Forecasting PVT properties of crude oil systems based on support vector machines modeling scheme

Emad A. El-Sebakhy *

Medical Artificial Intelligence (MEDai), Inc., an Elsevier Company Millenia Park One, 4901 Vineland Road, Suite 450 Orlando, Florida 32811, United States

A R T I C L E   I N F O

Article history:
Received 2 January 2007
Accepted 1 December 2008

Keywords:
Support Vector Machines Regression
feedforward neural networks
empirical correlations
PVT properties
Formation Volume Factor
bubble point pressure

A B S T R A C T

PVT properties are very important in the reservoir engineering computations. There are numerous approaches for predicting various PVT properties, namely, empirical correlations and computational intelligence schemes. The achievements of neural networks open the door to data mining modeling techniques to play a major role in petroleum industry. Unfortunately, the developed neural networks modeling schemes have many drawbacks and limitations as they were originally developed for certain ranges of reservoir fluid characteristics. This article proposes support vector machines a new intelligence framework for predicting the PVT properties of crude oil systems and solve most of the existing neural networks drawbacks. Both steps and training algorithms are briefly illustrated. A comparative study is carried out to compare support vector machines regression performance with the one of the neural networks, nonlinear regression, and different empirical correlation techniques. Results show that the performance of support vector machines is accurate, reliable, and outperforms most of the published correlations. This leads to a bright light of support vector machines modeling and we recommended for solving other oil and gas industry problems, such as, permeability and porosity prediction, identify liquid-holdup flow regimes, and other reservoir characterization.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Reservoir fluid properties are very important in petroleum engineering computations, such as, material balance calculations, well test analysis, reserve estimates, inflow performance calculations, and numerical reservoir simulations. The importance of accurate PVT data for material-balance calculations is well understood. Among those PVT properties is the bubble point pressure ($P_b$), Oil Formation Volume Factor ($B_o$), which is defined as the volume of reservoir oil that would produce one stock tank barrel. Precise prediction of $B_o$ is very important in reservoir and production computations. It is crucial that all calculations in reservoir performance, in production operations and design, and in formation evaluation be as good as the PVT properties used in these calculations. The economics of the process also depends on the accuracy of such properties.

The currently available PVT simulator predicts the physical properties of reservoir fluids with varying degree of accuracy based on the type of used model, the nature of fluid and the prevailing conditions. Nevertheless, they all exhibit the significant drawback of lacking the ability to forecast the quality of their answers. Ideally, the PVT properties are determined from laboratory studies on samples collected from the bottom of the wellbore or at the surface. Such experimental data are, however, very expensive to obtain. Therefore, the solution is to use equation of state, statistical regression, graphical techniques, and empirical correlations to predict these PVT properties. The development of correlations for PVT calculations has been the subject of extensive research, resulting in a large volume of publications. Several graphical and mathematical correlations for determining both $P_b$ and $B_o$ have been proposed during the last decade. These correlations are essentially based on the assumption that $P_b$ and $B_o$ are strong functions of the solution gas–oil ratio ($Rs$), the reservoir temperature ($T_r$), the gas specific gravity ($G_g$), and the oil specific gravity ($G_o$), as well as the oil ratio ($Bob$) and the solution gas–oil ratio ($Rs$). However, the equation of state involves numerous numerical computations that required knowing the detailed compositions of the reservoir fluids, which is expensive and time consuming. In addition, the successes of such correlations are not reliable in prediction and it depends mainly on the range of data at which they were originally developed and geographical area with similar fluid compositions and API oil gravity. Furthermore, PVT correlations are based on easily measured field data, such as, reservoir pressure, reservoir temperature, and oil and gas specific gravity.

Recently, artificial neural networks (ANNs) have been proposed for solving many problems in the oil and gas industry, including permeability and porosity prediction, identification of lithofacies types, seismic pattern recognition, prediction of PVT properties, estimating pressure drop in pipes and wells, and optimization of well production. The most popular neural networks in both machine learning and data mining communities are feedforward neural network (FFN) and multilayer perceptron (MLP), they have gained...
high popularity in oil and gas industry applications, Goda et al. (2003), Osman et al. (2001), Al-Marhoun and Osman (2002), Osman and Abdel-Aal (2002), and Osman and Al-Marhoun (2005). However, these neural networks modeling schemes suffer from a number of numerous drawbacks, such as, the limited ability to explicitly identify possible causal relationships and the time-consumer in the development of back-propagation algorithm, which lead to an overfitting problem and gets stuck at a local optimum of the cost function. In addition, the architectural parameters of ANN have to be guessed in advance, such as, number and size of hidden layers and the type of transfer function(s) for neurons in the various layers. Moreover, the training algorithm parameters were determined based on guessing initial random weights, learning rate, and momentum, see Castillo et al. (2002, 2006) for more details.

The main objective of this study is to investigate the capability of support vector machine regression (SVR) on modeling PVT properties of crude oil systems and solve some of the above neural networks limitations. The considerable amount of user intervention not only slows down model development, but also works against the principle of ‘letting the data speak’. In our research, we precisely investigate the capability of support vector machines regression in modeling both $P_b$ and $B_{ob}$ based on the kernel function scheme using worldwide published experimental PVT databases. Comparative studies are carried out to compare SVR performance with the one of the neural networks, nonlinear regression, and different empirical correlation techniques.

