Data mining in forecasting PVT correlations of crude oil systems based on Type1 fuzzy logic inference systems

Emad A. El-Sebakhy *

King Fahd University of Petroleum & Minerals, Information and Computer Science, College of Computer Sciences and Engineering, Dhahran 31261, Saudi Arabia

1. Introduction

Knowing both chemical and physical properties of formation water is very important in various reservoir engineering computations, especially in water flooding and production. Ideally, these properties should be obtained experimentally. On some occasions, these properties are neither available nor reliable; then, empirically derived correlations are used to predict brine pressure–volume–temperature (PVT) properties. These correlations offer an acceptable approximation of formation water properties. However, the success of such correlations in prediction depends mainly on the range of data at which they were originally developed. These correlations were developed using equation of state (EOS), linear/nonlinear statistical regression, or graphical techniques. The equation of state is based on knowing the detailed compositions of the reservoir fluids and the determination of such quantities is expensive and time consuming. The currently available PVT simulator predicts the physical properties of reservoir fluids with various degrees 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.

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. Ideally, these 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 costly to obtain, then the solution is to use the empirically derived correlations to predict PVT properties. Generally, each correlation was developed for a certain range of reservoir fluid characteristics and geographical area with similar fluid compositions, and API oil gravity. Therefore, the accuracy of such correlations is critical and not often known in advance. 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 be occupied by one stock tank barrel oil plus any dissolved gas at the bubble point pressure and reservoir temperature. Precise prediction of $B_o$ is very important in
reservoir and production computations. 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 \( B_{Ro} \) have been proposed during the last decade. These correlations are essentially based on the assumption that \( P_b \) and \( B_{Ro} \) 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 \( (G_o) \), see Standing (1974), Glazo (1980), Al-Marhoun (1992), Osman et al. (2001), Goda et al. (2003), El-Sebakhy et al. (2007a,b), and references therein for more details.

There are many empirical correlations for predicting PVT properties, most of them were developed using equations of state (EOS) or linear/nonlinear multiple regression or graphical techniques or feedforward neural networks (FFNs) (Osman et al., 2001; Goda et al., 2003). However, they often do not perform very accurately and suffer from a number of drawbacks, such as FFN is a black box modeling scheme that is based on the trial-and-error approach. In addition, FFN architectural parameters 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. Although acceptable results may be obtained with effort, it is obvious that potentially superior models can be overlooked.

Recently, El-Sebakhy (2008) investigated and developed an intelligence data mining scheme to identify and estimate the reservoir engineering properties based on support vector machine (SVM) to overcome some of the common problems in the standard neural networks. The results have shown that the performances of SVM were reliable and outperform the one of the existing empirical correlation techniques. Unfortunately, there are two critical issues with SVM regarding both high dimensional and uncertain situations, which are common in reservoir engineering problems and they should be taken into consideration during the implementation processes.

The main objective of this paper is to investigate both capability feasibility of Type1 neuro-fuzzy inference systems (ANFIS) in estimating the PVT properties of crude oil systems in both high dimensional and uncertain situations, specifically develop a new intelligence system framework for predicting both bubble point pressure and oil formation volume factor using different databases of four input parameters, namely solution gas–oil ratio \( (Rs) \), reservoir temperature \( (T_R) \), oil gravity (API), and gas relative density. The rest of this paper is organized as follows. Section 2 provides a brief literature review and related work. Section 3 provides both data acquisition and statistical quality measures. Section 4 consists of the adaptive neuro-fuzzy inference systems methodology and architecture. The experimental set-up is discussed in Section 5. Section 6 shows the performance of the approach by giving the experimental results.

2. Literature review

Last six decades, engineers realized the importance of developing and using empirical correlations for PVT properties. Studies carried out in this field resulted in the development of new correlations. El-Sebakhy (2008) provided a literature review regarding the estimation of reservoir engineering in details. Therefore, for the sake of space and simplicity, we are going to provide the reader with brief details only. The most popular empirical correlations were (i) Standing (1974): Standing empirical correlations for bubble point pressure and for oil formation volume factor, (ii) Glazo (1980): Glazo empirical correlation for formation volume factor, and (iii) Al-Marhoun (1992): Al-Marhoun for formation volume factors at, above, and below bubble point pressure worldwide database. The reader can consider other empirical correlations as well, see Al-Shammasi (1997) and El-Sebakhy et al. (2007a,b) for more details. In this paper, we only concentrate on the most common three empirical correlations, namely Standing (1974), Glazo (1980), and Al-Marhoun (1992), which are the most popular empirical correlations during the comparative studies.

