CI&CEQ
Chemical Industry & Chemical Engineering Quarterly

Available on line at
Association of the Chemical Engineers of Serbia *AChE*
www.ache.org.rs/CICEQ

Chem. Ind. Chem. Eng. Q. 21 (3) 379–390 (2015) CI&CEQ

SAEID SHOKRI[1]
MOHAMMAD TAGHI SADEGHI[1]
MAHDI AHMADI MARVAST[2]
SHANKAR NARASIMHAN[3]

[1]Department of Chemical Engineering, Iran University of Science and Technology (IUST), Tehran, Iran
[2]Process & Equipment Technology Development Division, Research Institute of Petroleum Industry (RIPI), Tehran, Iran
[3]Department of Chemical Engineering, IIT Madras, Chennai, India

SCIENTIFIC PAPER

UDC 66.094.522:004.932

DOI 10.2298/CICEQ140418039S

INTEGRATING PRINCIPAL COMPONENT ANALYSIS AND VECTOR QUANTIZATION WITH SUPPORT VECTOR REGRESSION FOR SULFUR CONTENT PREDICTION IN HYDRODESULFURIZATION PROCESS

Article Highlights

- Designing of a reliable data-driven soft sensor to predict sulfur content in HDS Process
- Integrating Principal Component Analysis (PCA) and Vector Quantization (VQ) with SVR
- Comparing between PCA and VQ methods on prediction accuracy of support vector regression model
- Improving prediction accuracy and computation time of SVR model by proposed approach

Abstract

An accurate prediction of sulfur content is very important for the proper operation and product quality control in hydrodesulfurization (HDS) process. For this purpose, a reliable data-driven soft sensor utilizing Support Vector Regression (SVR) was developed and the effects of integrating Vector Quantization (VQ) with Principle Component Analysis (PCA) were studied in the assessment of this soft sensor. First, in the pre-processing step the PCA and VQ techniques were used to reduce dimensions of the original input datasets. Then, the compressed datasets were used as input variables for the SVR model. Experimental data from the HDS setup were employed to validate the proposed integrated model. The integration of VQ/PCA techniques with SVR model was able to increase the prediction accuracy of SVR. The obtained results show that integrated technique (VQ-SVR) was better than (PCA-SVR) in prediction accuracy. Also, VQ decreased the sum of the training and test time of SVR model in comparison with PCA. For further evaluation, the performance of VQ-SVR model was also compared to that of SVR. The obtained results indicated that VQ-SVR model delivered the best satisfactory predicting performance (AARE = 0.0668 and R^2 = 0.995) in comparison with investigated models.

Keywords: Principal Component Analysis (PCA), Vector Quantization (VQ), Support Vector Regression (SVR), soft sensor, hydrodesulfurization (HDS) process.

HDS process is one of the key catalytic units that play a major role in most petroleum refineries. High concentration of sulfur in HDS product has a negative impact on the refining processes, health and the environment. To increase process performance, it would be necessary to manage the ultra low sulfur content in *final* product of this process. Hence, sulfur content prediction is very important for the proper operation of HDS units [1]. Furthermore, more stringent environmental restrictions give remarkable importance to have an accurate prediction of sulfur content in the HDS process [2]. For this purpose, either hardware on-line analyzers or analytical laboratory tests are used in HDS units. Hardware on-line

Correspondence: M.T. Sadeghi, Department of Chemical Engineering, Iran University of Science and Technology (IUST), Tehran, Iran.
E-mail: sadeghi@iust.ac.ir
Paper received: 18 April, 2014
Paper revised: 7 September, 2014
Paper accepted: 29 October, 2014

analyzers are too expensive. Moreover, operators and engineers find many problems such as calibration necessity, insufficient accuracy and long dead time when hardware sensors are used. Moreover, analytical laboratory tests are tedious and unreliable. Soft sensors (soft analyzers) are key technologies for managing high quality products when hardware process analyzers are not available. Soft sensors can also be applied as an alternative to laboratory tests.

A soft sensor is a predictive model that describes the relationship between the predicted process variables and the measured variables. A soft sensor model can be developed by using either model-driven approaches or data-driven approaches. Model-driven soft sensors are based on first principle mathematical models. First principle models describe the physical and chemical background of the process. They are often not available due to complexity of industrial processes. These models, obtained from the fundamental process knowledge, require a lot of process expert knowledge, effort and time to be developed. Data-driven models operate based on actual data measured within the operational plants, and describe the real process conditions [3]. Furthermore, process data have become widely available in modern industrial plants.

Data-driven soft sensors can be applied to the online estimation of product indices using process measurement data because they have become widely available in many chemical plants [4,5]. Unlike physical sensors, which directly measure the value of a variable, data driven-soft sensors measure the process variables whose direct measurements are associated with some technical problems. Therefore, soft sensors use the frequently sampled process variables such as temperature, pressure, flow rate, etc. to measure these hard to measure variables. In these processes, machine learning techniques are frequently used.

Support vector machines (SVM) are efficient machine learning techniques derived from statistical learning theory by Vapnik [6]. Compared with artificial neural networks (ANN), an SVM provides more reliable and better performance [7]. The SVM is used for classification and regression tasks. When applied for regression tasks, SVM is also called SVR [8].

The linear dependency between different variables in the dataset influences the generalization ability of the SVR model [9]. Moreover, due to encountering large datasets in process industries, the training time increases for SVR model. To tackle these problems, this paper uses the data compression techniques such as PCA or VQ. These methods can be used to generate a smaller training set with greatly reduce training time. Furthermore, integration of data compression techniques with SVR can strengthen the generalization ability of the SVR model and therefore increase prediction performance.

PCA has been used as a pre-processing step during SVR modeling to reduce dimensionality of the original multivariable dataset. The integrated PCA and machine learning methods showed good performance in various prediction fields, such as assessment of coronary artery diseases [10], forecasting greenhouse gas emissions [11] and predicting gasoline homogeneous charge compression ignition combustion behavior during transient operation [12]. VQ is a data compression method based on the principle of block coding. The purpose of using this technique is to simplify the training set and to reduce training time [13]. A few studies have been carried out that investigate the effect of using compression techniques with a number of machine learning algorithms. Moreover, no comparison has been carried out between the effects of VQ and PCA on the SVR model in literature. In this work PCA and VQ have been integrated with SVR to predict the sulfur content of the treated gas oil. To train and test SVR model, data collection is carried out from a HDS setup.

The main objectives of the present study are: 1) designing an accurate and reliable data-driven prediction model for sulfur content prediction in HDS process; 2) aplying a novel integrated technique using VQ/PCA and SVR model to increase the prediction performance; 3) comparing prediction accuracy and CT of both integrated models (VQ-SVR & PCA-SVR).