The potentiality prediction of enhanced oil recovery (EOR) is the basis of EOR potentiality analysis as well as the robust guarantee of the reliability of analysis results. In the light of statistical learning theory, establishing an EOR predictive model substantially falls within the problem of function approximation. According to Vapnik’s structural risk minimization principle, one should improve the generalization ability of learning machine, that is, a small error from an effective training set can guarantee a small error for the corresponding independent testing set. The up-to-date results from studies on statistical theory in recent years are firstly applied to EOR potentiality analysis. The improved backpropagation artificial neural network, and support vector machine are discussed. Comparative studies are carried out to compare SVR performance with the results of three distinct empirical correlations and feedforward neural networks modeling scheme. The results show that SVR outperforms most of the popular modeling schemes with reliable and efficient performance.

To demonstrate the usefulness of support vector machines regression as a new computational intelligence paradigm, the developed SVR calibration model was built using three published distinct PVT regressions as a new computational intelligence paradigm, the developed pressure for, we are able to build SVR model in estimating both challenge data with wide domain and uncertain distributions. There-
They used 90 data sets from 30 independent reservoirs in the Gulf of Suez to develop the correlations. The new correlations were tested against other Egyptian data in Saleh et al. (1987) and showed improvement over published correlations. The authors in Omar and Todd (1993) presented oil formation volume factor correlation, based on standing correlation model using data sets within 93 observations from Malaysian oil reservoirs. The authors in Kartoatmodjo and Schmidt (1994) used a global data bank to develop new correlations for all PVT properties using data of 740 different crude oil samples gathered from all over the world provided 5392 data sets for the correlation development. The authors in Almehaideb (1997) published a new set of correlations for UAE crude’s using 62 data sets from UAE reservoirs for measuring both bubble point pressure and oil formation volume factor. The bubble point pressure correlation, such as, Omar and Todd (1993) uses the oil formation volume factor as input in addition to oil gravity, gas gravity, solution gas oil ratio, and reservoir temperature. The authors in Sutton and Farshad (1990, 1991) published an evaluation for Gulf of Mexico crude oils using 285 data sets for gas-saturated oil and 134 data sets for under saturated oil representing 31 different crude oils and natural gas systems. The results show that Glaso correlation for oil formation volume factor perform the best for most of the data of the study. The authors in Petrosky and Farshad (1993) published a new correlation based on Gulf of Mexico crude oils and they reported that the best correlation for oil formation volume was Al-Marhoun correlation. The authors in McCain et al. (1991) published an evaluation of all reservoir properties correlations based on large global database and they recommended the Standing (1947) correlation for formation volume factor at and below bubble point pressure for future use.

The authors in Ghetto et al. (1994) performed a comprehensive study on PVT properties correlation based on 195 global data sets collected from the Mediterranean Basin, Africa, Middle East, and the North Sea reservoirs. They recommended both Vasquez and Beggs correlation for the oil formation volume factor. On the other hand, the author in Elsharkawy et al. (1994) evaluated PVT correlations for Kuwaiti crude oils using 44 samples; they concluded that the Standing correlation gave the best results for bubble point pressure while Al-Marhoun (1988) oil formation volume factor correlation performed satisfactory. The authors in Mahmood and Al-Marhoun (1996) presented an evaluation of PVT correlations for Pakistani crude oils, and used 166 data sets from 22 different crude samples for the evaluation. The author in Al-Marhoun (1992) reported that the oil formation volume factor correlation gave the best results. The bubble point pressure errors reported in this study, for all correlations, are among the highest reported in the literature. Furthermore, in Hanafy et al. (1997), the authors evaluate the most accurate correlation to apply to Egyptian crude oils for formation volume factor correlation based on Macary and El-Batanoney (1992). The results showed that the average absolute error of 4.9%, while Dokla and Osman (1992) showed 3.9%. Therefore, the study strongly supports the approach of developing a local correlation versus a global correlation.

An evaluation of all available oil formation volume factor correlations has been published in Al-Fattah and Al-Marhoun (1994) based on 674 data sets from published literature and they determined that Al-Marhoun (1992) correlation has the least error for global data set. In addition, they performed trend tests to evaluate the models’ physical behavior. Finally, Al-Shammasi (1997) evaluated the published correlations and neural network models for bubble point pressure and oil formation volume factor for accuracy and flexibility to represent hydrocarbon mixtures from different geographical locations worldwide. He presented a new correlation for bubble point pressure based on global data of 1661 published and 48 unpublished data sets. Also, he presented neural network models and compared their performance to numerical correlations. He concluded that statistical and trend performance analysis showed that some of the correlations violate the physical behavior of hydrocarbon fluid properties.

### 2.2. Forecasting PVT properties based on artificial neural networks

Artificial neural networks (ANNs) are parallel distributed information processing models that can recognize highly complex patterns within available data. In recent years, neural network have gained popularity in petroleum applications. Many authors discussed the applications of neural network in petroleum engineering such as Ali (1994), Elsharkawy (1998), Gharbi and Elsharkawy (1997a,b), Kumulyui and Daltaban (1994), Mohaghegh and Ameri (1994), Mohaghegh (1995), Mohaghegh (2000), and Varotsis et al. (1999). The most common widely used neural network in literature is known as the feedforward neural networks with backpropagation training algorithm, see Ali (1994), Duda et al. (2001), and Osman et al. (2001). This type of neural networks is excellent computational intelligence modeling scheme in both prediction and classification tasks. Few studies were carried out to model PVT properties using neural networks. Recently, feedforward neural network serves the petroleum industry in predicting PVT correlations; see the work of Gharbi and Elsharkawy (1997a,b), and Osman et al. (2001).

