Petroleum 6 (2020) 375–384

Contents lists available at ScienceDirect

Petroleum

journal homepage: http://www.keaipublishing.com/en/journals/petroleum

Transparent open-box learning network and artificial neural network predictions of bubble-point pressure compared

David A. Wood[a,*], Abouzar Choubineh[b]

[a] DWA Energy Limited, Lincoln, United Kingdom
[b] Petroleum University of Technology, Ahwaz, Iran

HIGHLIGHTS

- Transparent open-box (TOB) algorithm offers a fully auditable prediction tool.
- TOB complements neural-network algorithms in its prediction performance.
- TOB exploits spreadsheet functions, for example using Excel's Solver optimizers.
- TOB's non-linear capabilities provide credible bubble-point pressure predictions.
- TOB works less well with sparse data sets, consequently it avoids overfitting issues.

ARTICLE INFO

Keywords:
Learning network transparency
Learning network performance compared
Prediction of oil bubble point pressure
Over fitting data sets for prediction
Auditing machine learning predictions
TOB complements ANN

ABSTRACT

The transparent open box (TOB) learning network algorithm offers an alternative approach to the lack of transparency provided by most machine-learning algorithms. It provides the exact calculations and relationships among the underlying input variables of the datasets to which it is applied. It also has the capability to achieve credible and auditable levels of prediction accuracy to complex, non-linear datasets, typical of those encountered in the oil and gas sector, highlighting the potential for underfitting and overfitting. The algorithm is applied here to predict bubble-point pressure from a published PVT dataset of 166 data records involving four easy-to-measure variables (reservoir temperature, gas-oil ratio, oil gravity, gas density relative to air) with uneven, and in parts, sparse data coverage. The TOB network demonstrates high-prediction accuracy for this complex system, although it predictions applied to the full dataset are outperformed by an artificial neural network (ANN). However, the performance of the TOB algorithm reveals the risk of overfitting in the sparse areas of the dataset and achieves a prediction performance that matches the ANN algorithm where the underlying data population is adequate. The high levels of transparency and its inhibitions to overfitting enable the TOB learning network to provide complementary information about the underlying dataset to that provided by traditional machine learning algorithms. This makes them suitable for application in parallel with neural-network algorithms, to overcome their black-box tendencies, and for benchmarking the prediction performance of other machine learning algorithms.

1. Introduction

Machine learning algorithms are now essential and widely-used tools for predicting key variables for complex oil and gas systems with multiple influencing variables displaying highly irregular and/or non-linear relationships. Their application and diversity are growing [1]. However, one of the drawbacks to much of machine-learning algorithms is their lack of transparency, i.e., their inability to readily reveal the detailed calculations and correlations they apply in generating their predictions for each data record.

This leads to frustration and scepticism amongst users and those painstakingly generating precise experimental measurements and establishing auditable empirical formula to explain the relationships among their contributing variables. To many, such machine learning algorithms are merely black-box tools [2] unable to provide much insight to the underlying systems and datasets from which they are

Peer review under responsibility of Southwest Petroleum University.
* Corresponding author.
E-mail address: dw@dwasolutions.com (D.A. Wood).

https://doi.org/10.1016/j.petlm.2018.12.001
Received 11 May 2018; Received in revised form 3 December 2018; Accepted 7 December 2018
Available online 12 December 2018

making their predictions. Often the predictions from machine learning algorithms are supported with statistical demonstrations of very high levels of apparent accuracy (e.g., correlation coefficients R2 of > 0.99 for actual versus predicted datasets). Such claims are often dismissed, sometimes unjustifiably, because of data "over-fitting" issues.

Wood [3] has proposed a transparent open-box (TOB) learning-network algorithm that does not generate its predictions using neural networks nor by generating unauditable correlations. The TOB algorithm instead uses a matching, ranking and optimization technique in which all the weights applied and contributions to each prediction are revealed and auditable. This alternative approach to learning networks can provide reliable predictions of significant accuracy that are useful to compare with the less-transparent machine-learning approaches.

Machine learning algorithms now routinely applied to many complex oil and gas systems (with typically little associated transparency available in the predictions they make) are artificial neural networks (ANN) combined with various optimization algorithms, least squares support vector machine (LSSVM), Adaptive Neuro-Fuzzy Inference Systems (ANFIS) and radial basis function networks (RBFN). Some insight to the inner workings of these algorithms can be obtained by running a series of simulations to measure the significance of the impact each input variable has on its predictions across a specific data set. Such information is better than no insight, but it falls short of providing details of the exact calculations involved in each prediction. Such simulations lead to partial transparency, e.g., so-called "white boxes" [4].

On the other hand, the TOB algorithm can be applied to provide a much greater degree of transparency. Here we demonstrate its application and performance in the prediction of bubble-point pressure from a published pressure-volume-temperature (PVT) dataset and compare its performance with that of a trained ANN algorithm applied to the same dataset. Accurate PVT data is essential for many reservoir engineering calculations (e.g., oil reserves and resource estimates, well inflow and well-test analysis and reservoir simulations). These routinely require quantified input in terms of reservoir fluid properties including bubble point pressure, formation volume factor (FVF) at bubble point, gas-to-oil ratio (GOR). These fluid properties vary among geographic locations and crude oil types a need to be established by laboratory analysis, which is expensive and time consuming, or by predictive methods based upon easy-to-obtain variables such as temperature, gas gravity, GOR and oil density or API gravity.

Traditionally, the most common approach used to achieve predictions for key reservoir fluid properties is to apply PVT correlations. However, these correlations cannot be reliably applied in as universally-applicable models because they are based on limited data sets of crude oil from specific geographic locations. Consequently, the petroleum industry continuously endeavours to improve the accuracy of predicting reservoir-fluid properties and to establish reliable and universal methodologies for doing so that do not rely on correlations relevant to limited crude oil types. It is for this reason that we focus on applying the TOB methodology to predict bubble-point pressure, one of the most widely applied reservoir fluid properties in reservoir engineering calculations.

2. The transparent open-box learning network methodology

The TOB learning network methodology [3] consists of fourteen steps divided into two distinct stages (network construction and network optimization) described in summary in Fig. 1 and its detailed mathematical basis is described in Appendix 1. The matching and ranking process involved in steps 1 to 10 (stage 1) typically on its own generates predictions with impressive levels of accuracy from highly complex and non-linear distributions of underlying variables. The optimization process, conducted in steps 11 to 12, refines these predictions by varying the variables *Q* (the number of high-matching records to use in the predictions) and *wN* (weights applied independently to each of the N input variables to determine their relative contributions to the predictions) applied to the tuning subset. Step 13 applies the optimized network setting to the testing subset (consisting of data records not involved in the network building or tuning steps). Depending on the level of prediction accuracy achieved Step 14 decides on if and how the TOB network should be deployed or used to benchmark more complex machine learning algorithms. A key attribute of the method is that the calculations involved in predicting the dependent variable in each data record are fully accessible.