MATERIALS AND METHODS

Experimental setup

The main control index in product quality of hydrodesulfurization (HDS) process is the sulfur content. Accurate prediction of sulfur content plays an important role in this process. In this study, a HDS setup was used to obtain experimental datasets. The schematic diagram of the system is depicted in Figure 1. Gas oil stream is entered into the preheated section *via* a dozing pump. Then, preheated stream passes through a fixed bed reactor (trickle-bed), in which hydrogenation occurs. The selected catalyst is Co-Mo/Al_2O_3. The output liquid from the bottom of the reactor enters into a high pressure separator in which it is separated to treated gas oil and H_2S. H_2S is then absorbed by NaOH solution in caustic column vessel. The operating conditions and catalyst specifications of the mentioned setup are shown in Table 1.

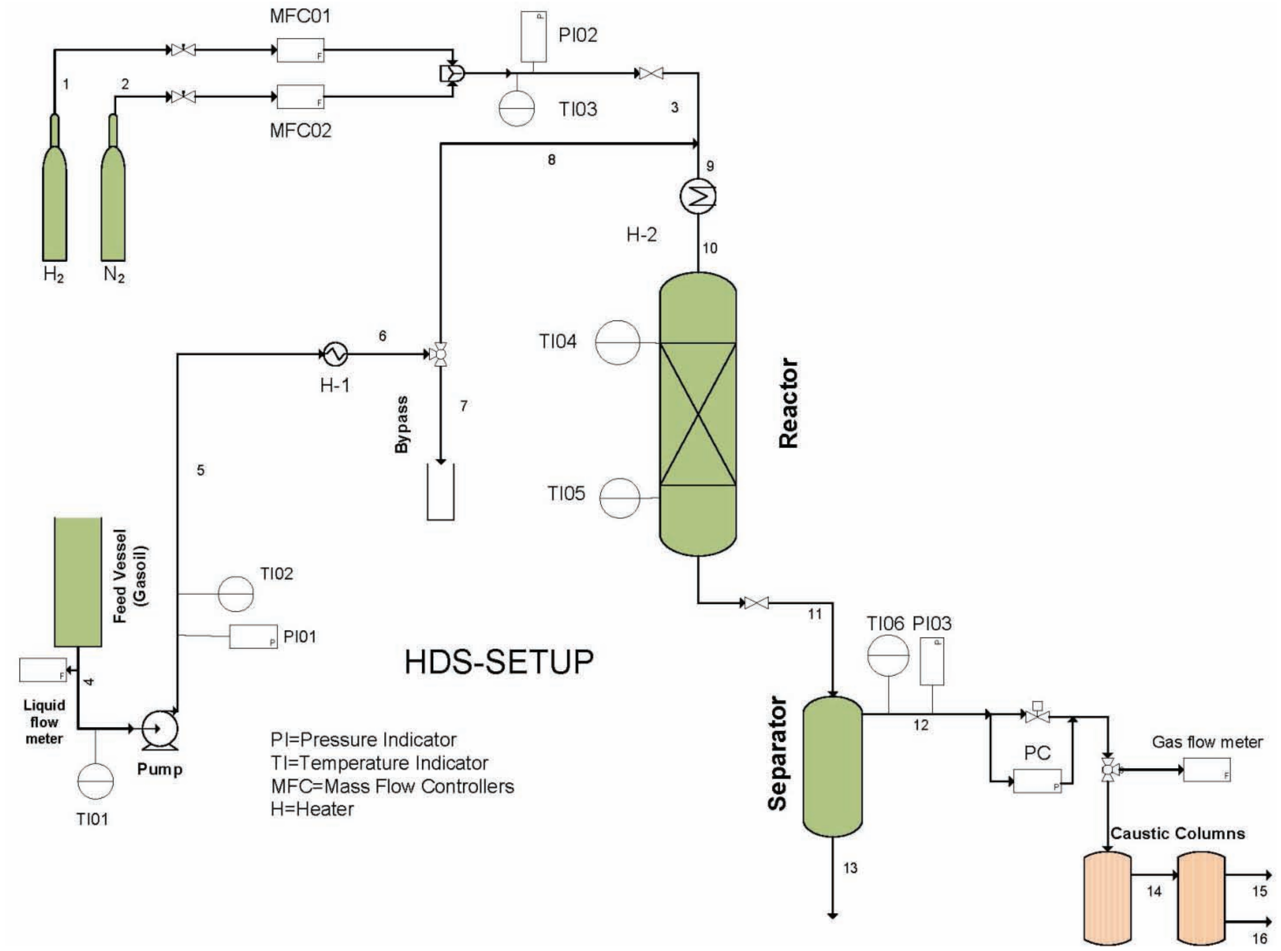

Figure 1. A schematic diagram of the diesel oil HDS setup.

In all experiments, the gas oil with total sulfur content of 7200 ppm was used for the feed stream. In this setup, the reactor diameter and the reactor length were 0.0127 and 0.63 m, respectively. Also, catalyst bed length of 0.11 m was selected.

Table 1. Setup specification

Component	Value
Reactor:	
Reactor diameter, m	0.0127
Reactor length, m	0.63
Catalyst bed length, m	0.11
Catalyst:	
Chemical composition, wt%, dry basis	
Cobalt	3.4
Molybdenum	13.6
Physical properties:	
Surface area, m^2/g	235
Pore volume, cm^3/g (H_2O)	0.53
Flat plate crush strength, N/cm (lb/mm)	200
Attrition index	99
Compacted bulk density, g/cm^3	0.72

In order to train and test the SVR model, a set of experiments were designed for the pilot. Experimental design for wide range of sulfur content was done. The minimum and maximum sulfur content in the treated gas oil products were found to be 10 and 4900 ppm, respectively. The HDS conversion varies with changes in the operating conditions. The independent operating parameters that affect the hydrogenation procedure are: inlet reactor temperature (T), reactor pressure (p), H_2/oil ratio and liquid flow rate (Q). In this work, these operating conditions are considered according to real condition of the refineries.

Therefore, the effect of these parameters was studied using the following range of values: T, 320, 337, 353 and 370 °C; p, 50, 60 and 70 bar; H_2/oil ratio, 85, 100, 120, 140 and 170 nm^3/m^3; liquid flow rate, 0.2, 0.23, 0.26, 0.29 and 0.32 cm^3/min. Only one of the above parameters was allowed to change in every test. The samples are collected based on 4 h of operation under nearly steady state conditions. A time interval of 4 h was required to reach the next steady state experimental condition. Change of any operating conditions was carried out continuously and incrementally with moderate slope within time intervals of 4 h to avoid a shock to the system and prevent catalyst deactivation.

