Journal of Process Control 22 (2012) 1122–1126

Contents lists available at SciVerse ScienceDirect

Journal of Process Control

journal homepage: www.elsevier.com/locate/jprocont

Inferential estimation of kerosene dry point in refineries with varying crudes

Chang Zhou[a], Qiyue Liu[a], Dexian Huang[a,c], Jie Zhang[b,*]

[a] Department of Automation, Tsinghua University, Beijing 100084, China
[b] School of Chemical Engineering and Advanced Materials, Newcastle University, Newcastle upon Tyne NE1 7RU, UK
[c] TNList, Beijing 100084, China

ARTICLE INFO

Article history:
Received 26 October 2010
Received in revised form 10 February 2012
Accepted 31 March 2012
Available online 23 April 2012

Keywords:
Bootstrap aggregated model
Classifier
Data-driven models
Inferential estimation
Crude distillation

ABSTRACT

A bootstrap aggregated model approach to the estimation of product quality in refineries with varying crudes is proposed in this paper. The varying crudes cause the relationship between process variables and product quality variables to change, which makes product quality estimation by soft-sensors a difficult problem. The essential idea in this paper is to build an inferential estimation model for each type of feed oil and use an on-line feed oil classifier to determine the feed oil type. Bootstrap aggregated neural networks are used in developing the on-line feed oil classifier and a bootstrap aggregated partial least square regression model is developed for each data group corresponding to each type of feed crude oil. The amount of training data in crude oil distillation is usually small and this brings difficulties for classification and estimation modelling. In order to enhance model reliability and robustness, bootstrap aggregated models are developed. The inferential estimation results of kerosene dry point on both simulated data and industrial data show that the proposed method can significantly improve the overall inferential estimation performance.

1. Introduction

Crude oil distillation is a primary process in petro-chemical industry and its operation determines the resource usage efficiency and economic benefits of refineries. In order to properly control refinery operations, it is essential that product quality measurements are available. Since most of the quality indexes can hardly be measured in real-time, various soft-sensor methods have been proposed to estimate these indexes using measurable process variables and have been successfully applied in practice [1–4]. As the mechanism of the relationship between the quality variables and the process variables is too complicated and not possible to know comprehensively, empirical models are often used in soft-sensor methods and have good performance in many applications [5–7].

However, soft-sensing in crude oil distillation with varying feedstock remains a difficult problem because the relationship between the easy measured process variables and the difficultly measured quality variables varies with the types of crude oil processed. The types of crude oil change with suppliers. Even the crude oil from the same supplier may also vary in the hydrocarbon content. Furthermore, many refineries operate with mixed sources of crude oil with varying blending ratios. For different crude oil, the relationship between process variables and quality variables is generally different.

One natural idea is to develop an inferential estimation model for each type of crude oil (each supplier or oil filed). But this is unpractical as crude oil from the same supplier or oil field may vary in the hydrocarbon content. A more practical solution is to build models, respectively, according to the type of refinery feed oil (not the crude oil that composes the feed oil). The blending ratio and the type of crude oil are often known and furthermore lab analysis of feed oil composition can be occasionally carried out, so it is possible to obtain data for different types of refinery feed oil.

This paper proposes a multiple model based inferential estimation system integrated with on-line refinery feed oil classification. An inferential estimation model is developed for each type of refinery feed oil. A classifier is used to classify on-line data and determine the type of feed oil currently used and then the corresponding model is chosen to estimate the quality variables. Using on-line measurements of process variables, the feed crude oil is classified into different types, and historical process operation data are also classified into these groups. A bootstrap aggregated partial least square (PLS) regression model is developed for each data group corresponding to each type of feed crude oil. Each model has a favourable predictive capability upon the same type of oil but low predictive accuracy upon other types. During on-line operation, the feed crude oil type is first estimated from the classifier using on-line process measurements and then the corresponding PLS model is invoked.

[*] Corresponding author. Tel.: +44 191 2227240; fax: +44 191 2225292.
E-mail address: jie.zhang@newcastle.ac.uk (J. Zhang).

