European Symposium on Computer Aided Process Engineering – 15
L. Puigjaner and A. Espuña (Editors)

Enhanced Modeling of an Industrial Fermentation Process through Data Fusion Techniques

Sophia Triadaphillou, Elaine Martin, Gary Montague[1], Paul Jeffkins[1], Sarah Stimpson[1], Alison Nordon[2]
Centre for Process Analytics and Control Technology
University of Newcastle, Newcastle upon Tyne, NE1 7RU, England
[1]GlaxoSmithKline, Worthing, England
[21]Centre for Process Analytics and Control Technology
University of Strathclyde, Glasgow, G1 1XN Scotland

Abstract
A novel strategy for the analysis and interpretation of spectral data from a fermentation process is considered. The interpretation is challenging as a consequence of the large number of correlated spectral measurements recorded from the process in which a complex series of biochemical reactions occur. A full spectral analysis using PLS is the standard interpretation strategy. However, within this paper an alternative method, Spectral Window Selection (SWS), is proposed, and compared with that of genetic algorithms. SWS is shown to provide a more robust calibration model. Furthermore its performance is hypothesised to be enhanced by multiple model bagging. This claim is investigated and proven. Finally an overall calibration model is compared with a local modelling approach. The methodologies are applied and compared on an industrial NIR data-set from an antibiotic production process.

Keywords: Data Fusion; Modelling; Fermentation Process; Industrial Application

1. Introduction

The large scale manufacture of pharmaceutical products is a highly competitive industry in which technological improvements can maintain fine business margins in the face of competition from those with lower manufacturing overheads. Processes in which pharmaceuticals are produced are particularly susceptible to large variability due to limited monitoring and control options. Previous research has demonstrated that the infrared spectral analysis of fermentation broth can provide on-line measurements of key concentrations throughout the duration of a batch but signal interpretation remains a challenge. Relating the spectra to the analyte of interest requires the construction of a robust calibration model. The traditional strategy is to apply projection to latent structures, PLS (Tosi et. al., 2003) utilising the full spectrum or else implement wavelength selection through genetic algorithms, GAs (Abrahamsson et. al., 2003) for example. An alternative approach is reported in this paper where a search strategy identifies a limited number of spectral windows (SWS) that are most descriptive of the concentrations of interest. The methodology is demonstrated by application to NIR spectral data generated from the routine operation of an industrial antibiotic production process. NIR spectroscopy was used as a result of recent successes in the determination of individual component concentrations in fermentation broth (Tamburini, et. al., 2003).

[1] Author to whom correspondence should be addressed:gary.montague@ncl.ac.uk

2. Wavelength Selection and Model Bagging

When developing a linear model for quantitative analysis of spectral data, prediction results can be affected by wavelengths that do not offer predictive information about the analyte of interest. Also absorbance ranges of different functional groups may overlap and many substances contained in the complex samples may contribute to signals across the complete spectral wavelength range. Wavelength selection is one approach to eliminating wavelengths where descriptive information is not present. Typical wavelength-selection approaches have focused on selecting individual wavelengths using methods such as genetic algorithms.
GAs' are a global search method that mimic biological evolution. GA's apply the principle of survival of the fittest to produce better approximations to the solution. Each member of the population is made up of a binary string which in this case serves to indicate whether a wavelength is selected or not. It is an iterative procedure and at each generation a portion of the population of solutions are selected with consideration of their fitness. The fitness is assessed through an objective function that characterises the performance of an individual member of the population. Once the individuals are chosen from the population, genetic operators are applied and the population is updated to produce the next generation. Further details of the GA methodology can be found in Goldberg (1989). Many significant drawbacks have been reported in the literature (McShane et al 1999): GAs tend to be slow to converge, they present a configuration challenge because of the adjustable factors (e.g. initial population, number of generations) that influence their outcome, and finally they can be biased by including wavelengths with a spurious correlation to the prediction property and the chosen wavelength subset may therefore not be appropriate for predicting future samples.

In this paper, a spectral window selection (SWS) algorithm is proposed where windows of wavelengths are chosen. The algorithm is based on that described in Hinchliffe et al. (2003). By constraining the spectra selection to a limited number of windows rather than allowing multiple individual wavelengths to be selected potentially improves the calibration model performance by preventing it becoming too specific to the training information. The steps in the algorithm are summarised in Figure 1. 'Bagging' has been proposed in the literature to improve the accuracy of models. It has proven to be successful mainly in classification and pattern recognition problems. It was originally introduced by Breiman (1996). The bagged prediction is generated as a weighted combination of the predictions from n individual models. Two methods were investigated for the calculation of the weights, mean and PLS for both the results from the GAs and SWS. For average bagging, each individual model is equally weighted and the mean of the predictions for each time point is calculated. For PLS bagging, linear PLS is used to attain a weighted average.