The root mean squared error (RMSE) and correlation coefficient (R^2) of the actual versus predicted values of the dependent variable are useful prediction-accuracy measures. These are used here to assess the performance of the TOB and that of other machine-learning algorithms.

The TOB learning network is relatively easy to setup and code. For small to mid-sized data sets it can use a spread-sheet platform. This enables it to use standard built-in optimizers (e.g. the generalized reduced gradient -GRG, and evolutionary optimizers of Excel's Solver optimization function). For large datasets (i.e. several thousand data records) the TOB can be readily coded with a customized optimizer in the standard mathematical coding languages (i.e., Octave, R, Python, MatLab etc.). For the prediction of bubble-point pressure from a dataset of 166 data records presented here, a hybrid approach is adopted, i.e., the TOB network is displayed in Excel with some calculations conducted VBA coding, but sufficient spreadsheet formula involved to utilize Excel's solver optimizer for steps 11 and 12.

3. Applying a TOB learning network to predict oil bubble point pressure (P_b)

Reservoir fluid properties, such as bubble-point pressure (P_b), derived by laboratory test from PVT datasets, provide key metrics in the calculation of recoverable resources of oil and gas from specific subsurface reservoirs. P_b is defined as the highest pressure at which the first gas bubble appears under the fixed temperature conditions. However, measuring P_b precisely by experimental methods in laboratories is time-consuming and expensive. To avoid such issues, many correlations relating bubble-point pressure to its underlying PVT metrics have been proposed to predict P_b. The work of Katz [5] and Standing [6] led the way, and, studies on larger datasets from oil fields across the globe [7,8] have expanded on these early correlations. Several empirical relationships for P_b derived using specific datasets [9,10], and related to equations of state [11], are now available.

Additionally, several studies have applied machine-learning (artificial intelligence) methods, especially neural-network models, to derive reliable prediction correlations for P_b from PVT data [12–17]. These learning networks facilitate the prediction of P_b from the more-easily measured input metrics, but typically do not provide details of the correlations involved in the predictions they generate.

The TOB learning-network is configured here to evaluate the published PVT dataset [18] for 22 bottom-hole samples of crude oils from Pakistan for the purposes of predicting P_b. An ANN model was recently applied [19] to this dataset that predicted P_b. This data set provides 166 data records from which P_b can be estimated (Table 1). For the TOB model, only four easy-to-measure variables (temperature- *T* (F), solution gas to oil ratio -*Rs* (scf/stb), specific gravity of gas - γ_g, and *API* gravity of the oil) are used as input metrics for each data record. We also construct and apply an ANN model to the same divisions of the 166 data records into training, tuning and testing subsets as used in the TOB model: i.e., training subset (99 records); tuning subset (34 records); and, testing subset (33 records). The roles of the three data subsets in the TOB methodology is explained in detail in Appendix 1 (also refer to steps 6 and 13 in Fig. 1). Specifically, the testing subset is extracted from the complete data set and not used in the tuning of the training

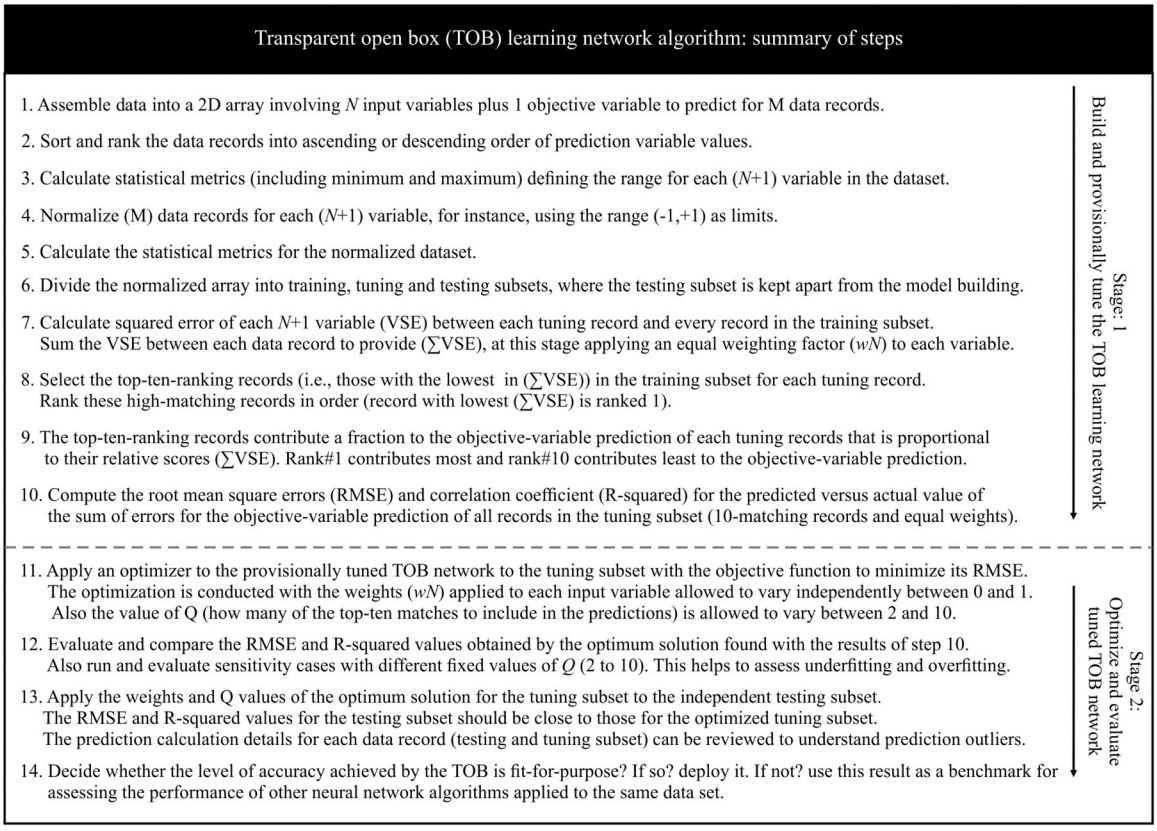

Fig. 1. Summary of the steps and stages involved in building a TOB learning network.

Table 1
PVT dataset for Pakistani crude oils (Al-Marhoun,1996) statistical summary of data record values to which the TOB learning network Wood [3] and an ANN model are applied to predict bubble-point pressure.

