CHEMICAL ENGINEERING TRANSACTIONS
VOL. 32, 2013
A publication of AIDIC
The Italian Association of Chemical Engineering
www.aidic.it/cet
Chief Editors: Sauro Pierucci, Jiří J. Klemeš

ISBN 978-88-95608-23-5; **ISSN** 1974-9791
DOI: 10.3303/CET1332207

Development of a Modelling Framework for NIR Spectroscopy Based on-line Analyzers using Dimensional Reduction Techniques and Genetic Programming

Tibor Kulcsar, Janos Abonyi*

Department of Process Engineering, University of Pannonia, P.O. Box 158, H-8201 Hungary
janos@abonyilab.com

The quality of fuels is determined by several properties (e.g. aromatic components, cloud point, flash point, density etc.). Measurements of these properties are mostly costly and time consuming that makes real-time control infeasible. An affordable option to perform real-time quality control is the application of Near Infra-Red (NIR) spectroscopy based on-line analyzers. Using this tool multiple measurements can be substituted by one measurement combined with a complex prediction model. The two dimensional mapping of the spectral space can be used for monitoring the operation of the production and for the validation of the analyzer. The mapping is based on two aggregates which are mathematical functions combining absorbance values at several wavelengths. We developed a genetic programming based approach to design these aggregates. Results related to the monitoring of diesel fuel blending process illustrate the applicability of the method.

1. Introduction

Sensor development is important task in chemical engineering (Zaouak, 2012). NIR is finding widespread use in the process industries. The technique can be applied to liquids, waxes, and solids without sample dilution. The NIR spectrum ranges from 800 to 2,500 nm and contains information about molecular structures. To estimate chemical properties based on the combinations of absorbance values of NIR spectra complex prediction models have to be developed. The most common analysis method is correlation of NIR absorbance values with physical and chemical properties through appropriate statistical treatments. This approach makes possible the measurement of several properties simultaneously from a single spectrum (Espinosa, 1994). There are several multivariate models and methods to support the prediction of product properties based on NIR spectra. There are parametric models (e.g. linear regression, multi-linear regression, Partial Least Squares regression - PLS) and nonparametric methods (e.g. k-NN, Neural Networks, Topological Near-Infrared Modelling - TOPNIR).
TOPNIR is widely used in the oil industry to estimate product properties (e.g. aromatic components, cloud point, flash point, density etc.) of products and process streams. TOPNIR performs a two dimensional mapping of the spectral space to visualize the operation regimes of the process. The aggregates are equations that combine absorbances measured at significant wavelengths. In ideal case aggregates reflect product properties. Since these properties can be dependent on different rages of the spectra each aggregate built up several wavelengths to contain enough information related to a certain chemical property (Descales, 2000).
Finding the proper model structure is a complex nonlinear optimization problem (Jain, 1997).. We present a Genetic Programming (GP) based algorithm to generate nonlinear aggregates Narendra, 1977). GP is a tree representation based symbolic optimization technique , Pudil, 1994). This representation is extremely flexible; trees can represent computer programs, mathematical equations or complete models of process systems (Madar, 2005).

Please cite this article as: Kulcsar T., Abonyi J., 2013, Development of a modeling framework for nir spectroscopy based on-line analyzers using dimensional reduction techniques and genetic programming, Chemical Engineering Transactions, 32, 1237-1242 DOI: 10.3303/CET1332207

GP is already applied for visualization of complex process data (Chemaly, 2001). In this work GP is used to find simple nonlinear functions by minimizing the distance preservation property of the mapping. The drawback of this approach is that since functions were not parameterized only simple mappings with approximate distance preserving properties were generated.

Since NIR spectra requires accurate and complex mapping we developed a much more sophisticated approach. The functions generated by GP are parameterized and a nonlinear parameter optimization step is embedded into the GP. Furthermore, instead of distance preserving measures the cost function is based on the neighbourhood preserving properties of the mapping since this measure is much closer reflects the goal of the application. The applicability of the proposed approach is confirmed through an industrial case study related to the product property estimation of a diesel fuel blending process.

The structure of the paper is the following. The basics of visualization of high dimensional data are introduced in Section 2. Section 3 details the proposed Genetic Programming based algorithm. The application example is given in Section 4. Finally, Section 5 draws some conclusions.

