J Petrol Explor Prod Technol
DOI 10.1007/s13202-017-0355-x

ORIGINAL PAPER - PRODUCTION ENGINEERING

Ensemble SVM for characterisation of crude oil viscosity

Munirudeen A. Oloso[1] · **Mohamed G. Hassan**[1] · **Mohamed B. Bader-El-Den**[2] · **James M. Buick**[1]

Received: 12 December 2016 / Accepted: 14 May 2017

Abstract This paper develops ensemble machine learning model for the prediction of dead oil, saturated and undersaturated viscosities. Easily acquired field data have been used as the input parameters for the machine learning process. Different functional forms for each property have been considered in the simulation. Prediction performance of the ensemble model is better than the compared commonly used correlations based on the error statistical analysis. This work also gives insight into the reliability and performance of different functional forms that have been used in the literature to formulate these viscosities. As the improved predictions of viscosity are always craved for, the developed ensemble support vector regression models could potentially replace the empirical correlation for viscosity prediction.

Keywords PVT · Dead oil · Bubble point · Empirical · Viscosity · Undersaturated · Black oil · Ensemble

List of symbols

APE	Average per cent error
AAPRE/E_a	Average absolute per cent error
RMSE	Root mean squared error
E_{max}	Maximum absolute per cent error
E_{min}	Minimum absolute per cent error
SD	Standard deviation of absolute per cent error
P	Reservoir pressure, psia
T	Reservoir temperature, ^{0}F
P_b	Bubble-point pressure, psia
γ_g	Dissolved gas relative density (air = 1)
γ_o	Oil gravity, stock tank oil relative density (water = 1)
γ_{API}	API gravity, ^{0}API
R_s	Solution gas/oil ratio, scf/STB
R_{sb}	Solution gas/oil ratio at bubble point, scf/STB
μ_o	Oil viscosity, cp
μ_{od}	Dead oil viscosity, cp
μ_{ob}	Saturated oil viscosity, cp
μ_{oa}	Undersaturated oil viscosity, cp

Introduction

Knowledge of oil pressure–volume–temperature (PVT) properties is of great interest to petroleum engineers as they are critical in performing most reservoir engineering studies. Viscosity is one of these PVT properties and it controls the fluid flow through the porous media. It is therefore important to be able to estimate crude oil viscosity at different stages of oil exploration. Empirical correlations, based on easily acquired field data, are usually employed to estimate dead oil, bubble-point and undersaturated viscosities. However, performance of these empirical correlations is not usually satisfactory and improved predictions are always sought.

✉ Munirudeen A. Oloso
munirudeen.oloso@port.ac.uk

Mohamed G. Hassan
mohamed.hassan@port.ac.uk

Mohamed B. Bader-El-Den
mohamed.bader@port.ac.uk

James M. Buick
james.buick@port.ac.uk

[1] School of Engineering, University of Portsmouth, Portsmouth, UK

[2] School of Computing, University of Portsmouth, Portsmouth, UK

Published online: 01 June 2017

In general, viscosity can be defined as the internal resistance to the flow of fluid. Crude oil viscosity is an important physical property that controls and influences the flow of oil through porous media and pipes (Ahmed 2010). It is also an important parameter when developing reservoir models to predict ultimate recovery, in designing enhanced oil recovery operations and when designing pipelines for effective fluid flow.

The viscosity of a liquid is related directly to the type and size of the molecules which make up the liquid (McCain Jr 1991). Crude oil viscosity can be categorised into three classes depending on the reservoir pressure, namely dead oil viscosity, saturated oil viscosity and undersaturated oil viscosity.

- Dead oil viscosity (μ_{od}) is the viscosity of the crude oil with no free gas at atmospheric pressure and temperature.
- Saturated/bubble-point oil viscosity (μ_{ob}) is the viscosity of the crude oil at the bubble-point pressure and the reservoir temperature.
- Undersaturated oil viscosity (μ_{oa}) is the viscosity of the crude oil at a pressure and temperature above the bubble-point pressure and reservoir temperature.

Oil viscosity can ideally be determined by laboratory experimentation. However, this is always costly and time demanding and a high technical speciality is required. The primary alternatives to this are the use of equations of states (EOS) and empirical correlations. Unfortunately, the EOS do require crude oil compositions which can only be determined through laboratory analysis; thus, they do not eliminate the requirement for laboratory analysis. This has paved way for the adoption of empirical correlations over a period of time. Likewise, some machine learning (ML) techniques have been used to improve the prediction of oil viscosity. However, stand-alone ML techniques or their hybrid systems can become stuck in local minimal, hindering the generalisation capability of such systems. However, this local minima problem can be addressed by ensemble systems (Dietterich 2000).

ML is the process of writing computer programs to optimise a performance criterion using example data or past experience (Alpaydin 2014). Learning involves creation of a system which applies past experience to analogous new situations. Learning can be in or through many forms; it can be through new knowledge acquisition, cognitive skills acquisition, effective representation of new knowledge or new fact discovery through observation and experimentation (Carbonell et al. 1983). Hybrid ML system involves fusion of two or more ML techniques with the aim of strengthening one another. An example is the fusion of genetic algorithm (GA) with support vector machines (SVMs). The GA is used to optimise the learning parameters of SVM. However, such hybrid system might have assumed a local minima since a given space of hypotheses must have been searched for a given data set. On the other hand, effectively constructed ensemble system of SVM can overcome this problem since it involves fusion of systems that should have been constructed on different spaces of hypotheses searched by the learning algorithm on the training data set (Dietterich 2000).

ML techniques usually utilise input variables similar to the empirical correlations. Mostly, petroleum fluid properties which are easily measured in the field and which have direct physical relationship with the target output are used as the correlating variables (Standing 1947; Chew and Connally 1959). Also, a trial and error method can be used to eliminate any correlating variable that does not improve the performance of the correlation significantly (Chew and Connally 1959). Pruning of the correlating variables, often referred to as feature selection, can be achieved by some statistical tools to select input variables that are used in the regression analysis to develop the empirical correlations or by common feature selection techniques such as neighbourhood component analysis (NCA), sequential feature selection and LASSO. NCA is an embedded and nonparametric feature selection method. NCA mainly learns the feature weights with the aim of minimising the objective function that measures the mean leave-one-out regression or classification loss over the given training data set (Goldberger et al. 2005; Yang et al. 2012). LASSO minimises the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. LASSO includes a penalty term that constrains the size of the estimated coefficients to produce some coefficients that are exactly zero in order to trim the selected features for prediction (Tibshirani 1996, 2011). Sequential feature selection in its basic form involves minimisation of an objective function over all feasible feature subsets by adding or removing features from a candidate subset while evaluating the criteria (Liu and Motoda 2007; Stańczyk and Jain 2015).

This paper develops ensemble ML model based on SVM to predict dead oil, saturated and undersaturated viscosities. For each property, different functional forms were explored to determine its best correlating variables. The prediction results of the ensemble models are compared with the most commonly used correlations in the petroleum industry.

Literature review

A review of empirical correlations and ML techniques that have been developed for viscosity predictions is explored in this section.

Viscosity correlations

A brief review of the available correlations in the literature for the estimation of crude oil viscosities is presented in tabular forms. The adopted functional form(s), API range, origin of the data sets and reported statistical errors are included in the review.

Dead oil viscosity correlation

Correlation for μ_{od} is usually developed with the API gravity (γ_{API}) and T as the independent variables.

$$\mu_{od} = f(\gamma_{API}, T) \quad (1)$$

γ_{API} is calculated using the specific gravity of an oil (γ_o) which is the ratio of oil density to that of water. Specific gravity for API is normally determined at 60 degrees Fahrenheit. It is thus given as:

$$\gamma_{API} = \frac{141.5}{\gamma_o} - 131.5 \quad (2)$$

Correlations for μ_{od} usually introduce large errors when applied to data sets which are different from the ones used to develop the original correlations. The difference in the results is related to the difference in the oil base (asphaltic, paraffinic or mixed base) (Labedi 1992).