The operating data was recorded minute-by-minute over this time periods. Therefore, the datasets were obtained with 240 data for each sensor (Q_{H2}, $Q_{\text{gas oil}}$, T_{preheat}, T_{in}, T_{out} and p). While the sulfur content in the product stream was obtained from laboratory tests once every 4 h. Hence, in twenty four hours, only 3 treated gas oil samples are collected. Therefore, 294 samples for 98 working days were collected. As there was only one data for sulfur in time intervals of 4 h, the values of other parameters were averaged within the time intervals and finally 294 datasets were selected.

The location of these measurements is shown in Figure 1 as follows: hydrogen flow rates (1); feed diesel flow rate (4); preheat temperature (6); temperature of the inlet stream to reactor (10); temperature of the outlet stream from reactor (11); reactor pressure (12).

In order to determine total sulfur content in HDS product after reaching steady state condition, the treated gas oil samples were analyzed by two methods: 1) ASTM D4294 and 2) ASTM D5453.

The ASTM D4294 method is a standard test method to determine the sulfur content in petroleum product by energy dispersive X-ray fluorescence spectrometry. It is capable of determining sulfur over a wide range of concentrations. This test method provides precise measurement of total sulfur in petroleum product with a minimum of a sample preparation. The applicable concentration range for this method is 0.015 to 5 mass%. To determine the sulfur content of less than this range the ASTM D5453 method is used. The ASTM D5453 is an ultra low sulfur analysis that uses ultraviolet fluorescence to determine the sulfur content in ultra low sulfur diesel. Liquid samples were collected for analysis every four hours. A typical analysis time is 5 min per sample after calibration and standardization.

Principal Component Analysis (PCA)

PCA is one of the multivariate statistical methods that are widely used to find a low dimensional representation of data matrix [14-16]. The PCA method is designed to transform a large set of interrelated independent variables into, uncorrelated new variables (axes), also known as principal components (PCs) [17,18]. In this method, the information of input variables will be presented without much loss of information [19,20]. This approach can be implemented as follows:

For a given p-dimensional dataset X, the m principal axes $T_1, T_2,..., T_m$, where $1 \le m \le p$, generally, $T_1, T_2,..., T_m$ can be determined by the m leading eigenvectors of the sample covariance matrix:

$$S = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^T(x_i-\mu) \tag{1}$$

where $x_i \in X$, μ is mean of the sample, N is the number of data points in the sample, so that:

$$ST_i = \lambda_i T_i,\ i \in 1,2,...,m \tag{2}$$

where λ_i is the i-th largest associated eigenvalue of the sample covariance matrix. The m principal components of a given observation vector $x_i \in X$ is given by:

$$y = \left[y_1, y_2, ..., y_m\right] = \left[T_1^T x, T_2^T x, ..., T_m^T x\right] = T^T x \tag{3}$$

where y is the vector containing the m principal components of x. Thus, each original data vector can be represented by its principal component vector with dimensionality m. The eigenvectors with the highest eigenvalues are projected into space. In mathematical terms, PCA involve the following major steps:

1) Standardization of the variables $X_1, X_2, ..., Xp$ by the Z matrix:

$$Z_{ij} = \frac{\left(x_{ij} - \bar{x}_j\right)}{s_j} \quad \text{for } i = 1,2,...,n \text{ and } j = 1,2,..,q \tag{4}$$

where $\bar{x}_j$ and s_j are, respectively, the mean and standard deviation of the generic variable x_j.

2) Calculation of the Kaisere-Meyere-Olkin (KMO): KMO index can be between 0 and 1. The index value 0.5 or more is considered determines the PCA can act efficiently [21]:

$$KMO = \frac{\sum\sum r_{ij}^2}{\sum\sum r_{ij}^2 + \sum\sum a_{ij}^2} \tag{5}$$

where r_{ij} is the correlation coefficient between variables i and j and a_{ij} is the partial correlation coefficient between them.

3) Calculation of the difference between the sample values and the means of input data set.

4) Calculation of the variance-covariance matrix.

5) Calculation of the eigenvalues and eigenvectors using covariance matrix.

6) Determination of principal components (PCs).

Vector quantization (VQ) method

Quantization is a survey process from an unlimited set of scalar or vector quantities by a limited set of scalar or vector quantities. Two types of quantization techniques exist: scalar quantization (SQ) and VQ. SQ deals with the quantization of samples on a sample by sample basis, while VQ deals with quantizing the samples in groups called vectors. VQ is a data compression method based on the principle of block coding. Using VQ, the training time for input parameters of predictive model is greatly reduced. The most influential gains here are the robustness of such systems [22].

The prediction speed is very important in soft sensors design. Therefore, In order to speed up the training time and reliability prediction of SVR model, the VQ technique is applied for data compression. The main goal of this method is to simplify the training set and increasing the prediction accuracy.

VQ method reduces the size of the training dataset. In Vector quantization the data is quantized in the form of contiguous blocks called vectors rather than individual samples. The VQ maps a vector **x** of *K*-dimensional in the vector space $\mathbf{R}^k$ to another vector **y** of *K*-dimensional that belongs to a finite set **C** (code book) of output vectors (code words) [23]. In this method *K*-dimensional input vectors are derived from input data $\{X\} = \{x_i: i = 1,2,...,N\}$. Data vectors are quantized into a finite set of *code words* $\{Y\} = \{y_j: j = 1,2,...,K\}$. Each vector **yj** is called a code vector or a codeword and the set of all the code words is called a code book where the overall distortion of the system should be minimized. The purpose of the generated code book is to provide a set of vectors which generate minimal distortion between the original vector and the quantized vector.

Therefore, VQ comprises of three stages: 1) code book generation, 2) vector encoding and 3) vector decoding. It works by encoding values from a multidimensional vector space into a finite set of values from a discrete subspace of lower dimension.

The generation of code book is the most important process that decides the performance of vector quantization. The aim of code book generation is to find code vectors (code book) for a given set of training vectors by minimizing the average pair-wise distance between the training vectors and their corresponding code words [24].

Each vector is compared with a collection of representative code vectors, $\hat{X}_i, i = 1,...,N_c$ taken from a previously generated code book. Best match code vector is chosen using a minimum distortion rule. To minimize the distortion, the following formula is used to determine the distance between two code words:

$$d(X,\hat{X}) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \tag{6}$$

where $d(X,\hat{X})$ denotes the distortion incurred in replacing the original vector X with the code vector $\hat{X}$.

