Petroleum Science 18 (2021) 1230–1239

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

Support vector regression modeling in recursive just-in-time learning framework for adaptive soft sensing of naphtha boiling point in crude distillation unit

Venkata Vijayan S, Hare Krishna Mohanta, Ajaya Kumar Pani*

Department of Chemical Engineering, Birla Institute of Technology and Science, Pilani, Rajasthan, 333031, India

ARTICLE INFO

Article history:
Received 8 January 2020
Accepted 26 January 2021
Available online 12 July 2021

Handling editor: Xiu-Qiu Peng

Keywords:
Adaptive soft sensor
Just in time learning
Regression
Support vector regression
Naphtha boiling point

ABSTRACT

Prediction of primary quality variables in real time with adaptation capability for varying process conditions is a critical task in process industries. This article focuses on the development of non-linear adaptive soft sensors for prediction of naphtha initial boiling point (IBP) and end boiling point (EBP) in crude distillation unit. In this work, adaptive inferential sensors with linear and non-linear local models are reported based on recursive just in time learning (JITL) approach. The different types of local models designed are locally weighted regression (LWR), multiple linear regression (MLR), partial least squares regression (PLS) and support vector regression (SVR). In addition to model development, the effect of relevant dataset size on model prediction accuracy and model computation time is also investigated. Results show that the JITL model based on support vector regression with iterative single data algorithm optimization (ISDA) local model (JITL-SVR:ISDA) yielded best prediction accuracy in reasonable computation time.

1. Introduction

Raw naphtha is a mid-stream liquid distillate, fractionated from atmospheric distillation unit. It is a mixture of alkanes, cycloalkanes and aromatics with carbon atoms range from 5 to 11 and majorly used as a solvent for elastomers, diluents for paints and varnishes, hydrogen production and blended with gasoline to form high octane fuels (Duchene et al., 2020). For different crude asset (Paraffinic, Naphthenic or Aromatic), the boiling range of fractional cut varies based on composition.

Initial boiling point (IBP) and End boiling point (EBP) are the two key indicators for naphtha quality which is obtained from laboratory analysis with significant time delay. Laboratory analysis based on EN ISO 3405 and ASTM D86 standards, provides the boiling range characteristics of different petroleum products under various conditions at atmospheric pressure. Moreover, online sensors e.g. gas chromatographs possess measurement delay. The sampling time varies from few minutes to several hours (Ujević et al., 2011). It is mandatory to maintain the quality of refinery products in which the properties of naphtha to be continuously monitored and controlled. Hence, there is a strong need for online monitoring of naphtha boiling point. In the absence of availability of any hardware sensor for the same, soft sensor can be a viable alternative.

Soft sensors or inferential sensors are data-driven intelligent software program (model) using statistical and/or artificial intelligence methods capable of translating information from measurable secondary variables (temperature, pressure, flow rates etc.) to predict primary variables (product quality). Use of inferential sensors helps to take faster and objective oriented decisions during practical difficulties associated with delay in measurements, unreliable measured variables due to drifts, fouling or accidental damage of process analyzers and manual errors in laboratory analysis.

A list of various soft sensors reported in literature for predicting naphtha quality parameters is presented in Table 1.

Literature survey reveals that though quite a few soft sensors for naphtha property estimation have been reported, most of these reported soft sensors are based on steady state analysis. Modern petroleum refineries are highly complex and also from time to time there will be changes in operating conditions and feedstock quality. Therefore, to maintain prediction accuracy, the soft sensor model

* Corresponding author.
E-mail address: akpani@pilani.bits-pilani.ac.in (A.K. Pani).

https://doi.org/10.1016/j.petsci.2021.07.001

Table 1
Literature review of soft sensors for estimation of naphtha property.

Author, Year	Output	Methods used
Dam and Saraf (2006)	IBP and EBP of heavy naphtha	Genetic Algorithm-Artificial Neural Networks
Macias-Hernandes et al. (2007)	Naphtha 95% cut point	Extended evolving Takagi-Sugeno fuzzy model
Yan (2008)	Naphtha 25% cut point	Modified nonlinear generalized ridge regression
Yan (2010)	Naphtha dry point	Hybrid artificial neural networks
Ujević et al (2011)	IBP and EBP of heavy naphtha	Multiple linear regression, Multilayer perceptron (MLP) and Radial basis function (RBF) neural networks
Wang et al. (2013)	Naphtha dry point	Backpropagation learning technique combining correlation pruning algorithm with multiple linear regression model
Shang et al. (2015)	Naphtha 100% cut point	Dynamic partial least square regression
Torgashov et al. (2018)	Desired cut 2 (mixture of naphtha and gasoline)	Static linear regression and dynamic finite impulse response model

should be designed so as to adapt with the changing process conditions. Unfortunately, adaptive soft sensors for naphtha boiling point prediction is rarely reported. Therefore, design of soft sensor for naphtha property prediction with adaptation capability will be a significant step towards effective implementation of intelligent and smart manufacturing concept in petroleum refinery.

The commonly used adaptive techniques are moving window (Kaneko and Funatsu, 2015; Liu et al., 2018), recursive principal component analysis and partial least squares regression (Poerio and Brown, 2018), just-in-time learning (Liu, 2017) and supervised ensemble (Shao and Tian, 2017) methods. Recently, Just-in-Time Learning method has gained significant attention among researchers due to its better prediction and implementation capability as compared to other methods. There has been significant progress in the past decade for JITL based adaptive soft sensor development in various chemical processes such as van de Vusse continuous stirred tank reactor (Cheng and Sen (2004); 2005), Tennessee Eastman process (Ge and Song, 2010), Batch Fermentation process (Liu et al., 2012), Debutanizer Column (Yuan et al., 2014) and Sulphur Recover Unit (Shao et al., 2015).

In this work, adaptive soft sensor based on JITL approach was designed to predict the initial and end boiling points (IBP & EBP) of heavy naphtha fractions from splitter bottom of CDU. In the proposed algorithm, a similarity index is initially computed based on the Euclidean distance between query data sample and each sample of the database. A relevant database is formed by extracting a fixed number of samples from the database which are most similar to the query data. Subsequently, the relevant data set is used to develop the predictive model and then the output is computed for a particular query data. After computation of the output, the query data is included and the most dissimilar object is removed from the database. This activity is repeated for each incoming query sample. In this way the database is recursively updated with every query sample.

The contribution of this article is three fold. First, to the best of authors' knowledge, adaptive soft sensing for prediction of naphtha boiling point is rarely reported in literature which is the main contribution of this work. Development of adaptive soft sensor based on JITL for prediction of naphtha boiling point in CDU finds better relevance in monitoring and implementation purpose in modern refineries.

Second, there is scope of exploring several local modeling techniques in JITL framework with better generalization and estimation capability. Different types of linear and non-linear local models were developed in this work in order to get the local model with best prediction accuracy. The linear models are: multiple linear regression (MLR), locally weighted regression (LWR) partial least square (PLS) regression models and the nonlinear model include support vector regression (SVR) model.