2. Topological Mapping for Visualization of High Dimensional Data

The goal of dimensionality reduction is to map a set of observations from a high-dimensional space (D) into a low-dimensional space $(d, d \ll D)$preserving as much of the intrinsic structure of the data as possible. Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$ be a set of the observed data, where $\boldsymbol{x}_i$ denotes the i-th observation $\left(x_i = \left[x_{i,1}, x_{i,2}, \dots, x_{i,D}\right]^T\right)$. Each data object is characterized by D dimensions, so $x_{i,j}$ yields the j-th $(j = 1,2, \dots, D)$ attribute of the i-th $(i = 1,2, \dots, N)$ data object. Dimensionality reduction techniques transform data set $\boldsymbol{X}$ into a new data set $\boldsymbol{Y}$ with dimensionality d $\left(\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_N\},\ y_i = \left[y_{i,1}, y_{i,2}, \dots, y_{i,d}\right]^T\right)$.

In the reduced space many data analysis tasks (e.g. classification, clustering, and image recognition) can be carried out faster than in the original data space.

As dimensional reduction methods are based on the preservation of the dissimilarities and/or the neighborhood relation of the objects, the numeral evaluation of the mappings aims to measure the realization of these principles. The neighborhood preservation of the mappings and the local and global mapping qualities can be measured by functions of trustworthiness and continuity. Kaski pointed out that every visualization method has to make a tradeoff between gaining good trustworthiness and preserving the continuity of the mapping (Kaski, 2003).

A projection is said to be trustworthy when the nearest neighbors of a point in the reduced space are also close in the original vector space. Let N be the number of the objects to be mapped, $U_k(i)$ be the set of points that are in the k size neighborhood of the sample i in the visualization display but not in the original data space. The measure of trustworthiness of visualization can be calculated in the following way:

$$M_1(k) = 1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^{N} \sum_{j \in U_k(i)} (r(i,j) - k) \tag{1}$$

where $r(i,j)$ denotes the ranking of the objects in input space. The projection onto a lower dimensional output space is said to be continuous when points near to each other in the original space are also nearby in the output space.

The measure of continuity of visualization is calculated by the following equation:

$$M_2(k) = 1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^{N} \sum_{j \in U_k(i)} (s(i,j) - k) \tag{2}$$

where $s(i,j)$ is the rank of the data sample i from $j$$in the output space, and $V_i(k)$ denotes the set of those data points that belong to the k-neighbors of data sample i in the original space, but not in the mapped space used for visualization. Both trustworthiness and continuity functions are function of the number of neighbors k. Usually, the qualitative measures of trustworthiness and continuity are calculated for $k = 1,2, \dots, k_{max}$, where k_{max} denotes the maximum number of the objects to be taken into account. At small values of parameter k the local reconstruction performance of the model can be tested, while at larger values of parameter k the global reconstruction is measured.

3. Visualization by use of Genetic Programming

Industrial applications require easily implementable, interpretable and accurate projections. Nonlinear functions (often referred as aggregates) are useful for this purpose. A pair of these functions realizes feature selection and transformation. Such mapping is used for the visualization and indexing of spectroscopic databases in the Topological Mapping using Aggregates (TOPNIR) modelling framework.
The two main forms of the aggregates are shown by equation (3).

$$y_1 = a_{1,0}\frac{a_{1,1}x_{1,1} * a_{1,2}x_{1,2}}{a_{1,3}x_{1,3} * a_{1,4}x_{1,4}}\ y_2 = a_{2,0}\frac{a_{2,1}x_{2,1} + a_{2,2}x_{2,2}}{a_{2,3}x_{2,3} * a_{2,4}x_{2,4}} \quad (3)$$

