Journal of Molecular Liquids xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Molecular Liquids

journal homepage: www.elsevier.com/locate/molliq

Toward a group contribution method for determination of speed of sound in saturated liquids

Farhad Gharagheizi [a], Poorandokht Ilani-Kashkouli [a], Amir H. Mohammadi [a,b,*], Deresh Ramjugernath [a,*]

[a] Thermodynamics Research Unit, School of Engineering, University of KwaZulu-Natal, Howard College Campus, King George V Avenue, Durban 4041, South Africa
[b] Institut de Recherche en Génie Chimique et Pétrolier (IRGCP), Paris Cedex, France

ARTICLE INFO

Article history:
Received 14 December 2013
Accepted 18 January 2014
Available online xxxx

Keywords:
Saturated liquid
Speed of sound
Group contribution
Support vector machine
ThermoData Engine

ABSTRACT

In this study, a group contribution model is presented for the estimation of the saturated liquid speed of sound of pure chemical compounds. A data set comprised of 1667 experimental data for 74 chemical compounds was extracted from the NIST ThermoData Engine and used to develop and test the model. The least squares support vector machine-group contribution (LSSVM-GC) model uses the occurrences of a set of 43 chemical substructures (to constitute a compound) in addition to temperature to represent/predict the saturated liquid speed of sound. The proposed model produces a low average absolute relative deviation (AARD) of less than 0.6% taking into consideration all 1667 experimental data values.

1. Introduction

Speed of sound is in an important property in thermodynamics. Like density, it can be measured with a high level of experimental accuracy, typically at least one order of magnitude greater than other thermodynamic quantities. Density has been conventionally used for the development of models for saturated pure liquids, however recently greater attention has been given to speed of sound measurements for the development of rigorous models capable of describing fluid properties over a wide range of temperatures and pressures. Quick and highly accurate measurement capabilities for the speed of sound make it a reliable quantity to measure in order to represent other thermodynamic properties with high precision. Almost all observable thermodynamic properties of a fluid phase can be directly obtained from the speed of sound by integration of partial differential equations which relate it to the other thermodynamic properties. This procedure of indirect determination of thermodynamic properties via speed of sound offers promising predictions over conventional direct approaches.

An important application of speed of sound is to determine the heat capacities of liquids along with $p-\rho-T$ data. This method may be a good alternative approach instead of the conventional calorimetric method. Furthermore, the heat-capacity ratio γ and the isentropic compressibility κ_s of pure liquids can be related to speed of sound. The latter offers a promising experimental route to measure these thermodynamic properties.

$$u^2 = \frac{1}{\rho\kappa_s} \tag{1}$$

$$u^2 = \frac{\gamma}{\rho\kappa_T} \tag{2}$$

where:

$$\kappa_S = \frac{1}{\rho}\left(\frac{\partial\rho}{\partial\rho}\right)_S \tag{3}$$

$$\kappa_T = \frac{1}{\rho}\left(\frac{\partial\rho}{\partial\rho}\right)_T \tag{4}$$

$$\gamma = \frac{C_p}{C_V}. \tag{5}$$

At high pressures where measuring $p-\rho-T$ data can generally be difficult, speed of sound measurements in liquids is probably of the greatest value [1].

Although there are some models that can be used just to estimate the speed of sound of single compounds [2–11], there is no general method for the estimation of the speed of sound of saturated liquids for a wide range of temperatures. A few available estimation approaches for speed of sound for a wide range of temperatures are restricted to mixtures and are based on equation of states [12].

* Corresponding authors.
E-mail addresses: a.h.m@irgcp.fr (A.H. Mohammadi), ramjuger@ukzn.ac.za (D. Ramjugernath).

0167-7322/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.molliq.2014.01.020

In this study a new group contribution method is proposed for the estimation of saturated liquid speed of sound of pure compounds. This model is based on the least squares support vector machine approach (LSSVM).

2. Methodology

2.1. Data preparation

In order to develop a reliable and accurate model, the use of a high quality data set is of great importance. For this purpose, the ThermoData Engine of the National Institute of Standard and Technology [13] was used in the present work to provide 1667 saturated liquid speed of sound data for 74 pure chemical compounds. The data are used to develop and test the model. Information about the name of the compounds and the original reference for each data point are presented as supplementary materials (see table mmc1).

As part of the method development, the database is divided into three parts: "training", "validation" and "test" sets. The "training" set is used to develop the model. The "validation" set is used to tune and optimize the model. The "test" set is used to check the capability of the model to predict speed of sound data which was not used in model derivation. In this study, K-means clustering is applied to select the training, the validation, and the test sets. K-means clustering is a method of cluster analysis, which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. According to a rule of thumb [14], nearly 80% (1335 data points), 10% (166 data points), and 10% (166 data points) of the database are assigned to the "training", the "validation", and the "test" sets, respectively.