0959-1524/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jprocont.2012.03.011

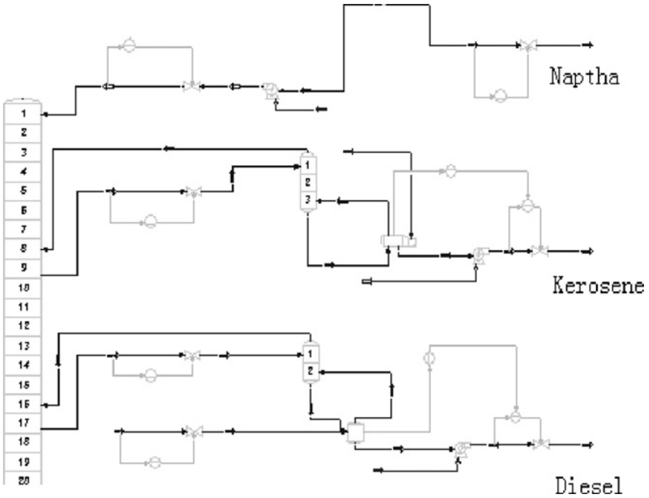

Fig. 1. A schematic diagram of an atmospheric distillation column.

In the crude oil classifier development, bootstrap aggregated neural networks are used. The inputs to the crude oil classifier are the ratios between product and feed rates. The reason to use ratios instead of product flow rates is that the flow rate of feed crude oil changes during refinery operations, so the ratios between products and feed can reflect the oil type better. Since the relationship between the classifier inputs and output is nonlinear, a nonlinear model has to be developed using process operational data. Through bootstrap re-sampling of the training datasets, multiple neural network models are developed based on bootstrap re-sampled datasets. A bootstrap aggregated neural network shows better accuracy and generalization capability than a single neural network which can be trapped in a local minimum or over-fit the training data during network training. The overall inferential estimation performance of the bootstrap aggregated PLS estimator integrated with feed crude oil classifier gives much better performance than various single PLS estimators.

The paper is organized as follows. Section 2 presents a simulated atmospheric distillation column in a refinery with varying feed crude oil. On-line crude oil type classification is presented in Section 3. Section 4 presents inferential estimation of kerosene dry point using bootstrap aggregated PLS models integrated with on-line crude oil classification. Section 5 shows how this method performs on real industrial data. Some concluding remarks are given in Section 6.

2. A simulated atmospheric distillation column in a refinery with varying crude oil feed

The techniques developed in this paper are first tested on a simulated refinery with varying feed crude oil. The simulation is carried out in the HYSYS environment. Fig. 1 shows a schematic diagram of an atmospheric distillation column which is one of the major units used in refineries. Crude oil contains many different types of hydrocarbon molecules and it is infeasible to simulate crude oil composition precisely. A set of hypothetical components are available in HYSYS. A type of crude oil can be approximated using a weighted linear combination of these hypothetical components. By changing the combination weights, different types of crude oil can be simulated. In this study, three kinds of crude oil are used as varying feed: with more light components, with more middle components, and with more heavy components, each of which is simulated by setting different assay values to make different hypothetical components that compose the crude oil. A MATLAB programme is used to change the operation condition in a random way to approximate the real industrial process and store the simulated process data automatically. During simulation, the product flows are carefully set to ensure that product quality constraints are met.

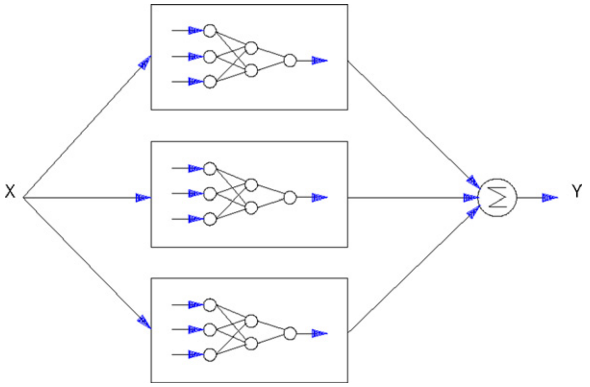

Fig. 2. A bootstrap aggregated neural network.