Finding optimal set of features $\{x_{i,j}\}$, the optimal model structure, and the optimal values of the parameters of these functions is a complex nonlinear optimisation problem. Our key idea is to represent these functions by trees and apply genetic programming for finding the optimal model structure. As it is shown in Fig 1. a population member in GP is a hierarchically structured tree representing a function as a set of operators and terminals. For example, the set of operators F can contain basic arithmetic operations: $F = \{+, -, *, /\}$; however, it may also include Boolean operators and conditional operators. In this work we only used arithmetic operations. The set of terminals T contains the arguments for the functions. For example $T = \{x_1, \dots x_n, p_j\}$ with x_i represents the elements of possible input variables and p_j represents the parameters. Now, a potential solution may be depicted as a rooted, labelled tree with ordered branches, using operations (internal nodes of the tree) from the function set and arguments (terminal nodes of the tree) from the terminal set.
Genetic Programming is an evolutionary algorithm. It works with a set of individuals (potential solutions), and these individuals form a generation. In every iteration the algorithm evaluates the individuals and selects the bets ones for reproduction according to their fitness values, generates new individuals by mutation, crossover and direct reproduction, and finally creates the new generation. The fitness function reflects the goodness of a potential solution which is proportional to the probability of the selection of the individual. In the current application the fitness function is based on the topology preserving property of the mapping:

$$fitness = M_1 M_2 = \frac{1}{N}\sum_{k=1}^{N} M_1(k) \sum_{k=1}^{N} M_2(k) \quad (4)$$

Parameters of the functions (aggregates) have huge impact to the mapping performance. The evaluations of the fitness of the models are performed at optimal parameter values. Therefore a nonlinear parameter optimization step is embedded into the GP. After GP generated the new population of model structures Sequential Quadratic Programming (SQP) calculates the optimal values of the parameters of these models.
The proposed approach has been implemented in MATLAB. The user should only define the high dimensional data that should be mapped, one aggregate function which optimal pair should be found by the optimization, and the set of the terminal nodes (the set of the variables of the model and set of the internal nodes - mathematical operators.

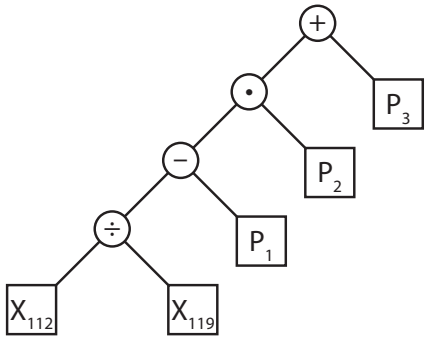

Figure 1: Binary tree representation of an aggregate equation. The terminal elements are the absorbance values and the constant parameters, the nodes are mathematical operands. Two aggregates are used in the same time to produce a 2D mapping of spectral space.

4. Visualization of a spectral database of a diesel fuel blending process

Increasing environmental restrictions and business competition have forced the refiners to look new directions to maintain their margins and competitiveness. "Fixed cost reduction" and "site modernization" are now a common language in the refineries. A key financial area in the refinery to look for savings is the fuels blending stage. Significant opportunities exist in modernizing the blending related controls and decision process (Espinosa, 1994).

Typical weaknesses of the blending process are the inaccuracy of blending indices and the difficulty in controlling set points for component flow rates. A discrepancy can occur if the usual proportion of base stocks is changed, or if a new base stock is introduced. Because of these sources of error, re- blending frequently is required to either bring the product within specifications or reduce giveaway.

BP introduced a NIR online analyzer to improve the blending process (Espinosa, 1994). The main goals for the NIR-based blending system were: (1) Minimization of quality giveaway, (2) Optimization of blend recipes, (3) Increase in blender flexibility by avoiding reblends (4) Reduction of future storage-capacity requirements for blended products. By minimizing property give away and re-blend rates, the benefits of such solution have been calculated up to 2 MUSD per year for a Gasoline or Diesel Blender for a medium size refinery (~200 000 bbl/d).

In the Danube Refinery of MOL three NIR applications (Gasoline Blending, Diesel blending and Reformer) are currently used. Based on earlier experiences, recommendations and advises from the vendor of TOPNIR (ABB) it was concluded that the success of an NIR project is largely dependent on:

- The accuracy and robustness of the modelling method,
- The stability of the spectrometer,
- The quality of the reference data on which the chemometrics models are built,
- The involvement of the End User in the project.

ABB NIR Hardware combined with TOPNIR modelling insures high accuracy and robustness of on-line predictions of the Hydrocarbon streams. A full suite of estimated properties for Diesel, Gasoline, Aviation fuel is available on-line every 2 min.

Due to the process changes (e.g. blending recipes, blending components quality) a continuous monitoring, analysis and model update is required to insure significant and continuous benefits. The huge amount of data generated by the applications can be used for continuous model development, e.g. for updating of blending equations in blending software and devices, technological feedback to producing units, reporting to planning and scheduling and support product development.