Determination of optimum relevant dataset size is an important design issue in adaptive soft sensor modeling in JITL framework. Unfortunately, there are few answers available in the literature. During development of the non-linear models, the effect of relevant dataset size and model hyper-parameter on prediction accuracy is investigated so as to choose the optimum model hyper-parameter. Rigorous grid search method is adopted for optimal relevant dataset size determination. Local models were developed from relevant dataset with varying sizes and the effect of size on prediction accuracy as well as on computation time was investigated and the optimum size of relevant data set was also determined. Performance of best model is further validated by conducting 4-plot analysis technique.

The article is organized as follows. Section 2 introduces the readers to the process of crude distillation in refineries, the production of naphtha and the importance of real time boiling point estimation for naphtha. The methodology adopted in this work for just-in-time based adaptive soft sensor development is presented in section 3. All results obtained during different model building steps are presented and critically analyzed in section 4. Finally, concluding remarks are presented in section 5.

2. Process description

Petroleum (also known as crude oil), a fossil fuel with complex hydrocarbon forms a chief source of energy. The steady increase in per-capita consumption of conventional petroleum and petroleum related products with simultaneous drop in exploration of oil wells shows an alarming trend towards major energy crisis. It is indispensable for every industries as well as individual, to safeguard the energy resources for the future generation. This warrants efficient and optimized process control for reducing the wastage of energy resources. The other agenda for process control is to avoid process variations and customer complaint of product. Process models can be potentially useful to design effective controllers. Petroleum undergoes three different stages (upstream, mid-stream and downstream) of process before approaching the end usage of consumers. The upstream process includes dewatering, desalting and desulphurization unit. The mid-stream stage comprises multi-component fractionation, hydro cracking, reforming and hydro treating. Finally, the downstream process includes petrochemical complex such as formation of fertilizers, polymers, dyes and pigments. The process described here, forms as small portion of mid-stream crude fractionation process, which consists of atmospheric distillation column and stripping unit with a preheater at the bottom section and condensing system at the top overhead section. The crude distillation unit (CDU) is the heart of the refinery. It is very essential to monitor and control the quality of products in CDU, to meet the product specification and be customer compliant

with reduced wastages and reasonable profit.

A schematic process flow diagram for CDU is presented in Fig. 1. The unstabilized naphtha (C_3–C_9) is subjected to stabilizer unit, separates LPG, liquefied petroleum gas (C_3, C_4) in the top and stabilized naphtha at the bottom. The stabilized naphtha (C_4–C_9) is then fed to splitter unit which then further separates light naphtha ($C_5 < 70\,^{\circ}C$) and heavy naphtha ($C_5 > 70^0$–$185\,^{\circ}C$). The temperature reading measured for the condensate first drop at the outlet of condenser is initial boiling point (IBP) and the upper temperature limit observed during the test is found as end boiling point (EBP).

It is noted that the end boiling point of splitter bottom fraction should not exceed 204 °C because the rate of deactivation of platinum catalyst increases while processing through catalytic reforming unit. Also, the initial boiling point of splitter bottom fraction should be maintained between 75^0-100 °C, so that it prevents the formation of precursor for undesirable benzene above this range in catalytic reforming unit. Therefore real time estimation of these parameters will be of great assistance in maintenance of naphtha quality. The estimation of naphtha boiling point (product quality) with marginable savings is the basis for this soft sensor development. For model development, the various input variables considered are: atmospheric column top temperature, splitter top temperature, splitter bottom temperature, splitter top pressure, reflux flow rate and splitter inlet temperature. The input variables were selected based on sensitivity analysis. Sensitivity analysis provides the information about most influential variables over response. The criterion for selection of suitable variable for model development is based on the sensitivity ratio (Rank order higher than one). Interested readers can refer Ujevic et al. (2011) for further details. The variables of interest in this work are initial boiling point and end boiling point of naphtha. These are quality parameters which are determined by laboratory analysis. Therefore, the sampling time varies from 1 h to several hours in different refineries. The different input and output variables considered in this work are presented in Table 2.

3. Methodology

The total datasets of 210 samples for IBP and 209 samples for EBP each with 6 inputs and 1 output were considered for this work (Ujević et al., 2011).

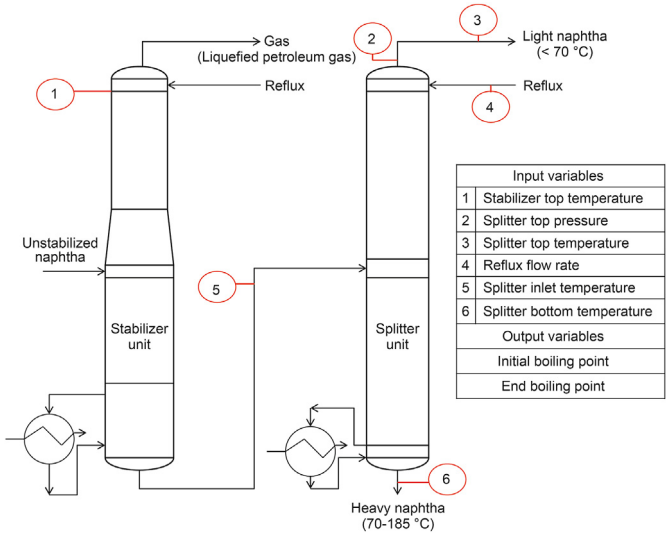

Fig. 1. Schematic process flow diagram of Crude Distillation Unit (Stabilizer and Stripper).

Table 2
Input-output process variables used for modeling.

Inputs	Outputs
Column top pressure	Initial Boiling Point & End Boiling Point
Splitter top pressure	
Splitter top temperature	
Reflux rate	
Splitter inlet temperature	
Splitter bottom temperature	

In offline global modeling, the model parameters developed from the particular training dataset is fixed. Once the online process data were measured, the output is delivered when it undergoes the sequence of prediction scheme. Here, the model parameters for the developed model were not changed for the particular process data. Because of this, the residual error (predicted − observed) for the predicted output seems very high. However, in just-in-time based local modeling framework, the online measured process data sample is compared with the stored data in database based on distance method. The nearest neighbor of the current process data are selected (relevant dataset) and the model is developed with the help of selected relevant dataset (RDS). In this way, the model parameters are updated for each incoming set of input data.

The database samples are selected in such a way that it must captures wide range of relevant process information. The regression surface can be created by using local regression equation based on linear and non-linear functions. The model for output is developed by using the 'n' number of closest neighborhood for incoming query. The neighborhood is selected based on distance weighting procedure. Higher weights are assigned to data points which are close to the query. Consequently, the farther ones acquire lower weight. Then, the weighted least squares were used to build a local function. Herewith, the effect of outlier in the incoming measurements from the real plant data can be minimized by using weighting method (Park and Han, 2000).

When the estimation of output is requested for a query data received from the process stream, the search for closest neighborhood point for the former is initiated in the database. After the search is over, a vector of closest neighborhood (for the incoming query point) is created by sorting each object of the database in ascending order from the query data, based on distance. The distance between the query data and an object of database may be computed based on Euclidean, Mahalanobis or Gaussian Mixture Model (Fan et al., 2014), combined distance and angle (Cheng and Chiu, 2004), combining Q statistic and Hotelling's T^2 from PCA (Fujiwara et al., 2009) and combining measure distance from support vector data decomposition and Hotelling's T^2 in Non-Gaussian JITL (Zeng et al., 2011).

