A computational intelligence scheme for prediction of interfacial tension between pure hydrocarbons and water

Mahdi Kalantari Meybodi¹, Amin Shokrollahi¹, Hossein Safari¹, Moonyong Lee², Alireza Bahadori³,∗∗

¹ Department of Petroleum Engineering, Petroleum University of Technology, Ahvaz, Iran
² Young Researchers and Elite Club, North Tehran Branch, Islamic Azad University, Tehran, Iran
³ School of Chemical Engineering, Yeungnam University, Gyeongsan, Republic of Korea
⁴ Southern Cross University, School of Environment, Science and Engineering, Lismore, NSW, Australia

ARTICLE INFO

Article history:
Received 9 July 2014
Received in revised form 12 December 2014
Accepted 3 January 2015
Available online 20 January 2015

Keywords:
Interfacial tension
Interface
Water
Hydrocarbon
Least squares support vector machine (LSSVM)

ABSTRACT

Interfacial tension plays a major role in many disciplines of science and engineering. Complex nature of this property has restricted most of the previous theoretical studies on thermophysical properties to bulk properties measured far from the interface. Considering the drawbacks and deficiencies of preexisting models, there is yet a huge interest in accurate determination of this property using a rather simple and more comprehensive modeling approach. In recent years, inductive machine learning algorithms have widely been applied in solving a variety of engineering problems. This study introduces least-square support vector machines (LS-SVM) approach as a viable and powerful tool for predicting the interfacial tension between pure hydrocarbon and water. Comparing the model to experimental data, an excellent agreement was observed yielding the overall squared correlation coefficient (R²) of 0.993. Proposed model was also found to outperform when compared to some previously presented multiple regression models. An outlier detection method was also introduced to determine the model applicability domain and diagnose the outliers in the gathered dataset. Results of this study indicate that the model can be applied in systems over temperature ranges of 45.40–890°C and pressure ranges of 0.1–300 MPa.

© 2015 The Institution of Chemical Engineers. Published by Elsevier B.V. All rights reserved.

1. Introduction

Thermophysical properties of fluids are very important for engineering and industrial purposes. As a crucially important thermo-physical property, interfacial tension plays a vital role in describing the interface characteristics. Interfacial tension and its behavior are more complex than other thermophysical properties because it deals with surface and bulk characteristics of the system whereas other thermophysical properties only deal with bulk characteristics of the system (Dandekar, 1994).

The complexity and intricacy of the underlying molecular mechanisms have restricted most of the investigations to bulk system properties and traditionally not much attention has been paid to quantification and modeling of interfacial tension in comparison to other thermophysical properties. When two immiscible or partially miscible fluids come into contact, lighter phase lays over heavier phase due to buoyancy.
effect. In this case, a thin layer with different properties than bulks of the fluids, separates the two phases.

This layer is referred to as the interface. In the bulk of each phase, a molecule is impacted by the same intermolecular interactions from all sides because of molecular similarity, so that the resultant of the forces acting on each molecule tends to zero. On the other side, each molecule on the interface experiences different intermolecular interactions from different sides because of different phases on each side of the interface having different chemical nature or molecular density which result in different intermolecular interactions (Aranda-Bravo et al., 2008).

As a result, overall quantity of interactive forces acting on each of interfacial molecules will be non-zero (Ling and He, 2012). The imbalance in the applied forces results in the attraction of the molecules toward the phase exercising stronger interactions.

This phenomenon is generally known as interfacial tension. In other word, interfacial tension could be thought of as a quantitative index indicating the molecular behavior at the interface (Dandekar, 1994). From another point of view, interfacial tension is defined as the change in Gibbs free energy per unit change of interfacial area at defined temperature, pressure and chemical composition and is expressed as below (Demond and Lindner, 1993):

$$\left( \frac{\partial G}{\partial A} \right)_{T,P,n} = \gamma$$

(1)

In Eq. (1), G is Gibbs free energy, A is interfacial area, T is temperature, P is pressure and n is chemical composition. IFT is usually expressed as the force per unit area and is reported in dyne/cm or mN/m.

IFT is an important parameter in chemical, petroleum and petrochemical science and engineering (Aranda-Bravo et al., 2008). It plays an important role in separation processes, multiphase flow, fluid transportation, formation and stability of emulsions and reservoir engineering processes like enhanced oil recovery processes, fluid saturation distribution, fluid contacts and recovery mechanisms (Aranda-Bravo et al., 2008; Bahramian and Danesh, 2004; Danesh, 1998; Jasper and Duncan, 1967; Kahl et al., 2003; Kim and Burgess, 2001).

IFT between pure hydrocarbons and water is very important for liquid-liquid extraction processes and hydrocarbon stream sweetening processes in petroleum and petrochemical industries (Aranda-Bravo et al., 2008; Kahl et al., 2003). The role of IFT has also been acknowledged in understanding hydrocarbon-water emulsion properties and behavior that usually occurs in environment (Demond and Lindner, 1993).

This property can also be implemented in modeling and theoretical works that are done on the hydrocarbon-water interface properties and behavior (Kahl et al., 2003). Therefore accurate quantification and determination of the IFT between pure hydrocarbons and water could be highly desirable. The most accurate way for determination of interfacial tension between two phases like hydrocarbon and water is to measure it experimentally but it is highly unlikely to measure it under all operating conditions. Besides, experimental measurement of IFT are often time consuming and expensive making it a difficult task when IFT value is required immediately especially in engineering calculations (Demond and Lindner, 1993).

In this connection, theoretical models and empirical correlations have been developed to predict IFT value between pure hydrocarbons and water from known phase parameters. Theoretical models for prediction of the interfacial tension such as those developed by Antonoff (Kim and Burgess, 2001), Donahue and Bartell (1952), Girfalco and Good (1957), Fowkes (1963), Pliskin and Treybal (1966), Paul and Chazal (1969), Fu et al. (1986), Boudh-Hir and Mansoori (1991), Zuo and Stenby (1998), Bahramian and Danesh (2004), Wang and Anderko (2013) are based on theory which in some cases provide only a rough estimate and thus are accompanied by some inaccuracies.