Some other correlations have included additional correlating variables such as average temperature, critical temperature, Watson characterisation factor (K_w), bubble-point pressure (P_b) and bubble-point gas/oil ratio (R_{sb}). Bergman and Sutton (2009) indicated that most of the correlations that use only γ_{API} and T usually have large errors and they are the least accurate compared to other methods that have additional correlating property. Alternative methods that could possibly give improved accuracy are the use of EOS or correlations that use crude oil compositions, though these are not usually available. Hence, the need to use simple methods that utilise easily acquired properties (γ_{API} and T).

Table 1 presents a concise review of some common correlations for μ_{od}.

Gas-saturated viscosity correlations

Gas-saturated viscosity (μ_{ob}) can be defined as the viscosity of the crude oil with dissolved gas, just above the μ_{od}, up to the bubble-point pressure at the reservoir temperature. The dissolved gas in crude oils reduces the observed value of the μ_{od}. Correlations for μ_{ob} are usually developed as a function of μ_{od} and gas oil ratio (R_s) or P.

$$\mu_{ob} = f(\mu_{od}, R_s) \quad (3)$$

Some other forms of correlations based on different input variables have also evolved for μ_{ob}. Table 2 presents some of these common correlations.

Undersaturated viscosity correlations

Beal (1946) was the first to develop a correlation for undersaturated μ_o and noted that the crude oil viscosity in this region increases proportionally with the increase in pressure. He used 52 data points of crude oil from the USA to develop the correlation and reported an E_a of 2.7% on the data set. Subsequently, different undersaturated viscosities with different correlating variables have been presented in the literature. Table 3 presents some of these correlations.

Machine learning for viscosity predictions

Viscosity prediction has also benefitted from the machine learning (ML) modelling capability. ML techniques that have been used for viscosity modelling include radial basis function neural network (RBFNN), artificial neural network (ANN), fuzzy logic (FL), functional networks (FN), genetic algorithm (GA), SVM and group method of data handling (GMDH) (Table 4).

Overview of ensemble models

Ensemble ML is a combination of multiple base models of classifiers or regressors for classification and regression problems, respectively. Each base model covers a different part of the input space or the complete input space. Although there is no defined taxonomy for building the ensemble models, some successful approaches and methodologies have been widely adopted (Dietterich 2000; Zhou 2012; Bader-El-Den and Gaber 2012; Perry et al. 2015; Bader-El-Den et al. 2016).

After generating a set of base learners, the ensemble method will then be formed by combining the base models or a subset of them based on defined criteria or algorithm to form a generalised prediction model. Three main benefits of the combination can be attributed to statistical, computational and representational issues (Dietterich 2000; Zhou 2012).

- Statistical Issue: there is always a large space of hypotheses for the base model to choose from. There is a chance that the base learning algorithm has not chosen the most efficient of these possible hypotheses. The combination approach tends to reduce the risk of choosing the wrong hypotheses for formulating the prediction models.

Table 1 Dead oil viscosity correlations

References	γ_{API} range	Region of data source	Data points	Comment
Beal (1946)	10.1–52.2	USA	98	APE = 24.2%. Performance of the correlation was tested by dividing the data sets into different temperature and API ranges
Beggs and Robinson (1975)	16–58	–	460	APE = −0.64%. The correlating variables were T and γ_o
Glasø (1980)	20.1–45.8	North Sea	38	The developed μ_{od} correlation is used in correcting for paraffinicity in order to adapt the correlation for different crude oils
Ng and Egbogah (1983)	5–58	–	394	The authors presented a modified correlation of Beggs and Robinson (1975) and also proposed a new μ_{od} correlation which included pour point temperature as a new correlating variable. Based on the 394 data points, the APE of 61, −5.13 and −4.3% were reported for the original Beggs–Robinson correlation, modified Beggs–Robinson correlation and the newly developed correlation, respectively
Twu (1985)	−4 to 93.1	–	563	APE = 7.85%
Bennison (1998)	–	North Sea	16	APE = 13%. The correlation was only recommended for API > 20 and T > 250 °F
Kartoatmodjo and Schmidt (1991)	14.4–59.0	Indonesia, North America, Middle east and Latin America	661	The correlation was derived using the functional form of Glaso's correlation. Sensitivity analysis of the correlation to reservoir temperature was performed
Labedi (1992)	32.2–48.0	Libya	91	APE = −2.61%. When the correlation by Beal (1946) was applied to the data set, a very poor result was observed. This was ascribed to the fact that the Beal's correlation was developed for light California crude oil
Petrosky Jr and Farshad (1998)	25.4–46.1	Gulf of Mexico	118	AAPRE = 12.38%
De Ghetto et al. (1995)	6–22.3	–	1200	Correlations were developed for heavy and extra heavy crude oils based on the correlation of (Egbogah and Ng 1990) with AAPRE of 30.3 and 41.8% for API < 10 and 10 < API ≤ 22.3, respectively
Elsharkawy and Alikhan (1999)	19.9–48	Middle East	254	AAPRE of 19.3% and correlation coefficient = 0.881
Elsharkawy et al. (2003)		Worldwide	361	An empirical correlation which predicts the entire viscosity curve was derived. It predicts from μ_{od} to the undersaturated μ_o. Different forms of the developed correlation were explored and the best result was obtained from the one with 8 variables, having average absolute deviation of 24.3%
Dindoruk and Christman (2004)	17.4–40	Gulf of Mexico	95	P_b and R_{sb} were introduced in the μ_{od} correlation along with the γ_{API} and T. It was stressed that these additional two properties allow the correlation to capture some of the characteristics of the oil type
Hossain et al. (2005)	7.1–21.8	–	–	The developed correlation was tested with 142, 42 and 23 data sets from Chevron, De Ghetto et al. (1995) and Kartoatmodjo and Schmidt (1994) with AAPRE of 28.8, 22.5 and 55.2% respectively.
Naseri et al. (2005)	17–44	Iran	472	250 data points were used to develop the correlations while 222 data points were used for testing and validation. AAPRE = 7.77% for the original regression data and 15.3% for the testing data set
Bergman and Sutton (2009)	0.45–135.9	Worldwide	9837	Accuracy of the correlation was reported on different temperature ranges. 1. γ_{API} = 5–80, T = 35–500 °F, AAPRE = 16.6% 2. γ_{API} = 5–80, T = 35–100 °F, AAPRE = 18.1% 3. γ_{API} = 5–80, T = 100–200 °F, AAPRE = 17.6% 4. γ_{API} = 5–80, T = 200–300 °F, AAPRE = 15.2%
El-hoshoudy et al. (2013)	21–52	Egypt	1000	AAPRE = 9.8855% and correlation coefficient = 0.9285
Alomair et al. (2014)	10–20	Kuwait	374/118	374 data points were used for developing the correlation while 118 data point were used for testing. For the training data set, AAPRE = 25.29% while for the testing data set AAPRE = 28.08%