3. Data and Spectral Measurements

The data considered is from an industrial pilot-plant scale fermentation process, which involves two stages, the seed and final stage. Biomass is grown in the seed stage and is then transferred to the final stage for the production of the desired product. The final stage is a fed batch process and lasts approximately 140 hours. Seven fermentations were carried out and NIR measurements were collected on-line from the final stage of the process. Product concentration in the broth was considered to be critical to the monitoring of the batch and in this paper is the analyte of interest. A further approach investigated to improve model robustness was local modelling as suggested by Arnold et. al (2001) to achieve an improvement in overall performance compared with a global

model. The local modelling approach was considered for this data set, with three regions of operation identified using mechanistic process knowledge.

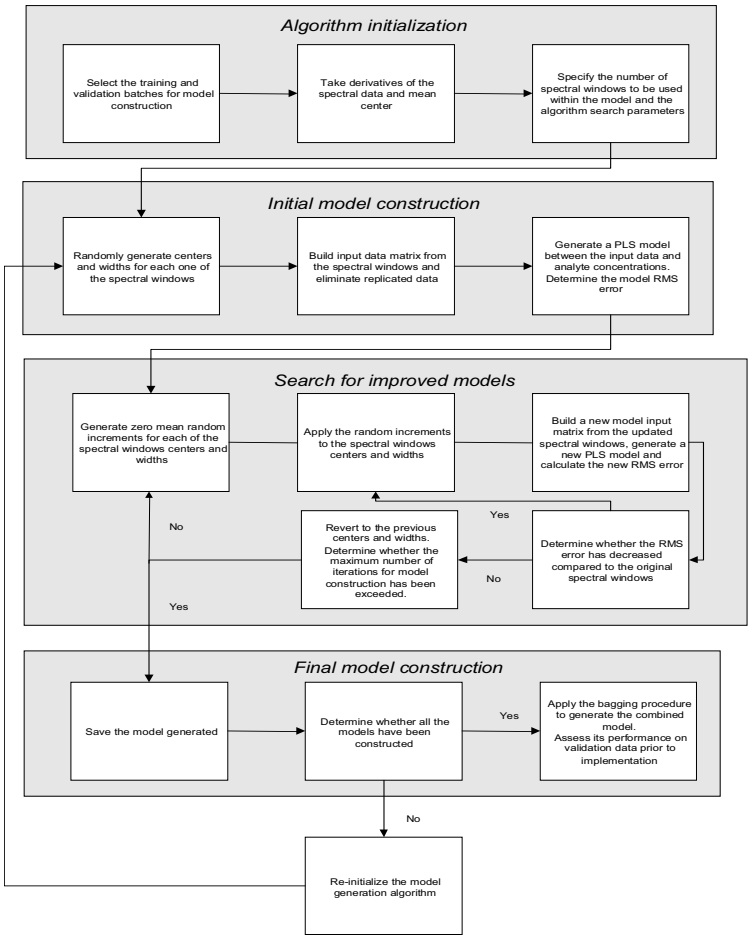

Figure 1. Flow diagram of the spectral window selection (SWS) algorithm

4. Application of the Methodology

Calibration model training was carried out utilising two thirds of the data with the remaining one-third being used to test the models. Several pre-treatments steps were required. The raw spectroscopic data for a batch can be seen in Figure 2. For this application, first derivatives were taken (Figure 3). The derivatives were calculated using Savitsky-Golay smoothing (Gory, 1990) for 11 points and a second order polynomial.

For the fitness function of the GA algorithm, the RMS prediction error was used in order to be consistent with the SWS method. The reproduction was performed with a single point crossover with probability 0.7 followed by mutation. A population of 100 individuals was used. 100 generations were performed with a generation gap of 0.5.