In this work we focus on the visualization of these spectral databases that can be effectively used for model development and monitoring. We used the proposed topology preserving mapping based cost function to select the best pairs of aggregates that reflects the best of the hidden structure of the spectral database of the Diesel oil blending process at MOL Ltd. Duna Refinery. There are 14 aggregates defined in the TOPWIN software used as a framework of the TOPNIR algorithm. Simultaneously two aggregates are used to give a two dimensional mapping of the spectral space. According to the proposed cost function the best pairs of aggregates are Naro and Parox represented by Eq (5) and Eq (6).

$$y_{Parox} = \left(\frac{x_{84}}{20 * x_{15} + x_{112}} - 0.0686\right) * 550 - 12.22 \quad (5)$$

$$y_{Naro} = \left(\frac{x_{112}}{x_{119}} - 1.2462\right) * 130 + 55 \quad (6)$$

where, x_i means the absorbance value at the i-th wavenumber in the range $(4776 - 4000)\left[\frac{1}{cm}\right]$.

As can be seen only four absorbance values among the 195 $x_1 \ldots x_{195}$ are used by these aggregates. We tried to increase the best 0.913 performance of this mapping by utilizing the proposed genetic programming methodology. The parameters of the algorithm are given in Table 1, while the results are summarized in Table 2.

The application of GP resulted the following two equations and mapping shown in Fig. 2:

$$y_{Genetic\,1} = 0.125 * x_{102} - 0.25 * x_{119} + 0.375 * x_{152} + 0.5 * x_{147} + 0.625 * x_{108} \quad (7)$$

$$y_{Genetic\,2} = 0.75 * \frac{(x_{34} - x_{191})}{x_{83}} - (0.875 * x_{61} - x_{83}) \quad (8)$$

Table 1: Parameters of GP in the application examples.

Parameter	Value
Population size	3. 50
4. Maximum number of evaluated individuals	5. 2,500
6. Type of selection	7. roulette-wheel
8. Type of mutation	9. point-mutation
10. Type of crossover	11. one-point (2 parents)
12. Type of replacement	13. elitist
14. Generation gap	15. 0-667
16. Probability of crossover	17. 0.5
18. Probability of mutation	19. 0.5
20. Probability of changing terminal - non-terminal nodes (vica versa) during mutation	21. 0.25

Table 2: Comparison of different mappings. The proposed GP based mapping gives excellent projection performance.

Mapping	Mapping Quality
Best aggregate pair from the original set	22. 0.91326
23. Simple genetic algorithm (breeding a pair for an existing aggregate)	24. 0.93954
25. Parallel genetic algorithm (breeding of two aggregate in the same time)	26. 0.94687