Dataset Used for Bubble Point Pressure (P_b) Prediction			
Dataset:166 data records	Min	Max	Mean
Reservoir temperature T (oF)	182	296	242
Solution gas to oil ratio R_s (scf/stb)	92	2496	500
Specific gravity of gas γ_g (Air = 1)	0.8253	3.4445	1.760
Oil Gravity (degrees API)	29.0	56.5	39.1
Bubble-point Pressure at bubble point P_b (psi)	79	4975	952

subset (TOB steps 7 to 12). The data records selected for the testing data set are not chosen randomly but spread intermittently and arbitrarily across the entire dependent variable value range covered by the full data set. The data records representing the maximum and minimum dependent-variable values for the full dataset are placed in the training subset defining the limits of the prediction range and are therefore not available to be selected for the testing subset.

The results listed for the TOB model (Table 2) show that it achieves its optimum P_b prediction performance (based on the difference between measured P_b values and predicted P_b values: RMSE = 132; R^2 = 0.9729) when applied to the full suite of data records (following stage 2 optimization) with settings of Q = 3 and weights wT = 0, wRs = 0.01025, $w\gamma_g$ = 1; $wAPI$ = 0. Sensitivity analysis (Table 2) applying different values of Q and different allocations of data records between the training, tuning and testing subset suggests that the TOB can generate acceptable levels of prediction performance when applied to this entire dataset. Varying the value of Q between 2 and 10 (Table 2) does not display a dramatic range of prediction outcomes (R^2 hovers around 0.97 for all values of Q in that range, and RMSE varies between 132 and 163 for that range of Q). The highest RMSE value is for Q = 2, which suggests at that value the data is being under-fitted. The inability of the TOB to provide a higher level of prediction accuracy than $R^2 \sim 0.97$, is due to the sparse data coverage for Pb values greater than about 1900 psi, as further analysis demonstrates.

Artificial neural networks (ANN) consist of neurons that are arranged in layers and highly interconnected to simplistically mimic the neural activities of the human brain. The ANN achieves learning by adjusting the weights of its neuron connections (i.e., synaptic weights) [20,21]. This adjustment is achieved by applying a learning algorithm to the ANN. A popular ANN, which is widely applied to many complex datasets, is the feed forward multilayer perceptron (MLP) processed with a back propagation learning algorithm. The theory and methodology of MLP is well established and documented [20,21] and applied to PVT data analysis for predicting petroleum reservoir fluid properties [15]. ANN theory and mathematical formulation is therefore not discussed further here.

The ANN model (MatLab-based), applied to the same subset configuration of data records (Fig. 2), involves two hidden layers (hidden layer 1 has 6 neurons; hidden layer 2 has 5 neurons). This ANN applies the following activation/transfer functions (purelin between input layer-hidden layer 1; logsig between hidden layer 1 and hidden layer 2; purelin between hidden layer 2- output layer). It is tuned with 1000 iterations of the applying a back-propagation algorithm optimizing mean squared error between measured and predicted P_b values. The mathematical relationships developed by the ANN algorithm, involving weighting, biases and transformation functions lead to complex correlations that should be expected, in most circumstances, to be able to out-perform the prediction accuracy of TOB when applied to complex data sets.

Table 2
Bubble point pressure (P_b) prediction performance of TOB learning network applied to the full 166-record data set.

Transparent Open Box (TOB) Learning Network Results and Variable Weightings in the Prediction of Bubble Point Pressure (Dataset with 166 records)

Variable Description	Variable Number	Pre-optimization Equal Weightings	Best Solution GRG Multi-start	Best Solution Solver Evolutionary Algorithm	Sensitivity Analysis with Q constrained to integers progressively from 10 to 2 (All cases runs with for the 34 records of the tuning subset with the Solver GRG optimizer configured in the same way)								
Q Constained to	Integer Constraints		2 to 10	2 to 10	10	9	8	7	6	5	4	3	2
Q selected for solution	Integer #		3	3	10	9	8	7	6	5	4	3	2
Prediction Performance of Optimum and Constrained Optimum Solutions Applied to the Tuning Subset (34 records: ~20.5% of total dataset)													
RMSE	psi	266.0	**132.0**	**132.3**	148.8	148.2	147.9	146.1	140.1	140.4	141.4	132.0	163.2
R2	fraction	0.9313	**0.9729**	**0.9731**	0.9684	0.9674	0.9672	0.9677	0.9706	0.9700	0.9712	0.9729	0.9706
Weightings (0 < = w < = 1) Applied to constrained optimum solutions for the tuning subset													
Temperature T	#1	0.500	**0**	**0**	0.00114	0.00429	0.00285	0.00815	0.00620	0.03387	0.01210	0	0.00222
Gas Oil Ratio Rs	#2	0.500	**0.01025**	**0.02335**	0.00114	0.00466	0.00236	0.00574	0.00396	0.01552	0.01422	0.01025	0.04203
Gas Gravity γg	#3	0.500	**1.00000**	**0.92236**	0.03128	0.13446	0.07416	0.17878	0.11861	0.91111	0.33014	1.00000	0.00044
Oil API Gravity	#4	0.500	**0**	**0.51568**	0.29323	1.00000	0.39958	0.74785	0.47630	0.06235	0.56736	0	0
Ratio of weight for T to weight for Rs			**0.00000**	**0.00000**	1.00000	0.92060	1.20763	1.41986	1.56566	2.18235	0.85091	0.00000	0.05282
Prediction Performance of Optimum Solution Variable Weightings and Q Value Applied to the Testing Subset (33 records: ~19.9% of total dataset)													
RMSE	psi		**291.4**										
R2	fraction		**0.9373**										

The value ranges and non-linear relationships between the input variables, *T, Rs,* γ_g*, API* and the dependent variable P_b are illustrated in Fig. 3a–d. As well as non-linearity, these figures highlight the irregularity of the spread of data records across the P_b range covered by the full suite of data records. There is reasonably dense data record cover for P_b values less than about 1900 psi and sparse data record cover for P_b greater than about 1900 psi. Solution gas-oil ratio (*Rs*) shows a reasonable positive correlation with P_b within the data record suite ($R^2 = 0.7595$). Oil gravity (API) and reservoir temperature show almost no correlation with P_b. Gas gravity (γ_g) has a poor negative correlation with P_b ($R^2 = 0.3798$). However, that negative relationship all but disappears for P_b values greater than about 1900 psi. *Rs* is, therefore, the input metric best correlated with P_b, and has the greatest discriminatory impact in the record-match selections established during stage 1 of the TOB network development.

