Calphad 60 (2018) 68–80

Contents lists available at ScienceDirect

Calphad

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

Genetic programming based models for prediction of vapor-liquid equilibrium

Veena Patil-Shinde[a,b,c], Sanjeev S. Tambe[a,*]

[a] Chemical Engineering and Process Development Division, CSIR-National Chemical Laboratory, Dr. Homi Bhabha Road, Pune 411008, India
[b] Chemical Engineering Department, Bharati Vidyapeeth University College of Engineering (BVUCOE), Pune 411043, India
[c] Savitribai Phule Pune University (SPPU), Pune, India

ARTICLE INFO

Keywords:
VLE
Genetic programming
Activity coefficient models

ABSTRACT

The design, operation, and control of chemical separation processes heavily rely on the knowledge of the vapor-liquid equilibrium (VLE). Often, conducting experiments to gain an insight into the separation behavior becomes tedious and expensive. Thus, standard thermodynamic models are used in the VLE prediction. Sometimes, exclusively data-driven models are also used in VLE prediction although this method too possesses drawbacks such as a trial and error approach in specifying the data-fitting function. For overcoming these difficulties, this paper employs a machine learning (ML) formalism namely "genetic programming (GP)" possessing certain attractive features for the VLE prediction. Specifically, three case studies have been performed wherein GP-based models have been developed using experimental data, for predicting the vapor phase composition of a ternary, and a group of non–ideal binary systems. The inputs to models consists of three pure component attributes (acentric factor, critical temperature, and critical pressure), and as many intensive thermodynamic parameters (liquid phase composition, pressure, and temperature). A comparison of the VLE prediction and generalization performance of the GP-based models with the corresponding standard thermodynamic models reveals that the former class of models possess either superior or closely comparable performance vis-a-vis thermodynamic models. Noteworthy features of this study are: (i) a single GP-based model can predict VLE of a group of binary systems, and (ii) applicability of a GP-based model trained on an alcohol-acetate series data for its higher homolog. The VLE modeling approach exemplified here can be gainfully extended to other ternary and non-ideal binary systems, and for designing corresponding experiments in different pressure and temperature ranges.

1. Introduction

An accurate prediction of the phase behavior of chemical species and their mixtures is essential for designing, optimizing, and controlling separation, and other unit operations in the chemical industry. Prediction of phase equilibrium characteristics such as composition, and partition coefficients at temperatures and pressures of practical interest using reliable models offers an attractive alternative to costly and time consuming experimental measurements [1]. Phase equilibrium, and in particular vapor-liquid equilibrium (VLE), is important in a number of process engineering tasks. Designing an effective, efficient, and economical separation scheme is essential since a lack of knowledge of VLE poses significant difficulties in chemical process design and development. It has been broadly recognized that a viable separation scheme is as important as good chemistry for the success of chemical processes on a commercial scale [2].

Conducting VLE experiments and precise measurements of the corresponding data are often a tedious, time-consuming, and expensive proposition; for highly reactive systems, the stated task becomes even more difficult and complex. For instance, it is not always feasible to carry out VLE experiments at all ranges of temperatures and pressures of practical interest [3]. To overcome this difficulty, mathematical models are developed for the prediction of VLE.

There exist two principal methods, namely, *phenomenological* and *empirical,* for modeling VLE. The phenomenological (also termed "mechanistic" or "first principles") approach includes thermodynamic models such as, *equation of state*, and *activity coefficient* models [4]. This approach needs a complete knowledge of the physico-chemical phenomena underlying the VLE. The prediction of VLE data by the conventional thermodynamic methods is tedious since it involves determination of various thermodynamic parameters, which is arbitrary in many ways, and in some cases also introduces significant inaccuracies [5]. For some of the components, determination of thermodynamic parameters such as *binary interaction parameter* (BIP) by

* Corresponding author.
E-mail address: ss.tambe@ncl.res.in (S.S. Tambe).

https://doi.org/10.1016/j.calphad.2017.11.002
Received 4 August 2017; Received in revised form 30 October 2017; Accepted 15 November 2017

itself can be an elaborate and time consuming exercise.

The second (i.e. empirical) approach to VLE modeling is exclusively data-driven and, therefore, can be employed in the absence of the detailed knowledge of the underlying physico-chemical phenomenon. It utilizes linear/nonlinear regression methods in formulating the models. A significant requirement of this approach is that the exact structure (form) of the linear/nonlinear data-fitting model needs to be specified unambiguously prior to the estimation of the unknown model parameters [6]. In the case of ideal systems exhibiting a linear VLE behavior, specifying the corresponding linear data-fitting function is relatively easy. However, VLE behavior of a large number of systems exhibits a nonlinear dependence on the operating parameters. In such cases, choosing an appropriate nonlinear data-fitting model, from numerous competing ones, becomes a daunting task. The above-stated difficulties in respect of both the thermodynamic and regression-based VLE modeling, necessitates exploration of alternative nonlinear modeling strategies.

The two machine learning (ML) based exclusively data-driven nonlinear modeling formalisms, namely, *Artificial neural networks* (ANNs) (see e.g., Bishop [7], Zurada [8], and Tambe et al. [9]), and *support vector regression* (SVR) (Vapnik [10], and Zaid [11]) are often used as alternatives to the regression based modeling. These methods have found numerous applications in the field of thermodynamics and prediction of transport properties. Table 1 reports a number of studies wherein ANNs and SVR have been employed in VLE predictions.