Table 2 Saturated viscosity correlations

References	γ_{API} range	Region of data source	Data points	Comment
Beal (1946)	15.8–45.7	USA	351	APE = 13.4%
Chew and Connally (1959)	NA	USA	457	The performance of the correlation was examined by using the confidence limit
Beggs and Robinson (1975)	16–58	–	2073	APE = −1.83%. The live oil viscosity was correlated as a function of μ_{od} and R_s
Khan et al. (1987)	14.3–44.6	Saudi Arabia	150/1691	A total of 150 and 1691 data points were used to develop viscosity correlations at and below P_b with AAPRE of 12.148 and 5.157%, respectively. The reported correlation coefficients were 0.953 and 0.994 for viscosity correlations at and below P_b, respectively
Kartoatmodjo and Schmidt (1991)	14.4–59.0	Indonesia, North America, Middle east and Latin America	5321	A total of 5321 data points were used to develop a μ_{ob} correlation. Similar sensitivity analysis as performed for the μ_{od} prediction was carried out
Khan et al. (1987)	21–49	Canada and Middle East	459	AAPRE = 4.91% and correlation coefficient = 0.9979
Labedi (1992)	32.2–48	Libya	91	μ_{ob} correlation was developed with the following independent variables: γ_{API}, μ_{od} and P_b, and the correlation's APE was −2.38%. Another correlation was developed for μ_o below P_b. For the μ_o correlation below P_b, a linear relationship between μ_o and P was established for the pressure range $P_b > P > 0.15P_b$. The slope of the linear relationship between μ_o and P was correlated with APE = 3.5%
Petrosky Jr and Farshad (1995)	25.4–46.1	Gulf of Mexico	864	AAPRE = 14.47%
De Ghetto et al. (1995)	6–22.3	–	1200	Correlations were developed for heavy and extra heavy crude oils based on correlation of Kartoatmodjo and Schmidt (1994) with AAPRE of 14.7 and 16.1% for API < 10 and 10 < API ≤ 22.3, respectively
Almehaideb (1997)	30.9–48.6	UAE	57	For μ_{ob} correlation, the reported AAPRE = 13% which is smaller than the results from all other compared correlations. Correlation coefficient = 0.9691
Hanafy et al. (1997)	17.8–47.7	Egypt	324	AAPRE = 19.1% and correlation coefficients = 0.91
Elsharkawy and Alikhan (1999)	19.9–48	Middle East	254	AAPRE of 18.6% and correlation coefficients = 0.978
Boukadi et al. (2002)	34.8–136	Middle East	32	22 data points were used to generate the correlation while the remaining was used for testing it. The correlation was developed for $P \leq P_b$. AAPRE = 35.5560%
Dindoruk and Christman (2004)	17.4–40	Gulf of Mexico	95	AAPRE = 13.2%
Naseri et al. (2005)	17–44	Iran	472	AAPRE = 16.4% for the original regression data and 26.3% for the testing data set
Hossain et al. (2005)	7.1–22.3	–	415	The saturated viscosity was developed as a function of μ_{od}. and R_s. The new correlation gave AAPRE of 53.2, 46.4 and 26.5% when applied to datasets from Kartoatmodjo and Schmidt (1994), De Ghetto et al. (1995) and Chevron respectively
Bergman and Sutton (2007)	6.0–61.7	Worldwide	12,474	The author proposed a new correlation as a result of the observed inconsistency in the behaviour of all the evaluated correlations. For the new correlation, AAPRE = 12.4%
Khamehchi et al. (2009)	33.4–124	–	94	Correlation coefficient = 0.98
El-hoshoudy et al. (2013)	21–52	Egypt	1000	AAPRE = 11.2281% and correlation coefficient = 0.9493
Ghorbani et al. (2016)	21.55–30.62	Iran	600	The developed correlations are fairly large with 36 and 42 coefficients for viscosity below bubble-point and saturated viscosity, respectively. It could be a bit cumbersome for field application. AAPRE of 3.77 and 0.01058% were reported for viscosity below bubble-point and saturated viscosity, respectively

Table 3 Undersaturated viscosity correlations

References	γ_{API} range	Region of data source	Data points	Comment
Beal (1946)	–	USA	52	APE = 2.7%
Vazquez and Beggs (1980)	–	Worldwide	6000	The correlation was developed without γ_{API} grouping unlike other correlations developed in the same paper
Khan et al. (1987)	14.3–44.6	Saudi Arabia	1503	AAPRE = 1.915% and correlation coefficients = 0.999
Kartoatmodjo and Schmidt (1991)	14.4–59.0	Indonesia, North America, Middle east and Latin America	3588	Sensitivity analysis based on grouping of P_b and R_s was carried out
Labedi (1992)	32.2–48	Libya	91	The general equation of a straight line was adopted in developing the correlation. The slope of the equation is the parameter that was correlated with APE = −3.1%
Petrosky Jr and Farshad (1995)	25.4–46.1	Gulf of Mexico	404	AAPRE = 2.91%
De Ghetto et al. (1995)	6–22.3	–	1200	Correlations were developed for heavy and extra heavy crude oils based on the correlations of Labedi (1992) for API < 10 and Kartoatmodjo and Schmidt (1994) for (10 < API ≤ 22.3) with AAPRE of 12.3 and 10.1%, respectively
Almehaideb (1997)	30.9–48.6	UAE	328	For μ_o correlation, AAPRE = 2.885% while the compared correlation of Vazquez and Beggs (1980) gave AAPRE of 8.58%
Elsharkawy and Alikhan (1999)	19.9–48	Middle East	254	AAPRE = 4.9% and correlation coefficient = 0.972
Dindoruk and Christman (2004)	17.4–40	Gulf of Mexico	93	AAPRE = 5.99%
Naseri et al. (2005)	17–44	Iran	472	AAPRE = 2.12% for the original regression data and 3.62% for the testing data set
Hossain et al. (2005)	7.1–22.3	–	390	AAPRE of 31.2, 38.9 and 56.3% were got for the data sets of Chevron and other previous works (De Ghetto and Villa 1994; Kartoatmodjo and Schmidt 1994)
Ghorbani et al. (2016)	21.55–37.62	Iran	600	The developed correlation has 36 coefficients which is very rare in the literature. AAPRE of 0.268% was reported for the correlation

Table 4 Machine learning techniques for viscosity prediction

References	γ_{API} range	Region of data source	Data points	Comment
Elsharkawy (1998)	20–45	–	–	RBFNN was used to predict viscosity across different ranges. Input parameters to the model were reservoir P, T, stock tank γ_o, and separator γ_g. AAPRE = 8.72% in testing
Elsharkwy and Gharbi (2001)	24.51–39.81	Kuwait	805	Four different ANN models were developed to predict the oil viscosity with AAPRE of 9.39, 12.17, 14 and 19.18% in testing. In total, 700 and 105 data points were used for training and testing the models, respectively
Ayoub et al. (2007)	29–43.8	Pakistan	99	ANN model was developed for viscosity below bubble point. AAPRE = 3.4%
Hajizadeh (2007)	–	Iran	89	GA was used to model oil viscosity with correlation coefficient of 0.9974. The region of reservoir viscosity covered by the data was not stated
Omole et al.(2009)	19–45.4	Nigeria	32	ANN model was developed for μ_{ob} with AAPRE of 6.781%
Oloso et al. (2009)	24.2–48	Middle East	99	SVM and FN models were developed to predict the entire viscosity curve. The learning parameters are the variables of the fitted viscosity curves. AAPRE of 8.5514% was reported for testing on 29 data points
Al-Marhoun et al. (2012)	–	Canada	100	Approach similar to (Khoukhi et al. 2011) was used in predicting the entire crude oil viscosity. Eight different ML techniques were explored in the work. A variant of FN gave the best performance
Ghorbani et al. (2014)	21.55–37.62	Iran	600+	365, 287 and 57 data points were used for developing hybrid models of GA and GMDH for viscosity below bubble point, μ_{oa} and μ_b with AAPRE of 13.57, 10.95 and 12.48%, respectively
Hemmati-Sarapardeh et al. (2016)	20–50	Worldwide	1497	Hybrid model of GA and SVM was used to predict μ_{od}. AAPRE of 17.17 was reported
Ghorbani et al. (2016)	21.55–37.62	Iran	600+	365, 287 and 57 data points were used for developing new multi-hybrid models with GA and GMDH for viscosity below bubble point, μ_{oa} and μ_b with AAPRE of 3.77, 0.268 and 0.01058%, respectively

- Computational Issue: ML algorithms usually involve searching for optimal parameters which may get stuck in local optima. Combination of different models reduces the risk of choosing a wrong local minimum.
- Representational Issue: In many ML problems, the true unknown hypothesis cannot be truly modelled in the hypothesis space. However, combination of different hypotheses may be able to form a more accurate representative function that learns the problem.

The most common ways of combining base models in ensemble modelling are averaging and voting. Averaging is the most fundamental and common combination method for numeric output (i.e. regression problem), while voting is a common combination method for nominal output (i.e. classification problem). Averaging can either be simple or weighted.

There are two main ensemble paradigms: sequential ensemble methods and parallel ensemble methods (Zhou 2012). Sequential ensemble methods are where the base learners are generated sequentially with *boosting* as a representative, while parallel ensemble methods are where the base learners are generated in parallel, with Bagging as a representative.

Bagging (Breiman 1996) which is also known as bootstrap aggregation involves training multiple models with training sets of data randomly drawn with replacement from the base training data sets. The training data sets for the base models are called bootstraps. Hence, bagging involves training different models with different samples and usually predictions are obtained by averaging the results of the different base models for a regression problem.

Boosting involves training and improving a weak learning algorithm into a strong one (Schapire 1990). In boosting, the training data set for each subsequent model increasingly focuses on instances wrongly predicted by the previous weaker model. ADABOOST (adaptive boosting algorithm) is one of the most used boosting algorithms which automatically adapts to the data given to it.