Al-Shammasi (1997, 2001) presented neural network models and compared their performance to numerical correlations. He concluded that statistical and trend performance analysis showed that some of the correlations violate the physical behavior of hydrocarbon fluid properties. In addition, he pointed out that the published neural network models missed major model parameters to be reproduced. He uses (2HL) neural networks (4-5-3-1) structure for predicting both properties: bubble point pressure and oil formation volume factor. He evaluates published correlations and neural-network models for bubble point pressure and oil formation volume factor for their accuracy and flexibility in representing hydrocarbon mixtures from different locations worldwide.

The authors of Gharbi and Elsharkawy (1997a,b) and Osman et al. (2001) carried out comparative studies between the feedforward neural networks performance and the four empirical correlations: Standing, Al-Marhoun, Glaso, and Vasquez and Beggs empirical correlation, see Al-Marhoun (1988), El-Sebakhy et al. (2007), and Osman et al. (2001) for more details about both the performance and the carried comparative studies results. In 1996, the authors in Gharbi and Elsharkawy (1997a,b) published neural network models for estimating bubble point pressure and oil formation volume factor for Middle East crude oils based on the neural system with log sigmoid activation function to estimate the PVT data for Middle East crude oil reservoirs, while in Gharbi and Elsharkawy (1997a,b), they developed a universal neural network for predicting PVT properties for any oil reservoir. In Gharbi and Elsharkawy (1997a,b), two neural networks are trained separately to estimate the bubble point pressure and oil formation volume factor, respectively. The input data were solution gas–oil ratio, reservoir temperature, oil gravity, and gas relative density. They used two hidden layers (2HL) neural networks: The first neural network, (4-8-4-2) to predict the bubble point pressure and the second neural network, (4-6-6-2) to predict the oil formation volume factor. Both neural networks were built using a data set of size 520 observations from Middle East area. The input data set is divided into a training set of 498 observations and a testing set of 22 observations.

The authors in Gharbi and Elsharkawy (1997a,b) follow the same criterion of Gharbi and Elsharkawy (1997a,b), but on large scale covering additional area: North and South America, North Sea, South East Asia, with the Middle East region. They developed a one hidden layer neural network using a database of size 5434 representing around 350 different crude oil systems. This database was divided into a training set with 5200 observations and a testing set with other 234 observations. The results of their comparative studies were shown that the FFN outperforms the conventional empirical correlation schemes in the prediction of PVT properties with reduction in the average absolute error and increasing in correlation coefficients. The reader can take a look at Al-Shammasi (1997) and El-Sebakhy et al. (2007) for more utilization of other type of neural networks in predicting the PVT properties, for instance, radial basis functions and abductive networks.
2.3. The most popular drawbacks of neural networks modeling scheme

Experience with neural networks has revealed a number of limitations for the technique. One such limitation is the complexity of the design space, Omar and Todd (1993). With no analytical guidance on the choice of many design parameters, the developer often follows an ad hoc, trial-and-error approach of manual exploration that naturally focuses on just a small region of the potential search space. Architectural parameters that have to be guessed a priori include the number and size of hidden layers and the type of transfer function(s) for neurons in the various layers. Learning algorithm parameters to be determined include initial weights, learning rate, and momentum. Although acceptable results may be obtained with effort, it is obvious that potentially superior models can be overlooked. The considerable amount of user intervention not only slows down model development, but also works against the principle of ‘letting the data speak’. To automate the design process, external optimization criteria, e.g. in the form of genetic algorithms, have been proposed, Petrosky and Farshad (1993). Over-fitting or poor network generalization with new data during actual use is another problem, Kartoatomdojo and Schmidt (1994). As training continues, fitting of the training data improves but performance of the network with new data previously unseen during training may deteriorate due to over-learning. A separate part of the training data is often reserved for monitoring such performance in order to determine when training should be stopped prior to complete convergence. However, this reduces the effective amount of data used for actual training and would be disadvantageous in many situations where good training data are often scarce. Network pruning algorithms have also been described to improve generalization Almehaideb (1997). The commonly used back-propagation training algorithm with a gradient descent approach to minimizing the error during training suffers from the local minima problem, which may prevent the synthesis of an optimum model19. Another problem is the opacity or black-box nature of neural network models. The associated lack of explanation capabilities is a handicap in many decision support applications such as medical diagnostics, where the user would usually like to know how the model came to a certain conclusion. Additional analysis is required to derive explanation facilities from neural network models; e.g. through rule extraction (Sutton and Farshad, 1991). Model parameters are buried in large weight matrices, making it difficult to gain insight into the modeled phenomenon or compare the model with available empirical or theoretical models. Information on the relative importance of the various inputs to the model is not readily available, which hampers efforts for model reduction by discarding less significant inputs. Additional processing using techniques such as the principal component analysis may be required for this purpose McCain (1991).

In this research, we proposed support vector machines regression to overcome these neural networks shortcomings and then we forecast PVT properties based on this new modeling scheme. The support vector machine modeling scheme is a new computational intelligence paradigm that is based statistical learning theory and structural risk minimization principle. Based on this principle, support vector machines achieve optimum network structure by striking a right balance between the empirical error and the Vapnik–Chervonenkis (VC) confidence interval, which is unlikely to generate local minima.

3. Support vector machines regression modeling scheme

Support vector machines regression is one of the most successful and effective algorithms in both machine learning and data mining communities. It has been widely used as a robust tool for classification and regression, see Vapnik (1995), Cortes and Vapnik (1995), El-Sebakhy (2004), El-Sebakhy et al. (2007), Cortes and Vapnik (1995), and Osuna et al. (1997). It has been found to be very robust in many applications, for example in the field of optical character recognition, text categorization, and face detection in images, Joachims (1997). The high generalization ability of SVR is ensured by special properties of the optimal hyperplane that maximizes the distance to training examples in a high dimensional feature space.