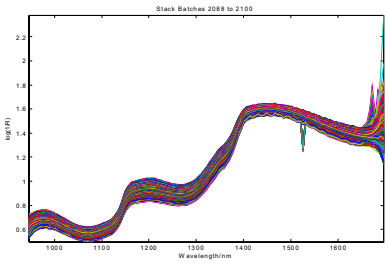

Figure 2: Raw NIR spectra

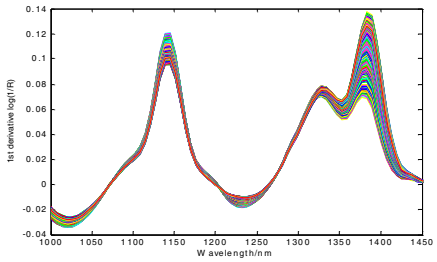

Figure 3: First derivative spectra

5. Results and Discussion

The results from the various approaches for the prediction of the product concentration are summarised in Table 1 and Table 2. Table 1 summarises the RMS errors for global modelling both without SWS, i.e. applying PLS to the complete spectra, and through the application of SWS and Average and PLS bagging. The results indicate that if a global modelling approach is implemented there are some benefits to be gained from the wavelength selection algorithm when average bagging is applied. PLS bagging tends to produce better fits to the training data but the models become too specific and consequently does not perform as well on the test data set.

Table 1. Results for global modelling of the product concentration

	Without SWS	With SWS	
		Average Bag	PLS Bag
Linear PLS for Training	0.056	0.068	0.045
Linear PLS for Testing	0.092	0.059	0.096

Table 2 reports the performance of the local models constructed following the application of SWS and GAs to the test data set The training data results are not presented due to space limitation. The enhanced performance of the local model approach in terms of the RMS error is a consequence of limiting the range over which the calibration model is constructed. The local model strategy (SWS with average bagging) outperforms both global modelling and strategy involving GAs followed by PLS in all but the first time interval where the results are comparable.

Table 2. RMS for the quality variable from the NIR spectra for the testing data set

	Time Interval 1		Time Interval 2		Time Interval 3	
	PLS Bag	Average Bag	PLS Bag	Average Bag	PLS Bag	AverageBag
SWS with Linear PLS	0.045	0.049	0.059	0.048	0.095	0.058
GAs with Linear PLS	0.045	0.043	0.067	0.069	0.177	0.139
PLS (without wavelength selection)	0.042		0.059		0.084	

The importance of bagging can be observed in Figure 4 where the results of the thirty individual model errors are presented. Most notably, the RMS error for the bagged model (presented in Table 1) is lower than that for the individual model errors justifying the bagging strategy.

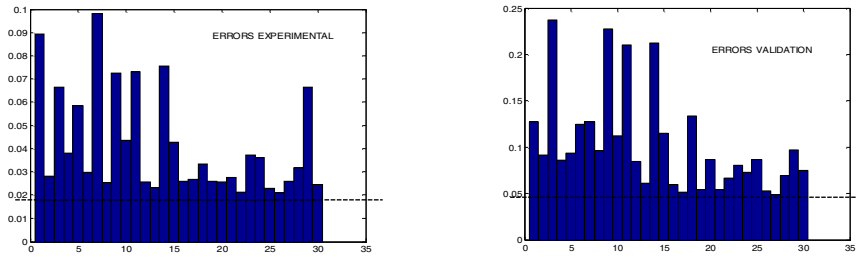

Figure 4. Errors for the 30 models for the first time interval for the standard batches for the experimental and the testing data set, ----- RMS error after PLS Bagging

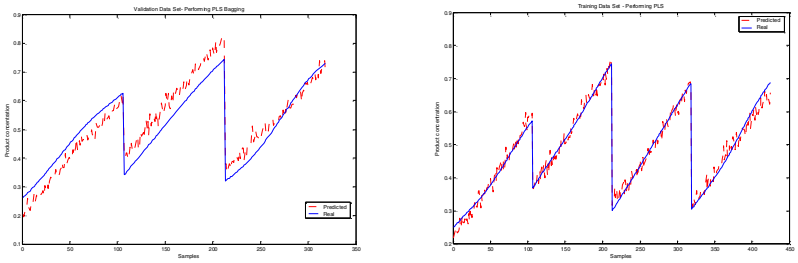

Figure 5a. Experimental results for 4 batches.

Figure 5b. Testing data results for 3 batches.

