ELSEVIER

Journal of Molecular Structure: THEOCHEM 719 (2005) 119–127

THEO CHEM

www.elsevier.com/locate/theochem

Support vector regression based QSPR for the prediction of some physicochemical properties of alkyl benzenes

Shansheng Yang[a], Wencong Lu[a,*], Nianyi Chen[a], Qiannan Hu[b]

[a]*Department of Chemistry, College of Science, Shanghai University, Shanghai 200436, China*
[b]*Research Center of Modernization of Chinese Herb Medicine, College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, China*

Received 11 July 2004; accepted 7 October 2004

Abstract

Physicochemical properties of alkyl benzenes are essential to separate pure component from alkyl benzene mixture. Support vector regression (SVR), a novel powerful machine learning technology based on statistical learning theory (SLT), integrated with topological indices was applied to the prediction of five physicochemical properties of alkyl benzenes including the normal boiling point (bp), enthalpy of vaporization at the boiling point (H_{vb}), critical temperature (T_c), critical pressure (P_c), and critical volume (V_c). In a benchmark test, SVR models for bp, H_{vb}, T_c, P_c, and V_c were compared with several modeling techniques currently used in this field. The prediction accuracy of the model was discussed on the basis of the leave-one-out cross-validation. The results show that the prediction accuracy of SVR model was higher than those of back propagation artificial neural network (BP ANN) and partial least squares (PLS) methods.

Keywords: Support vector regression; Topological index; Quantitative structure-property relationship; Alkyl benzene

1. Introduction

Physicochemical properties of pure components and mixtures are closely related to the separation process principles, since differences in the values of properties among the components of the mixture to be separated are exploited by the separation techniques. A large quantity of physicochemical property values is available in literatures and data banks. However, only a few compounds related to fine chemistry and often only one method for the estimation of each physicochemical property are available in some data banks. The approach known as QSPR (quantitative structure-property relationship) can be used to solve the problem of the prediction of physicochemical property of the special compound.

QSPR model employs chemometrics methods (quantitative regression methods) to correlate various physicochemical properties of compounds with their molecular descriptors derived from molecular structure alone. Some linear regression methods such as multiple linear regression (MLR) and partial least squares (PLS) regression are commonly used in QSPR analysis [1–3]. It has been reported that artificial neural network (ANN) was successfully used in QSPR [4,5]. However, ANN may give rise to overfitting problems (i.e. may lead to good performance in fitting but poor performance in prediction) in treating finite, multivariate data set.

Recently, Vapnik and his coworkers have worked out a new theory called statistical learning theory (SLT) and a new computational method called support vector machine (SVM) [6,7]. It has been shown that SVM has two distinct features. First, it is often associated to the physical meaning of the data and hence, it is easy to interpret. Second, it requires only small amount of training samples. According to some literatures, SVM has successfully applied to many topics, such as drug design [8], combinatorial chemistry [9], and prediction of protein structure [10].

* Corresponding author. Tel.: +86 21 6613 3513; fax: +86 21 6613 4275.
E-mail address: wclu@mail.shu.edu.cn (W. Lu).

0166-1280/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.theochem.2004.10.060

The goal of this work was to establish QSPR models of the normal boiling point (bp), enthalpy of vaporization at the boiling point (H_{vb}), critical temperature (T_c), critical pressure (P_c), and critical volume (V_c) of alkyl benzenes based on support vector regression (SVR) that can be used to predict the physicochemical properties from their molecular descriptors and to show the flexible modeling ability of SVR.

