QSAR Study of Capacity Factors by Quantum Chemical Descriptors and Using PLS and LS-SVM Methods

S. Jameh-Bozorgi*, A. Niazi
Chemistry Department, Arak Branch, Islamic Azad University, Arak, Iran

Abstract

Introduction: quantitative structure-activity relationship (QSAR) study is one of best chemometrics methods for prediction of chemical and biological properties of various compounds.

Aim: A quantitative structure-activity relationship (QSAR) study is suggested for the prediction of capacity factors ($\log k'$) of 25 substances as solutes to two different stationary phases (polyethylene–silica and polyethylene–alumina) were analyzed to their quantum chemical descriptors and related to their retention behavior as expressed by the logarithms of their capacity factors ($\log k'$).

Material and Method: Ab initio theory was used to calculate some quantum chemical descriptors including electrostatic potentials and local charges at each atom, HOMO and LUMO energies, etc. Modeling of the log $k'$ as a function of molecular structures was established by means of the partial least squares (PLS) and least squares support vector machines (LS-SVM). These models were applied for the prediction of the capacity factors, which were not in the modeling procedure.

Results: The resulted models showed high prediction ability with root mean square error of prediction of 6.5621 and 0.4960 for PLS and LS-SVM, respectively.

Conclusion: this research showed that LS-SVM method has a very good ability for prediction of capacity factors ($\log k'$) for benzene derivatives.

Keywords: Capacity factor; QSAR; Ab initio; PLS; LS-SVM.

Introduction

The main aim of QSAR studies is to establish an empirical rule or function relating the structural descriptors of compounds under investigation to bioactivities. This rule of function is then utilized to predict the same bioactivities of the compounds not involved in the training set from their structural descriptors. Whether the bioactivities can be predicted with satisfactory accuracy depends to a great extent on the performance of the applied multivariate data analysis method, provided the property being predicted is related to the descriptors.

Among the investigation of QSAR, one of the most important factors affecting the quality of the model is the method to build the model. Many multivariate data analysis methods such as multiple linear regression (MLR), partial least squares (PLS)\cite{1-3} and artificial neural network (ANN) \cite{4} have been used in QSAR studies. MLR, as most commonly used chemometrics method, has been extensively applied to QSAR investigations. However, the practical usefulness of MLR in QSAR studies is rather limited, as it provides relatively poor

*Corresponding author
accuracy. ANN offers satisfactory accuracy in most cases but tends to overfit the training data. The support vector machine (SVM) is a popular algorithm developed from the machine learning community. Due to its advantages and remarkable generalization performance over other methods, SVM has attracted attention and gained extensive applications.[5-7] As a simplification of traditional of SVM, Suykens and Vandewalle[5] have proposed the use of least-squares SVM (LS-SVM). LS-SVM encompasses similar advantages as SVM, but its additional advantage is that it requires solving a set of only linear equations (linear programming), which is much easier and computationally more simple.

Theory

Theory of LS-SVM has also been described clearly by Suykens and J. Vandewalle[5] and application of LS-SVM in quantification[7-9] reported by some of the workers. So, we will only briefly describe the theory of LS-SVM. The LS-SVM[5] is capable of dealing with linear and nonlinear multivariate calibration and resolves multivariate calibration problems in a relatively a fast way. In LS-SVM a linear estimation is done in kernel-induced feature space \((y = \mathbf{w}^T \phi(x) + b)\). As in SVM, it is necessary to minimize a cost function \((C)\) containing a penalized regression error, as follow:

\[
C = \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{1}{2} \gamma \sum_{i=1}^{N} e_i^2
\]

such that:

\[
y_i = \mathbf{w}^T \phi(x_i) + b + e_i
\]

for all \(i = 1,\ldots,N\), where \(\phi\) denotes the feature map.