3. Crude oil classification using bootstrap aggregated neural networks

3.1. Bootstrap aggregated neural networks

Neural networks have been shown to be capable of approximating any continuous nonlinear functions [8] and have been applied to nonlinear process modelling [9,10]. A key requirement for the successful practical implementation of neural network models is good model reliability and robustness. Several techniques have been developed to improve neural network generalization capability, such as regularization [11], early stopping [12], Bayesian learning [13], training with both dynamic and static process data [14], and combining multiple networks [7,15,16]. Among these techniques, combining multiple networks is a very promising approach to improving model predictions on unseen data.

Fig. 2 shows a bootstrap aggregated neural network where several neural networks are developed to model the same relationship and are combined together. The individual networks are developed on data sets obtained from bootstrap re-sampling of the original training data [17]. Earlier studies show that an advantage of stacked neural networks is that they can not only give better generalization performance than single neural networks, but also provide model prediction confidence measures [18]. The aggregated neural network output is given by:

$$f(X) = \sum_{i=1}^{n} w_i f_i(X) \quad (1)$$

where $f(X)$ is the aggregated neural network predictor, $f_i(X)$ is the ith neural network, w_i is the aggregating weight for combining the ith neural network, n is the number of neural networks, and X is a vector of neural network inputs. Note that Eq. (1) is for regression or predictive modelling. For classification, majority voting can be used as the aggregated neural network output. Majority voting can be implemented as Eq. (2) where the median is used as the aggregated neural network output.

$$f(X) = \text{median}\{f_i(X)\}, \quad i = 1, 2, \ldots, n \quad (2)$$

3.2. Feed crude oil classification

Using on-line measurements of process variables, the feed crude oil is classified into one of the three types: with more light components, with more middle components, and with more heavy

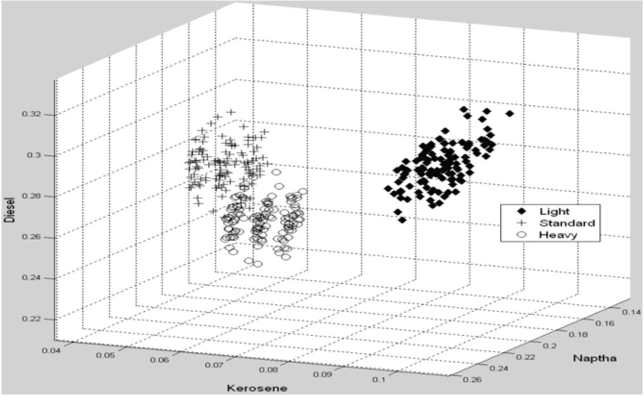

Fig. 3. Relationship between feed crude types and the ratios between products and feed.

components. As different crude oil produces different amount of products, so the product flow rates have direct reflection of oil types. Since the flow rates of feed crude oil changes during refinery operations, the ratios between product and feed rates should better reflect the oil types.

Fig. 3 shows the relationship between feed crude types and the ratios between products and feed. There are three products: kerosene, naphtha and diesel. Each coordinate axes represents the ratio between the flow rate of one product and the flow rate of the feed. Fig. 3 indicates that the three ratios could be used in classifying the feed crude oil. In practical refinery operations, the blending rates and the type of crude oil are often known and, furthermore, lab analysis of crude oil composition can be occasionally carried out, so the data for building crude oil classification models are generally available. In this study, 2190 samples of data were generated from simulation. The data were divided into training data set (25%), data set (25%), and unseen validation data set (50%). In neural network model development, the network is trained on the training data and the testing data set is used for network structure (number of hidden neurons) determination and early stopping in training (to avoid over-fitting). The final developed neural network is then evaluated on the unseen validation data.

A linear classifier was first developed and the classification accuracy on the unseen validation data is given in Table 1. It can be seen that the classification accuracy is not very high for the middle and heavy oil. This is due to the fact that the two classes are not linearly separable as indicated in Fig. 3. Thus a nonlinear classifier needs to be developed.