2.2. Development of new group contributions

The first step in developing a group contribution model is to determine which set of chemical substructures can be most efficient in describing the desired property under consideration.

In the next step, it is required to select the most efficient subset of substructures which can describe the desired property with highest accuracy. It should be noted that great attention should be paid when choosing the most efficient subset of chemical substructures to avoid selecting those substructures that have no or low contribution to the desired property. The latter makes the problem of developing a group contribution model difficult. Therefore, there is a requirement to develop new strategies. Another obstacle in developing a group contribution model is to determine how the desired property can be related to the most efficient subset of chemical substructures. In most cases the process of introducing the most efficient subset of chemical substructures and the developing the model are considered as two separate steps:

1. Introducing the most efficient chemical substructures according to some rule of thumb, some basic findings about how to decompose a chemical structure to its basic units, or their effectiveness on other physical properties.
2. Determining the contribution of each chemical substructure on the desired property using a multivariate linear regression method (in some cases, it is possible to develop some nonlinear correlations based on the linear contribution of each chemical substructures.)

In order to avoid choosing irrelevant chemical substructures, a subset variable selection strategy was employed. Also, nonlinear regression methods were implemented to develop the group contribution model for cases in which multivariate linear regression methods give poor results.

As a result, a combination of sequential search algorithm as a subset selection technique and least squares support vector machine (LSSVM) as a non-linear regression method is implemented in this study.

The sequential search [15] algorithm uses the following steps to select the most efficient and optimum number of chemical substructures [15]:

I.) Each chemical substructure present in the model is excluded, one at a time, and is replaced by all the excluded chemical substructures, one at a time. For each replacement, the value of the statistical parameter which is optimized is calculated and stored.

II.) After all the substitutions are performed, if the best optimized parameter obtained during the replacement procedure is better than the previous value, the old model is substituted by the best new model. In this case, the procedure is iteratively repeated for the new model. Otherwise, if the best optimized parameter is worse than the previous value, the procedure ends and a new model among the initially selected models is taken into consideration.

Just after selection of the most efficient subset of chemical structures, the least squares method is used to develop a non-linear model. Although the ANN-based models have generally proven to provide high accuracy [16], they may have the disadvantages of non-reproducibility of results, partly as a result of random initialization of the networks and variation of the stopping criteria during optimization. The support vector machine (SVM) is a well-known strategy developed from the machine-learning community. The SVM is considered as a non-probabilistic binary linear classifier. The following criteria indicate most of the advantages of the SVM-based methods over the traditional methods based on the ANNs [17–19,14]:

1. Greater probability for convergence to the global optimum;
2. Normally find a solution that can be quickly obtained by a standard algorithm;
3. No need to determine the network topology in advance, which can be automatically determined as the training process ends;
4. Over-fitting complications are less probable in SVM schemes;
5. No requirements for choosing the number of hidden nodes;
6. Acceptable generalization performance;
7. Fewer adjustable parameters;
8. They require convex optimization procedures.
9. The model has only two parameters.

Support Vector Machines (SVM) has been preliminarily proposed for classification problems utilizing the hyperplanes to define decision boundaries between the experimental data points of different classes. In this regression technique, the experimental data points from the input space are mapped into a high dimensional or even infinite dimensional feasible region (margin) using a definite function, in which a maximal separating plane is constructed. The objective of a SVM calculation step is to find the optimum hyper-plane, from which the distance to all of the experimental data points is minimum.

For the set of data with the form of $x = \{(x_1, y_1),\ldots,(x_n, y_n)\}$, the SVM method uses the following relation to express the separating plane in the input space of the problem [17–19]:

$$f(x) = w^T x + b. \tag{6}$$

Subject to the following constraint when the data of two classes are separable:

$$\begin{cases} f(x_i) \geq 1 & \text{if } y_i = +1 \\ f(x_i) \leq -1 & \text{if } y_i = -1 \end{cases} \tag{7}$$

where, x denotes the input vector of parameters of the model; w represents the nonlinear function that maps the input space to a high-dimension feature space and performs linear regression; y denotes the outputs; b is the bias term; and superscript T denotes the transpose matrix. Support Vectors (SVs) are those points that satisfy the aforementioned constraints [18,19,14].

Please cite this article as: F. Gharagheizi, et al., Journal of Molecular Liquids (2014), http://dx.doi.org/10.1016/j.molliq.2014.01.020