Last decade, standard neural networks have gained popularity in oil and gas real-world industry applications. Many authors discussed the applications of neural network in petroleum engineering, namely Ali (1994), Kumoluyi and Daltaban (1994), Mohaghegh and Ameri, 1994, Mohaghegh (1995, 2000), Charbi and Elsharkawy (1997, 1997a), Al-Shammasi (1997, 2001), Elsharkawy (1998), Varotsis et al. (1999), Osman et al. (2001), Goda et al. (2003), and El-Sebakhy et al. (2007a,b). The most common widely used neural network in literature is known as the feedforward neural network with backpropagation training algorithm, which is an excellent computational intelligence modeling scheme in both prediction and classification tasks. Few studies have been carried out to model PVT properties using neural networks. El-Sebakhy (2008) identified reservoir engineering properties based on the use of support vector machine and handled the over fitting and local minima in the standard neural networks. As discussed above, there are two critical issues with SVM regarding both high dimensional and uncertain situations, which are common in reservoir engineering problems. Comparative studies were carried out to evaluate the performance of neural networks to numerical correlations and statistical regression models for bubble point pressure and oil formation volume factor for their accuracy and flexibility in representing hydrocarbon mixtures from different locations worldwide. The performance results were explained in details within El-Sebakhy et al. (2007a,b) and El-Sebakhy (2008), and references therein.

3. Data acquisition and statistical quality measures

3.1. The acquired databases

To demonstrate the usefulness of Type1 fuzzy modeling scheme, the developed calibration model is based on three distinct databases: (i) a database with 160 observations and (ii) a database with 283 observations will be used to predict both \( P_b \) and \( B_{Ro} \), and (iii) a world wide database with 782 observations. The complete databases have been utilized before in distinct published research articles, and the details of these databases were explained in El-Sebakhy (2008). To evaluate the performance of each Type1 fuzzy inference system, feedforward neural network with backpropagation learning scheme, and the most common three empirical correlations in literature using the above defined three distinct databases. We use the stratified criterion to divide the provided database by selecting 70% for building the calibration Type1 fuzzy 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 1000 times to have a fair partition through the entire process operation.

During the implementation, the user should be aware of the input domain values to make sure that the values fall in a natural domain. This step called the quality control step is important to ensure reliable results at the end. The most common domains for the input/output variables, gas–oil ratio, API oil gravity, relative gas density, reservoir temperature; bubble point pressure, and oil formation volume factor that are used in the both input and output
layers of modeling schemes for PVT analysis are explained in details within El-Sebakhy (2008).

In this implementation process, we utilize the three databases provided in Al-Marhoun and Osman (2002) and Osman and Abdel-Aal (2002) and Osman and Al-Marhoun (2005) in both internal and external validation process and cross-validation process to evaluate its accuracy and trend stability. We investigate the capability of the established calibration neuro-fuzzy Type1 inference relationships to forecast both bubble point pressure and the oil formation volume factor for new unseen databases based on the same four input parameters, namely solution gas-oil ratio, reservoir temperature, oil gravity, and gas relative density. For both internal and external validation processes, different quality control and statistical measures were calculated to compare between the new intelligence framework, the feedforward neural networks with a backpropagation learning algorithm and sigmoid activation function, and the most popular empirical correlations (Standing, Al-Marhoun, and Glazo empirical correlation) performance. We repeat the same process with the other two databases as well. The obtained results of the entire process are shown in Tables 1–6.