For the purpose of comparison, a single neural network based classifier is also developed. The neural network used is a single hidden layer feedforward neural network. Hidden and output layer neurons use the sigmoidal activation function. The neural network has 3 inputs which are the ratios between the 3 products and feed. The network has 6 hidden neurons determined through cross validation. The network has 3 output neurons corresponding to the 3 types of feed crude oil. An output of 1 indicates that corresponding type of crude oil presents while an output of 0 indicates that corresponding type of crude oil does not present. Table 1 shows the classification accuracy of different classifiers. It can be seen that bootstrap aggregated neural network classifier gives the best classification accuracy while the linear classifier gives the worst performance.

Table 1
Classification accuracy on testing data.

	Light	Middle	Heavy	All
Linear	100%	88.22%	97.81%	95.34%
Single network	100%	89.04%	99.18%	96.07%
Aggregated network	100%	90.96%	98.63%	96.53%

Table 2
Network combination schemes.

1	Single neural network
2	Median of 20 neural networks
3	Median of 10 neural networks with better performance on training data
4	Average of 20 neural networks
5	Average of 10 neural networks with better performance on training data

In order to demonstrate the robustness of bootstrap aggregated neural networks, five different network combination schemes listed in Table 2 were studied. The experiments were repeated 20 times with different bootstrap re-samples generated. Fig. 4 shows the classification accuracy and their 95% confidence bounds. Fig. 4 clearly indicates that bootstrap aggregated neural networks give much accurate and reliable (with narrower confidence bounds) classifications. From Fig. 4, it also can be seen that the 3rd to the 5th network combination schemes have similar performance, both in accuracy and in reliability (narrower confidence bounds). Thus any of the combination schemes 3–5 could be used.

4. Inferential estimation with bootstrap aggregated PLS models

4.1. Bootstrap aggregated PLS models

In the practical implementation of inferential estimation models, the model robustness is the most important factor that has significant impact on the entire accuracy of kerosene dry point estimation in refineries with varying crudes. The amount of training data from a refinery is usually small as the kerosene dry point is infrequently measured through lab analysis. Models developed with limited data have the tendency to over-fit the data. The classifier may misclassify some data and, furthermore, the training data may not contain all the potential oil types in future refinery operations.

Due to the strong correlations among the process variables in a distillation column, multiple linear regression is not appropriate in

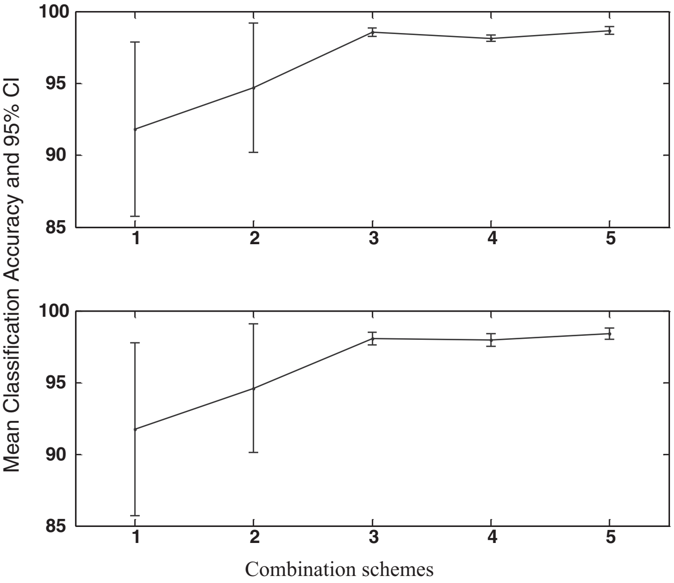

Fig. 4. Classification accuracies and their 95% confidence bounds on training data (top) and unseen validaiton data (bottom).

Table 3
RMSE on unseen validation data.