Proposed ensemble model

An ensemble model based on SVM regression has been developed. The steps for the algorithm will be given and discussed.

Ensemble support vector regression

The version of SVM that is used for regression problem is known as support vector regression (SVR). SVM is a statistical machine learning method that generates input–output mapping functions from a set of training data. It uses the principle of structural risk minimisation, seeking to minimise the upper bound of the generalisation error rather than just minimising the training error. In a simple pattern recognition problem, SVM uses a linear separating hyperplane to create a classifier with a maximal margin. When the input cannot be linearly transformed (e.g. complex classification problem or regression problem), SVM first nonlinearly transforms the input space into a higher-dimensional feature space. The transformation is achieved by using nonlinear mapping functions which are generally referred to as kernel functions. Typical kernel functions include RBF, Gaussian and polynomial functions. The steps for creating the SVR ensemble model are highlighted in Algorithm 1. Ensemble pruning has been performed using E_a. It is observed that similar performance is achieved when root mean squared error (RMSE) is used for the pruning. Ensemble pruning is basically the determination and selection of the final base models that will form part of the ensemble model.

The stratification process of selecting the sample input data ensures that random rows are selected. The four main parameters that control each SVR model are C, k, λ and ε. "C", the penalty factor, is the trade-off between achieving minimal training error and the complexity of the model. If it is too large, there is a high penalty for non-separable points which may result in overfitting. If it is too small, there may be underfitting (Alpaydin 2014). The options for the kernel, k, have been limited based on a preliminary experimentation on the data set. Based on preliminary investigation, λ assumes the value of ε in each iteration.

The developed ensemble SVR has n numbers of based models which are selected from the simulated SVR models based on E_a ranking. The base SVR models are ranked based on the values of E_a to form an ensemble SVR model which is henceforth referred to as Ensemble_SVR_APRE. For analysis and error sensitivity test, RMSE is also used for pruning based on the same algorithm to generate another model which is called Ensemble_SVR_RMSE.

This innovative way of creating ensemble models will also give us the opportunity to compare the two error evaluation criteria, E_a and RMSE, as there is no consensus on which of these two error evaluating criteria is the best (Chai and Draxler 2014).

Algorithm 1: Ensemble support vector machine regression

1. Select x data sets as 70% of the entire data sets (X) using stratification and the corresponding y from the output (Y)
2. Iterate for $C = 1$ to N
3. Iterate for kernel, $k \rightarrow \{RBF, Gaussian, polynomial\}$
4. Iterate for $\varepsilon \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$
5. Compute each SVR model $F(C, k, \varepsilon)$
6. Evaluate each SVR model using E_a
7. Continue for the next ε
8. Continue for the next k
9. Continue until $C = N$
10. Give ranks to the SVR model based on E_a
11. Choose the best n models based on their ranks to form the ensemble models based on E_a ranking
12. Predict the testing target Y from the testing input X using the n base SVM models
13. Compute each ensemble output $\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i$, where $\hat{Y}_i$ is the predicted target by the ith SVR base model

Implementation

This section focuses on data acquisition, feature extractions, simulation of the ensemble algorithm and statistical evaluation of the prediction results. To focus on the results of different functional forms for predicting oil viscosity, the ensemble SVR will only be compared with empirical correlations. Recently, the advantages of ensemble SVM compared to stand-alone SVM have been discussed (Oloso et al. 2016).

Data sets and input features selection

A total of 286 data points were available for the μ_{od} and μ_{ob} simulations. Among these, only about 250 data points have been used as other rows have missing values. For the μ_{oa} simulation, 910 data points were used. A statistical summary of the data is presented in Appendix A. In each case, approximately 70% of the data has been used for training the models and 30% for testing. A stratification process which involves random selection of non-sequential rows is used to divide the data into training and testing sets. Before the simulation exercise, some common feature extraction techniques were used to examine the inputs that are likely to be mostly correlated and influential for each of the desired output. This would have possibly reduced the dimension of the input matrix. However, no consensus was reached among the methods.

Experimental work

For each of the PVT properties considered in this paper, more than one functional form, that is, combination of correlating variables has been used in the literature. Ini-

tially, some feature selection methods (such as NCA, LASSO and sequential feature selections) were investigated. However, all the investigated functional forms have been implemented for the ensemble SVM to allow fair comparison with the empirical correlations. Also, the listed feature selection techniques favoured different input variables.

A. Investigated functional forms for μ_{od}

$$\mu_{od} = f(\gamma_{API}, T)$$

$$\mu_{od} = f(\gamma_{API}, T, R_{sb})$$

$$\mu_{od} = f(\gamma_{API}, T, P_b)$$

$$\mu_{od} = f(\gamma_{API}, T, R_{sb}, P_b)$$

B. Investigated functional forms for μ_{ob}

$$\mu_{ob} = f(\gamma_g, R_s, \gamma_o, T)$$

$$\mu_{ob} = f(\gamma_g, R_s, \gamma_{API}, T)$$

$$\mu_{ob} = f(\gamma_{API}, \mu_{od}, P_b)$$

$$\mu_{ob} = f(\mu_{od}, R_s)$$

C. Investigated functional forms for μ_{oa}

$$\mu_{oa} = f(\mu_{ob}, P_b, P)$$

$$\mu_{oa} = f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$$

$$\mu_{oa} = f(\mu_{ob}, \mu_{od}, P_b, P)$$

The proposed ensemble SVM model is simulated for all these functional forms and the results are compared with the available empirical correlations in the literature that utilise these functional forms.

During experimentation, it was noted that the results of the two ensemble SVR models are essentially the same. The ranks of the base models using RMSE and E_a for ranking may not be the same, but the ordering of the samples is almost always the same. In other words, the results of both Ensemble_SVR_APRE and Ensemble_SVR_RMSE are essentially the same. Hence, results of only Ensemble_SVR_APRE model are used and reported. Henceforth, the model will simply also be referred to as ensemble SVR or ensemble SVM.

Three error statistical criteria are primarily used to evaluate the performances of the simulated ensemble SVR and the compared empirical correlations. These are RMSE, E_a and maximum absolute error (E_{max}). The best model is expected to give the lowest values across these three parameters or two. E_{max} has been chosen as the third criterion to eliminate any tie between two models when both RMSE and E_a are not minimum for a particular model. It should also be noted that a model with minimum E_{max} is likely to have good prediction across the data points than the one with higher value.

Comparison with the previous ML studies for viscosity prediction

Commonly, an ML model for predicting oil viscosity assumes a particular functional form based on some empirical correlations. That is, the selected input variables in the ML model are similar to some correlations (Elsharkwy and Gharbi 2001; Omole et al. 2009). Contrary to this, different functional forms are selected for dead oil, saturated and undersaturated viscosities.

A novel approach was introduced in (Khoukhi et al. 2011) to predict the entire viscosity by training the parameters of the curve and bubble-point viscosity. However, the caveat to this method is its dependent on oil compositions which cannot be determined easily on the field, limiting the potential adoption of such methods for industrial application. Also, other works on viscosity prediction a stand-alone ML technique or hybrid systems have mainly adopted a single functional form (Elsharkwy and Gharbi 2001; Ghorbani et al. 2016; Hemmati-Sarapardeh et al. 2016). This paper aims to solve the problem of local minima by using ensemble model rather than a stand-alone SVM and the problem of preferential adoption of a single functional form by using different functional forms found in the literature for the prediction of oil viscosity.

Results and discussion

The simulation results for all the given functional forms for each of the three investigated PVT properties are presented. The developed ensemble SVR model clearly gives better performances than the compared empirical correlations in estimating the three viscosity variables.

Experimental results for μ_{od}

The results for the ensemble SVR in modelling μ_{od} using all the four stated functional forms are shown in Table 5. It is noticed that the functional form that gives the best result is $f(\gamma_{API}, T)$ with RMSE = 0.38784, E_a = 10.31983 and E_{max} = 29.1723. This result is followed by the functional form that incorporates R_{sb} as the additional correlating variable. However, it is important to note that the additional variable has not essentially improved the simulation results.