Figs. 4 and 5 compare the P_b predictions generated by the TOB and ANN models applied to the full suite of dataset records distributed in the same way between training, tuning and testing subsets. Both models produce predictions of P_b to a high degree of accuracy. For the tuning (Fig. 4) and testing (Fig. 5) subsets, as expected, the ANN yields a higher correlation coefficient and a lower RMSE value, for predicted versus measured data, than predictions produced by the TOB model. The lower level of prediction accuracy achieved by the TOB model is primarily due to the sparsity of data points in the full dataset for P_b values above 2000 psi. There are only 15 data records in the entire data set with P_b values greater than 1900 psi. Once these 15 data records are distributed between training, tuning and testing subsets it means that data coverage of the P_b interval from 1900 psi to 4975 psi is very limited in each subset. Whereas the ANN model can derive credible correlations from sparse data sets, the TOB network, because it does not work be generating correlations, is not able to do so.

The complex correlations generated within the ANN algorithm tends to perform particularly well with sparse and clustered datasets, such as the one evaluated here. As the TOB algorithm does not generate correlations as part of its prediction methodology it is unable to extrapolate with accuracy outside of the well-covered data ranges. This is because the TOB methodology is based on closely matching test data records with other records within the network and then adjusting the contributions of the best matching data records to the predictions by variably weighting the input variables. To illustrate this point, the TOB algorithm is applied to the portion of the dataset that is well covered by data records (i.e., $P_b < 1900$ psi) by removing the 15 data records with $P_b > 1900$ psi. Table 3 and Figs. 6 and 7 show the prediction performance results for the TOB applied to that dataset consisting of 151 data records (split: training subset 89 records; tuning subset 31 records; testing subset 31 records). In this data range the prediction accuracy achieved by the TOB, more closely matches that of the ANN, as seen by comparing Fig. 6 with Fig. 4 (right graph) and Fig. 7 with Fig. 5 (right graph).

The results listed for the TOB model (Table 3) applied to the reduced dataset (151 data records with $P_b < 1990$ psi) show that it achieves its optimum P_b prediction performance for the tuning subset (based on the difference between measured P_b values and predicted P_b values: RMSE = 68; $R^2 = 0.9805$). This is achieved following stage 2 optimization with settings of Q = 6 and weights $wT = 0.07038$, $wRs = 0.87741$, $w\gamma_g = 0.21069$; $wAPI = 0$. Sensitivity analysis (Table 2) applying different values of Q and different allocations of data records between the training, tuning and testing subset suggests that the TOB can generate acceptable levels of prediction performance when applied to this more densely populated portion of the dataset.

Varying the value of Q for the tuning subset between 2 and 10 (Table 3) for this reduced dataset, reveals that R^2 hovers around 0.98 for all values of Q of 4 and above in that range (with RMSE varying between about 68 and 77 for that range of Q). The higher RMSE values for Q = 2 and Q = 3, and much lower R^2 values, suggests at that for those Q values the data is being under-fitted. The TOB learning network

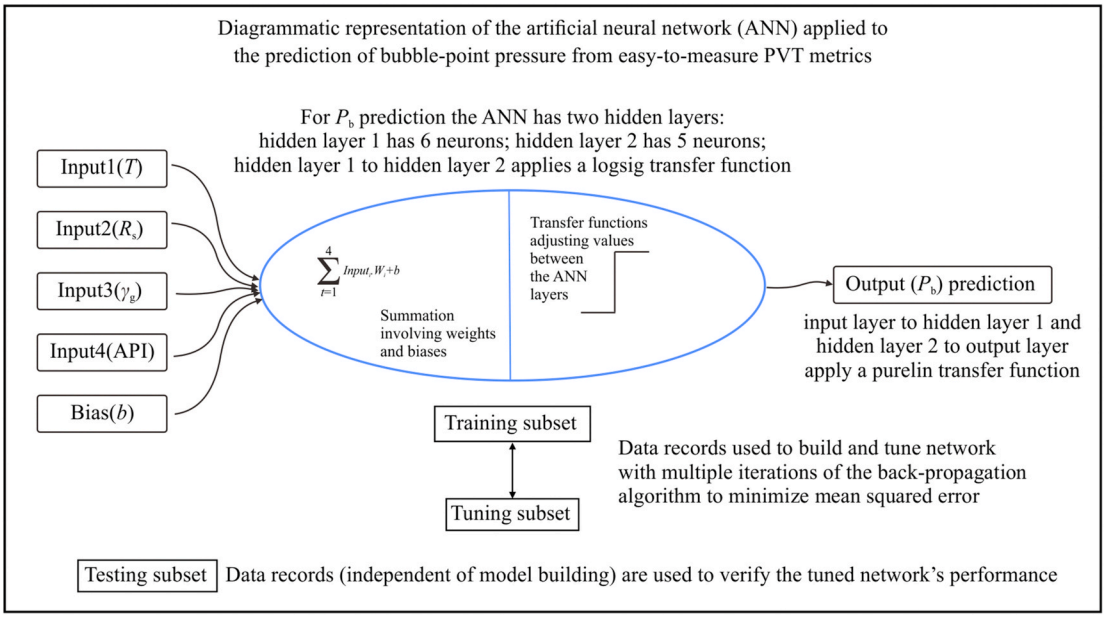

Fig. 2. Structure and methodology applied to the ANN model used to compare performance with the TOB learning network applied to the same P_b dataset.

clearly provides a higher level of prediction accuracy (RMSE < ~70; R^2~0.98) due to the dense spread of data records in this data subset limited to P_b < 1900 psi.

The inability of the TOB methodology to generate highly accurate predictions in sparse data regions, means that is less prone to overfitting than those machine-learning algorithms generating complex correlations to develop their predictions. The TOB relies on surrounding data points with close matches to the data record being predicted rather than on complex correlations establish by neural-network-type algorithms. The more-widely spaced the surrounding data points, and the more non-linear and clustered the relationship between the input variables, the less precise are the TOB's predictions are likely to be. This