Table 1
Testing results (Al-Marhoun, 1988; El-Sebakhy et al., 2007a,b; Osman et al., 2001 data): statistical quality measures when estimating $E_a$.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_a$</td>
<td>−0.17</td>
<td>1.8186</td>
<td>−0.115</td>
<td>0.3024</td>
<td>0.1501</td>
</tr>
<tr>
<td>$E_A$</td>
<td>2.724</td>
<td>3.374</td>
<td>2.205</td>
<td>1.789</td>
<td>1.322</td>
</tr>
<tr>
<td>$E_{min}$</td>
<td>0.008</td>
<td>0.003</td>
<td>0.003</td>
<td>0.008</td>
<td>0.002</td>
</tr>
<tr>
<td>$E_{max}$</td>
<td>20.18</td>
<td>17.776</td>
<td>13.179</td>
<td>11.775</td>
<td>7.4513</td>
</tr>
<tr>
<td>$SD$</td>
<td>2.5823</td>
<td>2.673</td>
<td>1.2842</td>
<td>0.8935</td>
<td>0.7876</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.974</td>
<td>0.972</td>
<td>0.981</td>
<td>0.988</td>
<td>0.997</td>
</tr>
</tbody>
</table>

Table 2
Testing results (Al-Marhoun, 1988; El-Sebakhy et al., 2007a,b; Osman et al., 2001 data): statistical quality measures when estimating $E_b$.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_b$</td>
<td>67.6</td>
<td>−1.616</td>
<td>8.008</td>
<td>8.129</td>
<td>7.432</td>
</tr>
<tr>
<td>$E_A$</td>
<td>67.73</td>
<td>18.52</td>
<td>20.01</td>
<td>21.02</td>
<td>14.22</td>
</tr>
<tr>
<td>$E_{min}$</td>
<td>0.162</td>
<td>0.1056</td>
<td>0.0254</td>
<td>0.0182</td>
<td>0.0093</td>
</tr>
<tr>
<td>$E_{max}$</td>
<td>102.08</td>
<td>138.96</td>
<td>190.12</td>
<td>145.29</td>
<td>101.23</td>
</tr>
<tr>
<td>$SD$</td>
<td>25.159</td>
<td>25.171</td>
<td>12.839</td>
<td>14.871</td>
<td>11.337</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.867</td>
<td>0.945</td>
<td>0.906</td>
<td>0.943</td>
<td>0.952</td>
</tr>
</tbody>
</table>

Table 3
Testing results (Al-Marhoun and Osman, 2002; Osman and Abdel-Aal, 2002; Osman and Al-Marhoun, 2005 dataset): statistical quality measures when estimating $R_b$.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_b$</td>
<td>−1.054</td>
<td>0.4538</td>
<td>−0.392</td>
<td>0.217</td>
<td>0.016</td>
</tr>
<tr>
<td>$E_A$</td>
<td>1.6833</td>
<td>1.7865</td>
<td>0.8451</td>
<td>0.5116</td>
<td>0.3247</td>
</tr>
<tr>
<td>$E_{min}$</td>
<td>0.066</td>
<td>0.0062</td>
<td>0.0003</td>
<td>0.0006</td>
<td>0.0001</td>
</tr>
<tr>
<td>$E_{max}$</td>
<td>7.7997</td>
<td>7.3839</td>
<td>3.5546</td>
<td>2.6001</td>
<td>2.3365</td>
</tr>
<tr>
<td>$SD$</td>
<td>2.1021</td>
<td>2.1662</td>
<td>1.1029</td>
<td>0.6626</td>
<td>0.3856</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.9947</td>
<td>0.992</td>
<td>0.9972</td>
<td>0.9977</td>
<td>0.997</td>
</tr>
</tbody>
</table>

3.2. The most common statistical quality measures

To compare the performance and accuracy of the new intelligence framework to other empirical correlations, statistical error analysis and quality measures are performed. The most common statistical quality measures that are utilized in both petroleum engineering and data mining journals were, namely the average percent relative error ($E_r$), average absolute percent relative error ($E_A$), minimum and maximum absolute percent error ($E_{min}$ and $E_{max}$) root-mean-square errors ($E_{rms}$), standard deviation ($SD$), and correlation coefficient ($R^2$), see Duda et al. (2001) and Osman et al. (2001) for more details about their corresponding mathematical formulæ.