3.1. Background and overview

Recent years, there have been plenty of researches on support vector machines. An overview can be found in Vapnik (1995), Burges (1998), Schölkopf and Smola (2002), and Kobayashi and Komaki (2006) developed support vector machine (SVM) modeling scheme, which is a novel learning machine based on statistical learning theory, and which adheres to the principle of structural risk minimization seeking to minimize an upper bound of the generalization error, rather than minimize the training error. This induction principle is based on the bounding of the generalization error by the sum of the training error and a confidence interval term depending on the VC dimension. Based on this principle, SVM achieves an optimum network structure by striking a right balance between the empirical error and the VC-confidence interval. This balance eventually leads to better generalization performance than other neural network models Peng et al. (2004).

Originally, SVM has been developed to solve pattern recognition problems. However, with the introduction of Vapnik’s ε-insensitive loss function, SVM has been extended to solve nonlinear regression estimation problems, such as new techniques known as support vector regression (SVR), which have been shown to exhibit excellent performance. The performance of SVR is based on the predefined parameters (so-called hyper-parameters). Therefore, to construct a perfect SVR forecasting model, SVR’s parameters must be set carefully. Recently, SVR has emerged as an alternative and powerful technique to predict a complex nonlinear relationship problem. It has achieved great success in both academic and industrial platforms due to its many attractive features and promising generalization performance.

3.2. The architecture of support vector machines regression

Recently, a regression version of SVM has emerged as an alternative and powerful technique to solve regression problems by introducing an alternative loss function. In this subsection, a brief description of SVR is given. Detailed descriptions of SVR can be found in Vapnik (1998) and El-Sebakhy (2004). Generally, the SVR formulation follows the principle of structural risk minimization, seeking to minimize an upper bound of the generalization error rather than minimize the prediction error on the training set. This feature makes the SVM with a greater potential to generalize the input–output relationship learnt during its training phase for making good predictions, for new input data. The SVR maps the input data x into a high-dimensional feature space F by nonlinear mapping, to yield and solve a linear regression problem in this feature space as it is shown in Fig. 1.

The regression approximation estimates a function according to a given data, \( G = \{(x_i, y_i) : \mathbb{R}^p \rightarrow \mathbb{R} \}_{i=1}^n \), where \( x_i \) denotes the input vector; \( y_i \) denotes the output (target) value and \( n \) denotes the total number of data patterns. The modeling aim is to build a decision function, \( \hat{y} = f(x) \) that accurately predicts the outputs \( (y_i) \) corresponding to a new set of input–output examples, \( \{(x_i, y_i)\} \). Using mathematical notation, the linear approximation function (in the feature space) is approximated using the following function:

\[
  f(x) = (\omega^T \varphi(x) + b), \quad \varphi : \mathbb{R}^p \rightarrow \mathbb{F}; \quad \omega, b \in \mathbb{F},
\]

where \( \omega \) and \( b \) are coefficients; \( \varphi(x) \) denotes the high-dimensional feature space, which is nonlinearly mapped from the input space \( x \). Therefore, the linear relationship in the high-dimensional feature space responds to nonlinear relationship in the low-dimension input space, disregarding the inner product computation between \( \omega \) and \( \varphi(x) \) in the high-dimensional feature space. Correspondingly, the
original optimization problem involving nonlinear regression is transformed into finding the flattest function in the feature space \( f \), and not in the input space, \( x \). The unknown parameters \( \omega \) and \( b \) in Eq. (1) are estimated by the training set, \( G \).

SVR performs linear regression in the high-dimensional feature space by \( \varepsilon \)-insensitive loss. At the same time, to prevent over-fitting, thereby improving the generalization capability, following regularized functional involving summation of the empirical risk and a complexity term \( \frac{1}{2} \| \omega \|^2 \), is minimized. The coefficients \( \omega \) and \( b \) can thus be estimated by minimizing the regularized risk function.

\[
R_{\text{SVR}}(\omega) = R_{\text{emp}} + \frac{1}{2} \| \omega \|^2 = C \sum_{i=1}^{n} L_i(y_i, \hat{y}_i) + \frac{1}{2} \| \omega \|^2, \tag{2}
\]

where \( R_{\text{SVR}} \) and \( R_{\text{emp}} \) represent the regression and empirical risks, respectively; \( \| \omega \|^2 \) denotes the Euclidean norm and \( C \) denotes a cost function measuring the empirical risk. In the regularized risk function given by Eq. (2), the regression risk (test set error), \( R_{\text{SVR}} \), is the possible error committed by the function \( f \) in predicting the output corresponding to a test example input vector.

\[
L_i(y_i, \hat{y}_i) = \begin{cases} 
|y_i - \hat{y}_i| - \varepsilon, & \text{if } |y_i - \hat{y}_i| > \varepsilon \\
0, & \text{Otherwise}
\end{cases} \tag{3}
\]

In Eq. (2), the first term \( C/n \sum_{i=1}^{n} L_i(y_i, \hat{y}_i) \) denotes the empirical error (termed "training set error"), which is estimated by the \( \varepsilon \)-insensitive loss function in Eq. (3). The loss function is introduced to obtain sufficient samples of the decision function in Eq. (1) by using fewer data points. The second item \( \frac{1}{2} \| \omega \|^2 \) is the regularization term. The regularized constant \( C \) calculates the penalty when an error occurs, by determining the trade-off between the empirical risk and the regularization term, which represents the ability of prediction for regression. Raising the value of \( C \) increases the significance of the empirical risk relative to the regularization term. The penalty is acceptable only if the fitting error is larger than \( \varepsilon \). The \( \varepsilon \)-insensitive loss function is employed to stabilize estimation. In other words, the \( \varepsilon \)-insensitive loss function can reduce the noise. Thus, \( \varepsilon \) can be viewed as a tube size equivalent to the approximation accuracy in training data as it is shown in Fig. 2. In the empirical analysis, \( C \) and \( \varepsilon \) are the parameters selected by users.

To estimate \( \omega \) and \( b \), we introduce the positive slack variables \( \xi_i \) and \( \xi_i^+ \), then according to Fig. 2, the sizes of the stated excess positive and negative deviations are represented by \( \xi_i \) and \( \xi_i^+ \), respectively. The slack variables assume non-zero values outside the \([-\varepsilon, \varepsilon]\) region. The SVR fits \( f(x) \) to the data such that, (i) the training error is minimized by minimizing \( \xi_i \) and \( \xi_i^+ \) and (ii) \( \| \omega \|^2 / 2 \) is minimized to raise the flatness of \( f(x) \), or to penalize excessively complex fitting functions. Thus, SVR is formulated as minimization of the following functional:

\[
\text{Minimize } R_{\text{SVR}}(\omega, C) = \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{n} L_i(\xi_i + \xi_i^+), \tag{4}
\]

Subjected to \( \{ y_i - \omega \phi(x_i) - b \xi_i + \xi_i^+ \} \leq \varepsilon, \) where \( \xi_i \) and \( \xi_i^+ \) denote slack variables that measure the error of the up and down sides, respectively. The above formulae indicate that increasing \( C \) decreases the corresponding \( \xi_i \) and \( \xi_i^+ \) in the same constructed function \( f(x) \), thereby reducing the error resulting from the corresponding data points. Finally, by introducing Lagrange multipliers and exploiting the optimality constraints, the decision function given by Eq. (1) has the following explicit form (Vapnik, 1995, 1998):

\[
f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{n} \left( \alpha_i - \alpha_i^* \right) K(x,x_i) + b, \tag{5}
\]

Where the parameters \( \alpha_i \) and \( \alpha_i^* \) are called the lagrangian multipliers in Eq. (5), which satisfy the equalities \( \alpha_i \alpha_i^* = 0, \alpha_i, \alpha_i^* \geq 0 \), where \( i = 1,2,...,n \), see Vapnik (1998), Cristianini and Shawe-Taylor (2000), Peng et al. (2004), and El-Sebakhy et al. (2007). The term \( K(x,x_i) \) in Eq. (5) is defined as kernel function, where the value of kernel function equals the inner product of two vectors \( x \) and \( x_i \) in the feature space \( \phi(x) \) and \( \phi(x_i) \), meaning that \( K(x,x_i) = \phi(x) \cdot \phi(x_i) \). The kernel function is intended to handle any dimension feature space without the need to calculate \( \phi(x) \) accurately. If any function can satisfy Mercer’s condition, it can be employed as a kernel function, Cortes and Vapnik (1995) and Vapnik (1998). The typical examples of kernel function are the polynomial kernel \( K(x,y) = (x \cdot y + 1)^d \) and the Gaussian kernel \( K(x,y) = \exp[-(x-y)^2 / 2\sigma^2] \). In these equations, \( d \) represents the degree of the polynomial kernel, and \( \sigma^2 \) represents bandwidth of the Gaussian kernel. These parameters must be selected accurately, since they determine the structure of high-dimensional feature space and govern the complexity of the final solution.

4. Data acquisition and statistical quality measures

4.1. The acquired data

The implementation studies were achieved based on three databases drawn from three distinct published research articles.

1. The first database was drawn from the article of Al-Marhoun (1988). This database has 160 data drawn from 69 Middle Eastern reservoirs, which published correlations for estimating bubble point pressure and oil formation volume factor for Middle East oils.
2. The second database was drawn from articles by Al-Marhoun and Osman (2002), Osman and Abdel-Aal (2002) and Osman and Al-Marhoun (2005). This database has 283 data points collected from different Saudi fields to predict the bubble point pressure, and the oil formation volume factor at the bubble-point pressure for Saudi crude oils. The models were based on neural networks with 142 training set to build FFN calibration model to predict the bubble point pressure, and the oil formation volume factor, 71 to cross-validate the relationships established during the training process and adjust the calculated weights, and the remaining 70 to test the model to evaluate its accuracy. The results show that the developed Bob model provides better predictions and higher accuracy than the published empirical correlations.

3. The third database was obtained from the works of Goda et al. (2003) and Osman et al. (2001), where the authors used feedforward learning scheme with log sigmoid transfer function in order to estimate the formation volume factor at the bubble point pressure. This database contains 782 observations after deleting the redundant 21 observations from the actual 803 data points. This data set is gathered from Malaysia, Middle East, Gulf of Mexico, and Colombia. The authors in Goda et al. (2003) and Osman et al. (2001) designed a one hidden layer (1HL) feedforward neural network (4–5–1) with the backpropagation learning algorithm using four input neurons covering the input data of gas–oil ratio, API oil gravity, relative gas density, and reservoir temperature, one hidden layer with five neurons, and single neuron for the formation volume factor in the output layer.

To evaluate the performance of each SVR, FFN, and the used three empirical correlations modeling scheme using the above defined three distinct databases. The entire database is divided using the stratified criterion. Therefore, we use 70% of the data for building the SVR learning model (internal validation) and 30% of the data for testing/validation (external validation or cross-validation criterion). We repeat both internal and external validation processes for training and cross validation check.

