3] M. R. Robinson, R. Eaton, D. Haaland, G. Koepp, E. Thomas, B. Stallard, and P. Robinson, "Noninvasive glucose monitoring in diabetic patients: a preliminary evaluation," *Clinical chemistry,* vol. 38, p. 1618, 1992.

[4] J. Chen and X. Wang, "A new approach to near-infrared spectral data analysis using independent component analysis," *Journal of Chemical Information and Computer Sciences,* vol. 41, pp. 992-1001, 2001.

[5] D. M. Haaland, M. R. Robinson, G. W. Koepp, E. V. Thomas, and R. P. Eaton, "Reagentless near-infrared determination of glucose in whole blood using multivariate calibration," *Applied Spectroscopy,* vol. 46, pp. 1575-1578, 1992.

[6] H. Martens and T. Naes, *Multivariate calibration*: John Wiley & Sons Inc, 1992.

[7] T. Næs and H. Martens, "Principal component regression in NIR analysis: viewpoints, background details and selection of components," *Journal of Chemometrics,* vol. 2, pp. 155-167, 1988.

[8] H. Martens, T. Karstang, and T. Næs, "Improved selectivity in spectroscopy by multivariate calibration," *Journal of Chemometrics,* vol. 1, pp. 201-219, 1987.

[9] M. Kano and M. Ogawa, "The state of the art in chemical process control in Japan: Good practice and questionnaire survey," *Journal of Process Control,* vol. 20, pp. 969-982, 2010.

[10] C. G. Atkeson, A. W. Moore, and S. Schaal, "Locally weighted learning," *Artificial intelligence review,* vol. 11, pp. 11-73, 1997.

[11] W. S. Cleveland and S. J. Devlin, "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of the American Statistical Association,* vol. 83, pp. 596-610, 1988.

[12] D. Pérez-Marín, A. Garrido-Varo, and J. E. Guerrero, "Non-linear regression methods in NIRS quantitative analysis," *Talanta,* vol. 72, pp. 28-42, 2007.

[13] S. Kim, M. Kano, H. Nakagawa, and S. Hasebe, "Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection," *International Journal of Pharmaceutics,* vol. 421, pp. 269-274, 2011.

[14] Savitzky, A., Golay, M., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627–1639

[14] B. M. Wise, N. B. Gallagher, R. Bro, and J. M. Shaver, "PLS Toolbox 3.0," *Manson, WA: Eigenvector Research Inc,* vol. 171, 2003.

[15] http://wiki.eigenvector.com/index.php?title=Using_Cross-Validation