After the computation of distance, a similarity index vector is created. The initial step in getting similarity index involves computation of weighting function corresponding to each distance according to Eqn. (1) given below.

$$w_i = \sqrt{K\left(\frac{d(x_q, x_i)}{h}\right)} \quad [1]$$

w_i – weighting function; d – distance between query sample and each sample of the database; K – Kernel function; h – Bandwidth of kernel function.

The commonly used kernel functions for weighting purpose are linear, Gaussian (or radial basis function) and polynomial. Among these, the Gaussian function maps non-linear complex features effectively so as to minimize the residual error due to under-fitting

problems by bias addition(Wang et al., 2016).

$$K(d) = e^{-d^2} \quad [2]$$

It may be noted that, when the distance from the query point to neighborhood point decreases, the weight function, w_i increases. Based on this, the similarity index (S_i) was computed between query data and its neighborhood in the database (Cheng and Chiu, 2004).

3.1. Similarity index (S_i) calculation

The similarity index is calculated by finding the distance between query data and its closest neighborhood in the database by means of Euclidean distance (Ge and Song, 2010).

$$S_i = \sqrt{e^{-d^2(x_q, x_i)}} \quad [3]$$

Where, $d^2(x_q, x_i) = (x_q - x_i)^T (x_q - x_i)$; $i = 1, 2, \ldots n$.

The similarity index was sorted down in descending order and converted into a diagonal matrix (by multiplying with identity matrix whose size is analogous to similarity index matrix). Then the dominant diagonal matrix (weighting matrix W) is multiplied with the database samples for model development. Then the input data matrix is subjected to singular value decomposition with query data point to get the model parameter (regression coefficient) values. In this work, three linear (MLR, LWR and PLS) and two non-linear (SVR) local models were developed. Based on the type of local model, the just-in-time based adaptive soft sensors will hereafter be named as JITL-MLR, JITL-LWR, JITL-PLS and JITL-SVR. Regression coefficients for linear local models are computed according to the equations given below.

Multiple linear regression:

$$\beta = \left(\phi^T W \phi\right)^{-1} \phi^T W y \quad [4]$$

Locally weighted regression:

$$\beta = \left(Z^T Z\right)^{-1} Z^T v \quad [5]$$

Partial least squares regression:

$$\beta = W_{pls}\left(P_L^T W_{pls}\right)^{-1} q_L^T \quad [6]$$

Where, ϕ and y were relevant dataset and its corresponding output data matrices, W – Weighting Matrix; β – Model parameter; $Z = W\phi$; $v = Wy$; W_{pls} – Weight matrix (for PLS model); P_L – Loading matrix of input variables; q_L – Loading vector for output variables.

The final Just-in-Time learning model equation for the prediction of Initial and End boiling point is found to be the predicted output (y_q) obtained directly by multiplying the query data point with regression coefficient value.

$$y_q = \beta x_q \quad [7]$$

It may be noted that during model development, variables with higher magnitudes may dominate over variables with lower magnitudes. In order to avoid this undesirable influence, all variable values were normalized using z-score normalization technique according to the equation given below:

$$S = (x_i - \bar{x}) / s_x \quad [8]$$

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})}{n-1}} \quad [9]$$

S – Scaled data; x_i – Sample; $\bar{x}$ – Sample mean; s_x – Sample standard deviation; n – Total number of samples.

The local model is discarded after prediction of output for a particular query sample and for the next query sample a new local model is constructed.

In JITL technique, the database may be fixed or may be continuously updated. In most of the reported techniques the database is continuously updated by including the query sample without any removal of data from database. This leads to an increase in database size with time and consequently, the computation time for each sample will increase with time (this is because, initially, distance of the query sample from each sample of the database has to be computed). For updation of database, several procedures were reported in literature. It is very important that the data having good information must be included and those with poor information must be excluded from the database (Kaneko and Funatsu, 2014). Several updation techniques were used by researchers in the past: absolute estimation error (Kansha and Chiu, 2009), data monitoring index (Kaneko and Funatsu, 2014) and hybrid similarity index (Jin et al., 2014) was proposed to remove the most similar samples using new sample from the database. In this work, we have followed a technique of database updation where, the latest query data is included in the database and one sample already present in the database and which has least similarity with query data is removed. In this manner, the database size is fixed while ensuring continuous updation of the database recursively.

Fig. 2 explains the JITL algorithm adopted in this work for adaptive soft sensor design. It may further be stated that the data used in this work is free from noise and outliers. However, in cases where there is possibility of presence of outliers, JITL finds a local region from the database to incoming query data and regression is applied around the region of interest. Hence, thereby the local weighting concept helps to lessen the noisy non-local region data for model development (Wang et al., 2016). Further, prior to application of JITL algorithm some univariate or multivariate outlier detection algorithms (data preprocessing) can be applied on the incoming sample vector (with some predefined threshold value) to determine whether the incoming set of measurements contains

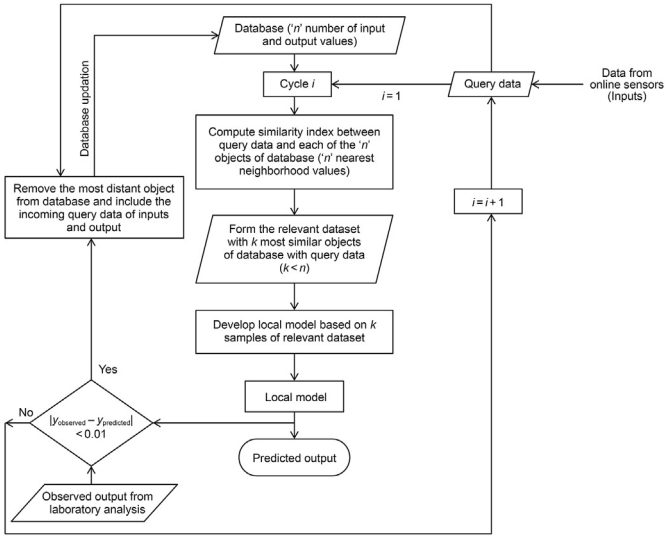

Fig. 2. Flow chart of the Just in Time Learning Technique adopted in this work.

any outlying value or not by combining data preprocessing algorithm (Chen et al., 2014) prior to JITL application.

4. Results and discussion

The prediction efficiency of all models was evaluated by computing the statistical parameters of mean absolute error (*MAE*) and correlation coefficient, *CC* (or Pearson coefficient, *R*). The *CC* provides the quantification of dependency of one variable with the other variable for a multivariate data distribution while, *MAE* gives the measure of respective deviation of predicted output from observed data in the form of average of error in absolute value (Rogina et al., 2011; Yuge et al., 2018). The mean absolute error and correlation coefficient is given by

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y-\hat{y}| \quad [10]$$

$$CC = \frac{\Sigma(y-\overline{y})\left(\hat{y}-\overline{\hat{y}}\right)}{\sqrt{\sum(y-\overline{y})^2\sum\left(\hat{y}-\overline{\hat{y}}\right)^2}} \quad [11]$$

y is the actual observed value, $\hat{y}$ represents the model predicted value, $\overline{y}$ is the observed mean, $\overline{\hat{y}}$ is the predicted mean and n is the total number of samples taken for prediction of output. The higher the CC value close to unity and the lower the MAE value close to zero, the better is the predictive performance of the developed model.