As it is shown below in the empirical study section, the results show that Type1 neuro-fuzzy inference intelligence system scheme is faster and more stable than both empirical correlations and other forecasting schemes reported in the petroleum engineering literatures. Moreover, the new data mining modeling scheme outperforms both feedforward neural network and all the most common existing correlations models in terms of root-mean-squared error, absolute average percent error, standard deviation, and correlation coefficient.
4. Neuro-fuzzy inference systems

Fuzzy logic is an application of recognized softcomputing techniques. It is a design method that can be effectively applied to problems that, because of complex, nonlinear, or ambiguous models, cannot be easily solved using traditional engineering analytical techniques. Fuzzy theory is a theoretical framework having fuzzy sets and fuzzy logic as its core; it started with the fuzziness concept and its expression in the form of fuzzy (LeCun et al., 1995), Liu et al. (2003), and McCain et al. (1998). In linguistic models, both the antecedent and the consequent are fuzzy sets, while in the TSK model the antecedent consists of fuzzy sets but the consequence is made up of linear equations. Fuzzy relational equation models aim at building the fuzzy relation matrices according to the input–output process data determined. We are going to focus on the use of the neuro-fuzzy systems with the TSK model for predicting the PVT correlations of crude oil systems, because of TSK needs less rules and its parameters can be estimated from numerical data using optimization methods such as least-square algorithms, see Abdulrahem et al. (2007), LeCun et al. (1995), and Liu et al. (2003).

4.1. Adaptive neuro-fuzzy inference systems

The neuro-fuzzy inference system Type1 is a hybrid forecasting/classification framework, which learns the rules and membership functions from data. It is a network of nodes and directional links. Associated with the network is a learning rule, for instance, backpropagation. It is called adaptive because some or all of the nodes have parameters that affect the output of the node. These networks are learning a relationship between inputs and outputs. This type of networks cover number of different approaches, namely Mamdani-type and the Takagi–Sugeno–Kang type, see Abdulrahem et al. (2007), Amabeoku et al. (2005), Cuddy (1998), Hambalek and Reinaldo (2003), Jong-Se and Kim (2004), McCain et al. (1998), and Standing (1974) for more details.

The basic architecture of a Type1 FLS with crisp inputs and output is shown in Fig. 1. The TSK fuzzy modeling method was proposed by Takagi and Sugeno as a framework for generating fuzzy if then rules from numerical data. A TSK fuzzy model consists of a set of fuzzy rules, each describing a local linear input–output relationship:

\[ y_i = \sum_{p=1}^{P} a_ip x_i + \cdots + b_n \quad i = 1, 2, \ldots, n; \]

where \( y_i \) is the output of the TSK fuzzy model for a crisp input vector \( x = (x_1, \ldots, x_p) \), \( a_ip = A_ip \) are the model consequent parameters. The final global output of the TSK fuzzy model for a crisp input vector \( x = (x_1, \ldots, x_p) \) is calculated using fuzzy mean-weight formula \( \bar{y} = \frac{\sum_{i=1}^{n} \hat{y}_i / \sum_{i=1}^{n} \hat{y}_i} {\sum_{i=1}^{n} \hat{y}_i} \), where \( \hat{y}_i \) represents the degrees of firing (DOF) of the fuzzy rule that is defined as \( \hat{y}_i = \min(\mu_{A_1}(x_1), \ldots, \mu_{A_p}(x_p)) \). The construction of TSK fuzzy model from numerical data proceeds in three steps: fuzzy clustering, setting the membership functions, and parameter estimation, see Al-Marhoun and Osman (2002), El-Sebakhy et al. (2007a,b), Osman and Abdel-Aal (2002) and Osman and Al-Marhoun (2005). The most common architecture of Type1 neuro-fuzzy inference systems in literature is shown in Fig. 2.
4.2. Fuzzy clustering and partitioning based on Gustafson–Kessel scheme

Fuzzy clustering partitioning of the input–output space is performed in the first step, using the selected clustering method. The clustering method utilizes training data, \( D = \{ (x_0, \ldots, x_p; y) \} \), \( i = 1, \ldots, n \) of \( n \) input vectors of dimension \( p \) and one output. Each obtained cluster represents a certain operating region of the system, where input–output data values are highly concentrated. The learning data, divided into these information clusters, are then interpreted as rules. The most popular fuzzy clustering methods in the machine learning and data mining literature are fuzzy c-means (FCM) and Gustafson–Kessel (GK), see Liu et al. (2003) and Taghavi (2005) for more details.