Table 6 shows the performance of some common empirical correlations. Correlation of Naseri et al. (2005) gives the best results among these μ_{od} correlations, followed by the correlation of Beal (1946). The additional correlating parameters in the correlation of Dindoruk and Christman (2004) have not improved its results compared to others.

Table 5 Performance of the ensemble SVR for the functional forms for μ_{od}

Inputs	RMSE	E_a	E_{min}	E_{max}	SD
γ_{API}, T	**0.387841**	**10.31983**	**0.693987**	**29.17233**	**0.070013**
γ_{API}, T, R_{sb}	0.373729	13.46401	0.540048	42.76276	1.952884
γ_{API}, T, P	0.413581	11.47461	0.369173	44.49372	1.998002
γ_{API}, T, R_{sb}, P	0.448936	11.27015	0.515526	40.92204	1.612681

Statistical measures for the best functional form are shown in bold

Table 6 Performance of μ_{od} empirical correlations

Correlation method	Inputs	RMSE	E_a	E_{min}	E_{max}	SD
Beal (1946)	γ_{API}, T	0.632632	69.64603	30.41812	87.02458	2.065349
Beggs and Robinson (1975)	γ_{API}, T	1.441646	170.5397	43.91294	316.8668	16.30285
Glasø (1980)	γ_{API}, T	0.988133	122.7806	21.80753	229.3247	18.38978
Dindoruk and Christman (2004)	γ_{API}, T, P_b, R_{sb}	0.866464	111.7765	13.43371	213.3762	16.38136
Naseri et al. (2005)	**γ_{API}, T**	**0.438342**	**30.84382**	**0.192923**	**103.9341**	**11.82092**
Kartoatmodjo and Schmidt (1991)	γ_{API}, T	0.887462	101.5805	7.035991	213.2981	17.51733
Petrosky Jr and Farshad (1995)	γ_{API}, T	1.113549	147.6598	39.49423	246.1332	19.75435
Labedi (1992)	γ_{API}, T	2.211061	285.9099	57.93207	445.5494	43.08376
Elsharkawy and Alikhan (1999)	γ_{API}, T	1.710592	225.0537	79.80017	351.5815	27.45034

Statistical measures for the best correlation are shown in bold

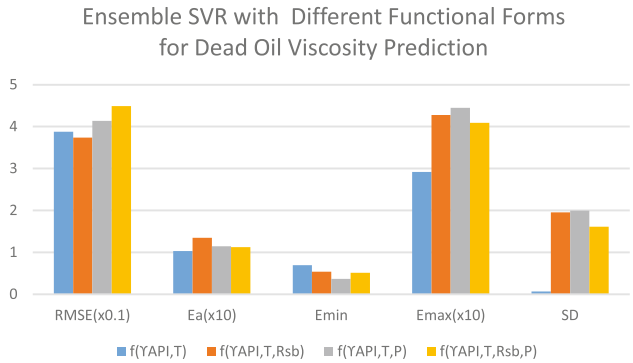

Fig. 1 Results of different functional forms for μ_{od} prediction with ensemble SVR

Comparing the ensemble SVR model with the listed empirical correlations, the results of all the functional forms simulated by the ensemble SVR model are better than the results of all the empirical correlations in Table 6. Meanwhile, the same functional form $f(\gamma_{API}, T)$ gives the best result for both the ensemble model and the empirical correlation. It is noted that the correlation of Naseri et al. (2005) has lower RMSE than the ensemble simulation for $f(\gamma_{API}, T, R_{sb}, P_b)$, but the latter has both lower E_a and E_{max}. Hence, the ensemble SVR model with functional form $f(\gamma_{API}, T, R_{sb}, P_b)$ is better than the leading correlation method of Naseri et al. (2005). The results in Table 5 compared to Table 6 show that the ensemble SVR has better strength to model the uncertainties of μ_{od} with overall best result from ensemble SVR with the functional form $f(\gamma_{API}, T)$. Figure 1 gives a graphical comparison of the ensemble SVR results with different functional forms for μ_{od} prediction.

Experimental results for μ_{ob}

Results of the ensemble SVR model for predicting μ_{ob} based on the previously stated four functional forms are given in Table 7. Among all the investigated functional forms for μ_{ob}, the best result is given by $f(\gamma_{API}, \mu_{od}, P_b)$ with RMSE = 0.063275, E_a = 7.036263 and

Table 7 Performance of the ensemble SVR for the functional forms for μ_{ob}

Inputs	RMSE	E_a	E_{min}	E_{max}	SD
μ_{od}, R_s	0.089441	9.34957	0.250187	49.06225	0.098373
γ_g, R_s, T, γ_{API}	0.143147	11.53513	1.37779	33.28361	0.104083
γ_{API}, μ_{od}, P_b	**0.063275**	**7.036263**	**0.268153**	**23.35724**	**1.446165**
γ_g, R_s, T, γ_o	0.105525	12.79756	0.030471	43.23039	2.174104

Statistical measures for the best functional form are shown in bold

Table 8 Performance of μ_{ob} empirical correlations

Correlation method	Correlating variables	RMSE	E_a	E_{min}	E_{max}	SD
Chew and Connally (1959)	$\boldsymbol{\mu_{od}}$, $\boldsymbol{R_s}$	**0.090067**	**9.278021**	**0.034077**	**30.27401**	**2.241977**
Al-Khafaji et al. (1987)	μ_{od}, R_s	0.093328	9.813704	0.574959	32.96323	2.380144
Khan et al. (1987)	γ_g, R_s, T, γ_o	0.112715	10.95839	0.133493	39.01657	1.792076
Dindoruk and Christman (2004)	μ_{od}, R_s	0.128336	14.44536	0.238722	41.73317	2.425695
Elsharkawy and Alikhan (1999)	μ_{od}, R_s	0.146096	15.38488	0.649525	36.19894	2.01021
Beggs and Robinson (1975)	μ_{od}, R_s	0.234077	23.9186	3.099475	42.13191	2.027648
Labedi (1992)	μ_{od}, γ_{API}, P_b	0.245091	25.19667	2.490099	64.07069	1.844393
Almehaideb (1997)	γ_g, R_s, T, γ_{API}	0.324045	34.14632	0.662503	51.31924	1.35145

Statistical measures for the best correlation are shown in bold

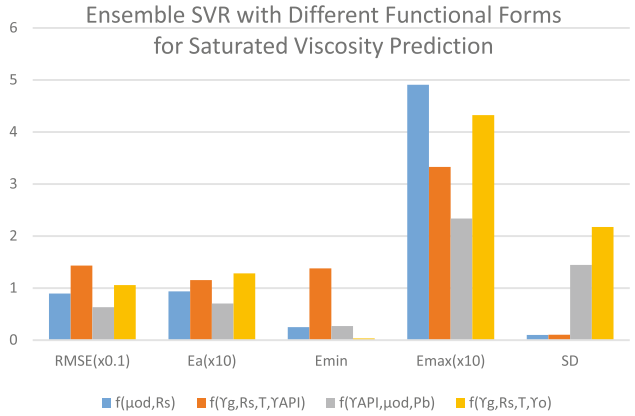

Fig. 2 Results of different functional forms for μ_{ob} prediction with ensemble SVR

$E_{max} = 23.35724$. This is followed by the results of the functional form $f(\mu_{od}, R_s)$. From Table 4, poorer performances are displayed by the functional forms which include T.

Results of some empirical correlations for μ_{ob} using the four functional forms are italicized in Table 8. The correlation of Chew and Connally (1959) which uses the functional form $f(\mu_{od}, R_s)$ gives the best performance among the empirical correlations for μ_{ob} with RMSE = 0.090067, $E_a = 9.278021$ and $E_{max} = 30.27401$. The second best result among the empirical correlations is given by the correlation of Al-Khafaji et al. (1987) which also uses the functional form of $f(\mu_{od}, R_s)$.

Clearly, the results show that the ensemble SVR with the functional form $f(\gamma_{API}, \mu_{od}, P_b)$ is the best in modelling the uncertainty in μ_{ob} as it has given the lowest values of RMSE, E_a and E_{max}. This outperforms all the empirical correlations in Table 8. Figure 2 gives a graphical comparison of the ensemble SVR results with different functional forms for μ_{ob} prediction.