Support Vector Regression (SVR)

The SVR is a computational tool that has recently received much attention in soft sensor design due to its successes in building nonlinear data-driven models [26]. SVR has more popularity over ANN due to having many attractive features and promising empirical performance. The main difference between conventional ANN and SVR lies in the risk minimization principle. Conventional ANN implement the empirical risk minimization (ERM) principle to minimize the error on the training data, while SVR models are based on the Structural Risk Minimization (SRM) principle which equips them with greater potential to generalize. Therefore the most important features of SVR are: 1) excellent generalization capability, 2) solving the high-dimension problems, 3) avoiding from local minima and the over fitting phenomenon, 4) does not require to determine network topology in advance and 5) needs fewer a priori-determined parameters than ANN. These aspects of SVR make it a more generalizable tool, permitting more robust prediction despite a small number of learning samples [27-29].

SVR is utilized to determine a nonlinear relation of the form $y = f(x)$ between the vectors of observation x and the desired y from a given set of training samples. A number of cost functions such as the Laplacian, Huber's, Gaussian, and ε-insensitive can be used for the SVR formulation. Among these, the robust ε-insensitive loss function (L_ε) is more common [30, 31]:

$$L_\varepsilon(f(x)-y) = \begin{cases} \left|f(x)-y\right| - \varepsilon & \left|f(x)-y\right| \geq \varepsilon \\ 0 & \text{Otherwise} \end{cases} \tag{7}$$

where ε is a precision parameter representing the radius of the tube located around the regression function, $f(x)$ (Figure 2).The goal in using the ε-insensitive loss function is to find a function that fits the current training data with a deviation less than or equal to ε. C and ε are user-defined parameters in the empirical analysis. A penalty parameter $C > 0$ is a parameter determining the trade-off between generalization ability and accuracy in the training data, while the parameter ε defines the degree of tolerance to errors. The optimization problem can be reformulated as:

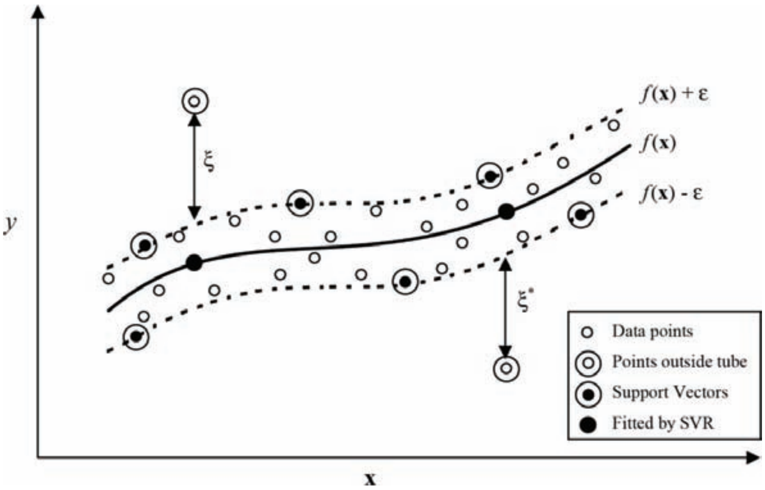

Figure 2. A schematic diagram of SVR using ε-sensitive loss function (with permission from the publisher [33]).

$$\mathrm{Min}\frac{1}{2}\left\|w\right\|^2 + C\sum_{i=1}^{l}\left(\xi_i^- + \xi_i^+\right) \tag{8}$$

That is subject to the constraints given below:

$$y = \begin{cases} y_i - \left(\left\langle w, x_i \right\rangle + b\right) \le \varepsilon + \xi_i \\ \left(\left\langle w, x_i \right\rangle + b\right) - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$

The positive slack variables ξ and ξ^* represent the distance from actual values to the corresponding boundary values of the ε-tube. The nonlinear model involves the inner products of feature vectors in a high dimensional feature space. For SVR based models, four different kernel functions, including linear, quadratic, Gaussian, and polynomial are used. Generally, the application of Gaussian function is shown to yield a better prediction performance [32]:

$$K(x, y) = \exp\left(-\frac{\left\|x_i - y_j\right\|^2}{2\sigma^2}\right) \tag{9}$$

In order to build a SVR model efficiently, the SVR parameters must be specified carefully. These parameters include the kernel function, regularization parameter C, bandwidth of the kernel function (σ^2) and the tube size of ε-insensitive loss function (ε).

Model development

Figure 3 represents the structure of the proposed method. The proposed system consists of two stages. One is the development a pre-processing step based on the VQ or PCA. The other is the SVR model and estimation the trained model.

In the first stage, experimental data taken from the setup were divided into two distinct sets including training and testing data. From a total of 294 experimental data, 240 data were used for training, and 54 for the testing set. Then, the PCA/VQ technique was implemented on the training and testing data.

In the second stage, the *k* fold cross-validation technique was employed to solve the over fitting problem of the training data. The training dataset was randomly partitioned into *k* subsets (folds) of approximately equal size. Next, *k*-1 subsets were used for training the model with the selected set of parameters while the model performance was measured by the only remaining subset (validation dataset). This procedure was repeated *k* times in a way that each subset was used as a validation subset once the others performed the role of training dataset. Finally, the overall model generalization ability for each set of parameters was estimated by averaging the performance values obtained over the *k* trails. In this paper, *AARE* is selected as the performance criteria whereas a 5-fold cross-validation is utilized to evaluate the performance of the model. Meanwhile, the input space is transformed into feature space by means of the Radial Basis Function (RBF) kernel.

In this study, the LIBSVM package was employed for developing SVR model. The implementation was carried out in MATLAB 7.10 simulation software [34]. The experimental results were obtained using a personal computer equipped with Intel (R) Core (TM) 2 CPU (3.0 GHz) and 3.25 GB of RAM.