Let \( x_j \) be the input vector for the \( j \)th observation over different samples, and let \( v_i \) be the \( i \)th cluster centroid (prototype). Then a typical distance norm between \( x_j \) and \( v_i \) is \( D_i^2 = (x_j - v_i)^T A (x_j - v_i) \), where \( A \) is a symmetric and positive definite matrix and \( V = \{ v_1, v_3, \ldots, v_n \} \) is a vector of the centroids of the fuzzy clusters \( C_1, C_2, \ldots, C_n \). Use of the matrix \( A \) makes it possible for each cluster to adapt the distance norm to the geometrical structure of the data at each iteration. Therefore, different norms can be induced by the choice of the matrix \( A \). The Euclidean norm is induced when \( A = I \), where \( I \) is an identity matrix. The Mahalonobis distance (norm) is induced when \( A = M^{-1} \), where \( M^{-1} \) is the inverse of the covariance matrix of patterns in the system. Although many clustering methods have been studied in the literature, a common limitation of conventional methods is to use a fixed distance norm for finding clusters; this fixed norm imposes a fixed geometrical structure and finds clusters of that shape even if they are not present. The Euclidean norm-based methods find only spherical shape of clusters and the Mahalanobis norm-based methods find only ellipsoidal ones even if those shapes of clusters are not present in a data set. Based on the norm-inducing matrices, the objective of the GK method is obtained by minimizing the function \( J_m \), that is,

\[
J_m(U, V, A : X) = \sum_{i=1}^{k} \sum_{j=1}^{n} (\mu_{ij})^m D_i^2, \\
\text{where } A = (A_1, A_2, \ldots, A_k) \text{ is a } k\text{-tuple of the norm-inducing matrices, } U, \text{ where } U = (\mu_{ij})_{k \times n} \text{ is a fuzzy partition matrix of } X \text{ satisfying the following constraints:}
\]

- \( \mu_{ij} \in [0,1], 1 \leq i \leq k, 1 \leq j \leq n \), \( \sum_{i=1}^{k} \mu_{ij} = 1 \) for \( 1 \leq j \leq n \) and
- \( 0 < \sum_{i=1}^{n} \mu_{ij} \leq n, 1 \leq i \leq k \), where \( m \in [1, \infty) \) is a weighting exponent that controls the membership degree \( \mu_{ij} \) of each data point \( x_j \) to the cluster \( C_i \).

The choice of appropriate \( m \) value is of importance because the final clusters may vary depending on the \( m \) value selected. As \( m \to 1 \), \( J_1 \) produces a hard partition where \( \mu_{ij} \in \{0,1\} \). As \( m \) approaches infinity, \( J_{\infty} \) produces a maximum fuzzy partition where \( \mu_{ij} = 1/c \).

To obtain a feasible solution by minimizing \( J_m \) the additional constraint is required for \( A_n \), that is,

\[
det(A_i) = \rho_i, \quad \rho_i > 0, 1 \leq i \leq k,
\]

where \( \rho_i \) is a cluster volume for each cluster. This constraint guarantees that \( A_i \) is positive definite, indicating that \( A_i \) can be varied to find the optimal shape of cluster with its volume fixed. Using the Lagrange multiplier method, minimization of the function \( J_m \) with respect to \( A_i \), then we obtain \( A_i = \lfloor \rho_i det(F_i) \rfloor^{1/p} F_i^{-1} \), where the fuzzy covariance matrix of cluster \( C_i \) is defined as

\[
F_i = \lfloor \sum_{j=1}^{n} (\mu_{ij})^m (x_j - v_i)(x_j - v_i)^T \rfloor^{1/m}.
\]

The set of fuzzy covariance matrices is represented as a \( k \)-tuple of \( F = (F_1, F_2, \ldots, F_k) \). Generally, to show that \( A_i \) satisfies the constraint of a symmetric and positive definite matrix, assume that \( x \in R^p \) is any linearly independent data points \( x \in R^p \) in the data set. Then, the matrices \( x x^T \) and \( x x^T \) are symmetric and positive semi-definite and also their weighted sum \( F_i \) and hence \( A_i \) is symmetric and positive definite, see Bezdek et al. (2005) and Dumitrescu et al. (2000) for more details. There is no general agreement on what value to use for \( \rho_i \); without any prior knowledge, a rule of thumb is that many investigators use \( 1 \) for \( \rho_i \) in practice, see Bezdek et al. (2005).

By using the GK clustering algorithm with \( K \) clusters on the data set \( D \), we compute the fuzzy partition matrix, \( U \). The process of this fuzzy clustering is an iterative process as it is shown below: For data set, \( D \), choose the number of clusters \( 1 < K < n \), the weighting exponent \( m > 1 \) and the termination tolerant \( e > 0 \). Initialize the partition matrix \( U^{(0)} \) randomly.