In addition to ANNs and SVR, the field of machine learning (ML) comprises a purely data-driven modeling strategy, namely *genetic programming* (GP). The GP technique has been used, for instance, in estimation of solvent activity in polymer solutions [12], prediction of °API gravity of crude oils [13], process identification [14], gasification performance prediction [15], and prediction of Kovats retention indices [16]. Since it possesses several attractive characteristics, in this study, the GP formalism has been utilized for developing data-driven models predicting the vapor phase composition of industrially relevant ternary, and a group of binary mixtures. An exhaustive literature search indicates that this is the first instance wherein GP has been used in VLE prediction. The three specific VLE modeling case studies performed here are listed in Table 2. In all, four GP-based models have been developed; the inputs and outputs pertaining to these models are given in Table 3.

In this study, all four GP-based models were developed using experimental VLE data reported in DECHEMA data series [30,31] and the same data were used in developing the corresponding standard activity coefficient models. These are, Wilson [32,33], Van Laar [34], Non-random two-liquid (NRTL) [35], and Universal Quasi-chemical (UNIQUAC) [36] models. The prediction and generalization performance of the GP-based models was rigorously compared with that of the above-stated thermodynamic models. The results of this comparison indicate that the GP-based models possess comparable or better VLE prediction ability vis-a-vis conventional thermodynamic models. In general, the GP models are of lower complexity and, therefore, convenient to deploy in a practical setting.

The remainder of this paper is structured as follows. Various thermodynamic models available for the VLE prediction are described briefly in Section 2 titled "Phase equilibria modeling." Next, in Section 3 titled "ML-based modeling formalism" an overview of the GP method is presented. Section 4 provides details of the data used in the GP-based modeling of vapor phase composition. Next, in Section 5, titled "Result and discussion," three case studies are presented wherein GP-based models have been developed for (i) ternary system (Section 5.2), (ii) a group of three non-ideal binary systems (Section 5.3), and (iii) a group of three non-ideal binary systems (Section 5.4). Additionally, Section 5 also provides results of a comparison of the prediction and generalization performance of the four GP–based models with their thermodynamic counterparts as also the fine-tuned genetic programming-Marquardt (GP–Marquardt) models. Finally, "Conclusion" section summarizes the principal findings of the study.

2. Phase equilibria modeling

2.1. Activity coefficient models

A number of methods such as, regular solution theory, universal functional activity coefficient (UNIFAC) [37], and analytical solution of groups (ASOG) [38] is available for the VLE prediction; however, none of these strategies can be regarded as an accurate predictor [33]. Thus, the stated methods are used only when no experimental VLE data are available for the system of interest. The notable features of thermodynamic VLE models are given below [33,39]:

- For moderately non-ideal systems, all major models (Van Laar, two constant Margules, Wilson, UNIQUAC, and NRTL) perform comparably well.
- For mixtures of very different species, such as, polar or related compounds (e.g. alcohols and other oxy hydrocarbons), the two-parameter VLE models, namely, Van Laar, two constant Margules, UNIQUAC, and Wilson equation, are preferred over the three parameter NRTL equation.
- For non-polar solvents (e.g. hydrocarbons), the Wilson, UNIQUAC, and NRTL models have been found to make superior predictions than the Van Laar and two-parameter Margules equations.
- The NRTL and UNIQUAC equations are useful whereas the Wilson equation is inapplicable for species that are dissimilar and are only partially soluble to form two liquid phases.

2.2. Equation of state models

A commonly employed method for predicting thermodynamic properties of fluids, mixtures of fluids, and solids is "*equation of state* (EoS)". It is an efficient tool for calculating also the phase equilibrium of systems in pure or mixture form. The EoS is widely used in theoretical and practical studies involving chemical process design, petroleum industry, reservoir fluids, etc. The Van der Waals equation of state [40] was the first equation to predict vapor-liquid coexistence. Later, the Redlich-Kwong equation of state [41] improved the accuracy of the Van der Waals equation by proposing temperature dependence for the attractive term. Soave [42], and Peng and Robinson [43] proposed additional modifications to the Redlich-Kwong equation for improving the accuracy of prediction of vapor pressure, liquid density, and equilibrium ratios. Numerous equations of state have been proposed in the literature with an empirical, semi-empirical, or theoretical basis. Some notable comprehensive reviews on EOS are by Martin [44], Anderko [45], and Sengers et al. [46].

Since design, operation, and control of a large number of industrial chemical processes are based on the predictions of VLE, and thermodynamic property models, it is at most important that they are simple to use, robust, require minimum number of inputs, and capable of accurate predictions.

3. ML-based modeling formalism

Genetic programming used in this study for VLE modeling is a machine learning formalism belonging to a class of methods termed "Evolutionary algorithms." A brief overview of ML and Evolutionary algorithms is presented below.

Machine learning (ML): It is the development and study of computer algorithms, which are capable of learning for improving their own performance in doing a task on the basis of their previous experience [47]. ML comprises algorithms from various computer science subfields such as artificial/computational intelligence, and evolutionary algorithms. An important application of ML algorithms is function approximation, which is of interest to this study. It allows computers to