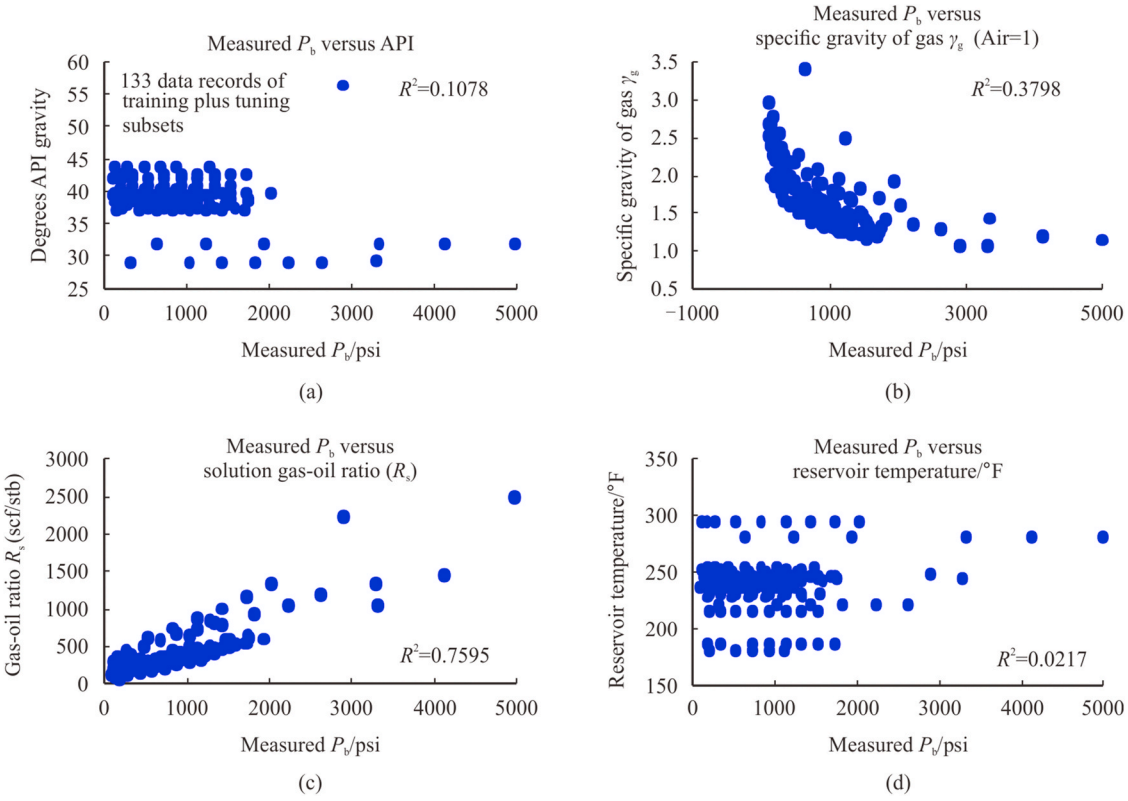

Fig. 3. (A to D). T, R_s, γ_g, and API relationships with Pb in the training and tuning subsets used for the TOB network application.

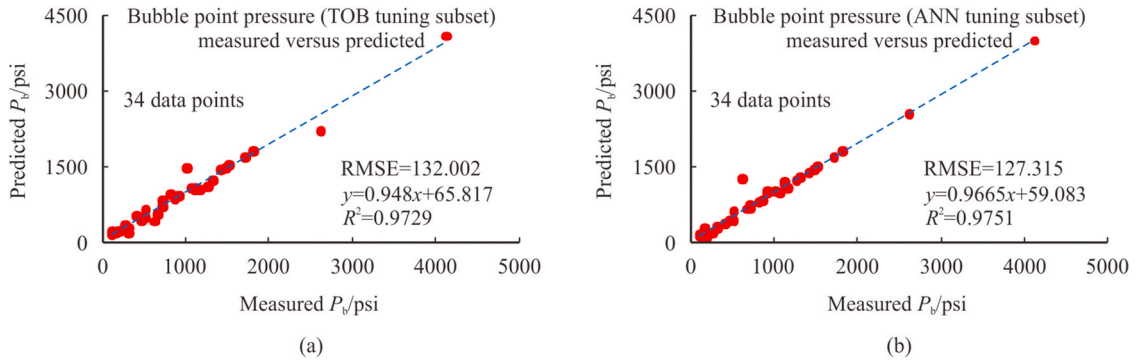

Fig. 4. Predicted versus measured bubble-point pressure for the tuning subset (34 records). 99 records in the tuning subset.

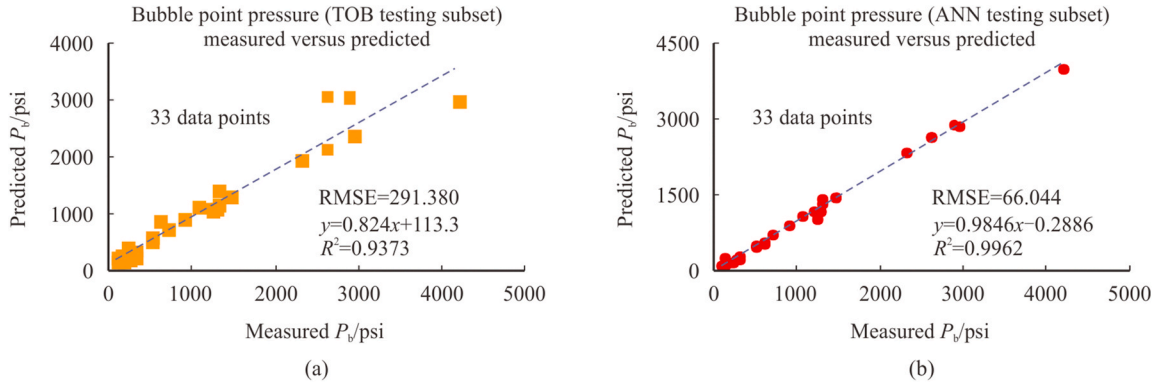

Fig. 5. Predicted versus measured bubble point pressure for the testing subset (33 records). 99 records in the tuning subset.

shortcoming of the TOB can be viewed as an attribute, in that it inhibits the algorithm from over-fitting sparse data sets.

The lower R^2 value achieved by the TOB algorithm than the ANN algorithm applied to the full dataset draws attention to that sparsely populated region of the dataset. Although, the ANN prediction accuracy achieved $R^2 > 0.99$ (Fig. 5, right graph) there is a strong possibility that it could not be relied upon to predict with such accuracy when applied to a more extensive data set in the P_b region > 1900 psi. The fact that the ANN correlations are based on such few data records (i.e., just 15 data records in the P_b region > 1900 psi) runs the risk that it has probably over-fitted the limited data available. Although the ANN tuned network performs well in predicting the few records in the testing subset in that range, its reliability when applied to more data records for that range remains in doubt. The fact that the ANN model is unable to reveal the details of the calculations involved in the prediction of each data record reinforces that doubt.

In contrast to the ANN algorithm, the TOB algorithm is fully transparent. This means that each of its prediction can be easily accessed, audited and assessed for its suitability. This attribute is particularly useful when evaluating prediction outliers. Furthermore, the TOB's Q factor enables sensitivity analysis that can reveal the degree to which its prediction assumptions are overfitting or underfitting the underlying training subset. This means that over-fitting and under-fitting are more easily avoided using the TOB network.

However, the TOB algorithm clearly has limitations in its feasible applications. As demonstrated by the data set evaluated here the TOB algorithm cannot provide highly-accurate prediction performance for clustered or sparsely populated data sets. Moreover, its methodology restricts it providing predictions that are within the minimum and maximum dependent-variable range covered by the training subset. It is unable to meaningfully extrapolate beyond that range because it does not develop correlations between its input variables, but rather relies on matches with data records in the training subset.