Some of these models may even require some advanced phase properties in their calculations like: partial molar surface area (Bahramian and Danesh, 2004), van der Waals surface area (Fu et al., 1986) and adsorption equilibrium constants (Wang and Anderko, 2013) which in turn introduce more complexity and makes them less popular in engineering calculations.

Some empirical correlations have also been developed for prediction of interfacial tension between pure hydrocarbons and water. They appear to be much simpler than most of the aforementioned theoretical models and mostly require some
routine fluid properties like density of the phases, hydrocarbon critical temperature and system temperature. These empirical correlations are easy to implement approaches for prediction of IFT between hydrocarbons and water; nevertheless, they are prone to producing relatively off range values in some cases mainly because of their empirical nature. In the following, some of the more popular empirical correlations for the prediction of the IFT between pure hydrocarbons and water are presented.

Firoozabadi and Ramey (1988) measured the interfacial tension of some hydrocarbons and water and they found that by plotting a group that they called IFT function versus density difference between water and hydrocarbon, data lays on a single curve. Accordingly, they provided a graphical correlation that could predict interfacial tension between pure hydrocarbon and water using system temperature, hydrocarbon critical temperature and the density difference between hydrocarbon and water. They defined their IFT function as:

$$\text{IFT function} = \left(\frac{\gamma_{hw}}{\rho_w - \rho_h}\right) \left(\frac{T}{T_{c,h}}\right)^{0.3125}$$ \hspace{1cm} (2)

They investigated the applicability of their graphical correlation from methane to n-dodecane.

Later, Danesh (1998) provided the mathematical form of the graphical correlation of Firoozabadi and Ramey (1988) as:

$$\gamma_{hw} = 111(\rho_w - \rho_h)^{1.024} \left(\frac{T_R}{T_{c,h}}\right)^{-1.25} \hspace{1cm} (3)$$

After that, Sutton (2006) tried to better fit data of Firoozabadi and Ramey (1988) to the curve that had been represented by them. He proposed a new correlation as:

$$\gamma_{hw} = \left(\frac{1.58(\rho_w - \rho_h) + 1.76}{(T_R/T_{c,h})^4}\right)$$ \hspace{1cm} (4)

The newest correlation has been proposed by Sutton (2009). This correlation can predict the interfacial tension between pure hydrocarbons and water more accurately than previous correlations. Also his model can better predict the effect of temperature on interfacial tension between pure hydrocarbons and water in comparison to previous correlations. This correlation is represented as follows:

$$\gamma_{hw} = \left(\frac{1.53988(\rho_w - \rho_h) + 2.08339}{0.821978 - 0.001837857 x_1 + 0.000001340167 x_2^2}\right)^{3.6667} \hspace{1cm} (5)$$

Based on our previous discussions on merits and demerits of other modeling and correlation approaches, there is still a huge interest in developing more accurate and simple models capable of providing precise IFT values.

In the last decade, computer based methods including neural networks (Ahmadi et al., 2013a,b; Kumar, 2009; Majidi et al., 2014; Roosta et al., 2011; Tatar et al., 2013; Zendehboudi et al., 2013), neuro-fuzzy models (Chang and Chang, 2006; Kajehl et al., 2009) and least squares support vector machines (LSSVM) (Chamkani et al., 2013; Hemmati-Saraparadel et al., 2014; Safari et al., 2014; Shokrollahi et al., 2013; Bahadori et al., 2008) have been used for better prediction of thermophysical properties with higher accuracy, more rapidly and less expensively.

LSSVM strength in prediction of non-linear and complex phenomena makes it appealing to use attractive for using in engineering and especially petroleum engineering where large number of non-linear problems might exist (Arabloo et al., 2013; Farasat et al., 2013). To the best of our knowledge, this is probably the first time that LSSVM is applied for prediction of interfacial tension between pure hydrocarbons and water. The main aim of this work is to provide a powerful, fast and more accurate model based on LSSVM algorithm for interfacial tension prediction between pure hydrocarbons and water.

The objectives of this work is first to collect a large database of interfacial tension values of pure hydrocarbons against water which contains large number of components and wide ranges of pressure and temperature. Then this database is used to develop a LSSVM based model for accurate prediction of interfacial tension between pure hydrocarbons and water using commonly known properties of hydrocarbons and water. Afterwards, the accuracy and efficiency of the proposed model is investigated and compared to other previous correlations. Moreover, the reliability of the model for different conditions of pressure and temperature is investigated.

2. Least squares support vector machine

Recently, support vector machines have been introduced as powerful and robust tools for solving a variety of complex problems from pattern classification to nonlinear function approximation. Assuming that we have a set of experimental data, where $x_i$ denote the input patterns and $y_i$ the output patterns, SVM maps the input parameters into a higher dimensional feature space via implementing using a nonlinear mapping procedure.

In fact, primary goal of SVM is to find the optimum hyper-plane that has the minimum distance from all the experimental data (Cristianini and Shawe-Taylor, 2000; Suykens and Vandewalle, 1999). In the literatures (Haifeng and Dejin, 2005; Smola and Schölkopf, 2004) there is a comprehensive study about the procedures and formulations of nonlinear mapping in SVM.

When we are dealing with a function approximation or regression problem, support vectors (SVs) in SVM methods are founded by solving a convex optimization problem which will be converted to a quadratic programming (QP) situation subjected to inequality constraints. Solving the regression problems using the SVM methods are associated with a high computational burden due to constraints optimization programming which is implemented in these methods (Haifeng and Dejin, 2005). Thus, using SVM for function approximation with a wide range of data is very time and memory consuming due to employed optimization processes (Haifeng and Dejin, 2005).