The models for IBP and EBP are developed separately. 59 objects (58 objects for EBP) were extracted from the available 210 objects (209 objects for EBP soft sensor) using Kennard-Stone algorithm (Kennard and Stone, 1969). These 59 objects are used as query data. The remaining 151 objects are used as database. When a query data is received by the JITL algorithm, distance based similarity values are computed between the query data and each of the 151 objects of the database. Subsequently, a certain number of samples which are closest to the query data are taken as relevant dataset. The local JITL models were developed from this relevant dataset. This is for the reason that, the samples of relevant dataset are most similar to the query data. The developed model is used to predict the output for the query data. Subsequently, the query data is included in the database and the most dissimilar object (based on the previously computed similarity index) is discarded from the database. The program for just-in-time (JIT) architecture and simulation of all the models were developed on MATLAB® (*R2018b*).

4.1. Determination of optimum relevant dataset size for linear local model

An open issue in JITL based approach is to determine the optimum size of relevant dataset. In order to determine the optimum size, local linear models (MLR, LWR, PLS) were developed based on different number of relevant dataset size ranging from 10 to 150 and the prediction results for 59 objects (58 for EBP) of query data were determined. The dependency of model prediction accuracy on the relevant dataset size is shown in Fig. 3.

In the above Figs. PLS (3) and PLS (4) corresponds to the PLS local models with three and four latent variables respectively (It may be recalled that the number of actual input variables is six). A common perception is: the more is the size of the data used for modeling, the better will be the model performance. This may be true in steady state soft sensor design where one model developed from a set of offline data is subsequently used for predicting outputs for all unknown inputs. However, this concept may not be always true in case of adaptive soft sensor design using local modeling concept where a model is built to predict output for just one set of input. A deep in performance around RDS 30 in Fig. 3(b) indicates existence of some optimum RDS size.

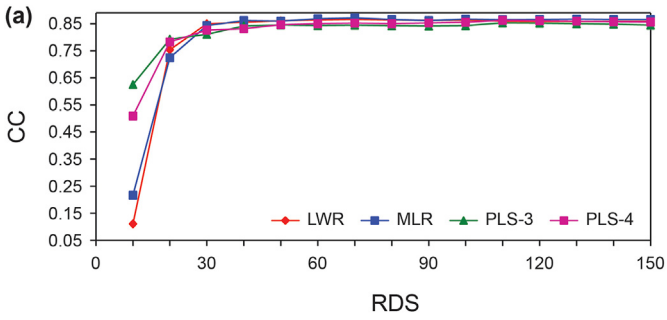

Fig. 3 (a). Prediction accuracy (CC) Vs relevant dataset size for naphtha initial boiling point (Linear local models).

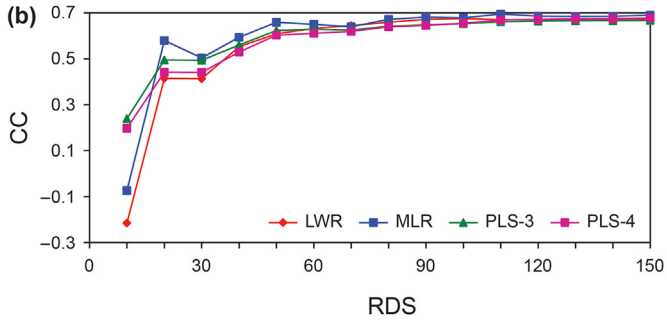

Fig. 3 (b). Prediction accuracy (CC) Vs relevant dataset size for naphtha end boiling point (Linear local models).

Both Fig. 3 (a) and 3 (b) shows that there is negligible improvement in prediction accuracy beyond a dataset size of 50. Therefore, for linear models, the optimum relevant dataset size was decided to be 50 (k = 50; refer algorithm in Fig. 2). Further, for any dataset size, performance of JITL-LWR is found to be better than that of JITL-MLR, JITL-PLS (3) and JITL-PLS (4). Also, performance values for IBP prediction are better than that for EBP prediction.

4.2. Design of non-linear local models

In addition to the linear local models (MLR, LWR and PLS), two non-linear local model was also developed in this work: support vector regression with sequential minimal optimization (SMO) and support vector regression with iterative single data algorithm optimization (ISDA) model. These adaptive soft sensors are mentioned as: JITL-SVR:SMO and JITL-SVR:ISDA.

Support vector regression (SVR) models have become attractive alternatives to neural network models for non linear processes. SVR model develops the relationship by projecting the predictor variables into the high dimensional space for solving convex quadratic optimization problems (Cortes and Vapnik, 1995; Qian et al., 2018; Vapnik, 1999; Zhong et al., 2010). In SVR, two optimization approaches sequential minimal optimization (SMO) algorithm and iterative single data algorithm (ISDA) were used to develop the model. SMO (Platt, 1999) is based on the rule of second-order iterative selection algorithm uses two lagrangian multipliers as a reference to solve optimization problem faster than the existing quadratic programming in SVR. ISDA (Kecman et al., 2005) works by classical Gauss-Seidel iterative algorithm updating the single Lagrangian multiplier every-time for a huge datasets to converge

rapidly. Here, the selection of optimum value of loss function (ε) for modeling applications is very important to select the best performing models with reasonable prediction (Shokri et al., 2015; Yan et al., 2004). There is no unanimous formula for determination of hyperparameter values (loss function in this case). There is a range of methods starting from grid search to different evolutionary techniques (such as genetic algorithm, ant colony optimization, particle swarm optimization etc.) used by researchers to determine the hyperparameter values resulting in best prediction accuracy. One such interesting work was observed in Shokri et al. (2014), who proposed a soft optimization technique based on hybrid meta-heuristic approach, which preset the hyper-parameter settings in advance before model development. Grid search method is rigorous and cumbersome but yields reasonably better optimum values like other proposed techniques in literature because the model is tested over a wide possible range of values (Pani and Mohanta, 2014). Therefore in this work the grid search method is adopted to determine the optimum loss function value.

For both JITL-SVR:SMO and JITL-SVR:ISDA, the prediction performance was tested for ε value ranging from 0.001 to 1.5 for both IBP and EBP. The precise results are presented in Fig. 4.

Other results showing effects of ε and relevant dataset size (RDS) on mean absolute error (MAE) are provided in the supplementary file. At 50 RDS, the mean error reached minimum value and started to climb higher by increasing the epsilon (ε) value.

The CC value starts decreasing while increasing the loss function value after 0.05 for IBP prediction. For EBP prediction, model performances start to degrade beyond an ε value of 0.9. Therefore, The optimum ε value was found to be 0.05 for IBP and 0.9 for EBP in JITL (Refer Fig. 4).

4.3. Effect of relevant dataset (RDS) size for non-linear local model

It may be noted that though, the results for non-linear local model based soft sensor are only presented for RDS size of 50, models with different RDS size were developed and tested (Interested readers can refer the supplementary file for details). Usually, model accuracy improves when the dataset size used for modeling increases. In this work, the local models are developed from the relevant dataset. Performance of linear local models in fact improves as the relevant dataset size increases and remained constant after some optimum size of 50 (Refer Fig. 3a and b). Irrespective of linear or non-linear local model, it was observed that at low size of RDS, the accuracy improves with increase in size of RDS. However, at higher values of RDS size, the accuracy shows negligible improvement with increase in RDS size. Low RDS size results in under fitting (low accuracy) and higher RDS size leads to increase in computation load. Further, the best result of JITL-SVR model was found to be better than JITL-linear (MLR, LWR and PLS) models.