Step 1: Compute cluster means:

\[
\mu_{ij} = \frac{\sum_{i=1}^{n} (\mu_{ij}^{(k-1)})^m x_i}{\sum_{i=1}^{n} (\mu_{ij}^{(k-1)})^m}, \quad k = 1, \ldots, K,
\]

Fig. 2. Adaptive neuro-fuzzy architecture for a two-rule Sugeno system.
where $\mu_{ik}$ is the triangular/Gaussian bell membership function defined above, $\mu(x_i)$.

Step 2: Compute cluster covariance:

$$F_{ik} = \sum_{i=1}^{n}(\mu_{ik}^{(l-1)})^p(x_k - \mu_{ik}^{(l)})^2, \quad k = 1, \ldots, K.$$ 

Step 3: Compute the distances:

$$D_{ik}^2 = |\text{det}(F_{ik})|^{(1/2(n+1)p)} / F_{ik}, \quad i = 1, \ldots, n,$$

where $H_k = [x_k - \mu_{ik}^{(l-1)}]$ for $k = 1, \ldots, K$.

Step 4: Update the partition matrix $\mu_{ik}$ as follows:

$$\mu_{ik}^{(l)} = \begin{cases} \frac{1}{\sum_{i=1}^{n}D_{ik}D_{ik'}}^2 / \sum_{i=1}^{n}C_{ik} & \text{if } D_{ik} > 0, \\ 0 & \text{if } D_{ik} > 0 \\text{and } \mu_{ik}^{(l)} \in [0, 1] \text{ with } \sum_{i=1}^{n}C_{ik} = 1. \end{cases}$$

for $k = 1, \ldots, K$; $i = 1, \ldots, n$ until $\|F^{(l)} - U^{(l-1)}\| < \epsilon$, where $D_{ik}$ is the clustering distance defined in Step 3 and $\mu_{ik}$ is the triangular/Gaussian bell membership function defined above.

At the end of the iterative procedure, the membership values, $\mu_{ik}$, and cluster centers, $V_k$, are obtained. The detected fuzzy clusters in the input–output product space give information on how the data points are structured in the input space. This information, which is captured in the cluster centers and eigen values of the fuzzy covariance matrices, is projected into the input axes to induce the antecedent fuzzy sets. If $v_1, \ldots, v_n$ are the input space coordinates of the ith cluster center, then the antecedent fuzzy sets of the TSK model are defined by the triangular membership as $\mu_{ik}(x_i) = \text{Max}\{0, 1 - |x_i - V_{ik}| / b_{ik}\}$; $k = 1, \ldots, K$, with the center coordinates $v_{ik}$ and the parameters $b_{ik}$ controlling, respectively, the mean and the spread of the membership function (Amahenok et al., 2005; LeCun et al., 1995; McCaig et al., 1998). Finally, the parameters are estimated using the least-square approximation.

Let $X$ denote the matrix whose $i$th row is the input vector $x_i$ and let $Y$ denote the column vector with $y_i$ as its $i$th component. Let $W_{ik}(X)$ denote the $n \times n$ real diagonal matrix that represents normalized firing strength of the $k$th rule for the $i$th observation or sample, $W_{ik}(X) = \beta_{ik}(x_i)^{\sum_{k=1}^{K}\beta_{ik}}$, where $k = 1, \ldots, K$; $i = 1, \ldots, n$. Suppose that $\Theta = \{a_0, \ldots, a_n\}$ denote the vector of consequent parameters of the $i$th rule. In order to estimate the offset term, $a_0$, a unitary column $l$ is appended to the matrix, $X$, to produce the extended matrix $X_0 = [X, \ldots, 1]$. Therefore, the unknown parameters $\Theta_i$ are calculated via least-square criterion, $\Theta_i = [X_0^T \cdot W_{ik} \cdot X_0]^{-1} \cdot X_0^T \cdot W_{ik} \cdot y_{ik}$. The end, the output of Type1 neuro-fuzzy logic inference systems model is approximated by $\mu(x) = X \cdot \Theta$.