In this study, the 382 data points, 267 were used to build the calibration model, the remaining 115 to cross-validate the relationships established during the training and testing process to evaluate its accuracy and trend stability. For the testing data, the statistical summary of the investigated quality measures corresponding to SVR modeling scheme, neural networks, and the most popular empirical correlations, such as, Standing, Al-Marhoun, and Glaso empirical correlation using the above distinct data sets. The prediction performance of both bubble point pressure and Oil Formation Volume Factor were shown in Tables 1–6, respectively.

### Table 1
Testing results (Osman et al., 2001 and El-Sebakhy et al., 2007 data): statistical quality measures when estimate \( B_o \)

<table>
<thead>
<tr>
<th>Correlation</th>
<th>( E_r )</th>
<th>( E_a )</th>
<th>( E_{min} )</th>
<th>( E_{max} )</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standing (1947)</td>
<td>-0.170</td>
<td>2.724</td>
<td>0.008</td>
<td>20.180</td>
<td>0.974</td>
</tr>
<tr>
<td>Glaso (1980)</td>
<td>1.8186</td>
<td>3.374</td>
<td>0.003</td>
<td>17.776</td>
<td>0.972</td>
</tr>
<tr>
<td>Al-Marhoun (1992)</td>
<td>-0.115</td>
<td>2.205</td>
<td>0.003</td>
<td>13.179</td>
<td>0.981</td>
</tr>
<tr>
<td>ANN System</td>
<td>0.3024</td>
<td>1.789</td>
<td>0.008</td>
<td>11.775</td>
<td>0.980</td>
</tr>
<tr>
<td>Support Vector Machine Regression</td>
<td>0.18</td>
<td>1.37</td>
<td>0.002</td>
<td>7.721</td>
<td>0.984</td>
</tr>
</tbody>
</table>

During the implementation, the user should be aware of the input domain values to make sure that the input data values fall in a natural domain. This step called the quality control and it is really an important step to have very accurate and reliable results at the end. The following is the most common domains for the input/output variables, gas–oil ratio, API oil gravity, relative gas density, bubble point pressure, and oil formation volume factor that are used in the both input and output layers of modeling schemes for PVT analysis:

- Gas oil ratio with range between 26 and 1602, scf/stb.
- Oil formation volume factor which varies from 1.032 to 1.997, bbl/stb.
- Bubble point pressure, starting from 130, ending with 3573, psia.
- Reservoir temperature with its range from 74 °F to 240 °F.
- API gravity which changes between 19.4 and 44.6.
- Gas relative density, change from 0.744 to 1.367.

### Table 2
Testing results (Osman et al., 2001 and El-Sebakhy et al., 2007 data): statistical quality measures when estimate \( B_o \)

<table>
<thead>
<tr>
<th>Correlation</th>
<th>( E_r )</th>
<th>( E_a )</th>
<th>( E_{min} )</th>
<th>( E_{max} )</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standing (1947)</td>
<td>-0.170</td>
<td>2.724</td>
<td>0.008</td>
<td>20.180</td>
<td>0.974</td>
</tr>
<tr>
<td>Glaso (1980)</td>
<td>1.8186</td>
<td>3.374</td>
<td>0.003</td>
<td>17.776</td>
<td>0.972</td>
</tr>
<tr>
<td>Al-Marhoun (1992)</td>
<td>-0.115</td>
<td>2.205</td>
<td>0.003</td>
<td>13.179</td>
<td>0.981</td>
</tr>
<tr>
<td>ANN System</td>
<td>0.3024</td>
<td>1.789</td>
<td>0.008</td>
<td>11.775</td>
<td>0.980</td>
</tr>
<tr>
<td>Support Vector Machine Regression</td>
<td>0.18</td>
<td>1.37</td>
<td>0.002</td>
<td>7.721</td>
<td>0.984</td>
</tr>
</tbody>
</table>

### Table 3
Testing results (Al-Marhoun and Osman, 2002 and Osman and Abdel-Aal, 2002 dataset): statistical quality measures when estimate \( B_o \)

<table>
<thead>
<tr>
<th>Correlation</th>
<th>( E_r )</th>
<th>( E_a )</th>
<th>( E_{min} )</th>
<th>( E_{max} )</th>
<th>( SD )</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standing (1947)</td>
<td>-1.054</td>
<td>1.6833</td>
<td>0.066</td>
<td>7.7997</td>
<td>2.0121</td>
<td>0.9947</td>
</tr>
<tr>
<td>Glaso (1980)</td>
<td>0.4538</td>
<td>1.7865</td>
<td>0.0062</td>
<td>7.3839</td>
<td>2.1662</td>
<td>0.9920</td>
</tr>
<tr>
<td>Al-Marhoun (1992)</td>
<td>-0.392</td>
<td>0.8451</td>
<td>0.0003</td>
<td>3.5546</td>
<td>1.0299</td>
<td>0.9972</td>
</tr>
<tr>
<td>ANN System</td>
<td>0.217</td>
<td>0.5116</td>
<td>0.0061</td>
<td>2.5601</td>
<td>0.6626</td>
<td>0.9577</td>
</tr>
<tr>
<td>Support Vector Machine Regression</td>
<td>0.096</td>
<td>0.353</td>
<td>0.0013</td>
<td>2.5835</td>
<td>0.4743</td>
<td>0.997</td>
</tr>
</tbody>
</table>

### Table 4
Testing results (Al-Marhoun and Osman, 2002 and Osman and Abdel-Aal, 2002 dataset): statistical quality measures when estimate \( B_o \)