Suykens and Vandewalle (1999) have developed a modified version of SVM, named least squares support vector machine (LSSVM) in which it is tried to reduce the complexity of the ordinary SVM and increasing its convergence speed. In LSSVM learning method, equality constrains are used instead of inequality ones that are used in ordinary SVM (Haifeng and Dejin, 2005; Suykens and Vandewalle, 1999). Using the equality constrains in LSSVM, make its learning process to include a system of linear equations that can be solved iteratively (Gharagheizi et al., 2011; Suykens and Vandewalle,
1999). Regarding these reasons, LSSVM is more favorable when we are dealing with a large scale problem in which the learning process must be done based on a wide dataset and time and accuracy are important. Generally, LSSVM optimization problem is formulated and expressed as follows (Suykens and Vandewalle, 1999):

\[
\min J(w, \varepsilon) = \frac{1}{2} \|w\|^2 + \frac{1}{2} \mu \sum_{i=1}^{n} \varepsilon_i^2
\]

subjected to the following linear constraints:

\[
y_k = \langle w, x_k \rangle + b + \varepsilon_k \quad k = 1, 2, \ldots, n
\]

In Eq. (7), \( \varepsilon_k \) represent error variables, is regularization constant, \( g(x) \) is the mapping function, \( w \) and \( b \) are weight vectors and bias terms respectively and superscript \( t \) denotes the transpose operator. Incorporating the linear constraint into the objective function previously introduced in Eq. (6) provides (Suykens and Vandewalle, 1999):

\[
L_{\text{LSSVM}} = \frac{1}{2} \|w\|^2 + \frac{1}{2} \mu \sum_{k=1}^{n} \varepsilon_i^2 - \sum_{k=1}^{n} \beta_k \left( \langle w, x_k \rangle + b + \varepsilon_k \right)
\]

with Lagrangian multipliers. According to the Lagrangian Multipliers method, the following conditions are required for optimization:

\[
\begin{align*}
\frac{\partial L_{\text{LSSVM}}}{\partial \varepsilon_k} &= 0 \quad \Rightarrow \quad \sum_{k=1}^{n} \beta_k = 0 \\
\frac{\partial L_{\text{LSSVM}}}{\partial w} &= 0 \quad \Rightarrow \quad w = \sum_{k=1}^{n} \beta_k g(x_k) \\
\frac{\partial L_{\text{LSSVM}}}{\partial \beta_k} &= 0 \quad \Rightarrow \quad \langle w, x_k \rangle + b + \varepsilon_k - y_k = 0 \quad (k = 1, \ldots, n) \\
\frac{\partial L_{\text{LSSVM}}}{\partial \beta_k} &= 0 \quad \Rightarrow \quad \beta_k = \mu \varepsilon_k
\end{align*}
\]

In LSSVM algorithm, if a linear regression is assumed between dependent and independent parameters, Eq. (8) can be rewritten as (Suykens and Vandewalle, 1999):

\[
y = \sum_{i=1}^{n} \beta_k x_i x + b
\]

Eq. (10) is only applicable for linear regression problems. For using Eq. (10) in nonlinear regression problems, Kernel function may be introduced in it as follows:

\[
y = \sum_{i=1}^{n} \beta_k K(x, x_i) + b
\]

where \( K(x, x_i) \) is the Kernel function obtained from inner product of vectors \( g(x) \) and \( g(x_i) \) in the feasible margin as is represented below:

\[
K(x, x_i) = \langle g(x), g(x_i) \rangle
\]

However, several types of Kernel functions are available and have been used in learning process of SVM methods but in LSSVM learning process, Gaussian Radial Basis Function (RBF) Kernel is more interesting to use which is expressed as (Hai Feng and Dejin, 2005):

\[
K(x, x_i) = \exp \left( \frac{|x - x_i|^2}{\sigma^2} \right)
\]

where \( \sigma^2 \) is squared bandwidth that is optimized during the training process by an external optimization technique.

The mean squared error (MSE) between the experimental data and LSSVM predicted values is defined as (Bahadori and Vuthaluru, 2009):

\[
\text{MSE} = \frac{\sum_{i=1}^{N} (Q_{\text{predi}} - Q_{\text{expi}})^2}{N}
\]

where \( Q \) represents IFT, \( N \) is the number of objects in training dataset and subscripts predi and expi denote the predicted and experimental IFT values, respectively.

The LSSVM algorithm that is used in this study to model the IFT data has been developed by Suykens et al. (2002) and Suykens and Vandewalle (1999), \( \mu \) and \( \sigma^2 \) which are the model parameters controlling the model accuracy and convergence are optimized by coupled simulated annealing (CSA) algorithm which is used during learning process to improve the model efficiency.

3. Coupled simulated annealing (CSA)

Simulated annealing (SA) is a probative search method usually applied for combinatorial optimization problems. The method was initially introduced by Metropolis et al. (1953), and was popularized by Kirkpatrick et al. (1983) afterwards. The physical process of annealing is the motivation for developing this method. In the annealing process, liquid metal which is produced by heating a solid metal, is cooled slowly enough in the way that all metal crystals finally attain the lowest minimum inner energy. Like this process, SA method gradually converges to the global optimum solution of the problem and prevent the wrong selection of local optimum instead of global optimum (Fabian, 1997).

The SA optimization method that is employed in this study is coupled simulated annealing (CSA) method proposed by Xavier-de-Souza et al. (2010) to improve the quality of SA optimization process. The concept of CSA is motivated by the coupled local minimizers (CLM) in which multiple gradient descent optimizers are used instead of multi-start gradient descent for optimization problem. In CSA optimization process, a group of individual SA processes is coupled by a term in acceptance probability function. CSA is faster and more decisive than SA in reaching convergence. The coupling is a function of the current costs of all the individual SA processes (Xavier-de-Souza et al., 2010).