For real applications, it will be difficult to select different

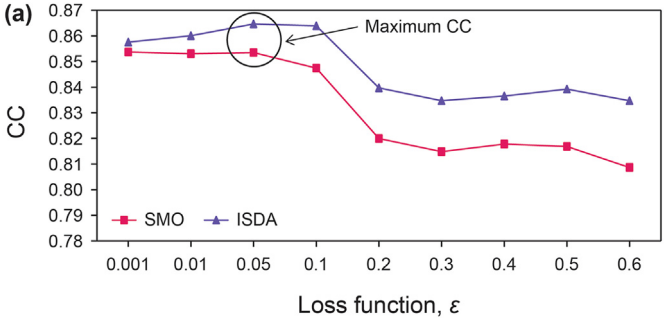

Fig. 4 (a). CC Vs loss function (ε) for IBP (RDS-50) [for SVR local model].

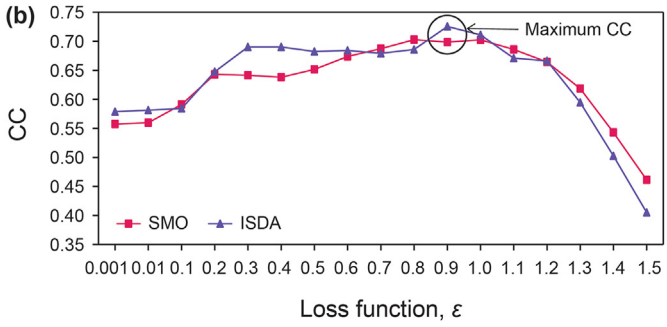

Fig. 4 (b). CC Vs loss function (ε) for EBP (RDS-50) [for SVR local model].

relevant dataset size and/or constituent model parametric values along with computation complications. For better model generalization, we propose an optimum relevant dataset size of 50 at which all models have maximum or close to maximum accuracy.

4.4. Effect of relevant dataset size (RDS) on model computation time for linear local model

In static soft sensors, the model parameters are determined during offline model development and the parameters remain same during their online use. However, in adaptive soft sensors, the model parameters are computed for each query data object. Therefore, for a static soft sensor, the model computation time involves the time for simulation of the already developed model when an input data set is supplied to the soft sensor. In adaptive soft sensors, computation time includes time required for model parameter computation and model simulation for each incoming input vector. Further, in case of JITL based adaptive soft sensor, the computation time involves computation of similarity of the query sample with each sample of the dataset, preparation of the relevant dataset, development of the local model (model identification) and finally simulation of the developed model with the query sample. Unless properly designed, a model possessing good accuracy may not be useful due to computational complexity. Therefore, an important issue for adaptive soft sensors is the model computation time. The computation time was determined using MATLAB® '*timeit*' function. Here, the computation time for every model was found in triplicates and average of those triplicates was used for comparisons to minimize the computational error.

The variations in computation time are between 0.008 and 0.0087 s. If we consider three places after decimal, then the effect of dataset size on computation time is insignificant. The effect is in the order of 10^{-4} s. This computation time is acceptable for online implementation of the algorithm because the online sensors monitoring the secondary variables are expected to have sampling time higher than this computation time. Hence it can be concluded that a relevant dataset size of 50 can be used in subsequent non-linear model development.

4.5. Computation time for non-linear local models

Average computation time for the JITL soft sensors based on non-linear local models (JITL-SVR:SMO and JITL-SVR:ISDA) are determined for a relevant dataset size of 50. The average computation time required for the JITL-SVR:SMO model varies from 0.0125 to 0.0148 s and 0.0124–0.0153 s in JITL-SVR:ISDA model. The average computation time required to predict the required output for non-linear local model based JITL soft sensor is higher than that based on linear local models (This may be due to more complexity involved in case of non-linear models).

Table 3
JITL based adaptive soft sensor performance for naphtha IBP prediction for optimum relevant dataset size.

Model type	Optimum relevant dataset size	Correlation coefficient (CC)	Mean Absolute Error (MAE), in ^{0}C
JITL-LWR	50	0.86	2.45
JITL-MLR	50	0.86	2.44
JITL-PLS (3)	50	0.85	2.70
JITL-PLS (4)	50	0.85	2.55
JITL-SVR-ISDA	50 ($\varepsilon = 0.05$)	0.86	2.58
JITL-SVR-SMO	50 ($\varepsilon = 0.05$)	0.85	2.68

Table 4
JITL based adaptive soft sensor performance for naphtha EBP prediction for optimum relevant dataset size.

Model type	Optimum relevant dataset size	Correlation coefficient (CC)	Mean Absolute Error (MAE), in ^{0}C
JITL-LWR	50	0.61	3.44
JITL-MLR	50	0.66	3.38
JITL-PLS (3)	50	0.62	3.47
JITL-PLS (4)	50	0.60	3.45
JITL-SVR-ISDA	50 ($\varepsilon = 0.9$)	0.73	2.87
JITL-SVR-SMO	50 ($\varepsilon = 0.9$)	0.70	3.15

In Tables 3 and 4, the prediction results by the optimum models (relevant dataset size = 50) of different types are presented for naphtha IBP and EBP respectively.

Figs. 5 and 6 show the prediction results by the models mentioned in Tables 3 and 4

The average computation time in case of any model for both IBP and EBP are very close to each other (as can be noticed in Fig. 7, the overlapping of two lines). Further, computation time for non-linear models is higher than that for linear local models. This may be due to the fact that higher degree of complexity is involved in development of the non-linear local model at each sampling instance. Though, the computation time was slightly higher, it is still well below the hardware sensor sampling rate which makes it suitable for online implementation. Therefore, the choice of whether to go for linear or non-linear local model is left to the user to decide. Both models are computationally acceptable. Both models almost give the same performance for IBP as well. However, in addition to IBP, if we also want reasonable accuracy for EBP prediction, then non-linear local model is to be preferred. Considering model accuracy and model computation time as the criteria for model selection it can be concluded that JITL-SVR:ISDA has the best prediction accuracy for both naphtha IBP and EBP prediction.

4.6. Further model validation of JITL-SVR:ISDA

The residual analysis of JITL-SVR:ISDA adaptive soft sensor is carried out by four-plot analysis technique (Fortuna et al., 2007; Pani and Mohanta, 2016). The results are shown in the following Fig. 8.

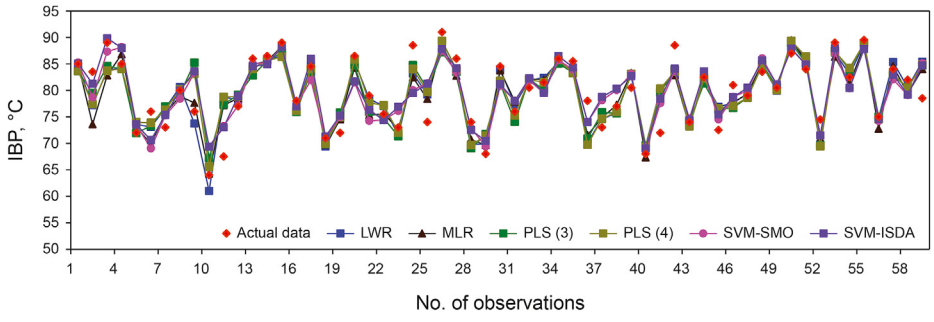

Fig. 5. Actual and predicted values of Naphtha initial boiling point (IBP) by JITL based adaptive soft sensors using different local models.

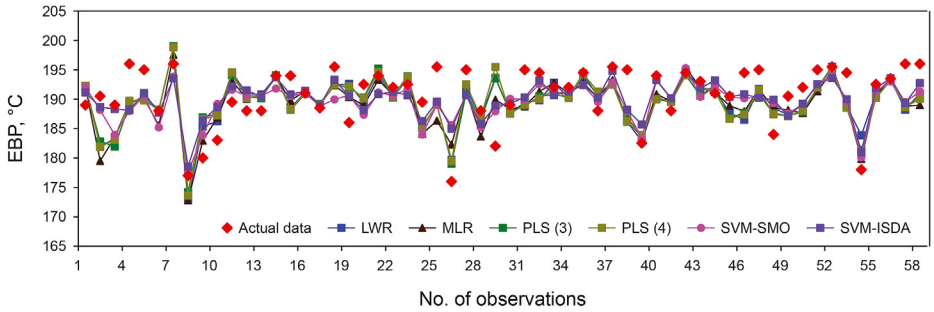

Fig. 6. Actual and predicted values of Naphtha end boiling point (EBP) by JITL based adaptive soft sensors using different local models. Further, in Fig. 7, we present the computation time for the different models at their optimum parameters.

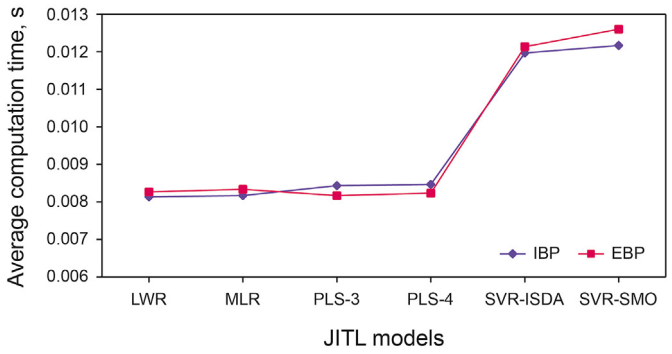

Fig. 7. Average computation time for JITL based adaptive soft sensors using different local models.

Four plots are used for residual analysis to identify the distribution of penalty for predicted variables over the observed output. It includes run sequence plot, lag plot, histogram and normal probability plot. Run sequence plot provides the information about residual trend around the mean while propagating through the time ordered observations. The trend may be increasing, decreasing or constant variance trend. The ideal trend observed to be constant variance trend, as the error variance seems no process drift as moving though the large number of time varying observations. Lag plot shows the graph of sequence of observations against the one time lagged sequence of observations. This gives the information about the spread of error variance around the mean. This is also observed to be random in nature and therefore not dependent on time. Histogram represents, how the error variance is normally distributed around the mean. The assumption of normal distribution is found to be true, when the histogram shape is symmetric and bell shaped. The normal probability plot shows the distribution of ordered error variance against the probability percentile is normally distributed. The linear graph indicates that the residual is normally distributed.

The prediction performance and computation time of non-linear models at different RDS sizes are provided as supplementary files for interested readers.

5. Conclusion

Initial and end boiling points of naphtha are important process parameters which need to be maintained at specified values for ensuring better process performance in petroleum refineries. Hence, adaptive soft sensors have huge scope of relevance to ensure quality of naphtha in fractionation process. In this work, just in time learning concept is applied to develop adaptive soft sensors for online monitoring of naphtha IBP and EBP. During the course of model development various aspects of JITL modeling are investigated. These include effect of relevant dataset size on model performance and model computation time. Linear and non-linear local models were designed. The best correlation coefficient values by any linear model are 0.86 for IBP and 0.66 for EBP (JITL-MLR) with a model computation time of 0.0083 s. The correlation coefficient values for JITL-SVR:ISDA are 0.86 for IBP and 0.73 for EBP with a model computation time of 0.012 s. The improvement in IBP prediction was insignificant. However, there is an improvement of more than 10% for EBP prediction in case of the proposed non-linear JITL model. After performance analysis of various models, it was observed that support vector regression model produced better prediction accuracy than multiple linear regression, partial least squares regression and locally weighted regression models. The performance was further validated using residual errors by four plot analysis. The criteria chosen for good modeling are, better generalization capability and low computational time. An error margin of about ±3 °C temperature is achieved by the JITL-SVR model for prediction of IBP and EBP. Reasonably low computation

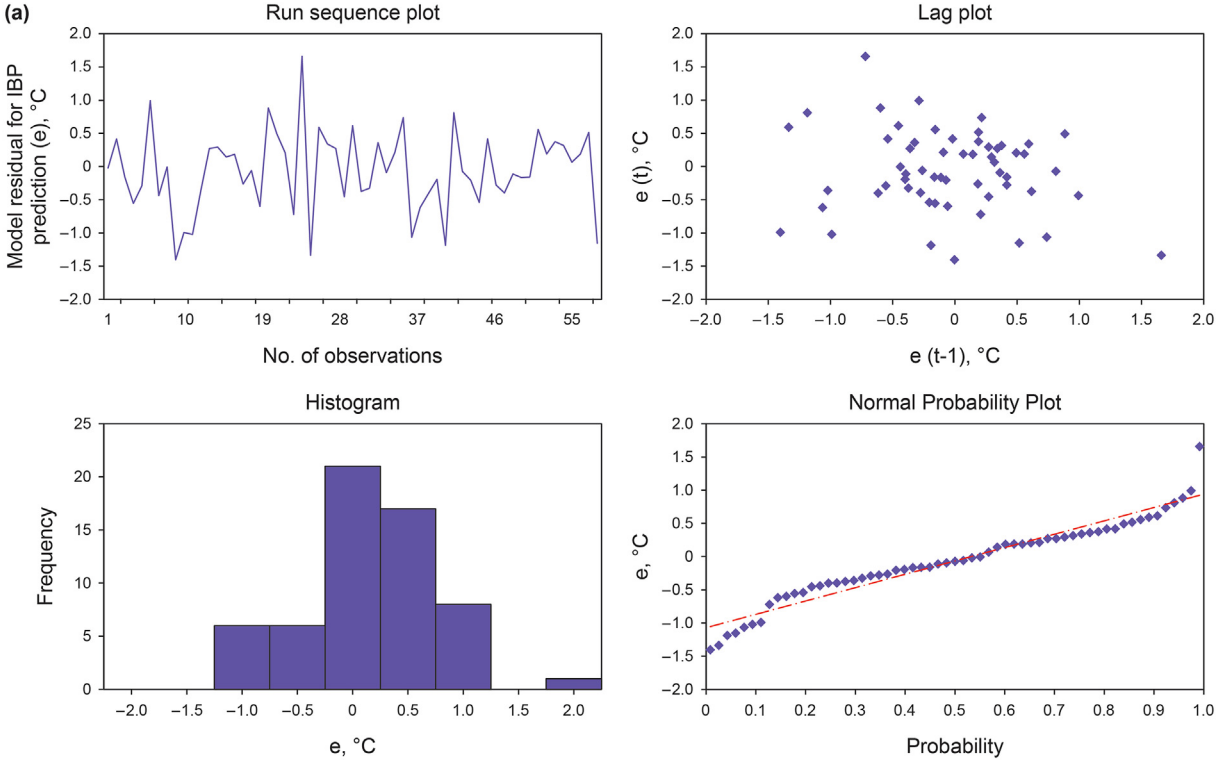

Fig. 8 (a). JITL-SVR:ISDA model validation for IBP prediction.