The main steps of model development were as follows:

1. Dividing data into two parts, namely training and test data;
2. Applying cross validation technique (the 5-fold cross validation technique was used);
3. Selecting support vector machine (the ε-SVR model was used);

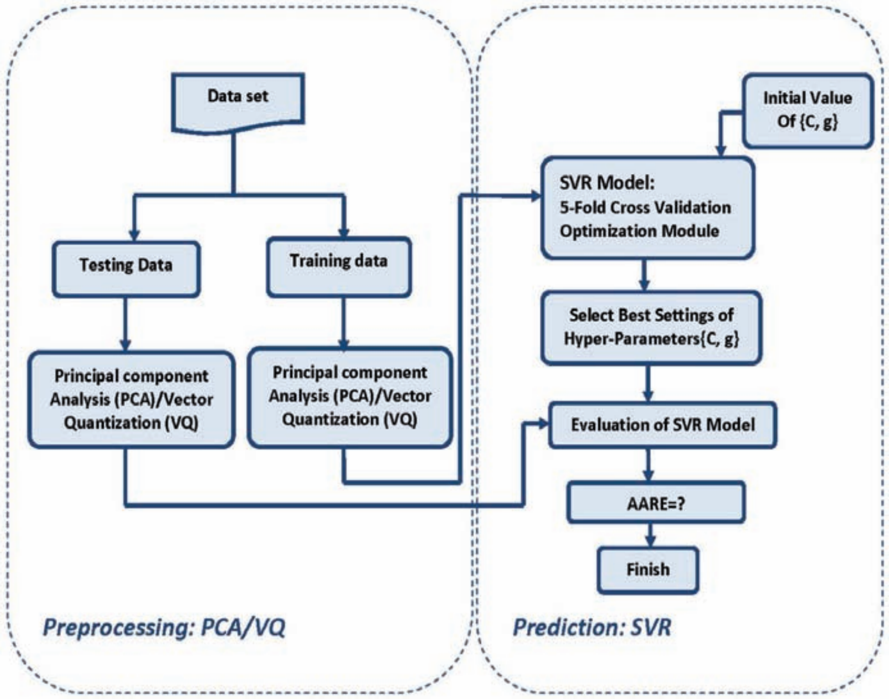
Data set
Testing Data
Training data
Principal component Analysis (PCA)/Vector Quantization (VQ)
Principal component Analysis (PCA)/Vector Quantization (VQ)
Initial Value Of {C, g}
SVR Model: 5-Fold Cross Validation Optimization Module
Select Best Settings of Hyper-Parameters{C, g}
Evaluation of SVR Model
AARE=?
Finish
Preprocessing: PCA/VQ
Prediction: SVR

Figure 3. Block diagram of the proposed integrated model and SVR.

4. Selecting the type of core kernel (Gaussian kernel was used);

5. Optimizing model parameters (*C* and *g* (1/2σ^2) using pattern search (PS) algorithm;

6. Validating of model and prediction of the results.

Model comparison criteria

The performance criteria and their calculations for comparison of different approaches including average absolute relative error (*AARE*) and squared correlation coefficient (R^2) were applied:

$$AARE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(y_i-\hat{y}_i)}{y_i}\right| \tag{10}$$

$$R^2 = 1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}, \bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i \tag{11}$$

RESULTS AND DISCUSSIONS

Tuning parameters of SVR model

For the SVR model, the Grid Search Method (GSM) is the most common method to determine appropriate values of hyper parameters [35]. The GSM does not consider all the values for parameters in parameters space. Hence, it is unable to converge to the global optimum. The accuracy of the GSM depends on the parameter range in combination with the chosen interval size. Therefore, this method is quite time-consuming and depends on the selection of boundary parameters. Sensitivity analysis of the hyper parameters of the SVR model is shown in Figure 4.

In this figure, the *x* and *y*-axes are *l*og$_2$ *C* and log$_2$ *g*, respectively. The *z*-axis is the *AARE*. In this study, a typical wide range is selected that could cover a broad range of analysis with adequate incremental size which is not too big to reduce the accuracy and not too small to make CT too long. log$_2$ *C and* log$_2$ *g* varied within [-5,20] and [-30,20], respectively, with incremental sizes of 3 and 2. It can be seen that by changing the hyper parameters (*C*,*g*), the *AARE* varies in a wide range. Since ε has little effect on *ARRE* it is assumed to be 0.01. It shows that tuning of the hyper parameters greatly influences the prediction accuracy.

Since the accuracy of SVR model depends on a proper setting of SVR hyper-parameters [36], pattern search (PS) has been used in conjunction with SVR to find an optimum set of hyper parameters for SVR model [37].

The PS method is a class of direct search method to solve nonlinear optimization problems. The algorithm calculates the function values of a pattern and tries to find the minimum. For hyper-parameter optimization with PS algorithm the procedures was summarized as follows:

1. Parameters setting, set iteration *i* = 0.

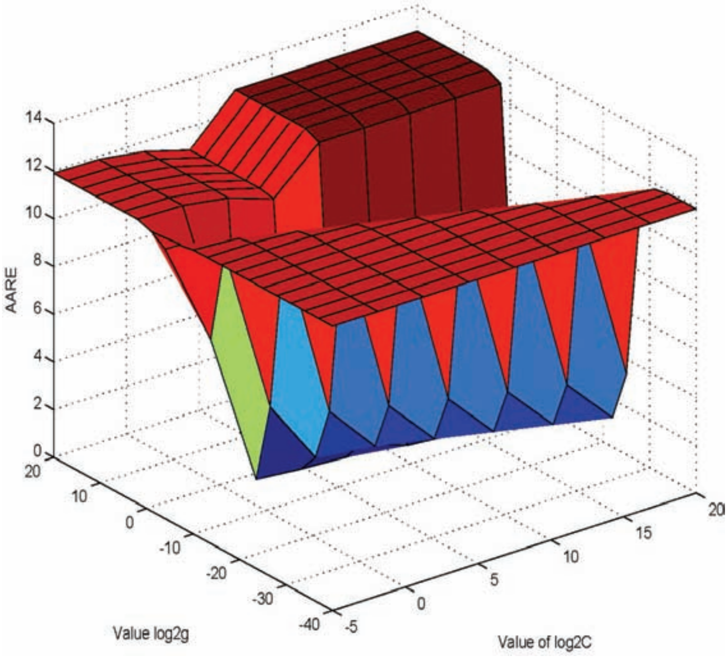

Figure 4. Sensitivity analysis of the hyper-parameters for SVR model.

2. Set iteration $i = i+1$

3. Model training: Hyper-parameter optimization.

4. Fitness definition and evaluation.

5. Termination: the evolutionary process proceeds until stopping criteria (maximum iterations predefined or the error accuracy of the fitness function is met). Otherwise, go to step 2.

The PS optimization procedure is a relatively fast alternative for the time consuming grid search approach. For choosing an overall optimal hyper-parameter, the *AARE* criterion was least for the test set. The optimal values of the C and g $(1/2\sigma^2)$ were obtained to be 20 and 0.017, respectively.

Integrated model results

This paper proposes a novel soft sensor model by integrating a data reduction technique (VQ or PCA) with SVR. The main idea is based on application of VQ or PCA for generation of a smaller training dataset. In our approach, these two techniques were used in the pre-processing step to make SVR model more effective. The major advantage of this approach is to train the model using the extracted low-dimensional datasets.

Before utilizing the PCA technique, The *KMO* index was applied to find out the applicability of PCA. The obtained value of the *KMO* index was 0.6978, which was above 0.5 and hence PCA implementation was feasible.