The TOB and ANN algorithms to some extent complement each other, with the strengths of one offsetting the weaknesses in the other. It is useful to use the TOB network to benchmark the performance of an ANN. Whereas, for complex datasets, the ANN should be able to outperform the prediction performance of the TOB methodology and/or extend its predictions to wider ranges of the dependent variable value, some uncertainty remains associated with the opaque underlying correlations established by ANN. Benchmarking an ANN's performance accuracy to that achieved by the TOB methodology helps to clarify this uncertainty to a degree. In some cases, the improvement in prediction accuracy offered by ANN may not justify sacrificing the ability to audit each individual prediction calculation (possible with TOB but not ANN). Moreover, by applying TOB and ANN methods better insight into regions of the dataset prone to overfitting are more likely to be revealed.

4. Conclusions

Transparency of prediction can be achieved for complex non-linear data sets can be readily achieved by applying the transparent open-box (TOB) learning network. There are both benefits and drawbacks to the degree of transparency it provides. The TOB algorithm calculates predictions differently from other machine-learning algorithms, by not

Table 3
Bubble point pressure (P_b) prediction performance of TOB learning network applied to the 151-record data set (excluding 15 data records with P_b greater than 1900 psi.

Transparent Open Box (TOB) Learning Network Results and Variable Weightings in the Prediction of Bubble Point Pressure (Dataset with 151 records −Pb < 1900 psi)

Variable Description	Variable Number	Pre-optimization Equal Weightings	Best Solution Solver GRG Multi-start	Best Solution Solver Evolutionary Algorithm	Sensitivity Analysis with Q constrained to integers progressively from 10 to 2 (All cases runs with for the 31 records of the tuning subset with the Solver GRG optimizer configured in the same way)								
Q Constained to	Integer Constraints		2 to 10	2 to 10	10	9	8	7	6	5	4	3	2
Q selected for solution	Integer #		6	6	10	9	8	7	6	5	4	3	2
Prediction Performance of Optimum and Constrained Optimum Solutions Applied to the Tuning Subset (31 records: ~20.5% of total dataset)													
RMSE	psi	89.1	**67.7**	**132.3**	76.2	73.9	73.0	69.3	67.7	71.0	77.3	144.7	142.0
R2	fraction	0.9687	**0.9805**	**0.9731**	0.9777	0.9777	0.9780	0.9677	0.9805	0.9778	0.9732	0.906	0.9103
Weightings (0 < = w < = 1) Applied to constrained optimum solutions for the tuning subset													
Temperature T	#1	0.500	**0.07038**	**0.07913**	0.11830	0.17762	0.07025	0.05970	0.08015	0.07766	0.17888	0	0.07018
Gas Oil Ratio Rs	#2	0.500	**0.87741**	**0.98709**	0.85147	0.99337	0.56238	0.71761	0.99915	1.00000	0.01422	0.95096	0.96261
Gas Gravity γg	#3	0.500	**0.21069**	**0.23696**	0.18225	0.36689	0.23190	0.23493	0.23992	0.17445	0.33014	0.07947	0.18972
Oil API Gravity	#4	0.500	**0**	**0**	0.02425	0	0	0	0	0	0.56736	0.02763	0
Ratio of weight for T to weight for Rs			**0.08021**	**0.08016**	0.13894	0.17880	0.12492	0.08320	0.08022	0.07766	12.57937	0.00000	0.07290
Prediction Performance of Optimum Solution Variable Weightings and Q Value Applied to the Testing Subset (31 records: ~20.5% of total dataset)													
RMSE	psi		**56.9**										
R2	fraction		**0.9864**										

Fig. 6. Predicted versus measured bubble point pressure for the tuning subset (31 records) applied to the 151-record dataset (excluding the 15 data records with Pb > 1900 psi).

Fig. 7. Predicted versus measured bubble point pressure for the testing subset (31 records) applied to the dataset (excluding the 15 data records with Pb > 1900 psi).

establishing correlations between the underlying input variables. Instead, it depends on closely matching (in stage 1) each data record with a specific number (from 2 to 10) of data records in the training subset. This enables the details of each prediction to be readily revealed and audited.

The TOB algorithm (stage 2) can be set up to utilize standard spreadsheet optimizers (e.g., Excel's Solver options), which simplifies its optimization process. Additionally, sensitivity analysis by varying its Q value helps to reveal conditions leading to underfitting or overfitting of the underlying data set. This makes it useful as a performance benchmark for ANN and other machine-learning algorithms. complex neural and fuzzy network algorithms.

Despite its advantages, the TOB learning network does not perform well with sparse or highly clustered datasets. Also, as it does not generate correlations between the underlying variables it cannot extrapolate predictions beyond the range of dependent variable values covered by its underlying training subset. In such conditions, it is better to deploy the TOB algorithm over limited data ranges for the dependent variable to be predicted, where multiple data records exist, and are relatively evenly distributed, rather than applying it to predict over ranges with only intermittent data coverage. For these reasons, it is considered justifiable to use the TOB and ANN algorithms together in some cases (i.e., running them in parallel), to generate predictions that can be compared, benchmarked, and at least in part audited and assessed for overfitting tendencies.

Future work is planned to apply TOB and ANN in parallel to a more extensive worldwide PVT database for crude oil and evaluating other crude oil characteristics (e.g. viscosity, chemical compositions) to refine and improve the prediction accuracy for bubble-point pressure. This approach would help to avoid the need for extensive and expensive laboratory analysis and/or the reliance on PVT correlations based on limited sample groups from specific geographic regions.

Appendix 1. TOB Methodology and Mathematical Formulation

The TOB learning network has been recently introduced and applied to a range of non-linear systems to provide predictions of credible accuracy [3,22]. It involves two distinct and sequential stages (1 and 2) involving fourteen calculation steps (Fig. 1). TOB Stage 1 develops lazy learning [23] and nearest neighbour [24] principles. It does so by applying very specific and evenly weighted mean square error (MSE) calculations on its input variables. TOB Stage 2 selects variable weightings using an optimizer. This approach provides a highly flexible and versatile weighting regime than typically associated with k-nearest neighbour classifiers [25].

TOB Stage 1 (data matching and provisional prediction)

Step 1: Set up a 2-D array of N input variables and one dependent variable to be predicted for each of M data records.
Step 2: Arrange the data records in a systematic order defined by the prediction variable's values (e.g. ascending or descending value order).
Step 3: Derive maximum and minimum values (and other standard statistics, such as mean and standard deviation) for all records in the dataset (Table 1).
Step 4: Normalize the data in the array so each variable spans a range from minus 1 to plus 1 $(-1,+1)$. This is achieved by using Eq. (A1)

$$X_i^* = 2*[(X_i - Xmin)/(Xmax-Xmin)]-1 \quad \text{(A1)}$$

Where:

X_i = variable X value for the ith data record
$Xmin$ = minimum value of variable X
$Xmax$ = maximum value of variable X

X_i^* is the normalized value of variable X for the ith data record.

Step 5. Generate statistical analysis of the normalized values to check that the variables are all correctly normalized.
Step 6. Distribute the data records between training, tuning and testing subsets. Sensitivity analysis is conducted to establish the optimum percentage of data records to allocate to each data subset. Firstly, the data records to be used for testing are extracted from the complete data set and placed to one side. Sensitivity analysis then helps to divide the remaining data records between the training and tuning subsets in proportions that achieve an acceptable prediction accuracy. For most data sets the training subset is likely to hold more than seventy-five percent of the data records. For large datasets of several thousand data records the sensitivity analysis often reveals that the training subset can be a much larger percentage without compromising prediction accuracy.
Step 7. The variable squared error (VSE) between each variable in the J data records of the tuning-data subset and the K data records in the training-data subset are calculated using Eq. (A2):

$$VSE(X)_{jk} = [X_k(tr) - X_j(tu)]^2 \quad \text{(A2)}$$

Where:

$X_k(tr)$ = variable X value for the kth training-subset data record
$X_j(tu)$ = variable X value for the jth tuning-subset data record
$VSE(X)_{jk}$ = squared error value for variable X for the jth tuning-subset data record versus the kth training-subset data record.
ΣVSE is then established as the sum of the VSE values for each variable for each data record match using Eq. (A3):

$$\sum VSE_{jk} = \sum_{n=1}^{n=N+1} VSE(Xn)_{jk} * (Wn) \quad \text{(A3)}$$

Where:

$VSE(Xn)_{jk}$ = squared error for variable Xn for the jth tuning-subset data record versus the kth training-subset data record.
$\sum VSE_{jk}$ = sum of the squared errors for all N + 1 variables for that data record match.
Wn = weight ($0 < Wn \leq 1$) applied VSE of each of the N + 1 variables involved. These weights are all set to the same values (e.g. 1) in TOB stage 1 to avoid any bias in the initial training of the prediction network.
Step 8. Select and rank (lowest in ΣVSE is ranked number 1) the top-Q-matching data records in the training subset for each tuning subset data record. Q = 10 is typically sufficient for TOB stage 1. However, Q could be adjusted to higher or lower values, if necessary, to improve prediction accuracy.
Step 9. The Q-selected training-subset data records (i.e. best matches) for the jth tuning-subset data record each contribute a fraction to the prediction of the dependent variable. That fraction is proportional to the relative ΣVSE scores of those Q records for the jth data record That fraction is calculated with Eq. (A4) to Eq. (A6) and

$$f_q = \sum VSE_{jq} \Big/ \left[\sum_{r=1}^{r=Q} \sum VSE_{jr} \right] \quad \text{(A4)}$$

Where:

q = qth top-ranking training-subset record for the jth tuning-subset data record.
f_q = fractional contribution of qth top-ranking records for the jth tuning-subset data record.

The constraint defined by Eq.(A5) applies the sum of the f values applied to each matching data record.

$$\sum_{q=1}^{q=Q} f_q = 1 \tag{A5}$$

The matching training-subset data record with the lowest $\sum VSE_{jk}$ value should contribute most to the dependent-variable prediction for the *j*th tuning-subset data record. To achieve this (1 - *f*) is the multiplier applied in Eq. (A6) to each of the Q top-matching records.

$$(X_{N+1})_j^{predicted} = \sum_{q=1}^{q=Q} [(X_{N+1})_q * (1 - f_q)] \tag{A6}$$

Where:

$(X_{N+1})_q$ = dependent variable for the *q*th data record in the training subset.
$(X_{N+1})_j^{predicted}$ = Stage −1 TOB predicted value for the dependent variable for the *j*th tuning-set data record.

This prediction is provisional because equal weights (*Wn*) are applied to the variables in TOB stage 1.

Step 10. Measures of statistical accuracy are calculated for the TOB stage 1 predictions. The measures used include: coefficient of determination (R^2); mean square error (MSE); and, root mean square error (RMSE). These are calculated with Eq. (A7) to Eq. (A9), respectively.

$$R^2 = 1 - \frac{\sum_{j=1}^{j=J} (X_j^{actual} - X_j^{predicted})^2}{\sum_{j=1}^{j=J} (X_{ave}^{actual} - X_j^{predicted})^2} \tag{A7}$$

$$MSE = \frac{1}{J} \sum_{j=1}^{j=J} (X_j^{actual} - X_j^{predicted})^2 \tag{A8}$$

$$RMSE = \sqrt{MSE} \tag{A9}$$

Where:

X_j = dependent variable (i.e. $(X_{N+1})_j$ in Eq. (A6)) for the *j*th tuning-subset data record;
X_j^{actual} = actual (or directly measured) value of the dependent variable for the *j*th tuning-subset data record;
$X_j^{predicted}$ = predicted value of the dependent variable for the *j*th tuning-subset data record
X_{ave}^{actual} = average actual value of the dependent variable for all *J* data records in the tuning subset.

TOB Stage 2 (optimization)

Step 11. Optimization is performed to minimize RMSE (Eq.(A9)) collectively for the J data records in the tuning subset. This is achieved by adjusting optimization control metrics while applying certain constraints.

The two optimization control metrics are:

(1) Varying the values applied to the N input-variable weights (W_n). Small non-zero values to weights applied to certain variables can and do have a significant impact on the accuracy of the predictions derived.
(2) Varying the number (Q) of top matching records in Eqs. (A4), (A5) and (A6). For most data sets: $2 \le Q < = 10$. The optimizer is allowed to select the best integer value of Q to minimize RMSE. It does this by systematically changing the value of Q in the three equations mentioned and by comparing the RMSE value for the predictions generated for each integer value of Q evaluated in the range $2 \le Q < = 10$. For examples, if Q is set to "4", the predictions for all of the tuning subset data records only use the top-4 matching records from the training subset related to each tuning subset record in making their predictions. In this way the optimization algorithm identified which value of Q leads to the most accurate predictions for the tuning subset as a whole.

Here, the GRG (Generalized Reduced Gradient) algorithm option of the standard "Solver" optimizer in Microsoft Excel (Frontline Solvers [26]) is used, in conjunction with visual basic for application (VBA) code, to conduct the optimization process. Other evolutionary optimizers could be applied to achieve similar outcomes. For mid-sized dataset calculating the TOB predictions in Excel facilitates the display all the intermediate calculations in a convenient format.

The top-matching data records in the training subset for each tuning-subset data record are carried forward from TOB stage 1 for selection by TOB stage 2. Eq. (A3) is re-evaluated by varying W_n in each iteration of the optimizer. Additionally, TOB stage-2 $\sum VSE_{jq}$ scores are derived with Eq. (4) by varying Q ($2 < Q \le 10$) in each iteration of the optimizer, contrasting with the fixed value of Q used in TOB stage 1.

Step 12. Calculate TOB stage-2 RMSE and R^2 values for the predictions provided by the optimum *Step 11* solution. Compare the TOB stage-2 predictions with the TOB stage-1 predictions to assess the prediction improvements achieved, if any. Running sensitivity analysis with different values of Q (i.e. Q = 2 to 10) often provides insight to potential underfitting or overfitting issues with the data set.
Step 13. Calculate TOB stage-1 and stage-2 predictions for the independent testing data subset using the optimum values established for Wn and Q in *Step 11*. Calculate and evaluate the RMSE and R^2 values for the predictions calculated for the testing data. Reviewing the intermediate steps in the calculations often provides useful insight to the variables that have the most influence on prediction accuracy (it is often not those with the

highest Wn values). It also helps perform outlier analysis (i.e., understanding why some data records lead to less-accurate predictions).
Step 14. Consider whether the prediction accuracy achieved by the method is sufficiently meaningful for it to be relied upon. Also, evaluate how its prediction accuracy compares with other machine-learning tools.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.petlm.2018.12.001.

References

[1] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Network. 61 (2015) 85–117.
[2] M. Heinert, Artificial neural networks – how to open the black boxes? in: A. Reiterer, U. Egly (Eds.), Application of Artificial Intelligence in Engineering Geodesy. Proceedings of AIEG December, vols. 42–62, 2008 978-3-9501492-4-1Vienna, Austria.
[3] D.A. Wood, A transparent Open-Box learning network provides insight to complex systems and a performance benchmark for more-opaque machine learning algorithms, Advances in Geo-Energy Research 2 (2) (2018) 148–162.
[4] S. Elkatatny, Z. Tariq, M. Mahmoud, Real time prediction of drilling fluid rheological properties using Artificial Neural Networks visible mathematical model (white box), J. Petrol. Sci. Eng. 146 (2016) 1202–1210.
[5] D.L. Katz, Prediction of shrinkage of crude oils. API, Drill Prod. Pract. (1942) 137–147.
[6] M.B. Standing, A Pressure–volume–temperature correlation for mixtures of California oils and gases. API, Drill Prod. Pract. (1947) 275–287.
[7] O. Glaso, Generalized pressure-volume-temperature correlations, J. Petrol. Technol. 32 (5) (1980) 785–795.
[8] M.A. Al-Marhoun, New correlation for formation volume factor of oil and gas mixtures, J. Can. Pet. Technol. 31 (1992) 22.
[9] M. Karimnezhad, M. Heidarian, M. Kamari, H. Jalalifar, A new empirical correlation for estimating bubble point oil formation volume factor, J. Nat. Gas Sci. Eng. 18 (2014) 329–335.
[10] A. Jarrahian, J. Moghadasi, E. Heidaryan, Empirical estimating of black oils bubble point (saturation) pressure, J. Petrol. Sci. Eng. 126 (2015) 69–77.
[11] G. Kim, J. Park, M. Lee, Bubble point measurement and high-pressure distillation column design for the environmentally benign separation of zirconium from hafnium for nuclear power reactor, Kor. J. Chem. Eng. 32 (1) (2015) 30–36.
[12] R.B. Gharbi, A.M. Elsharkawy, Neural-network model for estimating the PVT properties of Middle East crude oils, In: Paper SPE 37695 Presented at SPE Middle East Oil Show and Conference, Bahrain 15–18 (1997) March.
[13] N. Varotsis, V. Gaganis, J. Nighswander, P. Guieze, A novel noniterative method for the prediction of the PVT behavior of reservoir fluids, Paper SPE 56745 Presented at SPE Annual Technical Conference and Exhibition, Houston, Texas, 3-6 October, 1999.
[14] A.M. Malallah, R. Gharbi, M. Algharaib, Accurate estimation of the world crude oil PVT properties using graphical alternating conditional expectation, Energy Fuel. 20 (2006) 688–698.
[15] S. Dutta, J.P. Gupta, PVT correlations for Indian crude oil using artificial neural networks, J. Petrol. Sci. Eng. 72 (2010) 93–109.
[16] M.A. Al-Marhoun, S.S. Ali, A. Abdulraheem, S. Nizamuddin, A. Muhammadain, Prediction of bubble point pressure from composition of black oils using artificial neural network, Petrol. Sci. Technol. 32 (14) (2014) 1720–1728.
[17] M.A. Ahmadi, M. Pournik, S.R. Shadizadeh, Toward connectionist model for predicting bubble point pressure of crude oils: application of artificial intelligence, Petroleum 1 (4) (2014) 307–317.
[18] M.M. Mahmood, M.A. Al-Marhoun, Evaluation of empirically derived PVT properties for Pakistani crude oils, J. Petrol. Sci. Eng. 16 (1996) 275–290.
[19] M.H. Rammay, A. Abdulraheem, PVT correlations for Pakistani crude oils using artificial neural network, J Petrol Explor Prod Technol 7 (2017) 217–233.
[20] C.M. Bishop, Neural Networks for Pattern Recognition, second ed., Oxford University Press. U.K., 0198538642, 1995, p. 482.
[21] S. Haykin, Neural Networks: a Comprehensive Introduction, third ed., Pearson / Prentice Hall, New York, U.S.A., 1999, p. 906 ISBN-10:0-13-147139-2.
[22] D.A. Wood, A. Choubineh, B. Vaferi, Transparent open-box learning network provides auditable predictions: pool boiling heat transfer coefficient for alumina-water-based nanofluids, J. Therm. Anal. Calorim. (2018) Published online 12 September 2018 https://doi.org/10.1007/s10973-018-7722-9.
[23] M. Birattari, G. Bontempi, H. Bersini, Lazy Learning Meets the Recursive Least Squares Algorithm. Advances in Neural Information Processing Systems vol. 11, MIT Press, Cambridge, MA, 1999, pp. 375–381 1999.
[24] G.H. Chen, D. Shah, Explaining the success of nearest neighbor methods in prediction, Foundations and Trends R in Machine Learning 10 (5–6) (2018) 337–588.
[25] R. Samworth, Optimal weighted nearest neighbour classifiers, Ann. Stat. 40 (5) (2012) 2733–2763.
[26] Frontline Solvers. Standard Excel Solver - Limitations of Nonlinear Optimization (accessed December 2018) https://www.solver.com/standard-excel-solver-limitations-nonlinear-optimization.