In CSA method, individual SA processes share the information through both coupling term and acceptance probability function which is used to control general optimization attainment through optimization control parameters (Xavier-de-Souza et al., 2010). Although in traditional SA the acceptance probability of an uphill move is usually provided by Metropolis rule (Metropolis et al., 1953), other current solutions are also considered in CSA. This probability also depends on the solution costs via a coupling term \( \psi \) in state set \( S \), where \( S \) is the set of all possible solutions. \( \psi \) is generally believed to be a
<table>
<thead>
<tr>
<th>Reference</th>
<th>Hydrocarbon</th>
<th>Pressure range (MPa)</th>
<th>Temperature range (°C)</th>
<th>Density difference range (g/cm³)</th>
<th>IFT range (mN/m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harkins and Humphrey (1916)</td>
<td>C₆, benzene, toluene</td>
<td>0.1</td>
<td>509.67–563.67</td>
<td>0.11049–0.34227</td>
<td>33.22–49.54</td>
</tr>
<tr>
<td>Harkins et al. (1920)</td>
<td>C₆, C₈, benzene, toluene, m-xylene</td>
<td>0.1</td>
<td>527.67</td>
<td>0.11931–0.33885</td>
<td>35.03–51.25</td>
</tr>
<tr>
<td>Bartell et al. (1933)</td>
<td>Benzene, toluene, butylnaphthalene, propylbenzene</td>
<td>0.1</td>
<td>527.67–536.67</td>
<td>0.11931–0.22316</td>
<td>33.91–40.64</td>
</tr>
<tr>
<td>Carter and Jones (1934)</td>
<td>Benzene, cyclohexane</td>
<td>0.1</td>
<td>536.67</td>
<td>0.1204–0.223</td>
<td>36.64–47</td>
</tr>
<tr>
<td>Transue and Bartell (1942)</td>
<td>Benzene, toluene</td>
<td>0.1</td>
<td>536.67</td>
<td>0.12338–0.13481</td>
<td>33.9–35.7</td>
</tr>
<tr>
<td>Cupples (1947)</td>
<td>Benzene</td>
<td>0.1</td>
<td>536.67</td>
<td>0.12338</td>
<td>32.67</td>
</tr>
<tr>
<td>Hough et al. (1951)</td>
<td>C₁₀, benzene</td>
<td>0.1–102.04</td>
<td>534–740</td>
<td>0.69064–0.99676</td>
<td>25.5–75.5</td>
</tr>
<tr>
<td>Michaels and Hauser (1951)</td>
<td>C₁₀, benzene</td>
<td>0.1–70.1</td>
<td>533.52–728.73</td>
<td>0.14–20.8–284.85</td>
<td>24.72–46.93</td>
</tr>
<tr>
<td>Rose and Seyer (1951)</td>
<td>C₆, C₈, C₁₀, cyclohexane</td>
<td>0.1</td>
<td>527.67–617.67</td>
<td>0.21958–0.33885</td>
<td>49.09–51.25</td>
</tr>
<tr>
<td>Donahue and Bartell (1952)</td>
<td>C₅, C₆, C₇, C₈, cyclohexane, benzene, toluene, ethylbenzene</td>
<td>0.1</td>
<td>536.67</td>
<td>0.12400–0.37622</td>
<td>36.1–50.2</td>
</tr>
<tr>
<td>Hassan et al. (1953)</td>
<td>C₃, C₄, C₅, C₆, C₈, i-C₈, benzene</td>
<td>0.1–20.52</td>
<td>539.45–639.45</td>
<td>0.11780–0.54541</td>
<td>29–69.64</td>
</tr>
<tr>
<td>Shewmaker et al. (1954)</td>
<td>Benzene, toluene, ethylbenzene, propylbenzene, isopropylbenzene, butylbenzene</td>
<td>0.1</td>
<td>536.67</td>
<td>0.12365–0.14105</td>
<td>34–39.6</td>
</tr>
<tr>
<td>Jasper and Seitz (1959)</td>
<td>Toluene, ethylbenzene, propylbenzene, butylbenzene</td>
<td>0.1</td>
<td>527.67–635.67</td>
<td>0.13192–0.16347</td>
<td>29.2–35.8</td>
</tr>
<tr>
<td>Valentine and Heideger (1963)</td>
<td>Benzene</td>
<td>0.1</td>
<td>527.67–534.67</td>
<td>0.11927–0.12256</td>
<td>33.8–39.3</td>
</tr>
<tr>
<td>Aveyard and Haydon (1965)</td>
<td>C₅, C₆, C₇, C₈, C₁₀, C₁₂, C₁₄, C₁₆</td>
<td>0.1</td>
<td>454.40–559.17</td>
<td>0.2099–0.3715</td>
<td>49.24–53.77</td>
</tr>
<tr>
<td>Johnson and Dettre (1966)</td>
<td>C₅, C₆, C₇, C₈, C₁₀, C₁₂, C₁₆</td>
<td>0.1</td>
<td>535.67</td>
<td>0.25107–0.37585</td>
<td>50.3–53.3</td>