5. The empirical study, discussion and comparative studies

We have done the quality control check on all these data sets and remove redundant data and un-useful observations. To evaluate performance of each modeling scheme, the entire database is divided using the stratified criterion. Therefore, we use 70% of the data for building Type1 Fuzzy 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. Thus, the data were divided into two/three groups for training and the cross-validation check. Hence, 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 the model to evaluate its accuracy and trend stability. For the testing data, a statistical summary was used to investigate different quality measures corresponding to Type1 Fuzzy scheme, feedforward neural networks system, and the most popular empirical correlations in the literature to predict both bubble point pressure and oil formation volume factor.

Generally, after training Type1 fuzzy inference systems, 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 Type1 fuzzy model versus both the standard neural networks and the three commonly published empirical correlations, namely Standing, Al-Marhoun, and Claso 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) with a single hidden layer feedforward neural network based on the back propagation (BP) learning algorithm with both linear and sigmoid activation functions. The initial weights were generated randomly and the learning technique is achieved based on 1000 epoch or 0.001 goal error and 0.01 learning rate. Each layer contains neurons that are connected to all neurons in the neighboring layers. The connections have numerical values (weights) associated with them, which will be adjusted during the training phase. Training is completed when the network is able to predict the given output. For the two models, the first layer consists of four neurons representing the input values of reservoir temperature, solution gas-oil ratio, gas specific gravity and API oil gravity. The second (hidden) layer consists of seven neurons for the $P_0$ Model, and eight neurons for the $B_{06}$ model. The third layer contains one neuron representing the output values of either $P_0$ or $B_{06}$. Simplified schematic of the used neural networks for $P_0$ and $B_{06}$ models are illustrated in Al-Marhoun and Osman (2002) and Osman and Abdel-Aal (2002). It gives the ability to monitor the generalization performance of the network and prevent the network to over fit the training data based on repeating the computations for 1000 times and take the average over all runs.

We have used both triangular and Gaussian Bell membership functions during the implementation processes with both grid partition and subtractive clustering with radius 0.1 based on two different learning criteria, such as backpropagation and least squares. A combination of both least squares and back propagation gradient descent methods were used for training fuzzy inference system membership function parameters and is applied to emulate a given training data set. The resulting weights for the $B_{06}$ and $P_0$ 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 the $B_{06}$ and $P_0$ models as shown below.

5.2. Discussion and comparative studies

For the purpose of comparison, we use the same feature selection criterion and stratified sampling cross-validation schemes similar to El-Sebakhy (2008), namely the recursive feature elimination method. In order to evaluate the performance of the adaptive neuro-fuzzy inference systems for reservoir characteristics properties, the three above-mentioned benchmarks were utilized. Based on the obtained results within the external validation check (testing and validation), we summarized the results in Tables 1–6. The reader can observe from these results that Type1 Fuzzy intelligence modeling scheme outperforms both neural and the most common published empirical correlations, similar to the same analysis within El-Sebakhy (2008). In addition, this new intelligence framework scheme has
demonstrated high accuracy in predicting the $B_{ob}$ values with stable performances and achieved the lowest absolute percent relative errors, lowest minimum errors, lowest maximum errors, lowest root-mean-square errors, and the highest correlation coefficients among other correlations for the three distinct databases.

Figs. 3–5 illustrate six scatter plots of the predicted results versus the experimental data for both $P_b$ and $B_{ob}$ values using the provided 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 Type1 Fuzzy modeling scheme. The reader can compare theses patterns with the corresponding ones of the published neural networks modeling and common empirical correlations in Al-Marhoun and Osman (2002), Osman et al. (2001), and Osman and Abdel-Aal (2002).

Figs. 6 and 7 shown the relationship of the measurements of both absolute percent relative error ($E_A$) and correlation coefficient ($R^2$) for each utilized computational data mining schemes, namely Type1 fuzzy logic inference systems scheme, feedforward neural networks (the used computational intelligence schemes), and three empirical correlations. Each modeling scheme is represented by a symbol (point); a good forecasting scheme should have the highest correlation and
lowest absolute percent relative error. By looking at these graphs, we have observed for instance, in estimating \( B_{ob} \) based on the data set used in Abdulraheem et al. (2007), the symbol corresponding to Type1 fuzzy has the smallest absolute percent relative error, \( E_A = 1.3218\% \), the largest correlation coefficient, \( R^2 = 0.9970 \), and the smallest standard deviation, \( SD = 0.7876 \), while neural network is below Type1 fuzzy logic inference systems scheme with \( E_A = 1.7886\% \), \( R^2 = 0.9878 \), and \( SD = 0.89835 \). The other empirical correlations indicates higher error values with lower correlation coefficients, for instance, Al-Marhoun (1992) has \( E_A = 2.2053\% \), \( R^2 = 0.9806 \), and \( SD = 1.2842 \); Standing has \( E_A = 2.7238\% \), \( R^2 = 0.9743 \), and \( SD = 2.673 \); and Glaso Correlation with \( E_A = 3.3743\% \), \( R^2 = 0.9715 \), and \( SD = 2.583 \).