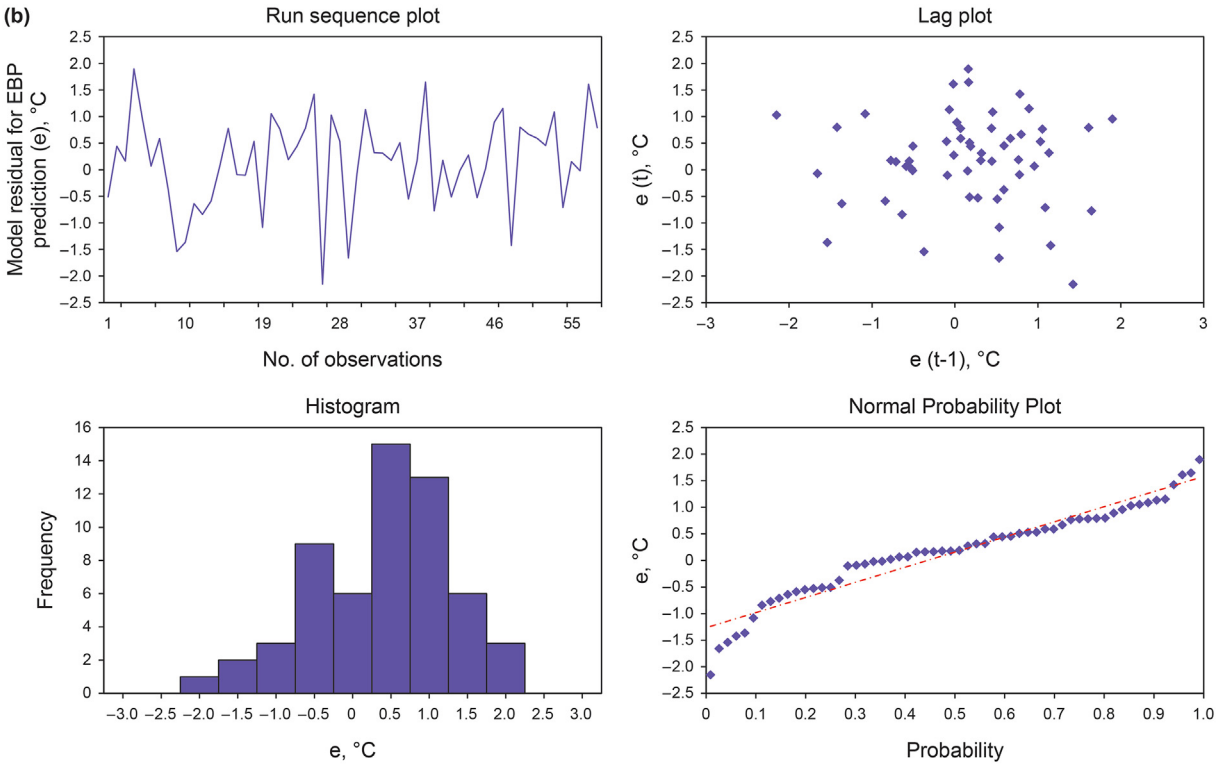

Fig. 8 (b). JITL-SVR:ISDA model validation for EBP prediction.

time of model simulation indicates that the proposed JITL-SVR model can be implemented online as adaptive soft sensor for continuous estimation of naphtha quality.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.petsci.2021.07.001.

References

Chen, M., Khare, S., Huang, B., 2014. A unified recursive just-in-time approach with industrial near infrared spectroscopy application. Chemometr. Intell. Lab. Syst 133–140. https://doi.org/10.1016/j.chemolab.2014.04.007.

Cheng, C., Sen, C.M., 2004. A new data-based methodology for nonlinear process modeling. Chem. Eng. Sci 2801–2810. https://doi.org/10.1016/j.ces.2004.04.020.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn 20 (3), 273–297. https://doi.org/10.1007/BF00994018.

Dam, M., Saraf, D.N., 2006. Design of neural networks using genetic algorithm for on-line property estimation of crude fractionator products. Comput. Chem. Eng 722–729. https://doi.org/10.1016/j.compchemeng.2005.12.001.

Duchene, P., Mencarelli, L., Pagot, A., 2020. Optimization approaches to the integrated system of catalytic reforming and isomerization processes in petroleum refinery. Comput. Chem. Eng 141. https://doi.org/10.1016/j.compchemeng.2020.107009, 107009.

Fan, M., Ge, Z., Song, Z., 2014. Adaptive Gaussian mixture model-based relevant sample selection for JITL soft sensor development. Ind. Eng. Chem. Res. 53 (51), 19979–19986. https://doi.org/10.1021/ie5029864.

Fortuna, L., Graziani, S., Alessandro Rizzo, M.G.X., 2007. Soft Sensors for Monitoring and Control of Industrial Processes. Springer London, London. https://doi.org/10.1007/978-1-84628-480-9.

Fujiwara, K., Kano, M., Hasebe, S., Takinami, A., 2009. Soft-sensor development using correlation-based just-in-time modeling. AIChE J 55 (7), 1754–1765. https://doi.org/10.1002/aic.11791.

Kaneko, H., Funatsu, K., 2014. Database monitoring index for adaptive soft sensors and the application to industrial process. AIChE J 60 (1), 160–169. https://doi.org/10.1002/aic.14260.

Ge, Z., Song, Z., 2010. A comparative study of just-in-time-learning based methods for online soft sensor modeling. Chemometr. Intell. Lab. Syst 104 (2), 306–317. https://doi.org/10.1016/j.chemolab.2010.09.008.

Jin, H., Chen, X., Yang, J., Wu, L., 2014. Adaptive soft sensor modeling framework based on just-in-time learning and kernel partial least squares regression for nonlinear multiphase batch processes. Comput. Chem. Eng 71, 77–93. https://doi.org/10.1016/j.compchemeng.2014.07.014.

Kaneko, H., Funatsu, K., 2015. Moving window and just-in-time soft sensor model based on time differences considering a small number of measurements. Ind. Eng. Chem. Res. 54 (2), 700–704. https://doi.org/10.1021/ie503962e.

Kansha, Y., Sen, C.M., 2009. Adaptive generalized predictive control based on JITL technique. J. Process Contr 1067–1072. https://doi.org/10.1016/j.jprocont.2009.04.002.

Kecman, V., Huang, T.-M., Vogt, M., 2005. Iterative Single Data Algorithm for Training Kernel Machines from Huge Data Sets: Theory and Performance, pp. 255–274. https://doi.org/10.1007/10984697_12.

Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11 (1), 137–148. https://doi.org/10.2307/1266770.

Liu, Y., 2017. Adaptive just-in-time and relevant vector machine based soft-sensors with adaptive differential evolution algorithms for parameter optimization. Chem. Eng. Sci 571–584. https://doi.org/10.1016/j.ces.2017.07.006.