2. Methodology

2.1. Support vector regression theory

SVM can be applied to regression by the introduction of an alternative loss function and the results appear to be very encouraging. In SVR, the basic idea is to map the data X into a higher-dimensional feature space F via a nonlinear mapping Φ and then to do linear regression in this space. Therefore, regression approximation addresses the problem of estimating a function based on a given data set $G = \{(\mathbf{x}_i; d_i)\}_{i=1}^{l}$ ($\mathbf{x}_i$ is input vector, d_i is the desired value). SVM approximates the function in the following form

$$y = \sum_{i=1}^{l} w_i \Phi_i(\mathbf{x}) + b \tag{1}$$

where $\{\Phi_i(\mathbf{x})\}_{i=1}^{l}$ is the set of mappings of input features, and $\{w_i\}_{i=1}^{l}$ and b are coefficients. They are estimated by minimizing the regularized risk function $R(C)$

$$R(C) = C\frac{1}{N}\sum_{i=1}^{N} L_\varepsilon(d_i, y_i) + \frac{1}{2}||\mathbf{w}||^2 \tag{2}$$

where

$$L_\varepsilon(d, y) = \begin{cases} |d - y| - \varepsilon & \text{for } |d - y| \geq \varepsilon \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

and ε is a prescribed parameter.

In Eq. (2), $C(1/N)\sum_i^N L_\varepsilon(d_i y_i)$ is the so-called empirical error (risk), which is measured by ε-insensitive loss function $L_\varepsilon(d, y)$, which indicates that it does not penalize errors below ε. The second term, $(1/2)||\mathbf{w}||^2$, is used as a measurement of function flatness. C is a regularized constant determining the tradeoff between the training error and the model flatness. Introduction of slack variables 'ξ' leads Eq. (2) to the following constrained function

$$\text{Max } R(\mathbf{w}, \xi^*) = \frac{1}{2}||\mathbf{w}||^2 + C^* \sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{4}$$

$$\text{s.t. } w\Phi(x_i) + b - d_i \leq \varepsilon + \xi_i, \quad d_i - w\Phi(x_i) - b_i \leq \varepsilon + \xi_i \quad \xi, \xi^i \geq 0 \tag{5}$$

Thus, decision function Eq. (1) becomes the following form

$$f(\mathbf{x}, \alpha_i, \alpha_i^*) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)K(\mathbf{x}, \mathbf{x}_i) + b \tag{6}$$

In Eq. (6), α_i,α_i* are the introduced Lagrange multipliers. They satisfy the equality $\alpha_i\alpha_i^* = 0$, $\alpha_i \geq 0$, $\alpha_i^* \geq 0$; $i = 1,\ldots,l$ and are obtained by maximizing the dual form of Eq. (4), which has the following form

$$\Phi(\alpha_i, \alpha_i^*) = \sum_{i=1}^{l} d_i(\alpha_i - \alpha_i^*) - \varepsilon\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\alpha_i, \alpha_j) \tag{7}$$

with the following constrains

$$0 \leq a_i \leq C \quad i = 1, \ldots, l \; 0 \leq a_i^* \leq C$$
$$i = 1, \ldots, l \sum_{i=1}^{l}(a_i - a_i^*) = 0 \tag{8}$$

Based on the Karush-Kuhn-Tucker (KKT) conditions of quadratic programming, only a number of coefficients ($\alpha - \alpha_i^*$) will assume nonzero values, and the data points associated with them could be referred to as support vectors.

In Eq. (6), $K(\mathbf{x}_i,\mathbf{x}_j)$ is the kernel function. The value is equal to the inner product of two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ in the feature space $\Phi(\mathbf{x})$. That is, $K(\mathbf{x}_i,\mathbf{x}_j) = \Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j)$. The elegance of using kernel function lied in the fact that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\Phi(\mathbf{x})$ explicitly. Any function that satisfies Mercer's condition can be used as the kernel function.

2.2. Implementation of SVR

The SVM software package including SVR was programmed referring to the literature [6]. The validation of the software has been tested in some applications in chemistry and chemical engineering [11–13]. All the computations were carried out on a Pentium IV computer with a 2.0 GHz processor.

3. Results and discussion

3.1. Data set

Table 1 lists a set of 47 alkyl benzenes studied and their experimental property values including the normal boiling point (bp), enthalpy of vaporization at the boiling point (H_{vb}), critical temperature (T_c), critical pressure (P_c), and critical volume (V_c), which are taken from the literatures [14–16].