Models	I		II		III	
	Single	Multiple	Single	Multiple	Single	Multiple
Light	2.04	1.92	45.39	34.86	24.41	27.83
Middle	6.03	4.95	2.39	2.24	13.77	12.87
Heavy	7.75	6.06	3.59	3.70	2.24	2.13
All	5.79	4.65	26.32	20.28	16.23	17.75

building the inferential estimation models. PLS is a powerful modelling technique for situations where the predictors are correlated [19]. The prediction performance of a PLS or PCR (principal component regression) model on unseen data is highly influenced by the number of latent variables or principal components retained in the model. The number of latent variables is typically selected through cross validation. The data for building a PLS model is divided into a training data set and a testing data set. A number of PLS models with different numbers of latent variables are developed on the training data and tested on the testing data set. The model gives the lowest errors on the testing data is considered to have the appropriate number of latent variables. When data for building a PLS model is limited, the selection of latent variables could be highly influenced by the way that the training and testing data sets are partitioned. This problem could be solved through using bootstrap aggregated PLS model. Bootstrap aggregated PCR or PLS models have been shown to be an effective way to obtain robust PCR or PLS models [20].

Three bootstrap aggregated PLS inferential estimation models, corresponding to the three types of feed crude oil, were developed. In the bootstrap aggregated PLS model, 20 re-sampled datasets are produced through bootstrap re-sampling of the training datasets. A PLS model is developed on each re-sampled data set. The bootstrap aggregated PLS model is the average of all the 20 PLS models.

The inferential estimation model uses the following 16 measured process variables as its inputs: top stage temperature, diesel temperature, AGO temperature, kerosene temperature, preheat crude flow rate, reflux temperature, feed temperature, five product and feed flow rate ratios, reflux ratio, and three middle draw heat ratios. These process variables have higher correlation with the kerosene dry point and are selected from more than 50 measured process variables.

Simulated data were obtained using the HYSIS simulation described in Section 2. These data were then corrupted with random noise to represent the industrial situation where measurement noises always present. The number of latent variables for each PLS model is obtained through cross validation.

For building each of the PLS models corresponding to the three types of feed crude oil, the data is divided into 3 sets: training data set of 50 samples, testing data set of 50 samples and unseen validation data set of 100 samples. The reason for using a small training data set is from a practical consideration that the laboratory analysis data for kerosene dry point is usually limited. The estimation model will be more valuable in industrial application if proved to perform well after being trained on small amount of training samples.

4.2. Results

For each of the 3 training data groups, a single PLS model and a bootstrap aggregated PLS model are developed. And the test data is classified into 3 types: light, middle and heavy. All the models are tested on each of the test data group.

Table 3 shows the root mean squared errors (RMSE) of the developed models on the unseen validation data, using both single PLS and bootstrap aggregated PLS. In Table 3, models I–III are developed using light oil data, middle oil data, and heavy oil data, respectively. The model type "multiple" means "bootstrap aggregated PLS model" and "single" means "single PLS model".

Table 4
Average classification accuracy over 50 experiments on the unseen validation data in the industrial case study.

	Oil I	Oil II	Oil III	Oil IV	All
Single network	71.4%	86.4%	79.3%	85.7%	79.1%
Aggregated network	82.10%	95.50%	82.80%	85.70%	87.2%

The results reveal that the idea of building models, respectively, on each class is proved to be efficient. It can be seen from Table 3 that when the models are tested on their corresponding data sets, much higher accuracy is obtained than the models are tested on other datasets. Table 3 shows that when the unseen validation data and the estimation model are from the same type of feed oil, the estimation accuracy is much better than others. This indicates that different types of oil have quite different relationship between the product quality variables and the process variables, and the idea to build models on each class of oil is justified.

When the models are applied to their corresponding types of crude oil, the bootstrap aggregated PLS models performs better than the single PLS models on all the 3 cases. When a model is used on other types of crude oil (e.g. model II used on Light dataset), the bootstrap aggregated PLS model still generally performs better in most cases compare to a single PLS model.

5. An industrial application

5.1. Comparison of the single and bootstrap aggregated classifiers

Process data from a refinery in China is divided into 3 parts: training data group with 200 samples, test data group with 70 samples, and unseen validation data group with 87 samples. The feed oil is mixed by two different types of crude oil with different blending ratios. Lab analysis of these crude oil shows that the blended oil can be divided into 4 classes which are named as Oil I to Oil IV. Note that not all different blending ratios will lead to different classes of crude oil.