The first part of this cost function is a weight decay which is used to regularize weight sizes and penalize large weights. Due to this regularization, the weights converge to similar value. Large weights deteriorate the generalization ability of the LS-SVM because they can cause excessive variance. The second part of Eq. (1) is the regression error for all training data. The parameter \(\gamma\), which has to be optimized by the user, gives the relative weight of this part as compared to the first part. The restriction supplied by Eq. (2) gives the definition of the regression error. Analyzing Eq. (1) and its restriction given by Eq. (2), it is possible to conclude that we have a typical problem of convex optimization[6] which can be solved by using the Lagrange multipliers method, as follow:

\[
L = \frac{1}{2} \|\mathbf{w}\|^2 + \gamma \sum_{i=1}^{N} e_i^2 - \sum_{i=1}^{N} \alpha \{\mathbf{w}^T \phi(x_i) + b + e_i - y_i\}
\]

where

\[
y_i = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \quad e_i = \begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix} \quad \text{and} \quad \alpha_i = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix}
\]

Obtaining the optimum, that is, carrying out \(\partial L(\mathbf{w}, b, e_i, \alpha_i)/\partial \mathbf{w}, \partial L(\mathbf{w}, b, e_i, \alpha_i)/\partial b, \partial L(\mathbf{w}, b, e_i, \alpha_i)/\partial e_i, \partial L(\mathbf{w}, b, e_i, \alpha_i)/\partial \alpha_i\), and setting all partial first derivatives to zero, it obtains the weights that are linear combinations of the training data are obtained:

\[
\frac{\partial L(\mathbf{w}, b, e_i, \alpha)}{\partial \mathbf{w}} = w - \sum_{i=1}^{N} \alpha_i \phi(x_i) = 0 \Rightarrow w = \sum_{i=1}^{N} \alpha_i \phi(x_i)
\]
then:
\[
\begin{align*}
    w &= \sum_{i=1}^{N} \alpha_i \phi(x_i) + \sum_{i=1}^{N} y_i \phi(x_i)
\end{align*}
\]
where a positive definite kernel is used as follows:
\[
K(x_i, x_j) = \phi(x_i)^T \phi(x_j)
\]

An important result of this approach is that the weights \(w\) can be written as linear combinations of the Lagrange multipliers with the corresponding data training \(x_i\). Putting the result of Eq. (6) into the original regression line \(y = w^T \phi(x) + b\), the following result is obtained:
\[
\begin{align*}
y &= \sum_{i=1}^{N} \alpha_i \phi(x_i)^T \phi(x) + b = \sum_{i=1}^{N} \alpha_i \phi(x_i)^T \phi(x) + b
\end{align*}
\]
for a point \(y_i\) to be evaluated it is:
\[
\begin{align*}
y_i &= \sum_{i=1}^{N} \alpha_i \phi(x_i)^T \phi(x_j) + b = \sum_{i=1}^{N} \alpha_i \phi(x_i) \phi(x_j) + b
\end{align*}
\]
The \(\alpha\) vector follows from solving a set of linear equation:
\[
M \begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix}
\]
where \(M\) is a square matrix given by:
\[
M = \begin{bmatrix} K + \frac{I}{\gamma} & 1_N \\ 1_N^T & 0 \end{bmatrix}
\]
where \(K\) denotes the kernel matrix with \(ijth\) element \(K = (x_i, x_j) = \phi(x_i)^T \phi(x_j)\) and \(I\) denotes the identity matrix \(N \times N\), \(1_N = [1 \quad \Lambda \quad 1]^T\). Hence, the solution is given by:
\[
M^{-1} \begin{bmatrix} y \\ 0 \end{bmatrix}
\]
As can be seen from Eqs. (11) and (12), usually all Lagrange multipliers (the support vectors) are nonzero, which means that all training objects contribute to the solution. In contrast with standard SVM the LS-SVM solution is usually not sparse. However, as described by Suykens and J. Vandewalle\(^{[5]}\) a sparse solution can be easily achieved via pruning or reduction techniques. Depending on the number of training data set either direct solvers can be used or iterative solve such as conjugate gradients methods (for large data sets), in both cases with numerically reliable methods.
In applications involving nonlinear regression it is enough to change the inner product \( \langle \phi(x_i), \phi(x_j) \rangle \) of Eq. (9) by a kernel function and the \( ij \)th element of matrix \( K \) equals \( K_{ij} = \phi(x_i)^T \phi(x_j) \). If this kernel function meets Mercer’s condition the kernel implicitly determines both a nonlinear mapping, \( x \rightarrow \phi(x) \) and the corresponding inner product \( \phi(x_i)^T \phi(x_j) \). This leads to the following nonlinear regression function:

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

for a point \( x_j \) to be evaluated it is:

\[
y_j = \sum_{i} \alpha_i K(x_i, x_j) + b
\]

The attainment of the kernel function is cumbersome and it will depend on each case. However, the kernel function more used is the radial basis function (RBF), \( \exp(-\|x_i - x_j\|^2 / 2\sigma^2) \), a simple Gaussian function, and polynomial functions \( \langle x_i, x_j \rangle^d \), where \( \sigma^2 \) is the width of the Gaussian function and \( d \) is the polynomial degree, which should be optimized by the user, to obtain the support vector. For \( \alpha \) of the RBF kernel and \( d \) of the polynomial kernel it should be stressed that it is very important to do a careful model selection of the tuning parameters, in combination with the regularization constant \( \gamma \), in order to achieve a good generalization model.

In the present paper, the LS-SVM method was applied in QSAR for modeling the relationship between the capacity factors of 25 substances as solutes to two different stationary phases (polyethylene–silica and polyethylene–alumina) were analyzed to their quantum chemical descriptors and related to their retention behavior as expressed by the logarithms of their capacity factors (log \( k' \)). Ab initio geometry optimization was performed at the B3LYP level, with a known basis set, 6-31++G**. Local charges, electrostatic potential, dipole moment, polarizability, HOMO-LUMO energies, hardness, softness, electronegativity and electrophilicity were calculated for each compound.

Materials and methods

Hardware and software

The computations were accomplished with microcomputer (CPU 3.0 GHz and RAM 1 Gb) with the Windows XP operating system and with Matlab (version 6.5, Mathwork, Inc.). The PLS evaluations were carried out by using the PLS program from PLS-Toolbox Version 2.0 for use with Matlab from Eigenvector Research Inc. The LS-SVM optimization and model results were obtained using the LS-SVM lab toolbox (Matlab/C Toolbox for Least-Squares Support Vector Machines).\(^2\) Hyperchem (version 6.03, Hyperchem, Inc.) and Gaussian 98 software\(^{10}\) were used for geometric optimization of the molecules and calculation of the quantum chemical descriptor.

Data set

The capacity factors (log \( k' \)) were reported by Y.L. Loukas.\(^{11}\) The structures of 25 substances as solutes to two different stationary phases (polyethylene–silica and polyethylene–alumina) and observed log \( k' \) are listed in Table 1. We randomly divided the 25 substance into two subsets, a training set of 20 compounds and a test set of 5 compounds.
Table 1 - Structures and observed log \( k' \) values of the examined solutes in two stationary phases.