Liu, Y., Gao, Z., Li, P., Wang, H., 2012. Just-in-time kernel learning with adaptive parameter selection for soft sensor modeling of batch processes. Ind. Eng. Chem. Res. 51 (11), 4313–4327. https://doi.org/10.1021/ie201650u.

Liu, Z., Ge, Z., Chen, G., Song, Z., 2018. Adaptive soft sensors for quality prediction under the framework of Bayesian network. Contr. Eng. Pract 72, 19–28. https://doi.org/10.1016/j.conengprac.2017.10.018.

Macias-Hernandez, J.J., Angelov, P., Zhou, X., 2007. Soft Sensor for Predicting Crude Oil Distillation Side Streams Using Evolving Takagi-Sugeno Fuzzy Models 2007 IEEE Int. Conf. Syst. Man Cybern. IEEE 3305–3310. https://doi.org/10.1109/ICSMC.2007.4413939.

Pani, A.K., Mohanta, H.K., 2014. Soft sensing of particle size in a grinding process: application of support vector regression, fuzzy inference and adaptive neuro fuzzy inference techniques for online monitoring of cement fineness. Powder Technol 264, 484–497. https://doi.org/10.1016/j.powtec.2014.05.051.

Pani, A.K., Mohanta, H.K., 2016. Online monitoring of cement clinker quality using multivariate statistics and Takagi-Sugeno fuzzy-inference technique. Contr. Eng. Pract 57, 1–17. https://doi.org/10.1016/j.conengprac.2016.08.011.

Park, S., Han, C., 2000. A nonlinear soft sensor based on multivariate smoothing procedure for quality estimation in distillation columns. Comput. Chem. Eng. 871–877. https://doi.org/10.1016/S0098-1354(00) 00343-00344.

Platt, J., 1999. Fast training of support vector machines using sequential minimal optimization. Adv. Kernel Methods -- Support Vector Learn. https://dl.acm.org/doi/10.5555/299094.299105.

Poerio, D.V., Brown, S.D., 2018. Highly-overlapped, recursive partial least squares soft sensor with state partitioning via local variable selection. Chemometr. Intell. Lab. Syst 175, 104–115. https://doi.org/10.1016/j.chemolab.2018.02.006.

Qian, K.R., He, Z.L., Liu, X.W., Chen, Y.Q., 2018. Intelligent prediction and integral

analysis of shale oil and gas sweet spots. Petrol. Sci. 15 (4), 744–755. https://doi.org/10.1007/s12182-018-0261-y.
Rogina, A., Šiško, I., Mohler, I., Ujević, Ž., Bolf, N., 2011. Soft sensor for continuous product quality estimation (in crude distillation unit). Chem. Eng. Res. Des 89 (10), 2070–2077. https://doi.org/10.1016/j.cherd.2011.01.003.
Shang, C., Gao, X., Yang, F., Lyu, W., Hunag, D., 2015. A comparative study on improved DPLS soft sensor models applied to a crude distillation unit Elsevier Ltd. IFAC Papers OnLine 28 (8), 234–239. https://doi.org/10.1016/j.ifacol.2015.08.187.
Shao, W., Tian, X., 2017. Semi-supervised selective ensemble learning based on distance to model for nonlinear soft sensor development. Neurocomputing 91–104. https://doi.org/10.1016/j.neucom.2016.10.005.
Shao, W., Tian, X., Wang, P., 2015. Supervised local and non-local structure preserving projections with application to just-in-time learning for adaptive soft sensor. Chin. J. Chem. Eng 23 (12), 1925–1934. https://doi.org/10.1016/j.cjche.2015.11.012.
Shokri, S., Sadeghi, M.T., Marvast, M.A., 2014. High reliability estimation of product quality using support vector regression and hybrid meta-heuristic algorithms. J Taiwan Inst Chem Eng 45 (5), 2225–2232. https://doi.org/10.1016/j.jtice.2014.04.016.
Shokri, S., Sadeghi, M.T., Marvast, M.A., Narasimhan, S., 2015. Improvement of the prediction performance of a soft sensor model based on support vector regression for production of ultra-low sulfur diesel. Petrol. Sci. 12 (1), 177–188. https://doi.org/10.1007/s12182-014-0010-9.
Torgashov, A., Goncharov, A., Zhuravlev, E., 2018. Evaluation of steady-state and dynamic soft sensors for industrial crude distillation unit under parametric constraints, 51. IFAC-PapersOnLine Elsevier B.V, pp. 566–571. https://doi.org/10.1016/j.ifacol.2018.09.364, 18.
Ujević, Ž., Mohler, I., Bolf, N., 2011. Soft sensors for splitter product property estimation in CDU. Chem. Eng. Commun 198 (12), 1566–1578. https://doi.org/10.1080/00986445.2011.556692.
Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE Trans. Neural Network 10 (5), 988–999. https://doi.org/10.1109/72.788640.
Wang, Y., Chen, C., Yan, X., 2013. Structure and weight optimization of neural network based on CPA-MLR and its application in naphtha dry point soft sensor. Neural Comput. Appl. 22, 75–82. https://doi.org/10.1007/s00521-012-1044-9.
Wang, J., Yu, L.C., Lai, K.R., Zhang, X., 2016. Locally weighted linear regression for crosslingual valence-arousal prediction of affective words. Neurocomputing 194, 271–278. https://doi.org/10.1016/j.neucom.2016.02.057.
Yan, X., 2008. Modified nonlinear generalized ridge regression and its application to develop naphtha cut point soft sensor. Comput. Chem. Eng 32 (3), 608–621. https://doi.org/10.1016/j.compchemeng.2007.04.011.
Yan, X., 2010. Hybrid artificial neural network based on BP-PLSR and its application indevelopment of soft sensors. Chemometr. Intell. Lab. Syst 103 (2), 152–159. https://doi.org/10.1016/j.chemolab.2010.07.002.
Yan, W., Shao, H., Wang, X., 2004. Soft sensing modeling based on support vector machine and Bayesian model selection. Comput. Chem. Eng 28 (8), 1489–1498. https://doi.org/10.1016/j.compchemeng.2003.11.004.
Yuan, X., Ge, Z., Song, Z., 2014. Locally weighted kernel principal component regression model for soft sensing of nonlinear time-variant processes. Ind. Eng. Chem. Res. 53 (35), 13736–13749. https://doi.org/10.1021/ie4041252.
Yuge, N., Tanaka, K., Kaneko, H., Funatsu, K., 2018. Selective use of adaptive models considering the prediction efficiencies. Ind. Eng. Chem. Res. 57 (42), 14286–14296. https://doi.org/10.1021/acs.iecr.8b01171.
Zeng, J., Xie, L., Gao, C., Sha, J., 2011. Soft sensor development using non-Gaussian Just-In-Time modeling. IEEE Conf. Decis. Control Eur. Control Conf. IEEE 5868–5873. https://doi.org/10.1109/CDC.2011.6160693.
Zhong, Y., Zhao, L., Liu, Z., Xu, Y., Li, R., 2010. Using a support vector machine method to predict the development indices of very high water cut oilfields. Petrol. Sci. 7 (3), 379–384. https://doi.org/10.1007/s12182-010-0081-1.