<table>
<thead>
<tr>
<th>Correlation</th>
<th>( E_r )</th>
<th>( E_a )</th>
<th>( E_{min} )</th>
<th>( E_{max} )</th>
<th>( SD )</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standing (1947)</td>
<td>-8.441</td>
<td>10.4562</td>
<td>0.2733</td>
<td>47.0213</td>
<td>11.841</td>
<td>0.8974</td>
</tr>
<tr>
<td>Glaso (1980)</td>
<td>-18.48</td>
<td>20.7569</td>
<td>0.0345</td>
<td>63.7634</td>
<td>16.160</td>
<td>0.9837</td>
</tr>
<tr>
<td>Al-Marhoun (1992)</td>
<td>0.941</td>
<td>8.1028</td>
<td>0.0935</td>
<td>38.0854</td>
<td>11.411</td>
<td>0.9905</td>
</tr>
<tr>
<td>ANN System</td>
<td>-0.222</td>
<td>5.8915</td>
<td>0.2037</td>
<td>38.1225</td>
<td>8.678</td>
<td>0.9930</td>
</tr>
<tr>
<td>Support Vector Machine Regression</td>
<td>1.054</td>
<td>10.4562</td>
<td>0.2733</td>
<td>47.0213</td>
<td>11.841</td>
<td>0.8974</td>
</tr>
</tbody>
</table>
error ($E_i$), average absolute percent relative error ($E_a$), minimum and maximum absolute percent error ($E_{\text{min}}$ and $E_{\text{max}}$) root mean square errors ($E_{\text{rms}}$), standard deviation (SD), and correlation coefficient ($R^2$), see Duda et al. (2001) and Osman et al. (2001) for more details about the corresponding equations of these statistical parameters. To demonstrate the usefulness of the SVR modeling scheme, the developed calibration model based on the three distinct databases (i) database with 160 observations and (ii) database with 283 observations will be used to predict both $P_b$ and $R_{ob}$, and (iii) the world wide database with 782 observations published in Goda et al. (2003) and Osman et al. (2001).

The results show that the support vector machines regression algorithm has both reliable and efficient performance. In addition, its performance outperforms the one of the most popular empirical correlation schemes and the standard feedforward neural networks in terms of root mean squared error, absolute average percent error, standard deviation, and correlation coefficient.

5. Empirical study

We have conducted the quality control check on all these data sets and clean it from redundant data and un-useful observations. To evaluate performance of each modeling scheme, entire database is divided using the stratified criterion. Therefore, we use 70% of the data for building SVR learning model (internal validation) and 30% of the data for testing/validation (external validation or cross-validation criterion). Both internal and external validation processes are repeated 1000 times. Therefore, data were divided into two/three groups for training and cross validation check. Therefore, of the 782 data points, 382 were used to train the neural network models, the remaining 200 to cross validate the relationships established during training process and 200 to test model to evaluate its accuracy and trend stability. For testing data, statistical summary to investigate different quality measures corresponding to SVR modeling scheme, feedforward neural networks system, and the most popular empirical correlations in literatures to predict both bubble point pressure and Oil Formation Volume Factor.

Generally, after training the SVR modeling system, the calibration model becomes ready for testing and evaluation using the cross-validation criterion. Comparative studies were carried out to compare the performance and accuracy of the new SVR model versus both the standard neural networks and the three common published empirical correlations, namely, Standing, Al-Mahroun, and Glaso empirical correlations.

5.1. Parameters initialization

In this study, we follow the same procedures in Al-Marhoun and Osman (2002), Osman et al. (2001), and Osman and Abdel-Aal (2002) and

<table>
<thead>
<tr>
<th>Correlation</th>
<th>$E_i$</th>
<th>$E_a$</th>
<th>$E_{\text{min}}$</th>
<th>$E_{\text{max}}$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standing (1947)</td>
<td>12.811</td>
<td>24.684</td>
<td>0.62334</td>
<td>59.038</td>
<td>0.8657</td>
</tr>
<tr>
<td>Glaso (1980)</td>
<td>-18.887</td>
<td>26.551</td>
<td>0.28067</td>
<td>98.78</td>
<td>0.9675</td>
</tr>
<tr>
<td>Al-Marhoun (1992)</td>
<td>5.023</td>
<td>8.9416</td>
<td>0.13115</td>
<td>87.989</td>
<td>0.9701</td>
</tr>
<tr>
<td>ANN System</td>
<td>4.9205</td>
<td>6.7495</td>
<td>0.16115</td>
<td>65.3839</td>
<td>0.9765</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>3.426</td>
<td>4.0757</td>
<td>0.128</td>
<td>53.219</td>
<td>0.9808</td>
</tr>
</tbody>
</table>

Table 5: Testing results (Osman et al., 2001 and Goda et al., 2003 dataset): statistical quality measures when estimate $P_b$.
we used three distinct kernel functions, namely, polynomial, sigmoid kernel, and Gaussian Bell kernel. In designing the support vector machine regression, the important parameters that will control its overall performance were initialized, such as, kernel = ‘poly’; kernel opt. = 5; epsilon = 0.01; lambda = 0.0000001; verbose = 0, and the constant C either 1 or 10 for simplicity. The cross-validation method used in this study utilized as a checking mechanism in the training algorithm to prevent both over fitting and complexity criterion based on the root-mean-squared errors threshold. The resulted weights for the Bob and Pb models are given below in different tables and graphs. Moreover, the relative importance of each input property are identified during the training process and given for Bob and Pb models as it is shown below.