Table 2 shows the variance distribution of PCs (PC1-PC6). It is clear from this table that the cumulative variance of PC1 to PC4 is 99.98%. In this study, the first four principal components have got more than 95% of the total cumulative variance. Therefore the first four principal components will be sufficient to develop the model and therefore they were selected as the main model parameters. It was observed that only PC5 and PC6 were insignificant within all variables. Moreover, Figure 5 shows Pareto Chart in MATLAB software in which only the first four PCs that have more than 95% of the cumulative distribution were depicted.

Table 2. Descriptive statistics of the created PCs

PCs	Variance	Variance proportion, %	Cumulative variance proportion, %
1	0.4549	47.2042	47.2042
2	0.2276	23.6176	70.8218
3	0.1549	16.0770	86.8988
4	0.1261	13.0878	99.9866
5	0.0001	0.0099	99.9965
6	0.0000	0.0034	100

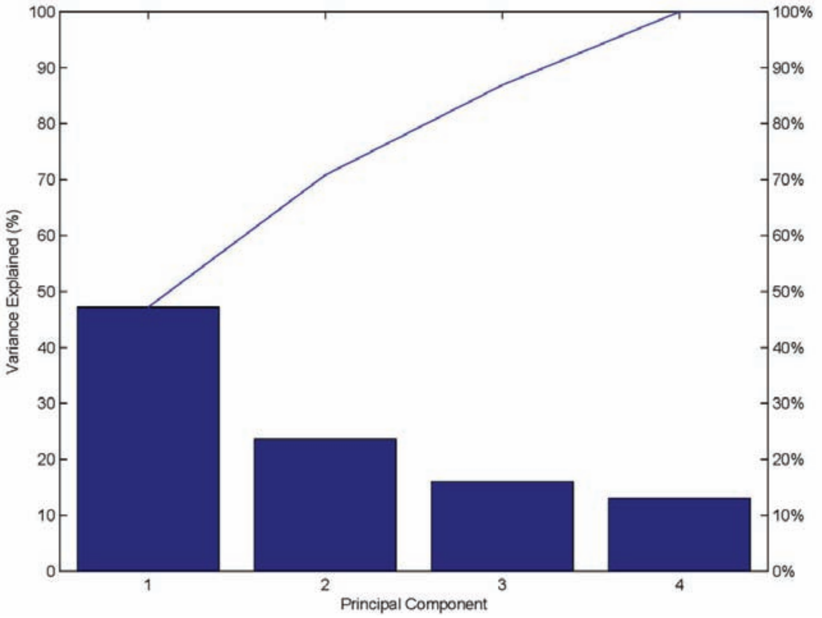

Figure 5. Pareto chart for the first four PCs (PC1-PC4).

The main contribution in this work was to implement VQ technique to reduce the dimension feature vector of the training dataset. This technique can be used to reduce the computation time (CT) for soft sensor model, thus soft sensor model was trained on a low-dimensional dense datasets. The aim of applying this method was to reduce the dimension of the training dataset in order to reduce the training time. Therefore, using VQ technique leads to higher accuracy for sulfur content prediction.

Typical results by integrated VQ-SVR method are shown in the Table 3. As can be seen in Table 3, integration of VQ with SVR model has good prediction accuracy for the treated gas-oil sulfur content in a wide range.

The proposed integrated approaches were compared with SVR. The validity of these methods was evaluated by the statistical parameters (*AARE* and R^2) in Table 4. Moreover, the parity plots for different optimization algorithms integrated with the SVR model are shown in Figure 6.

Comparison of three methods shows that the VQ-SVR model is more accurate and faster than PCA-SVR in predicting sulfur content. Also both approaches (VQ-SVR and PCA-SVR) exhibit high performance in accuracy and CT compared to the

Table 3. Typical input and output data for proposed integrated VQ-SVR method

Test No.	$Q_{\text{gas oil}}$ / 10^{-6} kg s^{-1}	Q_{H2} / 10^{-5} m^3 h^{-1}	p / kPa	T_{Preheat} / °C	T_{in} / °C	T_{out} / °C	S_{exp} / ppm	S_{pre} / ppm
1	4.56	3.55	5000	170.0	370.0	374.0	151	160
2	2.85	3.55	5000	120.8	320.0	324.3	3669	3392
3	3.71	3.55	5000	119.5	320.0	322.3	4362	3904
4	3.71	6.91	5000	145.6	345.0	347.0	1393	1176
5	3.71	11.3	5000	169.2	370.0	376.5	64	74
6	4.56	3.55	5000	120.3	320.0	322.2	4845	4469
7	4.56	3.55	5000	144.3	345.0	348.6	1948	1524
8	2.85	3.55	7000	119.4	320.0	324.2	3738	3644
9	2.85	3.55	7000	144.2	345.0	348.7	924	843
10	4.56	11.3	7000	144.4	345.0	347.4	2116	2063
11	4.56	11.3	7000	169.8	370.0	372.3	211	307
12	3.71	11.3	7000	119.7	320.0	323.2	4382	4360
13	3.71	6.91	7000	145.0	345.0	348.2	1546	1287
14	4.56	3.55	7000	145.9	345.0	347.7	2113	2473
15	2.85	11.3	7000	119.5	320.0	320.4	3681	3650

Table 4. Effect of VQ and PCA on predictive performance of SVR model

No.	Method	R^2		*AARE* / %		CT(s)
		Training data	Test data	Training data	Test data	
1	SVR	0.978	0.970	16.74	22.32	243
2	PCA-SVR	0.989	0.988	9.57	11.86	126
3	VQ-SVR	0.997	0.995	5.32	6.68	85