A single neural network classifier and a bootstrap aggregated neural network classifier are built on the training data group. To build the bootstrap aggregated neural network classifier, 20 replications of the training data were generated through bootstrap re-sampling with replacement. For each replication, an RBF neural network classifier is built. The final bootstrap aggregated classifier is the average of all the 20 classifiers.

The neural networks were trained using MATLAB Neural Network Toolbox. The inputs of the network are the three ratios as described in Section 3.

Both the single and the bootstrap aggregated classifiers are used to classify data on the unseen validation data group. In order to properly evaluate the performance, the experiments were repeated 50 times with different seeds of the random number generator in bootstrap re-sampling. Table 4 shows the average classification accuracy over the 50 experiments on the unseen validation data. From Table 4 it can be seen that, on all the unseen validation data, the classification accuracy of single network is 79.1% while that of the aggregated network is 87.2%. This means that the aggregated network classifier has 10.3% improvement over the single network classifier.

5.2. Inferential estimation

Using the bootstrap aggregated network as the online classifier, the test data and unseen validation data are then classified into 4

Table 5
RMSE on industrial unseen validation data.

Models	I		II		III		IV	
	Single	Multiple	Single	Multiple	Single	Multiple	Single	Multiple
Oil I	1.08	1.02	2.60	1.77	2.74	1.96	3.41	2.11
Oil II	1.70	1.66	1.39	1.22	2.35	1.96	4.81	2.28
Oil III	1.64	1.66	1.76	1.38	1.08	0.95	6.87	4.63
Oil IV	2.87	3.42	10.76	5.72	2.32	2.41	1.56	1.14
All	1.59	1.68	4.00	2.35	2.37	1.86	4.48	2.75

groups: Oil I to Oil IV. For every class of training data, a single PLS model and a bootstrap aggregated PLS model are built. The number of latent variables used in each PLS model is determined through cross validation.

Each of the 8 models is used to test on each of the 4 unseen validation data groups. The results are shown in Table 5, where “multiple” means “bootstrap aggregated PLS model” and “single” means “single PLS model”. Model I to model IV are, respectively, built from the 4 classes of training data, and Oil Ito Oil IV are the 4 classes of test data classified by the bootstrap aggregated neural network classifier.

As Table 5 shows, the results in the diagonal of the table have much smaller RMSE than the others. The results in the diagonal come from the situation that the on-line data and the estimation model are from the same type of feed oil. And the results besides the diagonal have shown relative poor accuracy, which reveals different types of oil have different relationship between the product quality variables and the process variables. The overall results indicate that, in the real industrial process, the inferential models built on each class of the feedstock will significantly improve the estimation accuracy with varying feed oil and the solution proposed in this paper is proved to be effective.

Another comparison is between the bootstrap aggregated PLS model and the single PLS model. Among the total 16 pairs of results, the bootstrap aggregated PLS model performs better than the single PLS model in 13 pairs, that means bootstrap aggregated PLS model has a chance of 81.3% to improve the estimation result. In the diagonal of the table multiple models are better in all the 4 cases, so bootstrap aggregated PLS model will perform significantly better on the whole when integrated with the on-line feed oil classifier.

Besides the results in the diagonal of the table, the bootstrap aggregated PLS models give even more significant improvement over the single PLS models in most of the cases. The reason is that single PLS models might over-fit the data, which leads to poor results when the oil type changes. The bootstrap aggregated PLS models have weaken this problem due to their multiplicity-aggregated internal structure. In industry processes, it is not easy to collect training data that contains all types of oil and the inferential estimation model will need to work on data different from the training data, so a robust estimation model is in great need. The over-fitting problem occurs heavily especially when the amount of training data is small, which happens to be the common thing in the industrial application. Bootstrap aggregated PLS model is proved to provide robust estimation results when the crudes are varying and the amount of training data is small.