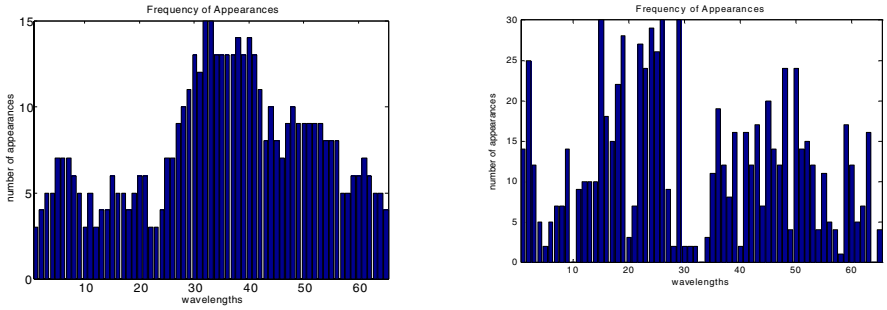

Figure 6: Frequency distribution of the wavelengths selected by the SWS (left) and GA (right).

As an example of typical model behaviour in Table 2, Figure 5 shows the results from Time Interval 1 from a SWS PLS bagged model. Thirty models were used to generate the final model. Multiple batches are concatenated and the large falls in product concentrations are the breaks between batches. It can be observed that an off-set exists which is most significant for the second validation batch. The issue of offset removal is discussed in the conclusions. Figure 6 shows the frequency distribution of the wavelengths selected by the SWS algorithm and the GA. The wavelengths in the region 30 to 40 were selected the most frequently. This range aligns closely with those identified by analytical chemistry specialists. The GA did not indicate any special critical regions and the important wavelengths were not selected preferentially.

6. Conclusions and Future Work

NIR spectroscopic techniques can potentially be used to measure and predict within a few minutes a number of critical components concentrations in an unprocessed

fermentation broth sample. This paper has demonstrated that the selection of informative spectral regions can improve the results by reducing the contribution of overall noise from those regions not containing relevant information on the analyte concentrations. In this paper a new method, SWS, in combination with bagging has been developed and compared with the traditional approach of GAs. A local modelling strategy was used to improve the accuracy of the prediction. Offsets in predictions based solely on spectral analysis still occur, particularly if the calibration region is large. The inclusion of other process measurements in a hybrid calibration model structure can potentially deliver more robust models and reduce the offsets. For example, the residuals of the calibration model can be related to other process information to 'explain' the model deviations and 'corrections' to spectral prediction can be made. This is an ongoing area of research (Triadaphillou et al., 2004).

References

Abrahamsson C., Johansson, J, Sparén A. and Lindgren, F. 2003. Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets, Chemometrics and Intelligent Laboratory Systems, 69, 1-2, 28: 3.

Arnold, A.S., Matheson, L, Harvey, L M., McNeil, B. 2001. Temporally segmented modelling: a route to improved bioprocess monitoring using near infrared spectroscopy?, Biotechnology Letters, 23: 143.

Breiman L. 1996. Bagging Predictors. Machine Learning Journal. 24(2):123-140.

Goldberg D.E. 1989. Genetic algorithms in search, optimization and machine learning, Addison Wesley.

Gory, P.A. 1990. General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method. Anal. Chem. 62:570.

Hinchliffe M, Montague GA, Willis M, Burke A. 2003. Correlating polymer resin and end-use properties to molecular-weight distribution. AIChE Journal. 49:2609.

Kornmann, H, Rhiel, M, Cannizzaro, C, Marison, I, von Stockar, U. 2003. Methodology for real-time, multianalyte monitoring of fermentations using an in-situ mid-infrared sensor. Biotechnology and Bioengineering, 82 (6): 702.

McShane M.J., Cameron, B.D., Cote, G.L., Motamedi, M., Spiegelman, C.H. 1999. A novel peak-hopping stepwise feature selection method with application to Raman spectroscopy. Analytica Chimica Acta 388:S251.

Tamburini E, Vaccari G, Tosi S, Trilli A. 2003. Near-Infrared Spectroscopy: A tool for monitoring submerged fermentation processes using an immersion optical-fiber probe. Applied Spectroscopy. 57(2).

Tosi, S., Rossi, M., Tamburini, E., Vaccari, G., Amaretti, A., Matteuzzi, D. 2003. Assessment of In-Line Near-Infrared Spectroscopy for Continuous Monitoring of Fermentation Processes', Biotechnology Progress, 19(6):1816.

Triadaphillou, S., Martin, E., Montague, G., Jeffkins, P., Stimpson, S., Nordon, A. 2004. Monitoring of a fermentation process through on-line spectroscopic data and the conjunction of spectroscopic and process data. Presented in BatchPro Symposium, Poros, Greece.

Acknowledgements

ST would like to acknowledge the EPSRC award, KNOWHOW and the EU project BATCHPRO for financial support. CPACT acknowledges the vendor company Clairet Scientific for the loan of spectroscopic instrumentation.