5.2. Discussion and comparative studies

One can investigate other common empirical correlations besides these chosen empirical correlations, see El-Sebakhy et al. (2007) and Osman et al. (2001) for more details about these empirical correlations mathematical formulas. The results of comparisons in the testing (external validation check were summarized in Tables 1–6, respectively. We observe from these results that the support vector machines modeling scheme outperforms both neural networks with backpropagation training algorithm and the most popular empirical correlations. The proposed model shown a high accuracy in predicting both Pb and Bob values with a stable performance, and achieved the lowest absolute percent relative error, lowest minimum error, lowest maximum error, lowest root mean square error, and the highest correlation coefficient among other correlations for the used three distinct data sets.

We draw a scatter plot of the absolute percent relative error (\(E_A\)) versus the correlation coefficient for all computational intelligence forecasting schemes and the most common empirical correlations. Each modeling scheme is represented by a symbol; the good forecasting scheme should appear in the upper left corner of the graph. Fig. 3 shows a scatter plot of \((E_A)\) versus \((R^2\) or \(r\)) for all modeling schemes that are used to determine Bob based on the data set used in (Osman et al., 2001).

We observe that the symbol corresponding to SVR scheme falls in the upper left corner with \(E_A = 1.368\%\) and \(r = 0.9884\), while neural network is below SVR with \(E_A = 1.7886\%\) and \(r = 0.9878\), and other correlations indicates higher error values with lower correlation coefficients, for instance, Al-Marhoun (1992) has \(E_A = 2.2053\%\) and \(r = 0.9806\); Standing (1947) has \(E_A = 2.7238\%\) and \(r = 0.9742\); and Glaso
(1980) correlation with $E_s = 3.3743\%$ and $r = 0.9715$. Fig. 4 shows the same graph type, but with $b_p$ instead, based on the data set used in Osman et al. (2001) using the same modeling schemes. We observe that the symbol corresponding to SVR scheme falls in the upper left corner with $E_b = 1.3688\%$ and $R^2 = 0.9884$, while neural network is below SVR with $E_b = 1.7886\%$ and $r = 0.9878$, and other correlations indicates higher error values with lower correlation coefficients, for instance, Al-Marhoun (1992) has $E_b = 2.2053\%$ and $r = 0.9806$; Standing (1947) has $E_b = 2.7238\%$ and $r = 0.9742$; and Glaso (1980) correlation with $E_b = 3.3743\%$ and $r = 0.9715$.

The same implementations process were repeated for the other data sets used in Al-Marhoun (1988, 1992), and Al-Marhoun and Osman (2002) and Osman and Abdel-Aal (2002), but for the sake of simplicity, we did not include it in this context.

Figs. 5–10 illustrate six scatter plots of the predicted results versus the experimental data for both $bpp$ and $B_{ob}$ values using the same three distinct data sets. These cross plots indicates the degree of agreement between the experimental and the predicted values based on the high quality performance of the SVR modeling scheme. The reader can compare the performance of theses patterns with the corresponding ones of the published neural networks modeling and the most common empirical coefficient correlations in (Al-Marhoun, 1992; Al-Marhoun and Osman, 2002), (Osman and Abdel-Aal, 2002), and (Osman et al., 2001). Finally, we conclude that developed SVR modeling scheme has better and reliable performance compared to the most published modeling schemes and empirical correlations.

The bottom line is that, the developed SVR modeling scheme outperforms both the standard feedforward neural networks and the most common published empirical correlations in predicting both $bpp$ and $B_{ob}$ using the four input variables: solution gas–oil ratio, reservoir temperature, oil gravity, and gas relative density.

6. Conclusion and recommendation

In this study, three distinct published data sets were used in investigating the capability of the SVR modeling scheme as a new framework for predicting the PVT properties of oil crude systems. Based on the obtained results and comparative studies, the conclusions and recommendations are drawn as:

A new computational intelligence modeling scheme based on the SVR scheme to predict both bubble point pressure and oil formation volume factor using the four input variables: solution gas–oil ratio, reservoir temperature, oil gravity, and gas relative density. As it is shown in the petroleum engineering communities, these two predicted properties were considered the most important PVT properties of oil crude systems.

The developed SVR modeling scheme outperforms both the standard feedforward neural networks and the most common published empirical correlations. Thus, the developed SVR modeling scheme has better, efficient, and reliable performance compared to the most published correlations. In addition, the developed SVR modeling scheme shown a high accuracy in predicting the $B_{ob}$ values with a stable performance, and achieved the lowest absolute percent relative error, lowest minimum error, lowest maximum error, lowest root mean square error, and the highest correlation coefficient among other correlations for the used three distinct data sets. Furthermore, the SVR modeling scheme is flexible, reliable, and shows bright future in implementing it for the oil and gas industry, especially permeability and porosity prediction, history matching, predicting the rock mechanics properties, flow regimes and liquid-holdup multiphase follow, and facies classification.

Nomenclature

- $B_{ob}$: OFVF at the bubble-point pressure, RB/STB
- $R_s$: on solution gas oil ratio, SCF/STB
- $T$: reservoir temperature, degrees Fahrenheit
- $\gamma_o$: oil relative density (water=1.0)
- $\gamma_g$: gas relative density (air=1.0)
- $E_r$: average percent relative error
- $E_i$: percent relative error
- $E_s$: average absolute percent relative error
- $E_{max}$: Maximum absolute percent relative error
- $E_{min}$: Minimum absolute percent relative error
- RMS: Root Mean Square error

Acknowledgement

The author wishes to thank King Fahd University of Petroleum and Minerals and Mansoura University for the facilities utilized to perform the present work and for their support. In addition, he is thanking both editor-in-chief and the reviewer's for their valuable comments and support.