Experimental results for μ_{oa}

Ensemble SVR experimental results for predicting μ_{oa} based on the considered three functional forms are shown

Table 9 Performance of the ensemble SVR for the functional forms for μ_{oa}

Inputs	RMSE	E_a	E_{min}	E_{max}	SD
$\boldsymbol{\mu_{ob}}$, $\boldsymbol{P_b}$, $\boldsymbol{P}$, $\boldsymbol{\mu_{od}}$, $\boldsymbol{\gamma_{API}}$	**0.016043**	**1.189452**	**0.003476**	**7.945602**	**0.267682**
μ_{ob}, P_b, P	0.017461	1.353631	0.011328	15.18432	0.410537
μ_{ob}, P_b, P, μ_{od}	0.015435	1.211996	0.004035	15.42525	0.310108

Statistical measures for the best functional form are shown in bold

Table 10 Performance of μ_{oa} empirical correlations

Correlation method	Correlating variables	RMSE	E_a	E_{min}	E_{max}	SD
Beal (1946)	μ_{ob}, P_b, P	0.026831	2.002863	0.015063	8.917567	0.311456
Vazquez and Beggs (1980)	μ_{ob}, P_b, P	0.076335	3.859566	0.005598	52.02459	0.639844
Labedi (1992)	$\boldsymbol{\mu_{ob}}$, $\boldsymbol{P_b}$, $\boldsymbol{P}$, $\boldsymbol{\mu_{od}}$, $\boldsymbol{\gamma_{API}}$	**0.022716**	**1.713268**	**0.001569**	**7.621104**	**0.064843**
Elsharkawy and Alikhan (1999)	μ_{ob}, P_b, P, μ_{od}	0.030771	2.474925	0.001203	16.15609	0.228751

Statistical measures for the best correlation are shown in bold

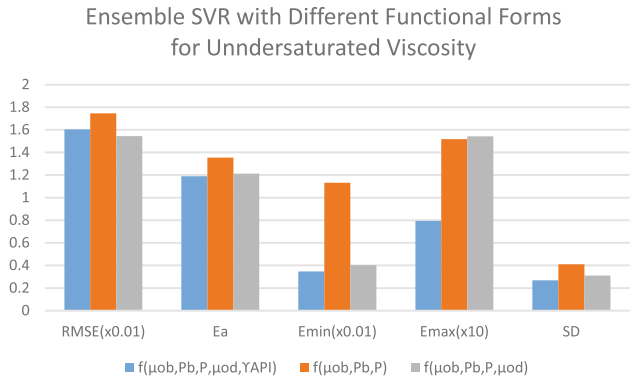

Fig. 3 Results of different functional forms for μ_{oa} prediction with ensemble SVR

in Table 9. The functional form $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$ has the best result with RMSE = 0.016043, $E_a = 1.189452$ and $E_{max} = 7.945602$. The second best performance is given by the functional form $f(\mu_{ob}, \mu_{od}, P_b, P)$.

Results of the investigated empirical correlations with different functional forms for modelling μ_{oa} are shown in Table 10. The correlation of Labedi (1992) gives the best performance among the empirical correlations with RMSE = 0.022716, $E_a = 1.713268$ and $E_{max} = 7.621104$. It is noted that additional correlating variables in the μ_{oa} modelling has improved the prediction results and the functional form of the best ensemble SVR model is the same as that of the best empirical correlation.

Similar to other two previous viscosity variables, μ_{od} and μ_{ob}, the ensemble SVR is again giving the best performance for μ_{oa} prediction. The overall best performance is given by the ensemble SVR of functional form $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$ with lowest values of RMSE, E_a and E_{max} among all the ensemble SVR models, and lowest RMSE and E_a among all methods. In fact, the results of other functional forms, $f(\mu_{ob}, P_b, P)$ and $f(\mu_{ob}, \mu_{od}, P_b, P)$ in modelling μ_{oa} are also better than all the empirical correlations since they have lower RMSE and E_a. This again shows consistency, reliability and good performing capability of the developed ensemble SVR in modelling the crude oil viscosity property. Figure 3 gives a graphical comparison of the ensemble SVR results with different functional forms for μ_{oa} prediction.

Conclusion

The following conclusions can be drawn from this work.

(1) This paper has presented a novel ensemble SVR model which uses E_a in ranking and building the final ensemble model. It was observed during experimentation that using RMSE for selecting the base models for the ensemble system also gives similar and consistent results.
(2) Different functional forms that are used for predicting μ_{od}, μ_{ob} and μ_{oa} have been investigated.
(3) In all cases, the ensemble SVR model gives the best results in predicting μ_{od}, μ_{ob} and μ_{oa} with the least statistical error values.
(4) For μ_{od} modelling, the best result is given by ensemble SVR with functional form $f(\gamma_{API}, T)$. This is an indication that additional correlating variable may not necessarily improve the performance of a model.
(5) For the μ_{ob} prediction, the best functional form for the ensemble SVR simulation is $f(\gamma_{API}, \mu_{od}, P_b)$. Among the investigated μ_{ob} correlations, Chew and Connally (1959) give the best performance and it is based on the functional form $f(\mu_{od}, R_s)$.
(6) For the μ_{oa} modelling, ensemble SVR with respect to all the three investigating functional forms, $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$, $f(\mu_{ob}, P_b, P)$ and $f(\mu_{ob}, \mu_{od}, P_b, P)$, gives better performance than all the compared empirical correlations. The overall best performance for μ_{oa} modelling is given by the ensemble SVR with functional form $f(\mu_{ob}, \mu_{od}, P_b, P, \gamma_{API})$.
(7) It can be noted that the errors are very high for the μ_{od} predictions from the empirical correlations. This has been noted by the previous works (Bergman and Sutton 2009). These are significantly reduced in the ensemble SVR model.
(8) Finally, it can be satisfactorily concluded that the ensemble SVR has better ability to model the uncertainties in the prediction of dead, saturated and undersaturated oil viscosity.

Acknowledgement Oloso Munirudeen thanks the Petroleum Technology Development Fund, Nigeria, for sponsoring his PhD research at the University of Portsmouth. The authors also thank GeoMark Research for supplying part of the PVT data used for this research. The authors also wish to thank the anonymous referees for their helpful comments.

Appendix 1

Statistical descriptions of the data sets used for this study are presented in Table 11.

Table 11 Data set for viscosity modelling

Variable	Minimum value	Maximum value
T	83	330
γ_{API}	15.5	49.5
R_s	25	2944
P_b	319	10,326
$P > P_b$	450	18,894.3
γ_g	0.85289	1.63131
μ_{od}	0.736	23.652
μ_{ob}	0.08	10
$\mu_o > \mu_{ob}$	0.09	11.5

Appendix 2

The PVT correlations evaluated in this study are given below.

Dead oil viscosity

Beal (1946)

$$\mu_{od} = z \times \left(360/(T + 200)\right)^X \tag{4}$$

where

$z = 0.32 + 1.8 \times 10^7/\gamma_{API}^{4.53}$,

$X = e^y$, and

$$y = 2.302585\left(0.43 + \frac{8.33}{\gamma_{API}}\right).$$

Beggs and Robinson (1975)

$$\ln(\ln(\mu_{od} + 1)) = a_1 + a_2\gamma_{API} + a_3 \ln T \tag{5}$$

where

$a_1 = 7.816432$, $a_2 = -0.04658$ and $a_3 = -1.163$

Glasø (1980)

$$\ln \mu_{od} = a_1 + a_2 \ln T + a_3 \ln(\ln(\gamma_{API}) + a_4(\ln T) \times \ln(\ln(\gamma_{API}) \tag{6}$$

where $a_1 = 54.5680543$, $a_2 = -7.1795304$, $a_3 = -36.447$ and $a_4 = 4.478879$

Kartoatmodjo and Schmidt (1991)

$$\mu_{od} = \left(\frac{16 \times 10^8}{T^{2.8177}}\right)(\log\gamma_{API})^X \tag{7}$$

where

$X = 5.7536\log T - 26.9718$.

Labedi (1992)

$$\ln \mu_{od} = a_1 + a_2 \ln\gamma_{API} + a_3 \ln T \tag{8}$$

where

$a_1 = 21.23904$; $a_2 = -4.7013$; $a_2 = -0.6739$

Petrosky Jr and Farshad (1995)

$$\mu_{od} = 2.3511 \times 10^7 T^{-2.10255}(\log\gamma_{API})^X \tag{9}$$

where

$X = 4.59388\log T - 22.82792$.