6. Conclusions

A bootstrap aggregated PLS model inferential estimation approach integrated with online crude classification is developed for kerosene dry point estimation with varying crudes. The basic idea of this solution is to build estimation models, respectively, on each type of oil, and choose the proper model through a classifier during on-line estimation. Bootstrap aggregation is used mainly to solve the problems caused by the small amount of training data, which is common in many industrial applications. Bootstrap aggregated neural networks are used to on-line classify the feed oil using the ratios between on-line measured product and feed rates, and gives better classification accuracy than a single neural network. A bootstrap aggregated PLS inferential estimation model is developed for each type of feed oil. The results demonstrate that, the accuracy of the models which are generated from the training data with the same type of the on-line data is much higher than those of other models. The bootstrap aggregated PLS also shows good robustness. The proposed method is tested on both simulated data and industrial data, and proved to be an efficient way for the estimation of product quality variables with varying crude.

Acknowledgements

The research is supported by the National Basic Research Program of China (No. 2012CB720500) and UK Department for Innovation, Universities and Skills under the UK/China Fellowship for Excellence programme.

References

[1] M. Dam, D.N. Saraf, Design of neural networks using genetic algorithm for on-line property estimation of crude fractionator products, Computers & Chemical Engineering 30 (4) (2006) 722–729.

[2] L. Fortuna, P. Giannone, S. Graziani, M.G. Xibilia, Virtual instruments based on stacked neural networks to improve product quality monitoring in a refinery, IEEE Transactions on Instrumentation and Measurement 56 (1) (2007) 95–101.

[3] Y. Li, Q. Li, H. Wang, N. Ma, Particle soft sensing based on LS-SVM and its application to a distillation column, in: Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06), Jinan, China, vol. 1, 1995, pp. 177–182.

[4] J. Zhang, Offset-free inferential feedback control of distillation composition based on PCR and PLS models, Chemical Engineering & Technology 29 (2006) 560–566.

[5] M. Kano, Y. Nakagawa, Data-based process monitoring, process control, and quality improvement: recent developments and applications in steel industry, Computers & Chemical Engineering 32 (1) (2008) 12–24.

[6] T. Mejdell, S. Skogestad, Estimation of distillation compositions from multiple temperature measurements using partial-least-squares regression, Industrial and Engineering Chemistry Research 30 (1991) 2543–2555.

[7] J. Zhang, A.J. Morris, E.B. Martin, C. Kiparissides, Inferential estimation of polymer quality using stacked neural networks, Computers & Chemical Engineering 21 (1997) s1025–s1030.

[8] G. Cybenko, Approximation by superposition of a sigmoidal function, Mathematics of Control Signals and Systems 2 (1989) 303–314.

[9] N.V. Bhat, T.J. McAvoy, Use of neural nets for dynamical modelling and control of chemical process systems, Computers and Chemical Engineering 14 (1990) 573–583.

[10] A.B. Bulsari (Ed.), Computer-Aided Chemical Engineering, vol. 6, Neural Networks for Chemical Engineers, Elsevier, Amsterdam, 1995.

[11] C. Bishop, Improving the generalisation properties of radial basis function neural networks, Neural Computation 13 (1991) 579–588.

[12] C. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Oxford, 1995.

[13] D.J.C. MacKay, Bayesian interpolation, Neural Computation 4 (1992) 415–447.

[14] J. Zhang, Developing robust neural network models by using both dynamic and static process operating data, Industrial and Engineering Chemistry Research 40 (2001) 234–241.

[15] D.V. Sridhar, R.C. Seagrave, E.B. Bartlett, Process modelling using stacked neural networks, AIChE Journal 42 (1996) 2529–2539.

[16] D.H. Wolpert, Stacked generalization, Neural Networks 5 (1992) 241–259.

[17] B. Efron, The Jackknife, The Bootstrap and Other Resampling Plans, Society for Industrial and Applied Mathematics, Philadelphia, 1982.

[18] J. Zhang, Developing robust non-linear models through bootstrap aggregated neural networks, Neurocomputing 25 (1999) 93–113.

[19] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Analytica Chimica Acta 185 (1986) 1–17.

[20] M. Ahmed, J. Zhang, Improved inferential feedback control through combining multiple PCR models, in: Proceedings of the 2003 IEEE International Symposium on Intelligent Control, Houston, TX, USA, 2003, pp. 878–883.