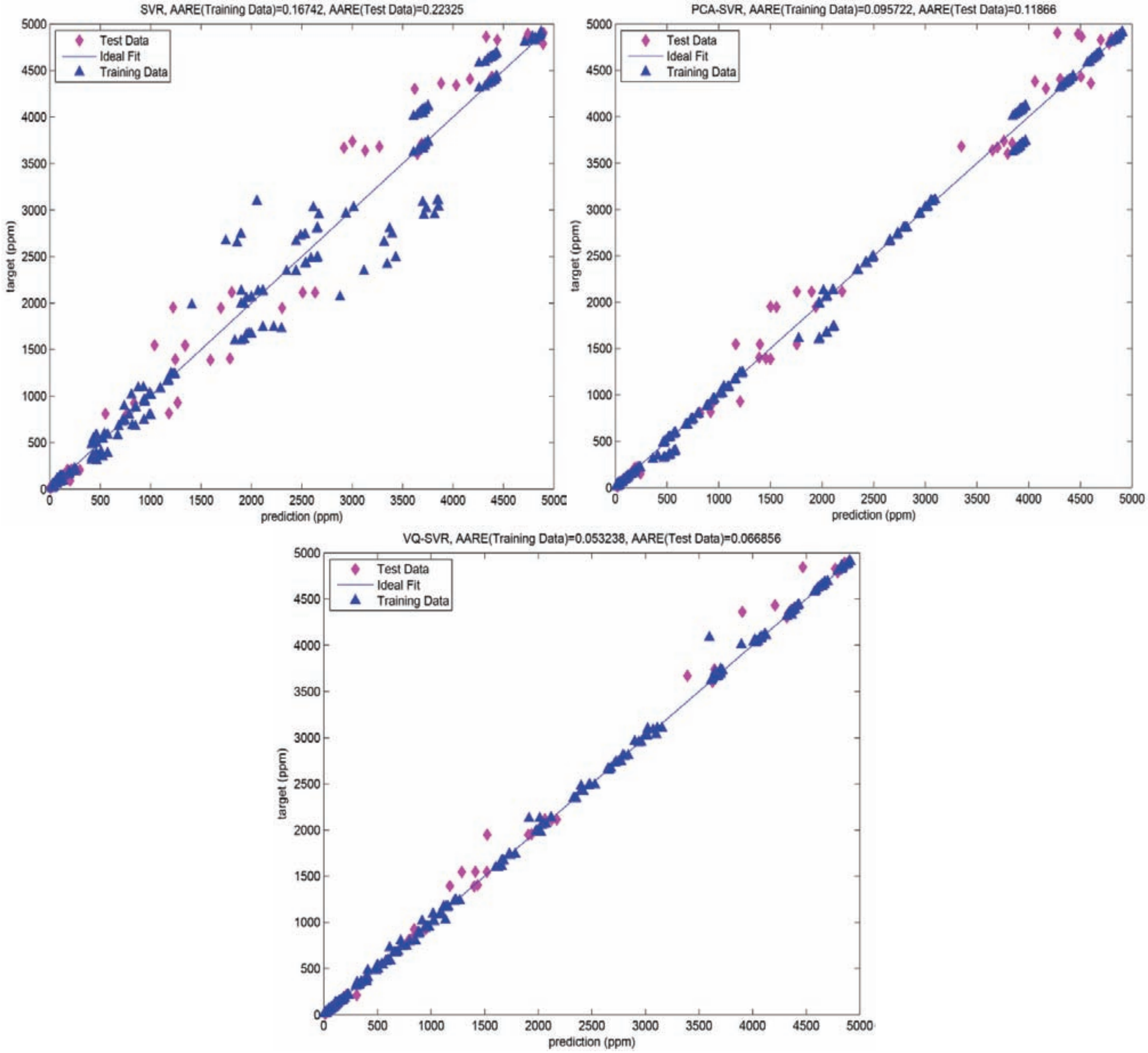

Figure 6. The parity plots for different models.

conventional algorithm based on SVR. The PCA reduced the dataset used as an input to SVR method and increased the accuracy. Obviously, the reduction of the input vector dimensions has resulted in the reduction of SVR size and hence has decreased the training and testing time. However, VQ performed greatly in accuracy prediction of sulfur content and training time of datasets for SVR model.

CONCLUSION

Determination of ultra low sulfur content of treated gas oil in HDS process is highly important to increase the productivity and efficiency of refinery operations. Sulfur content prediction by online hardware analyzers is mostly expensive with high maintenance cost. SVR is a new model developed for soft sensor applications in process engineering. However, the industrial applications have different approaches. Due to the huge size of industrial data sets used by soft sensors, training and validation time of SVR model has become a challenging issue. In this paper, an accurate and reliable data-driven soft sensor has been developed by means of a SVR integrated with a data compression technique (VQ/PCA) to predict the sulfur content in an industrial HDS process. The proposed integrated technique incorporated two stages: 1) the data compression stage and 2) the prediction stage. First, the PCA and VQ were applied to reduce the dimensionality of the dataset and then a SVR model was developed. In order to evaluate the performance of the proposed approach, a wide range of experimental data according to real condition of the refineries were taken from a HDS setup. Therefore, the results are generalizable to real processes in the refinery.

The performance of both VQ-SVR and PCA-SVR methods were compared with that of SVR. Some statistical criteria (*AARE*, R^2) were used to evaluate the prediction performance of models. Comparison of the results indicated that the best prediction accuracy could be obtained using VQ-SVR method, *i.e.*, highest R^2 (0.995) and lowest *AARE* (0.0668). Moreover, it was observed that VQ-SVR had a shorter CT in comparison with PCA-SVR.

Abbreviations

AARE	Average absolute relative error
ASTM	American Society for Testing and Materials
KMO	Kaisere-Meyere-Olkin
CT	Computation time
SVR	Support vector regression
PCA	Principal Component Analysis
HDS	Hydrodesulfurization
PS	Pattern Search
GSM	Grid search method
VQ	Vector quantizaion
RBF	Radial basis function

Nomenclature

a_{ij}	Partial correlation coefficient
X	Code vector
$\hat{y}_i$	The predicted value
y_i	The observed values
r_{ij}	Correlation coefficient
$K(x,y)$	Kernel function
S_{exp}	Sulfur content from laboratory tests
S_{pre}	Predicted sulfur content
C	Regularization parameter (hyper-parameter)
k	Subsets (folds)
exp_i	Actual values
pre_i	Predicted values
L_ε	Loss function
PCs	Principal component
W	Weight vector
R^2	Squared correlation coefficient
T_m	m principal axes
b_0	Intercept
Q_{H2}	Hydrogen flowrate
$Q_{\mathrm{Gas\ oil}}$	*Gas oil* flowrate

Greek symbols and subscripts

μ	Mean of the sample
σ	Width of kernel of radial basis function
ε	Precision parameter (hyper-parameter)
ξ_i, ξ_i^*	Slack variables
λ_{i}	i-th largest associated eigenvalue

REFERENCES

[1] E.A. Medina, J.I.P. Paredes, Math. Comput. Modell. **49** (2009) 207-214

[2] F.S. Mederos, J. Ancheyta, Appl. Catal., A **332** (2007) 8--21

[3] P. Kadlec, R. Grbic, B. Gabrys, Comput. Chem. Eng. **35** (2011) 1-24

[4] H. Kaneko, K. Funatsu, Ind. Eng. Chem. Res. **50** (2011) 10643-10651

[5] J. Ji, H. Wang, K. Chen, Y. Liu, N. Zhang, J. Yan, J. Taiwan Inst. Chem. Eng. **43** (2012) 67-76