Elsharkawy and Alikhan (1999)

$$\mu_{od} = 10^X - 1 \tag{10}$$

where

$X = 10^y$, (and)

$y = 2.16924 - 0.02525\gamma_{API} - 0.68875\log T$

Dindoruk and Christman (2004)

$$\mu_{od} = \frac{a_3 T^{a_4}(\log\gamma_{API})^A}{a_5 P_b^{a_6} + a_7 R_{sb}^{a_8}} \tag{11}$$

where

$A = a_1 \log T + a_2$.

Naseri et al. (2005)

Coefficient	Value
a_1	14.505357625
a_2	−44.868655416
a_3	9.36579e+09
a_4	−4.194017808
a_5	−3.1461171e−09
a_6	1.517652716
a_7	0.010433654
a_8	−0.000776880

$$\mu_{od} = 10^X \tag{12}$$

where

$X = 11.2699 - 4.2699\log\gamma_{API} - 2.052\log T$.

Saturated viscosity

Chew and Connally (1959)

$$\mu_{ob} = X\mu_{od}^Y \tag{13}$$

where

$X = a_1 + a_2 e^{a_3 R_s}$

$Y = a_4 + a_5 e^{a_6 R_s}$

Beggs and Robinson (1975)

$$\mu_{ob} = X \mu_{od}^{Y} \tag{14}$$

where

$X = a_1 (R_s + a_2)^{a_3}$;

$Y = a_4 (R_s + a_5)^{a_6}$

$a_1 = 10.715;\ a_2 = 100;\ a_3 = -0.515;$

$a_4 = 5.44;\ a_5 = 150;\ a_6 = -0.338.$

Al-Khafaji et al. (1987)

$$\mu_{ob} = A \mu_{od}^{B} \tag{15}$$

$A = 0.247 + 0.2824X + 0.5657X^2 - 0.4065X^3 + 0.0631X^4$;

$B = 0.894 + 0.0546X + 0.07667X^2 - 0.0736X^3 + 0.01008X^4$;

$X = \log R_s$;

Khan et al. (1987)

$$\mu_{ob} = 0.09 \gamma_g^{0.5} R_s^{\frac{1}{3}} \theta_r^{-4.5} (1 - \gamma_o)^{-3} \tag{16}$$

where

$$\theta_r = \frac{T + 459.67}{459.67}.$$

Labedi (1992)

$$\ln \mu_{ob} = a_1 + a_2 \gamma_{API} + a_3 \ln \mu_{od} + a_4 \ln P_b \tag{17}$$

where

$a_1 = 5.397259;\ a_2 = -0.081557;\ a_3 = 0.6447;\ a_4 = -0.426$

Almehaideb (1997)

$$\ln \mu_{ob} = 13.4 - 0.597627 \ln R_s - 0.941624 \ln T - 0.555208 \ln \gamma_g - 1.487449 \ln \gamma_{API} \tag{18}$$

Elsharkawy and Alikhan (1999)

$$\mu_{ob} = A(\mu_{od})^{B} \tag{19}$$

where

$A = 1241.932(R_s + 641.026)^{-1.12410}$;

$B = 1768.841(R_s + 1180.335)^{-1.06622}$ (.)

Dindoruk and Christman (2004)

$$\mu_{ob} = A \mu_{od}^{B} \tag{20}$$

where

$$A = \frac{a_1}{\exp(a_1 R_s)} + \frac{a_3 R_s^{a_4}}{\exp(a_5 R_s)};$$

$$B = \frac{a_6}{\exp(a_7 R_s)} + \frac{a_8 R_s^{a_9}}{\exp(a_{10} R_s)}.$$

Coefficient	Value
a_1	1
a_2	4.740729e−04
a_3	−1.023451e−02
a_4	6.600358e−01
a_5	1.075080e−03
a_6	1
a_7	−2.191172e−05
a_8	−1.660981e−01
a_9	4.233179e−01
a_{10}	−2.273945e−04

Undersaturated viscosity

Beal (1946)

$$\mu_{oa = \mu_{ob} + (P - P_b)\left(a_1 \mu_{ob}^{a_2} + a_3 \mu_{ob}^{a_4}\right)} \tag{21}$$

$a_1 = 24e-06;\ a_2 = 1.6;\ a_3 = 38e-6;\ a_4 = 0.56$

Vazquez and Beggs (1980)

$$\mu_{oa} = \mu_{ob} (P/P_b)^{m} \tag{22}$$

where

$\ln m = a_1 + a_2 P + a_3 \ln P;$

$a_1 = -10.55749;\ a_2 = -89.8e-06;\ a_3 = 1.187$

Labedi (1992)

$$\mu_o = \mu_{ob} + m(P - P_b) \tag{23}$$

where

$\ln m = a_1 + a_2 \gamma_{API} + a_3 \ln \mu_{od} + a_4 \ln P_b$

$a_1 = -5.728832;\ a_2 = -0.045361;\ a_3 = 0.9036;\ a_4 = -0.3849$

Elsharkawy and Alikhan (1999)

$$\mu_o = \mu_{ob} + 10^{-2.0771} (P - P_b) \mu_{od}^{1.19279} \mu_{ob}^{-0.40712} P_b^{-0.7941} \tag{24}$$

Appendix 3

Statistical measures for the performance analysis

Average per cent relative error

$$E_r = \frac{1}{n}\sum_{1}^{n} E_i \tag{25}$$

where

$$E_i = \left(\frac{X_{\text{exp}} - X_{\text{pred}}}{X_{\text{exp}}}\right)_i \times 100 \tag{26}$$

$i = 1, 2, \ldots, n$

Average absolute per cent relative error

$$E_{\text{a}} = \frac{1}{n}\sum_{1}^{n} |E_i| \tag{27}$$

Maximum absolute per cent relative error

$$E_{\max} = \max_i |E_i| \tag{28}$$

Standard deviation

$$\text{SD} = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(E_i - E_r)^2} \tag{29}$$

where

$$E_r = \frac{1}{n}\sum_{i=1}^{n} E_i.$$

Root mean squared

$$\text{RMSE} = \left[\frac{1}{n}\sum_{i=1}^{n} E_i\right]^{0.5} \tag{30}$$

References

Ahmed T (2010) Reservoir Engineering Handbook. Gulf Professional Publishing, Boston

Al-Khafaji AH, Abdul-Majeed GH, Hassoon SF (1987) Viscosity correlation for dead, live and undersaturated crude oils. J Pet Res 6:1–16

Al-Marhoun MA, Nizamuddin S, Raheem AAA et al (2012) Prediction of crude oil viscosity curve using artificial intelligence techniques. J Pet Sci Eng 86:111–117

Almehaideb RA (1997) Improved PVT correlations for UAE crude oils. In: Middle east oil show and conference, SPE, Bahrain, 15–18 March 1997

Alomair O, Elsharkawy A, Alkandari H (2014) A viscosity prediction model for Kuwaiti heavy crude oils at elevated temperatures. J Pet Sci Eng 120:102–110

Alpaydin E (2014) Introduction to machine learning. MIT press, Cambridge

Ayoub MA, Raja DM, Al-Marhoun MA (2007) Evaluation of below bubble point viscosity correlations & construction of a new neural network model. In: Asia pacific oil and gas conference and exhibition, SPE, Jakarta, Indonesia, 30 Oct–1 Nov 2007

Bader-El-Den M, Gaber M (2012) Garf: towards self-optimised random forests. In: Neural information processing. Springer, pp 506–515

Bader-El-Den M, Teitei E, Adda M (2016) Hierarchical classification for dealing with the Class imbalance problem. In: Neural Networks (IJCNN), 2016 International Joint Conference on IEEE, pp 3584–3591

Beal C (1946) The viscosity of air, water, natural gas, crude oil and its associated gases at oil field temperatures and pressures. Trans AIME 165:94–115

Beggs HD, Robinson JR (1975) Estimating the viscosity of crude oil systems. J Pet Technol 27:1140–1141

Bennison T (1998) Prediction of heavy oil viscosity. In: IBC heavy oil field development conference, London, 2–4 December 1998