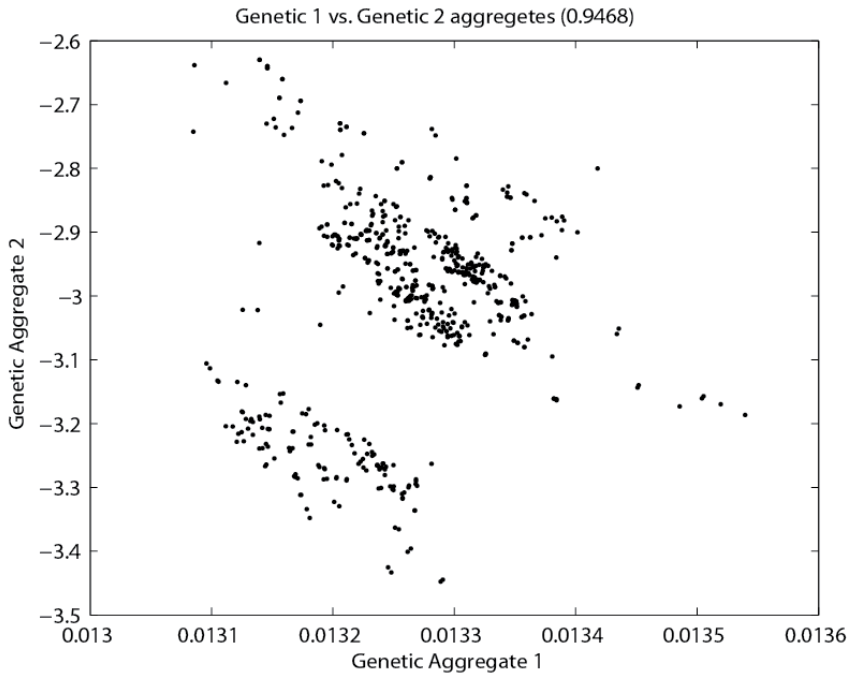

Figure 2: The results of simultaneously optimized aggregates produced by two independent populations after 50 generation with elitist strategy. The database contains samples from two different operating modes (summer and winter diesel) and the proposed mapping is able to separate these operating regimes. The resulted plot is informative for the monitoring of the technology.

5. Conclusion

In case of complex production systems the control of measured process values (e.g. temperature, pressure, flow rate) does not always ensure that unmeasured product properties will be in predefined ranges of production orders or standards. E.g. in oil-industry cetane index and sulphur content are not measured online and the frequencies of flash point, density, cold filter plugging point measurement are not enough for real time control. The objective of the development of software sensors and online analyzers is to support process control and monitoring by providing on-line information about these properties. Soft (software) sensors are especially useful in data fusion where measurements of different characteristics and dynamics are combined. The interaction of signals like temperatures, pressures - in our case absorption intensities - can be used for calculating new unmeasured quantities (like flash point, density etc.). These models can also be used for fault diagnosis and inferential control applications.
A widely used on-line measurement technique is near infrared spectroscopy. Topological modelling techniques are based on looking for similar spectra from a spectral database by nearest neighbourhood algorithms.
Development of these models cannot be a fully automatised process, human supervision and intervention is always needed. In practical data mining and process monitoring problems high-dimensional data has to be analyzed. In most of the cases it is very informative to map and visualize the hidden structure of complex data in a low-dimensional space. We presented a Genetic Programming (GP) based algorithm to generate nonlinear functions can be used for feature selection and transformation and applied to build an on-line spectroscopy based process monitoring system. We defined a novel cost function based on the topology preserving property of the mapping. The resulted tool was applied to design new aggregates for the TOPNIR modelling framework. The applicability of the proposed approach is confirmed through an industrial case study related to the product property estimation of a diesel fuel blending process.

Acknowledgement

This work was supported by the European Union and financed by the European Social Fund in the frame the TAMOP-4.2.2/B-10/1-2010-0025 and TAMOP-4.2.2/A-11/1/KONV-2012-0071 project. The work of Tibor Kucsar was also supported by the frames of TÁMOP 4.2.4. A/2-11-1-2012-0001 „National Excellence Program – Elaborating and operating an inland student and researcher personal support system” The project was subsidized by the European Union and co-financed by the European Social Fund.

Reference

Chemaly T. P., Aldrich C., Visualization of process data by use of evolutionary computation. Computers and Chemical Engineering, 25(9-10):1341-1349, 2001.
Descales B., Lambert D., Llinas J.R., Martens A., Osta S., Sanchez M., Bages S., Method for determining properties using near infra-red (NIR) spectroscopy, 2000. US6.070.128.
Espinosa, A., M. Sanches, On-line NIR analysis and advanced control improve gasoline blending, Oil and Gas Journal, 1994
Jain A.,Zongker D., Feature selection: evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 192:153-158, 1997.
Kaski S., Nikkila J., Oja M., Venna J., Tronen J., Castren E., Trustworthiness and metrics in visualizing similarity of gene expression. BMC Bioinformatics, 4(48), 2003.
Madar J., Abonyi J., Szeifert F., Genetic programming for the identification of nonlinear input-output models. Industrial and Engineering Chemistry Research, 44(9):3178-3186, 2005.
Narendra P., Fukunaga K., A branch and bound algorithm for feature subset selection. IEEE Transactions on Computers, C-269:917-922, 1977.
Pudil P., Novovicova J., Kittler J., Floating search methods in feature selection. Pattern Recognition Letters, 15(1):1119-1125, 1994.
Zaouak O, Daoud B. Fages M, Fanlo J. Aubert, B., High performance cost effective miniature sensor for continuous network monitoring of h2s, Chemical Engineering Transactions, Volume 30, 2012, Pages 325-330, 2012