Similarly, for the bubble point pressure, \( P_b \) based on the used data sets used in Al-Marhoun and Osman (2002), El-Sebakhy et al. (2007a, b), Goda et al. (2003), and Osman et al. (2001) and Osman and Abdel-Aal (2002), we observed that the symbol corresponding to Type1 Fuzzy scheme has the smallest absolute percent relative error, \( E_A = 14.224\% \), the largest correlation coefficient, \( R^2 = 0.9520 \), and \( SD = 11.337 \), while neural network is below Type1 Fuzzy with \( E_A = 21.017\% \), \( R^2 = 0.943 \), and \( SD = 14.871 \).
The other correlations indicate higher error values with lower correlation coefficients. Al-Marhoun (1992) has $E_R = 20.011\%$, $R^2 = 0.906$, and $SD = 12.839$; Standing has $E_R = 67.73\%$, $R^2 = 0.867$, and $SD = 25.171$; and Glaso empirical correlation with $E_R = 18.523\%$, $R^2 = 0.945$, and $SD = 25.159$. Overall computations, we observed that the new intelligence system framework has a reasonable values of the relative errors, $E_r$, especially, by looking at Tables 5 and 6. Type1 neuro-fuzzy system has the smallest $E_r$ values compared to the other published techniques. This indicator can be considered to say that the performance of the new framework (Type I fuzzy logic inference system) is the one with the least bias, the most reliable technique with the highest correlation coefficients.

The same implementations processes may be repeated for the other statistical quality measures, but for the sake of simplicity, we did not include it in this context. Finally, we conclude that Type1 fuzzy inference systems modeling scheme has a better and a reliable performance compared to other published modeling schemes and empirical correlations. It can be concluded that, Type1 fuzzy inference systems modeling scheme outperforms both the standard feedforward neural networks and the most common published empirical correlations in predicting both $P_b$ and $B_{ab}$ 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 databases were utilized to investigate the capabilities of Type1 fuzzy logic inference systems for predicting the PVT properties of oil crude systems, namely both bubble point pressure and oil formation volume factor. Based on the obtained results and comparative studies, we conclude that the developed Type1 fuzzy inference systems scheme showed a high accuracy in predicting the $B_{ab}$ values with a stable performance, and achieved the lowest absolute percent relative error, lowest minimum error, lowest maximum error, lowest RMSE, and the highest $R^2$ among other correlations for the used three distinct data sets. Based on the performance of the new framework (Type I fuzzy logic inference system) and on the GK clustering criterion, we demonstrate that the new framework avoids the risk of overfitting and complexity problems. Thus, there is no lack of robustness in predicting the one with the least bias. In addition, it outperforms the most common empirical correlations. Furthermore, Type1 fuzzy inference modeling scheme is flexible, reliable, and shows a bright future in implementing it for the oil and gas industry, especially for permeability and porosity prediction, feature selection, rock mechanics properties, flow regimes, and liquid-holdup multiphase follow, 3D seismic data, and facies classification.

Acknowledgements

The author wishes to thank the Editor, the associate editor, and referees for their careful reading of the manuscript and providing many helpful comments. He is also grateful to MEDai Inc. research and development team Orlando, FL, USA, and Mansoura University, Egypt, for the facilities utilized to perform the present work and for their support.

References


Al-Shammasi, A.A., 2001. A review of bubble point pressure and oil formation volume factor correlations. SPE Reservoir Evaluation and Engineering, 146–160. This paper (SPE 71302) was revised for publication of paper SPE 53185.


Taghavi, A.A., 2005. Improved permeability estimation through use of fuzzy logic in a carbonate reservoir from southwest Iran. Presented at SPE Middle East Oil and Gas Show and Conference, March 12–15, Bahrain.