Bergman DF, Sutton RP (2007) An update to viscosity correlations for gas-saturated crude oils. In: SPE annual technical conference and exhibition, SPE, Anaheim, California, U.S.A, 11–14 November 2007

Bergman DF, Sutton RP (2009) A consistent and accurate dead-oil-viscosity method. SPE Reserv Eval Eng 12:815–840

Boukadi FH, Bemani AS, Hashmi A (2002) PVT empirical models for saturated Omani crude oils. Pet Sci Technol 20:89–100

Breiman L (1996) Bagging predictors. Mach Learn 24:123–140

Carbonell J, Michalski R, Mitchell T (1983) An Overview of Machine Learning. In: Michalski RS, Carbonell JG, Mitchell TM (eds) Machine Learning. Springer, Berlin, pp 3–23

Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geosci Model Dev 7:1247–1250

Chew J, Connally CA (1959) A viscosity correlation for gas-saturated crude oils. Trans AIME 216:23–25

De Ghetto G, Villa M (1994) Reliability analysis on PVT correlations. In: European petroleum conference, SPE, London, 25–27 October 1994

De Ghetto G, Paone F, Villa M (1995) Pressure–volume–temperature correlations for heavy and extra heavy oils. In: SPE international heavy oil symposium, SPE, Calgary, Alberta, Canada, 19–21 June 1995

Dietterich TG (2000) Ensemble methods in machine learning. In: Kittler J, Roli F (eds) Multiple classifier systems. Springer, Berlin, pp 1–15

Dindoruk B, Christman PG (2004) PVT properties and viscosity correlations for Gulf of Mexico oils. SPE Reserv Eval Eng 7:427–437

Egbogah EO, Ng JT (1990) An improved temperature–viscosity correlation for crude oil systems. J Pet Sci Eng 4:197–200

El-hoshoudy AN, Farag AB, Ali OIM et al (2013) New correlations for prediction of viscosity and density of Egyptian oil reservoirs. Fuel 112:277–282

Elsharkawy AM (1998) Modeling the properties of crude oil and gas systems using RBF network. In: SPE Asia pacific oil and gas conference and exhibition, SPE, Perth, Australia, 12–14 October 1998

Elsharkawy AM, Alikhan AA (1999) Models for predicting the viscosity of middle east crude oils. Fuel 78:891–903

Elsharkwy AM, Gharbi RBC (2001) Comparing classical and neural regression techniques in modeling crude oil viscosity. Adv Eng Softw 32:215–224

Elsharkawy AM, Hassan SA, Hashim YSK, Fahim MA (2003) New compositional models for calculating the viscosity of crude oils. Ind Eng Chem Res 42:4132–4142

Ghorbani B, Ziabasharhagh M, Amidpour M (2014) A hybrid artificial neural network and genetic algorithm for predicting viscosity of Iranian crude oils. J Nat Gas Sci Eng 18:312–323

Ghorbani B, Hamedi M, Shirmohammadi R et al (2016) A novel multi-hybrid model for estimating optimal viscosity correlations of Iranian crude oil. J Pet Sci Eng 142:68–76

Glasø Ø (1980) Generalized pressure–volume–temperature correlations. JPT 32(5):785–795. SPE-8016-PA. doi:10.2118/8016-PA

Goldberger J, Hinton GE, Roweis ST, Salakhutdinov RR (2005) Neighbourhood Components Analysis. In: Saul LK, Weiss Y, Bottou L (eds) Advances in neural information processing systems 17. MIT Press, Cambridge, pp 513–520

Hajizadeh Y (2007) Intelligent prediction of reservoir fluid viscosity. In: Production and Operations Symposium,SPE, Oklahoma City, Oklahoma, U.S.A, 31 Mar–3 Apr 2007

Hanafy HH, Macary SM, ElNady YM et al (1997) A new approach for predicting the crude oil properties. In: SPE production operations symposium, SPE, Oklahoma City, Oklahoma, 9–11 March 1997

Hemmati-Sarapardeh A, Aminshahidy B, Pajouhandeh A et al (2016) A soft computing approach for the determination of crude oil viscosity: light and intermediate crude oil systems. J Taiwan Inst Chem Eng 59:1–10

Hossain MS, Sarica C, Zhang H-Q, et al (2005) Assessment and development of heavy oil viscosity correlations. In: SPE international thermal operations and heavy oil symposium, SPE, Calgary, Alberta, Canada, 1–3 November 2005

Kartoatmodjo T, Schmidt Z (1991) New correlations for crude oil physical properties. SPE paper 23556

Kartoatmodjo T, Schmidt Z (1994) Large data bank improves crude physical property correlations. Oil Gas J 92(27):27

Khamehchi E, Rashidi F, Rasouli H, Ebrahimian A (2009) Novel empirical correlations for estimation of bubble point pressure, saturated viscosity and gas solubility of crude oils. Pet Sci 6:86–90

Khan SA, Al-Marhoun MA, Duffuaa SO, Abu-Khamsin SA (1987) Viscosity correlations for Saudi Arabian crude oils. In: Middle east oil show, SPE, Bahrain, 7–10 March 1987

Khoukhi A, Oloso M, Elshafei M et al (2011) Support vector regression and functional networks for viscosity and gas/oil ratio curves estimation. Int J Comput Intell Appl 10:269–293. doi:10.1142/S1469026811003100

Labedi R (1992) Improved correlations for predicting the viscosity of light crudes. J Pet Sci Eng 8:221–234

Liu H, Motoda H (2007) Computational methods of feature selection. CRC Press, Boca Raton

McCain WD Jr (1991) Reservoir-Fluid property correlations-state of the art. SPE Reserv Eng 6:266–272

Naseri A, Nikazar M, Dehghani SAM (2005) A correlation approach for prediction of crude oil viscosities. J Pet Sci Eng 47:163–174

Ng JTH, Egbogah EO (1983) An improved temperature–viscosity correlation for crude oil systems. In: Annual technical meeting. Petroleum Society of Canada

Oloso MA, Khoukhi A, Abdulraheem A, Elshafei M (2009) Prediction of crude oil viscosity and gas/oil ratio curves using recent advances to neural networks. In: SPE/EAGE Reservoir Characterization and Simulation Conference, SPE, Abu Dhabi, UAE, 19–21 October 2009

Oloso MA, Hassan MG, Buick J, Bader-El-Den M (2016) Oil PVT characterisation using ensemble systems. In: 2016 International conference on machine learning and cybernetics (ICMLC). IEEE, pp 61–68

Omole O, Falode OA, Deng AD (2009) Prediction of Nigerian crude oil viscosity using artificial neural network. Pet Coal 151:181–188

Perry T, Bader-El-Den M, Cooper S (2015) Imbalanced classification using genetically optimized cost sensitive classifiers. In: Evolutionary computation (CEC), 2015 IEEE Congress on IEEE, pp 680–687

Petrosky GE Jr, Farshad FF (1995) Viscosity correlations for Gulf of Mexico crude oils. In: SPE production operations symposium, SPE, Oklahoma City, Oklahoma, 2–4 April 1995

Petrosky GE Jr, Farshad F (1998) Pressure–volume–temperature correlations for Gulf of Mexico crude oils. SPE Reserv Eval Eng 1:416–420

Schapire RE (1990) The strength of weak learnability. Mach Learn 5:197–227

Stańczyk U, Jain LC (2015) Feature selection for data and pattern recognition. Springer, Berlin

Standing MB (1947) A pressure–volume–temperature correlation for mixtures of California oils and gases. In: Drilling and production practice, American Petroleum Institute, New York

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 58:267–288

Tibshirani R (2011) Regression shrinkage and selection via the lasso: a retrospective. J R Stat Soc Ser B (Stat Methodol) 73:273–282

Twu CH (1985) Internally consistent correlation for predicting liquid viscosities of petroleum fractions. Ind Eng Chem Process Des Dev 24:1287–1293

Vazquez M, Beggs HD (1980) Correlations for fluid physical property prediction. J Pet Technol 32:968–970

Yang W, Wang K, Zuo W (2012) Neighborhood component feature selection for high-dimensional data. JCP 7:161–168

Zhou Z-H (2012) Ensemble methods: foundations and algorithms. CRC Press, Boca Raton