</tr>
<tr>
<td>Tamai and Kobayashi (1970)</td>
<td>C₅, cyclohexane</td>
<td>0.1</td>
<td>527.67</td>
<td>0.21943–0.33870</td>
<td>50.4–50.7</td>
</tr>
<tr>
<td>Jennings and Newman (1971)</td>
<td>C₁₀, C₁₀</td>
<td>10–81.63</td>
<td>714–990</td>
<td>0.18490–0.98487</td>
<td>31–64.2</td>
</tr>
<tr>
<td>Jones (1973)</td>
<td>C₁₀</td>
<td>0.1</td>
<td>518.67–590.67</td>
<td>0.26478–0.28252</td>
<td>48.1–51.6</td>
</tr>
<tr>
<td>Tornberg (1977)</td>
<td>C₆, benzene</td>
<td>0.1</td>
<td>527.67</td>
<td>0.11931–0.33885</td>
<td>33.35–50.71</td>
</tr>
<tr>
<td>Patterson and Ross (1979)</td>
<td>C₅, benzene</td>
<td>0.1</td>
<td>536.67</td>
<td>0.1190–0.3378</td>
<td>34–48.75</td>
</tr>
<tr>
<td>Motomura et al. (1983)</td>
<td>C₆, C₈, C₁₀, C₁₂, cyclohexane, benzene</td>
<td>0.1</td>
<td>536.67</td>
<td>0.12339–0.34228</td>
<td>34–51.85</td>
</tr>
<tr>
<td>Lögren et al. (1984)</td>
<td>C₁₀,Benzene</td>
<td>0.1</td>
<td>527.67</td>
<td>0.11927–0.26780</td>
<td>35–50.6</td>
</tr>
<tr>
<td>Mori et al. (1984)</td>
<td>C₅, i-C₅, C₆, benzene</td>
<td>0.1</td>
<td>518.67–599.67</td>
<td>0.11931–0.38324</td>
<td>34.05–49.46</td>
</tr>
<tr>
<td>Fu et al. (1986)</td>
<td>C₅, C₆, C₇, C₈, benzene, toluene, cyclohexane</td>
<td>0.1</td>
<td>527.4–545.4</td>
<td>0.11931–0.37612</td>
<td>33–50.2</td>
</tr>
<tr>
<td>Firoozabadi and Ramey (1988)</td>
<td>Cyclohexane, benzene</td>
<td>0.1–82.74</td>
<td>533.94–810.06</td>
<td>0.23383–0.28348</td>
<td>31.1–51.1</td>
</tr>
<tr>
<td>Motomura et al. (1988)</td>
<td>Cyclohexane, benzene</td>
<td>0.1–97.8</td>
<td>505.17–554.67</td>
<td>0.10618–0.29266</td>
<td>33.27–51.2</td>
</tr>
<tr>
<td>Backes et al. (1990)</td>
<td>C₅, C₆, benzene, toluene, cyclohexane, ethylbenzene, methylcyclohexane, m-xylene, o-xylene, p-xylene</td>
<td>0.1</td>
<td>509.67–563.67</td>
<td>0.12156–0.35068</td>
<td>33.8–51.9</td>
</tr>
<tr>
<td>Badakshan and Bakes (1990)</td>
<td>C₅, cyclohexane</td>
<td>0.1</td>
<td>529.2–536.4</td>
<td>0.22020–0.34217</td>
<td>51.23–52.3</td>
</tr>
<tr>
<td>Ikeda et al. (1992)</td>
<td>C₅</td>
<td>0.1–98.6</td>
<td>509.67–563.67</td>
<td>0.31251–0.35140</td>
<td>48.89–52.99</td>
</tr>
<tr>
<td>Rashidnia et al. (1992)</td>
<td>Cyclohexane</td>
<td>0.1</td>
<td>529.2</td>
<td>0.219</td>
<td>48</td>
</tr>
<tr>
<td>Demond and Lindner (1993)</td>
<td>C₅, C₁₀, C₁₂, i-C₈, m-C₈, C₁₀, C₁₂, C₁₄, C₁₆, benzene, toluene, m-xylene, o-xylene, p-xylene, ethylbenzene, cyclohexane, butylbenzene</td>
<td>0.1</td>
<td>527.67</td>
<td>0.11801–0.37793</td>
<td>35–53.3</td>
</tr>
<tr>
<td>Jańczuk et al. (1993)</td>
<td>C₅, C₇, C₈, C₉, C₁₀, C₁₁, C₁₂, C₁₃, C₁₄, C₁₅, C₁₆, benzene</td>
<td>0.1</td>
<td>527.67</td>
<td>0.11927–0.37246</td>
<td>33.62–51.39</td>
</tr>
<tr>
<td>Wiegand and Franck (1994)</td>
<td>C₁, C₁₂, C₁₀, toluene</td>
<td>0.1–300</td>
<td>531–851.4</td>
<td>0.09394–0.99642</td>
<td>16.66–71.24</td>
</tr>
<tr>
<td>Liggieri et al. (1995)</td>
<td>C₁</td>
<td>0.1</td>
<td>527.67</td>
<td>0.33885</td>
<td>50.5</td>
</tr>
<tr>
<td>Sachs and Meyn (1995)</td>
<td>C₁</td>
<td>0.44–46.88</td>
<td>536.76</td>
<td>0.74635–0.99437</td>
<td>51.83–71.32</td>
</tr>
<tr>
<td>Cai et al. (1996)</td>
<td>C₅, C₈, C₁₀, C₁₂, C₁₄, C₁₆</td>
<td>3.69–29.25</td>
<td>536.4–635.4</td>
<td>0.23390–0.36829</td>
<td>47.20–54.26</td>
</tr>
</tbody>
</table>
function of all costs of solution in $y$. The following relationship
gives the acceptance probability function in CSA, $A^*$:

$$A^*(\psi, x_i \rightarrow y_j) = \frac{\exp \left( \frac{E(x_i) - \max_{x_i \in A} E(x_i)}{T_k^*} \right)}{\psi}$$

(15)

where $T_k^*$ is the acceptance temperature, $x_i$ and $y_j$ represent individual solutions in and their corresponding probing solution, respectively. And coupling term, $\psi$, is defined as:

$$\psi = \sum_{\nu \in \mathcal{C}} \exp \left( \frac{E(x_i) - \min_{y_j \in \mathcal{D}} E(y_j)}{T_k^*} \right)$$

(16)

In this study, CSA-LSSVM method is used to model the hydrocarbon-water IFT in which the CSA optimization method is used to optimize the parameters in LSSVM method. Finally MSE, as given in Eq. (14), is selected as objective function of CSA-LSSVM which must be minimized during the search for the optimum of the problem.

4. Results and discussion

4.1. Data gathering

Comprehensiveness and validity of any model depends to a large extent on reliability and accuracy of the used data points during the development of the model. A review on literature shows that the main parameter that influence the pure hydrocarbon/water IFT are critical temperature of hydrocarbon compound, temperature, and density difference between water and hydrocarbon (Danesh, 1998; Firoozabadi and Ramey, 1988; Sutton, 2006, 2009). In order to construct the LSSVM model, an approximate number of 1213 experimental IFT data points were assembled from 31 different hydrocarbon/water mixtures reported in the open literature, the details of which have been provided in Table 1. According to this table, the employed dataset covers a wide range of experimental conditions. The overall details of the experimental data used for construction of LSSVM model are provided in Table 2.

4.2. LSSVM design

LSSVM algorithm proposed by Pelckmans et al. (2002) and Suykens and Vandewalle (1999) was used to model the hydrocarbon/water IFT. To accurately predict the hydrocarbon/water IFT, correlating variables were chosen as follows:

$$\text{IFT}_{\text{hydrocarbon/water}} = f \left( \frac{T_c}{T_c^- \text{hydrocarbon}}, T, \rho_{\text{hydrocarbon}}, \rho_{\text{water}} \right)$$

(17)

To best correlate the aforementioned variables, the following form of variables was considered to build the LSSVM model:

$$\text{IFT}_{\text{hydrocarbon/water}} = f \left( \frac{T}{T_c \text{hydrocarbon}}, \rho_{\text{water}} - \rho_{\text{hydrocarbon}} \right)$$

(18)

Most widely used kernel function, radial basis function (RBF), was applied to construct the model (Arabloo et al., 2013; Hemmati-Sarapardeh et al., 2014; Mohammadi et al., 2011;
4.3. Model accuracy

For checking the accuracy of developed LSSVM model, comparisons between predicted IFT values and corresponding experimental values from the prepared database are done using statistical and graphical means.

For graphical investigations, the agreement between predicted values using the proposed model and corresponding experimental values is shown in Fig. 1. It shows that the model reliably predicts the IFT between pure hydrocarbons and water with a very good coefficient of determination ($R^2$) of 0.993. For further verification, relative error is plotted as a function of experimental values in Fig. 2. It shows the concentration of data points around the zero error corresponding horizontal line which verifies the reliability and accuracy of the proposed model.

Statistical investigations are performed using statistical parameters like coefficient of determination ($R^2$), percent average relative error (ARE%), percent average absolute relative error (AARE%), root mean square error (RMSE), Maximum absolute relative error (%) and Minimum absolute relative error (%) for Train, Test, and the whole dataset and the results are reported in Table 3.

As both graphical and statistical investigations show, proposed model predicts the IFT between pure hydrocarbons and water efficiently with a very small deviation.

For performance comparison between the proposed model and other correlations, the same statistical and graphical investigations are also performed for Danesh correlation (Danesh, 1998) and Sutton’s correlations (Sutton, 2006, 2009) and the results are provided in Fig. 3 and Table 4.

As shown in Fig. 3, for all of the correlations especially Danesh correlation (Danesh, 1998), the deviations from the quadrant line increase as the IFT value increases. These higher deviations demonstrate their lower applicability for high IFT value prediction. However, the proposed model has a uniform and very small deviation from the quadrant line as shown in Fig. 1 revealing its accuracy for high IFT predictions.

Graphical comparisons of statistical parameters of the correlations and proposed model are represented in Fig. 4.

From Fig. 4 and Table 4, it is deduced that the proposed model has the highest coefficient of correlation ($R^2$) and lowest values of ARE, AARE and RMSE which show that the proposed model is a more effective and more accurate IFT predictive tool than other correlations. Also it is concluded that Danesh correlation (Danesh, 1998) gives less effective and more erroneous predictions in comparison with other correlations and

---

**Table 2 – Ranges of input/output variables used for developing and testing the model.**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature [°C]</td>
<td>454.40</td>
<td>890</td>
<td>598.31</td>
</tr>
<tr>
<td>Critical Temperature (hydrocarbon) [°C]</td>
<td>343.02</td>
<td>1299.60</td>
<td>903.02</td>
</tr>
<tr>
<td>Density difference (water/hydrocarbon) [g/cm³]</td>
<td>0.0939</td>
<td>0.9968</td>
<td>0.3671</td>
</tr>
<tr>
<td>IFT (water/hydrocarbon) [mN/m]</td>
<td>16.70</td>
<td>75.50</td>
<td>46.40</td>
</tr>
</tbody>
</table>
Fig. 3 – Agreement between predicted IFT values using correlations and experimental values: (a) Danesh (1998); (b) Sutton (2006); (c) Sutton (2009).

Fig. 4 – Statistical parameter comparison of proposed LSSVM model and other correlations: (a) $R^2$; (b) ARE; (c) AARE; (d) RMSE.
proposed model. In addition, as it is shown in Fig. 4b, Danesh correlation (Danesh, 1998) overestimates the IFT values and large discrepancies exist between the experimental IFT values and those predicted by this correlation.

Cumulative frequency of absolute relative error is shown in Fig. 5 as the percent of all used data versus percent absolute relative error for proposed model and other correlations. This shows that more than 90% of the predicted values using the proposed model have absolute relative errors of lower than 4% indicating the effectiveness of the proposed model.

4.4. Validity of developed model against pressure and temperature

One of the distinguishing features of the LSSVM model developed in this study is that it can accurately predict the interfacial tension between hydrocarbons and water at high pressure and high temperature conditions at which preexisting correlations usually fail. Figs. 6 and 7 show the degrees of accuracy for the proposed model as well as empirical correlations when used for IFT prediction at the entire pressure and temperature ranges, respectively.

Fig. 6 shows the average absolute relative error of proposed model and other correlations for several ranges of pressures. It shows that the proposed model predicts the IFT values reliably, with an acceptable degree of accuracy (less than 2% error) within all pressure ranges. It also shows that Danesh (1998) and Sutton (2006) correlation provides less accurate results especially at low pressures. Also it shows that the accuracy of Sutton (2009) correlation decreases as the pressure increases indicating that they might not be a good choice for IFT prediction at higher pressures.