[6] V.N. Vapnik, The nature of statistical learning theory, Springer, New York, 1995, p.93

[7] G. Liu, D. Zhou, H. Xu, C. Mei, Expert. Syst. Appl. **37** (2010) 2708-2713

[8] P. Niu, W. Zhang, Neurocomputing **78** (2012) 64-71

[9] H. Son, Ch. Kim, Ch. Kim, Automat. Constr. **27** (2012) 60--66

[10] I. Babaoglu, O. Findik, M. Bayrak, Expert. Syst. Appl. **37** (2010) 2182-2185

[11] D.Z. Antanasijevic, M.D. Ristic, A.A.P. Grujic, V.V. Pocajt, Int. J. Greenhouse Gas Con. **20** (2014) 244-253

[12] V.M. Janakiraman, X. Nguyen, D. Assanis, Appl. Soft Comput. **13** (2013) 2375-2389

[13] G.R. Lloyd, R.G. Brereton, R. Faria, J.C. Duncan, J. Chem. Inf. Model. **47** (2007) 1553-1563

[14] L.I. Smith, A tutorial on principle component analysis, http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf (2002)

[15] K. Polat, S. Gunes, Expert Syst. Appl. **34** (2008) 2039--2048

[16] M. Aminghafari, N. Cheze, J. Poggi, Comput. Stat. Data Anal. **50** (2006) 2381-2398

[17] I.T. Jolliffe, Principal Component Analysis, 2nd ed., Springer, Berlin, 2002, p.29

[18] C. Sarbu, H. F. Pop, Talanta **65** (2005) 1215-1220

[19] J.D. Wu, C.T. Liu, Expert Syst. Appl. **38** (2011) 14284--14289

[20] I. Lindsay, A. Smith, A tutorial on principal components analysis, http://kybele.psych.cornell.edu/ edelman/Psych--465-Spring-2003/PCA-tutorial (2002)

[21] S. Shrestha, F. Kazama, Environ. Modell. Software **22** (2007) 464-475

[22] K. Sayood, Introduction to Data Compression, Fourth ed., Morgan Kaufmann, University of Nebraska, 2012, p.295

[23] A. Gersho, R.M. Gray, Vector Quantization and Signal Compression, Kluwer Academic Publishers, Boston, MA, 1992, p.309

[24] M.H. Horng, Expert Syst. Appl. **39** (2012) 1078-1091

[25] J. Yin, Procedia Environ. Sci. **8** (2011) 173-178

[26] S.B. Chitralekha, S.L. Shah, Can. J. Chem. Eng. **88** (2010) 696-709

[27] K.Y. Chen, Reliab. Eng. Syst. Saf. **92** (2007) 423-432

[28] V. Vapnik, S. Golowich, A. Smola, Advances in Neural Information Processing Systems, Cambridge, MA, Vol. 9, 1997, pp. 281-287

[29] C. Bergeron, F. Cheriet, J. Ronsky, R. Zernicke, H. Labelle, Eng. Appl. Artif. Intel. **18** (2005) 973-983

[30] F. Si, C.E. Romero, Z. Ya, Z. Xu, R.L. Morey, B.N. Liebowitz, Fuel Process. Technol. **90** (2009) 56-66

[31] *S*. Zaidi, Chem. Eng. Sci. **69** (2012) 514-521

[32] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, Cambridge, 2000, p. 112

[33] *S*. Nandi, Y. Badhe, J. Lonari, U. Sridevi, B.S. Rao, S.S. Tambe, B.D. Kulkarni, Chem. Eng. J. **97** (2004) 115-129

[34] C.C. Chang, C.J. Lin, LIBSVM: A Library for Support Vector Machines, Software,http://www. csie.ntu.edu.tw//cjlin/libsvm/, (2001)

[35] B. Scholkopf, A.J. Smola, Learning with kernels, MIT Press, Cambridge, 2002, p.187

[36] S.K. Lahiri, N.M. Khalfe, Chem. Ind. Chem. Eng. Q. **15** (2009) 175-187

[37] M. Momma, K.P. Bennett, in Proceedings of the Second SIAM International Conference on Data Mining, Arlington, VA, USA (2002), pp. 261-274.

SAEID SHOKRI[1]
MOHAMMAD TAGHI SADEGHI[1]
MAHDI AHMADI MARVAST[2]
SHANKAR NARASIMHAN[3]

[1]Department of Chemical Engineering, Iran University of Science and Technology (IUST), Tehran, Iran
[2]Process & Equipment Technology Development Division, Research Institute of Petroleum Industry (RIPI), Tehran, Iran
[3]Department of Chemical Engineering, IIT Madras, Chennai, India

NAUČNI RAD

PREDVIĐANJE SADRŽAJA SUMPORA U PROCESU HIDROSULFURIZACIJE INTEGRACIJOM ANALIZE GLAVNIH KOMPONENTI I VEKTORSKE KVANTIZACIJE SA PODRŽANOM VEKTORSKOM REGRESIJOM

U procesu hidrodesulfurizacije (HDS) veooma je važno precizno predvideti sadržaj sumpora, kako bi se obezbedili pravilan rad i kontrola kvaliteta proizvoda. U tu svrhu, razvijen je pouzdan soft sensor (virtualni senzor) koji koristi podrzanu vektorsku regresiju (SVR). Proučavan je efekat integrisanja vektorske kvantizacije (VK) i analize glavnih komponenti (PCA) na izvršenje ovog sensora. Kao prvo, u fazi prethodne obrade, PCA i VQ tehnike su korišćene za smanjenje dimenzije originalnih ulaznih podataka. Nakon toga, komprimovani podaci su korišćeni kao ulazne promenljive za SVR model. Eksperimentalni podaci iz HDS koraka su iskorišćeni za validaciju predloženog integrisanog modela. Integracija VQ/PCA tehnike sa SVR modelom povećava tačnost predviđanja samog SVR modela. Dobijeni rezultati pokazuju da je integrisana tehnika (VQ-SVR) bolja od PCA-SVR u predviđanju tačnosti. Takođe, VQ smanjuje ukupno vreme obuke i testiranja SVR modela u poređenju sa PCA. Za dalju procenu, performanse VQ-SVR modela su poređene sa SVR modelom. Dobijeni rezultati ukazuju da VQ-SVR model daje najbolje performanse u predviđanju (AARE= 0,0668 and R^2= 0,995) u poređenju sa analiziranim modelima.

Ključne reči: analize glavnih komponenti, vektorska kvantizacija, podrzana vektorska regresija, soft senzor, proces hidrodesulfurizacije.