<table>
<thead>
<tr>
<th>Structure</th>
<th>log ( k' )</th>
<th>Structure</th>
<th>log ( k' )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hexylbenzene</td>
<td>4.560</td>
<td>Dibenzothiophene</td>
<td>3.041</td>
</tr>
<tr>
<td>1,3,5-Tris(1-methylethyl)benzene</td>
<td>4.887</td>
<td>Phenol</td>
<td>0.099</td>
</tr>
<tr>
<td>1,4-Dinitrobenzene</td>
<td>0.969</td>
<td>1,1,2,3,4,4-Hexachloro-1,3-butadiene</td>
<td>3.248</td>
</tr>
<tr>
<td>3-(Trifluoromethyl)phenol</td>
<td>0.975</td>
<td>1H-Indazole</td>
<td>0.822</td>
</tr>
<tr>
<td>3,5-Dichlorophenol</td>
<td>1.502</td>
<td>3,7-Dihydro-1,3,7-trimethyl-1-H-purine-2,6-dione</td>
<td>1.616</td>
</tr>
<tr>
<td>4-Hydroxybenzonitrile</td>
<td>0.396</td>
<td>4-Nitrobenzoic acid</td>
<td>-0.899</td>
</tr>
<tr>
<td>4-Iodophenol</td>
<td>1.174</td>
<td>1-Methyl-2-pyrrolidone</td>
<td>0.257</td>
</tr>
<tr>
<td>Methoxybenzene</td>
<td>0.835</td>
<td>Naphthalene</td>
<td>1.769</td>
</tr>
<tr>
<td>Benzamide</td>
<td>0.303</td>
<td>4-Chlorophenol</td>
<td>0.758</td>
</tr>
<tr>
<td>Benzene</td>
<td>0.584</td>
<td>Methylbenzene</td>
<td>1.027</td>
</tr>
<tr>
<td>Chlorobenzene</td>
<td>1.129</td>
<td>Piperazine</td>
<td>0.797</td>
</tr>
<tr>
<td>Cyclohexanone</td>
<td>0.337</td>
<td>Piperidine</td>
<td>0.574</td>
</tr>
<tr>
<td>Benzonitrile</td>
<td>0.705</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Quantum chemical descriptors calculation

The molecular structures of all the compounds (Table 1) were built with Hyperchem software for structural chemistry. Gaussian 98\(^\text{[10]}\) was operated to optimize with the 6-31++G** basis set for all atoms at the B3LYP level. No molecular symmetry constraint was applied; instead, full optimization of all bond lengths and angles was carried out at the B3LYP/6-31++G** level. The calculated descriptors for each molecule are summarized in Table 2. Local charges (LC) and electrostatic potential (EP) at each atom, highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies, molecular polarizabilities (MP) and molecular dipole moment (MDP) were calculated by Gaussian 98. Quantum chemical indices of hardness (\( \eta \)), softness (\( S \)), electronegativity (\( \chi \)), chemical potential (\( \mu \)) and electrophilicity (\( \omega \)).

Table 2 - The calculated quantum chemical descriptors used in this study.

<table>
<thead>
<tr>
<th>Descriptor name</th>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local charges</td>
<td>( LC_i )</td>
<td>The local charges at each atom</td>
</tr>
<tr>
<td>Electrostatic potential</td>
<td>( EP_i )</td>
<td>The electrostatic potential at each atom</td>
</tr>
<tr>
<td>Molecular polarizability</td>
<td>( MP )</td>
<td>Total molecular polarizability</td>
</tr>
<tr>
<td>Dipole moment</td>
<td>( DM )</td>
<td>Total molecular dipole moment</td>
</tr>
<tr>
<td>HOMO</td>
<td>( E_{HOMO} )</td>
<td>Highest occupied molecular orbital energy</td>
</tr>
<tr>
<td>LUMO</td>
<td>( E_{LUMO} )</td>
<td>Lowest unoccupied molecular orbital energy</td>
</tr>
<tr>
<td>Electronegativity</td>
<td>( \chi )</td>
<td>( -0.5 \cdot (E_{HOMO} - E_{LUMO}) )</td>
</tr>
<tr>
<td>Hardness</td>
<td>( \eta )</td>
<td>( 0.5 \cdot (E_{HOMO} + E_{LUMO}) )</td>
</tr>
<tr>
<td>Softness</td>
<td>( S )</td>
<td>( 1/\eta )</td>
</tr>
<tr>
<td>Electrophilicity</td>
<td>( \omega )</td>
<td>( \chi^2/2\eta )</td>
</tr>
</tbody>
</table>

Results and discussion

The calculated descriptors for each molecule are summarized in Table 1. The capacity factors of 25 specified compounds were randomly classified into a training set (20 data) and a prediction set (5 data). The data were centered to zero means and scaled to the unit variance. For the evaluation of the predictive ability of a different model, the root mean square error of prediction (RMSEP) and relative standard error of prediction (RSEP) can be used:
where $y_{\text{pred}}$ is the predicted log $k'$ using different model, $y_{\text{obs}}$ is the observed value of the log $k'$ and $n$ is the number of samples in the prediction set.