Fig. 7 also compares the average absolute relative errors of the model proposed in this study with empirical correlations over the entire temperature ranges. According to the results, the LSSVM model provides more accurate results (less than 3% error) within wide temperature ranges.

Based on the results obtained in this study, although the model of Sutton (2009) provides better predictions when compared to other empirical correlations, however it still fails at high pressure and temperature conditions. On the other hand, the LSSVM model was found to outperform these correlations by providing reasonably accurate results when used over the whole temperature and pressure ranges especially at extreme conditions.

Accordingly, LSSVM model equipped with CSA optimization algorithm is a very effective and highly accurate tool for nonlinear regressions, as it was implemented in this work for interfacial tension prediction between pure hydrocarbons and water.

4.5. Outliers in experimental data

To propose a valid and highly applicable model for IFT prediction, experimental data must be reliable and accurate. However, precise measurement of experimental IFT data is not always possible and usually some errors and faulty measured data exist in reported data. Occurrence of this problem in such large database that has been used in this study is possible. These erroneous data are usually different from overall bulk of the data which may be a threat to a successful modeling. Thus, recognizing these erroneous data from overall bulk

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No of data points</td>
<td>972</td>
<td>241</td>
<td>1213</td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.9944</td>
<td>0.9878</td>
<td>0.9931</td>
<td></td>
</tr>
<tr>
<td>ARE%</td>
<td>0.0722</td>
<td>−0.2636</td>
<td>0.00547</td>
<td></td>
</tr>
<tr>
<td>AARE%</td>
<td>1.29750</td>
<td>2.06471</td>
<td>1.44993</td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.96346</td>
<td>1.41906</td>
<td>1.06954</td>
<td></td>
</tr>
<tr>
<td>Maximum absolute relative error (%)</td>
<td>16.8363</td>
<td>17.59282</td>
<td>17.59282</td>
<td></td>
</tr>
<tr>
<td>Minimum absolute relative error (%)</td>
<td>0.00028</td>
<td>0.00061</td>
<td>0.00028</td>
<td></td>
</tr>
<tr>
<td>Absolute relative error (°)</td>
<td>148.7238</td>
<td>173.6699</td>
<td>55.16132</td>
<td></td>
</tr>
<tr>
<td>Minimum absolute relative error (%)</td>
<td>0.14325</td>
<td>0.00121</td>
<td>0.0134</td>
<td></td>
</tr>
</tbody>
</table>

Table 3 – Statistical parameters of the proposed LSSVM model for IFT prediction between pure hydrocarbons and water.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R²</td>
<td>0.719</td>
<td>0.7487</td>
<td>0.889</td>
<td>0.9931</td>
</tr>
<tr>
<td>ARE%</td>
<td>20.17135</td>
<td>0.30547</td>
<td>−0.54267</td>
<td>0.00547</td>
</tr>
<tr>
<td>AARE%</td>
<td>29.14639</td>
<td>12.05297</td>
<td>6.48074</td>
<td>1.44993</td>
</tr>
<tr>
<td>RMSE</td>
<td>17.50293</td>
<td>9.8465</td>
<td>4.44609</td>
<td>1.06954</td>
</tr>
<tr>
<td>Maximum absolute relative error (%)</td>
<td>148.7238</td>
<td>173.6699</td>
<td>55.16132</td>
<td></td>
</tr>
<tr>
<td>Minimum absolute relative error (%)</td>
<td>0.14325</td>
<td>0.00121</td>
<td>0.0134</td>
<td></td>
</tr>
</tbody>
</table>

Table 4 – Statistical parameters of other correlations and proposed LSSVM model for IFT prediction between pure hydrocarbons and water.
of the data is important for proposing an accurate, efficient, and reliable model.

In this study, the leverage value statistics method was used so as to successfully and efficiently determine the doubtful experimental data. In usage of leverage value statistics method, two parameters including residual values for all data points and Hat matrix are needed for calculations. Residual value for each data point is the deviations of proposed CSA-LSSVM model in predicting IFT values from experimental data. Hat matrix is composed of experimentally measured data and those values predicted by the model and generally, it is constructed as follows (Eslamimanesh et al., 2013; Goodall, 1993; Gramatica, 2007):

$$ H = X(X^t X)^{-1} X^t $$

(19)

In Eq. (19), $X$ is a two-dimensional matrix containing $m$ rows and $n$ columns where $m$ is the total number of used data and $n$ is the number of model parameters. The superscript $t$ in this equation also indicates the transpose operator.

The applicable region of the proposed model is demonstrated by the diagonal elements of Hat matrix. Outlier detection is usually conducted graphically by sketching the William’s plot using the calculated $H$ values from Eq. (19) (Eslamimanesh et al., 2013; Goodall, 1993; Gramatica, 2007; Mohammadi et al., 2012). William’s plot shows the relationship between standardized residuals and $H$ values as it is shown in Fig. 8 for the proposed model in this study. A warning leverage value, denoted by $H^*$ and represented by a green line in Fig. 8, is normally defined equal to $3(n+1)/m$ where $n$ is the number of input parameters and $m$ is the number of all data points used in model development. The acceptable data from this point of view must be laid in the window bounded between $H$ value equal to 0 to $H^*$ which is equal to 0.0099 in this study. Moreover, leverage value of 3 is generally selected as the cut-off value which means that the acceptable data are laid in the window bounded between standard deviations of −3 to +3 from the mean which are represented as two red lines in Fig. 8 (Eslamimanesh et al., 2013; Goodall, 1993; Gramatica, 2007; Mohammadi et al., 2012). According to the two mentioned
limits, the acceptable data must be applied to and . As it is shown in Fig. 8 most of the data lie in the mentioned ranges of H and R which indicates the high applicability, reliability, accuracy and statistical validity of the proposed model. According to the mentioned limits for H and R, suspected data may be categorized in two main groups including leverage points and regression outliers. The suspected data of the leverage point group can be categorized in two subgroups including good leverage and bad leverage. Good leverage points are the data points in which \( H^* \leq H \) and \(-3 \leq R \leq 3\). Although the good leverage points have high leverage values but they usually behave similar to the bulk of data and close to the line where most of the data are concentrated around \(-3 \leq R \leq 3\). Thus good leverage points do not invariably affect the correlation coefficient. Bad leverage points are those data in which \( R > 3 \) or \( R < -3 \) where H values are not a concern. These points are also called influential outliers. Even though influential points may possess high leverage values, but the concern about these points is that they do not behave similar to the bulk of data and they are far from the bulk of data. So these points strongly affect the slope and intercept of regression line and correlation coefficient. The last group of suspected data are regression outliers that may violate the acceptable ranges of H and R; nevertheless, apparently they have no influence on regression line except having high leverage values (Eslamimanesh et al., 2013; Goodall, 1993; Gramatica, 2007; Mohammadi et al., 2012).

As is shown in Fig. 8, total number of 18 data points were diagnosed of deviating from bulk IFT measurements and tagged as probable outlying points which their details are given in Table 5.

<table>
<thead>
<tr>
<th>No.</th>
<th>Reference</th>
<th>Hydrocarbon</th>
<th>Pressure (MPa)</th>
<th>Temperature (°R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tian et al. (1997)</td>
<td>C1</td>
<td>30</td>
<td>851.4</td>
</tr>
<tr>
<td>2</td>
<td>Aveyard and Haydon (1965)</td>
<td>C8</td>
<td>0.1</td>
<td>454.4</td>
</tr>
<tr>
<td>3</td>
<td>Jennings and Newman (1971)</td>
<td>C1</td>
<td>20</td>
<td>852</td>
</tr>
<tr>
<td>4</td>
<td>Michaels and Hauser (1951)</td>
<td>C10</td>
<td>40.1</td>
<td>534.06</td>
</tr>
<tr>
<td>5</td>
<td>Hassan et al. (1953)</td>
<td>C3</td>
<td>3.39</td>
<td>539.55</td>
</tr>
<tr>
<td>6</td>
<td>Michaels and Hauser (1951)</td>
<td>C10</td>
<td>20.1</td>
<td>534.06</td>
</tr>
<tr>
<td>7</td>
<td>Michaels and Hauser (1951)</td>
<td>C10</td>
<td>0.1</td>
<td>534.06</td>
</tr>
<tr>
<td>8</td>
<td>Donahue and Bartell (1952)</td>
<td>Benzene</td>
<td>0.1</td>
<td>536.67</td>
</tr>
<tr>
<td>9</td>
<td>Jennings and Newman (1971)</td>
<td>C1</td>
<td>40</td>
<td>852</td>
</tr>
<tr>
<td>10</td>
<td>Jennings and Newman (1971)</td>
<td>C1</td>
<td>30</td>
<td>852</td>
</tr>
<tr>
<td>11</td>
<td>Michaels and Hauser (1951)</td>
<td>C10</td>
<td>60.1</td>
<td>534.06</td>
</tr>
<tr>
<td>12</td>
<td>Jennings and Newman (1971)</td>
<td>C1</td>
<td>50</td>
<td>852</td>
</tr>
<tr>
<td>13</td>
<td>Kashefi (2012)</td>
<td>C1</td>
<td>20.94</td>
<td>851.67</td>
</tr>
<tr>
<td>14</td>
<td>Wiegand and Franck (1994)</td>
<td>C1</td>
<td>10</td>
<td>536.4</td>
</tr>
<tr>
<td>15</td>
<td>Jennings and Newman (1971)</td>
<td>C1</td>
<td>60</td>
<td>851.4</td>
</tr>
<tr>
<td>16</td>
<td>Wiegand and Franck (1994)</td>
<td>C10</td>
<td>60</td>
<td>851.4</td>
</tr>
<tr>
<td>17</td>
<td>Contras and Oteopeu (2000)</td>
<td>C10</td>
<td>0.1</td>
<td>534.87</td>
</tr>
<tr>
<td>18</td>
<td>Jennings and Newman (1971)</td>
<td>C1</td>
<td>60</td>
<td>852</td>
</tr>
</tbody>
</table>

5. Conclusion

Interfacial tension is a key parameter which is used in many fields of science and engineering. Due to Complex nature of this thermophysical property, modeling of this property is not simple. In this study, applicability of LSSVM modeling approach for modeling of water/hydrocarbon IFT was investigated for the first time. Statistical parameters (i.e. \( R^2 \), RMSE, AARE, and ARE) prove the superiority of the proposed model.
compared to the well-established empirical formulas in the literature for predicting water/hydrocarbon IFT over different conditions of pressure and temperature. Finally, the leverage statistical algorithm was used to evaluate the experimental data of water/hydrocarbon IFT as well as applicability domain of proposed model. The agglomeration of most data points between the ranges of $0 < H < H^*$ and $-3 < R < 3$ show that the proposed model is statistically valid for prediction of used values.

Appendix A. Instruction for utilizing the model

In this paper, a computer program using MATLAB is utilized to develop the proposed model. First, the LSSVM toolbox for MATLAB should be installe. Then, the directory of the toolbox should be inserted as the main directory in the MATLAB environment. Next, the model.mat file is dragged and dropped in the MATLAB workspace and then the model can be utilized as it is described here. For example we need to calculate the IFT for the system with below specifications:

- Decane/water system
- $T = 531 \, ^\circ\text{R}$
- $\Delta_p = 0.26561 \, \text{g/cm}^3$
- $\rho_c$ of decane $= 1112 \, \text{g/cm}^3$

The IFT of mentioned system is calculated simply applying the below command in the MATLAB workspace as follows:

```matlab
IFT = svm([trainX', trainY', type, gam, sig2, 'RBF_SVM', 'preprocess'], [alpha, b], [0.26561, 531], [1112]);
```

This command make the result of $51.60 \, \text{mN/m}$ where as the measured value is $51.94$ which shows the absolute relative error of about 0.658%.

References


