Support Vector regression applied on NIR spectroscopic data of agricultural products

J. A. Fernández Pierna[a], B. Lecler[a], J.P. Conzen[b], A. Niemoeller[b], V. Baeten[a] and P. Dardenne[a]*

[a] Walloon Agricultural Research Centre (CRA-W), Quality of Agricultural products Department, Chaussée de Namur 24, 5030; Gembloux, Belgium
[b] BRUKER OPTIK GmbH, NIR & Process Technology, Rudolf-Plank-Str. 27, 76275 Ettlingen, Germany
*E-mail : fernandez@cra.wallonie.be

Introduction

In most of the countries that produce important amounts of food, feed, feed ingredients, fresh silages or soil products, regulations and laws exist about the chemical composition of these products. Normally these regulations lay down some kind of limits (minimum water content, etc) in the final product in order to guarantee that it meets their legitimate expectations and fulfils the good manufacturing practice. However, manufacturers want to produce at minimal costs and try to arrange their formulations in a way, that the chemical composition of the products is approaching those limiting values. In order to do so, the chemical composition of the raw materials must be known. This requires considerable analytical methods, which are expensive and require the use of chemical reactive.

NIRS is the most widely used non-destructive technology in the food and feed industries and official control method to determine different qualitative parameters. The high throughput of the method, the capacity to determine in one single analysis large panoply of parameters and the possibility to build network of spectrometers made this technique very attractive for the food and feed sectors. The combination of NIR spectroscopy and chemometrics is a good alternative to the analytical techniques needed to determine those parameters and criteria.

The objective of this work is to present the chemometric method, **Support Vector Regression** as a good alternative for model construction and prediction of several properties of different agricultural products. The aim is also to compare its performance to different more classical supervised regression methods.

Experimental part

NIR spectral data (measured in the range 1100:2500 nm and a resolution of 2 nm) with reference values of different properties for different products were used:

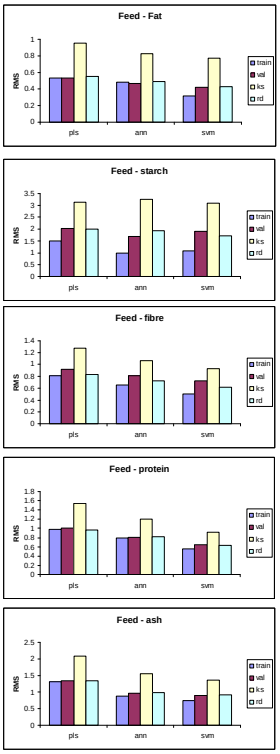
- **Feed**: Ash, Fat, Fibre, Starch, Protein (28676 spectra)
- **Feed Ingredients**: Ash, Fat, Fibre, Protein (26652 spectra)
- **Fresh Silages**: Dry Matter, Fibre, Protein (1035 spectra)
- **Soils**: CEC (Cation Exchange Capacity, COT_SK (Organic Carbon), N_Kj (Nitrogen) (1625 spectra)

The spectral data were pre-processed by the Standard Normal Variate transform followed by detrend and 1st derivative Savitzky-Golay treatment (15,2,1) in order to remove the scattering effects and to smooth the spectra.
A regression model for each data set and for each property is determined, which can then be applied to predict new (unknown) samples. A full analysis including diagnostics, feature reduction/selection, modelling and validation of models has been performed.

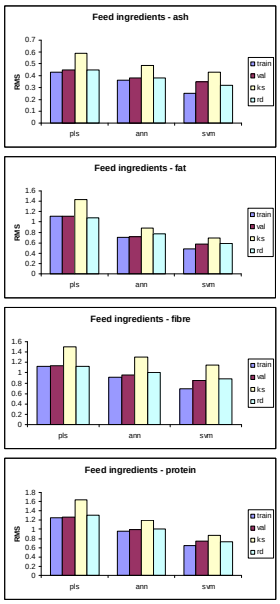
For the application of all the different regression methods, the data set is split into three subsets. The one used for model construction constitute the *training set* (80% of the samples). The models obtained using this training set are then applied to *two test sets* (10% of the samples each). These two test sets have been selected in two different ways: by using the duplex design proposed by Snee[4] (*ks set*) and randomly (*rd set*). Samples selected by the duplex method cover the whole range in the PC space. The random selection contains less extreme samples than the test selected by duplex.

Chemometric models have been constructed using Partial Least Squares, Artificial Neural Networks and Support Vector Regression.

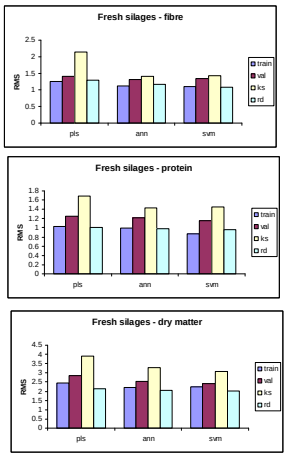
FEED



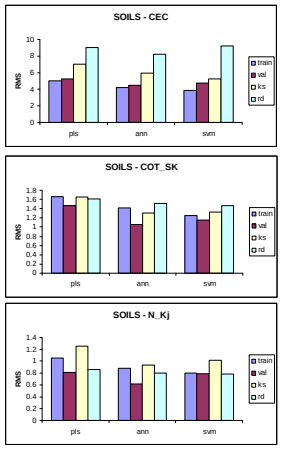
FEED INGREDIENTS



FRESH SILAGES



SOILS



Conclusion

In most cases, PLS gives poor results, but it can be due to the fact that the relationship between the predicted values and the actual concentration is not linear. This can be checked by visual examination of the method response versus the analyte concentration (not shown). This usually works well but it is subjective and open to different interpretations.

SVM gives similar results than ANN but in most of the cases its prediction ability is higher than the other two methods (lower RMS Errors).

SVM gives in general better results not only for clear non-linear cases, but also is performing quite well in the case of linear models.

References

1 'Partial Least-squares Regression: A Tutorial'. P. Geladi and B. R. Kowalski, Analytica Chimica Acta, 185 (1986), 1-17.
2 'Artificial neural networks in multivariate calibration ' T. Naes, K. Kvaal, T. Isaksson and C. Miller, J. Near Infrared Spectrosc. 1 (1993), 1-11.
3 'Least-squares support vector machines for chemometrics: an introduction and evaluation'. R.P. Cogdill and P. Dardenne. Journal of Near Infrared Spectroscopy 12 (1) (2004), 93-100.
4 'Validation of Regression Models: Methods and Example'. R.D. Snee, Technometrics, 19 (1977), 415-428

Walloon Agricultural Research Centre