Modeling by PLS and LS-SVM

The multivariate calibration is a power tool for modeling, because it extracts more information from the data and allows building more robust models. \cite{11,12} According to capacity factors data (Table 1), data randomly classified to training and prediction sets. The PLS model was run. The optimum latent variable to be included in the calibration model was determined by computing the PRESS by cross-validated models.\cite{13} According to Haaland suggestion, \cite{13} the optimum number of factor selected. In Fig. 1, the PRESS obtained by optimized the training set of the descriptor data with PLS model is shown.

LS-SVM was performed with radial basis function (RBF) as a kernel functions. In the model development using LS-SVM and RBF kernel, $\gamma$ and $\sigma^2$ parameters were a manageable task, similar to the process employed to select the number of factors for PLS models, but in this case for a two-dimensional problem. For each combination of $\gamma$ and $\sigma^2$ parameters, root mean square error of cross-validation (RMSECV) was calculated and the optimum parameters were selected produced the smaller RMSECV. In Table 3 is shown the optimum latent variables, $\gamma$ and $\sigma^2$ parameters for the PLS and LS-SVM, using the calibration sets for 20 capacity factors data.

Prediction of log $k'$

The predictive ability of these methods (PLS and LS-SVM) were determined using 5 capacity factors data (their structure are given in Table 1). The results obtained by PLS and LS-SVM methods are listed in Table 3. Table 3 also shows RMSEP, RSEP and the percentage error for prediction of log $k'$. As can be seen, the percentage error was also quite acceptable for LS-SVM. Good results were achieved in LS-SVM model with percentage error ranges from -0.85 to 0.75 % for log $k'$. 
Table 3- Actual and predicted of log $k'$ values using PLS and LS-SVM models.

<table>
<thead>
<tr>
<th>Substance</th>
<th>Actual log $k'$</th>
<th>Predicted log $k'$</th>
<th>PLS Error (%)</th>
<th>LS-SVM Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-Iodophenol</td>
<td>1.174</td>
<td>1.089</td>
<td>7.24</td>
<td>1.172</td>
</tr>
<tr>
<td>Piperazine</td>
<td>0.797</td>
<td>0.723</td>
<td>9.28</td>
<td>0.791</td>
</tr>
<tr>
<td>Benzonitrile</td>
<td>0.705</td>
<td>0.735</td>
<td>-4.25</td>
<td>0.711</td>
</tr>
<tr>
<td>1,4-Dinitrobenzene</td>
<td>0.969</td>
<td>1.051</td>
<td>-8.46</td>
<td>0.971</td>
</tr>
<tr>
<td>Naphthalene</td>
<td>1.769</td>
<td>1.689</td>
<td>4.52</td>
<td>1.760</td>
</tr>
<tr>
<td>LV*</td>
<td></td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRESS</td>
<td></td>
<td>0.0761</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\gamma$</td>
<td></td>
<td>90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td></td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSEP</td>
<td>0.0731</td>
<td>0.0057</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RSEP (%)</td>
<td>6.5621</td>
<td>0.4960</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The plots of the predicted log $k'$ versus actual values are shown in Fig. 2 for each model (line equations and $R^2$ values are also shown). The correlation coefficients ($R^2$) for LS-SVM model were better than other models and close to one. Also, it is possible see that LS-SVM presents excellent prediction abilities when compared with other regression.

Fig 2 - Plots of predicted versus actual log $k'$ PLS and LS-SVM.

Conclusion

Chemometrics methods such as, PLS and LS-SVM were established to predict the log $k'$ of some substances as solutes to two different stationary phases (polyethylene–silica and polyethylene–alumina) were analyzed to their quantum chemical descriptors and related to their retention behavior as expressed by the logarithms of their capacity factors (log $k'$). The quantum chemical descriptors concerning all the molecular properties and those of individual atoms in the molecule were found to be important factors controlling the log $k'$. In this study, the results obtained by LS-SVM are compared with results obtained by PLS. The results show LS-SVM is more powerful in prediction of log $k'$ than PLS.
References: