IAU-ARAK

J. Iran. Chem. Res. 4 (2011) 287-290

Journal of the Iranian Chemical Research

www.iau-jicr.com

QSAR study of retention index of different alkanes and alkenes using different chemometrics methods

Mehrana Motiee *

Young Researchers Club, Islamic Azad University, Arak Branch, Arak, Iran

Received 6 May 2011; received in revised form 25 August 2011; accepted 10 September 2011

Abstract

An important property that has been extensively studied in quantitative structure activity relationship (QSAR) is the chromatographic retention index. QSAR study is suggested for the prediction of retention index of alkanes and alkenes compounds. Modeling of the retention index of alkanes and alkenes compounds as a function of molecular structures was established by different chemometrics methods. These models were applied for the prediction of the retention index of these compounds, which were not in the modeling procedure. In the present study, the PLS and LS-SVM methods were applied in QSAR for modeling the relationship between the retention index 179 alkanes and alkenes compounds by using structural molecular descriptors.

Keywords: QSAR; Retention index; Alkane; Alkene; PLS; LS-SVM.

1. Introduction

Among the investigation of QSAR/QSPR, 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) and artificial neural network (ANN) have been used in QSAR studies [1-3]. 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 [2-6]. In the present study, the PLS and LS-SVM methods were applied in QSPR for modeling the relationship between the retention index of 179 alkanes and alkenes compounds.

2. Materials and computational methods

The retention indices of 179 alkanes and alkenes were obtained from the literature [7-10]. The QSRR model for the estimation of the retention indices of various alkanes and alkenes compounds is established in the following steps: the molecular structure input and generation of the files containing the chemical structures is stored in a computer-readable format; quantum mechanics geometry is optimized with a semi-empirical (AM1) method; structural descriptors

* Corresponding author. Tel. & fax: +98 861 3670017.
E-mail address: mehrana.motiee@yahoo.com

are computed; and the structural-retention index model is generated by the chemometrics methods and statistical analysis.

2.1. Computer hardware and software

The computations were made with an Intel 3.0 (1 Gb RAM) microcomputer 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 from Eigenvector Research Inc. The LS-SVM optimization and model results were obtained using the LS-SVM lab Toolbox. ChemDraw Ultra version 9.0 (Chem Office 2005, Cambridge Soft Corporation) software was used to draw the molecular structures and optimization by the AM1. Descriptors were calculated utilizing Dragon software (Milano Chemometrics and QSAR research group).

3. Results and discussion

In order to detect the homogeneities in the data set and identify possible outlier and cluster, principal component analysis (PCA), was performed within the calculated descriptors space for the whole data set. An important feature is that the obtained PCs are uncorrelated, and they can be used to derived scores which can be used to display most of the original variations in a smaller number of dimensions. Fig. 1 shows the distribution of compounds over the two first components. As can be seen from Fig. 1, there is not a clear clustering between compounds. The retention index of 179 specified alkanes and alkenes were randomly classified into a training set (154 retention index data) and a prediction set (25 retention index data). The data were centered to zero means and scled to the unit variance.

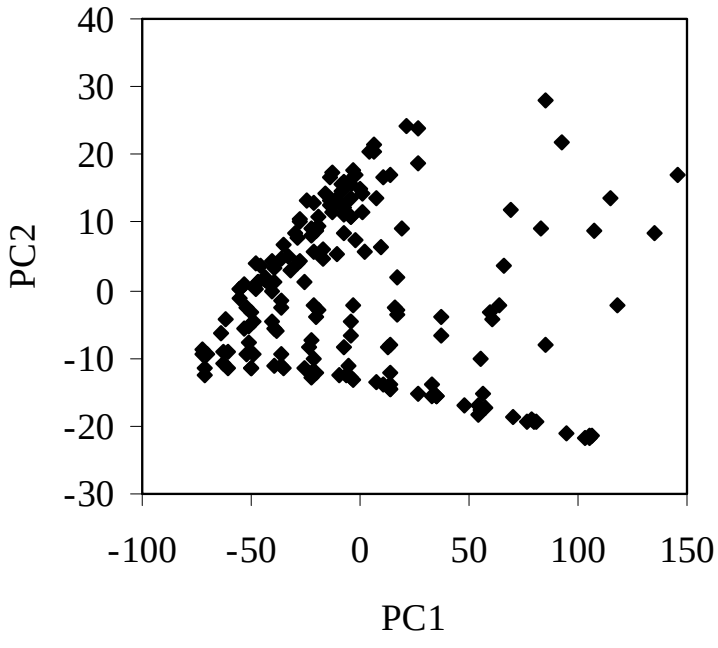

Fig. 1. Principal component analysis of the structural descriptors for the data set.

3.1. PLS and LS-SVM analysis

The factor-analytical multivariate calibration is a powerful tool for modeling, because it extractive more information from the data and allows building more robust models. The optimum number of factors to be included in the calibration model was determined by

computing PRESS from cross-validated molded using high number factors according to Haaland suggestion [11]. The optimum number was resulted 8 (PRESS = 0.056) for PLS modeling.

LS-SVM was performed with radial basis function (RBF) as kernel function [12]. In the model development using LS-SVM and RBF kernel, γ and δ^2 parameters were a manageable task, similar to the process employed to select the number of factors for PLS model, but in this case for a two-dimensional problem. These parameters were optimized generating models with values of γ and δ^2 in the range of 1-1000 with adequate increments. The optimum γ and δ^2 are 250 and 320, respectively.

3.2. Prediction of retention index

The proposed methods were successfully applied to prediction of retention index for several alkanes and alkenes. The results obtained by PLS and LS-SVM for prediction set is shown in Fig. 2. Fig. 2 plots of the predicted retention index versus experimental values [7]. Linear equations and R^2 are also shown on the Fig. 2. The correlation coefficient (R^2) for LS-SVM model were better that other models and close to one.

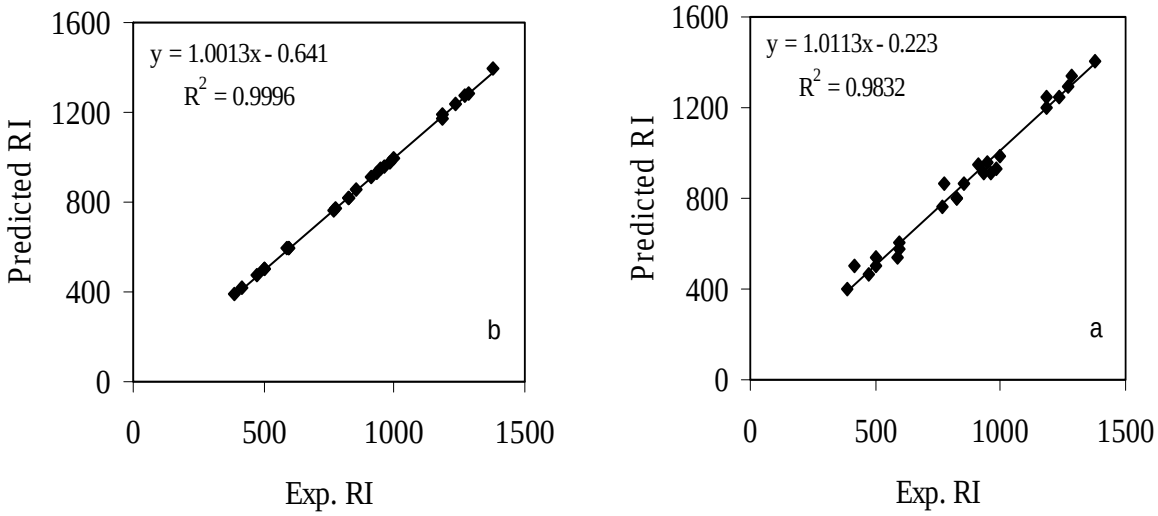

Fig. 2. Plots of predicted retention index versus experimental retention index for alkanes and alkenes in the prediction set.

The statistical parameters obtained by PLS and LS-SVM methods are listed in Table 1. Table 1 show RMSEP (root mean squares error of prediction), RSEP (relative standard error of prediction) and the range of percentage errors for prediction of retention index of alkanes and alkenes compounds. As can be seen, the percentage error was also quite acceptable for LS-SVM. Also, it is possible to see that LS-SVM presents excellent prediction abilities when compared with PLS regression.

Table 1
Comparison of the statistical parameters by different QSRR models for prediction of retention index**.**

Method	RMSEP	RSEP (%)	Percentage error range
PLS	0.1421	13.2642	-4.51 to +6.83
LS-SVM	0.0011	0.4102	-0.05 to 0.08

4. Conclusion

LS-SVM was established to predict the retention index of some alkanes and alkenes. A suitable model with high statistical quality and low prediction errors was obtained. The structural descriptors (descriptors obtained by Dragon software) concerning all the molecular properties and those of individual atoms in the molecule were found to be important factors controlling the retention index behavior. The results show that, LS-SVM is more powerful in prediction to retention index of cited compounds than PLS.

Acknowledgements

The authors gratefully acknowledge the support to this work from Yong Researcher Club, Islamic Azad University, Arak Branch.

References

[1] A. Niazi, S. Jameh-Bozorghi, D. Nori-Shargh, J. Hazard. Mat. 151 (2008) 603-609.
[2] A. Niazi, S. Jameh-Bozorghi, D. Nori-Shargh, Chin. Chem. Lett. 18 (2007) 621-624.
[3] B. Hemmateenejad, M.A. Safarpour, F. Taghavi, J. Mol. Struc. 635 (2003) 183-190.
[4] A. Niazi, J. Ghasemi, A. Yazdanipour, Spectrochim. Acta Part A 68 (2007) 523-530.
[5] A. Niazi, M. Goodarzi, A. Yazdanipour, J. Braz. Chem. Soc. 19 (2008) 536-542.
[6] C. Cortes, V. Vapnik, Mach. Learn. 20 (1995) 273-275.
[7] N. Bosnjak, Z. Michalic, N. Trinajstic. J. Chromatogr. A 540 (1991) 430-440.
[8] R. Chretien, E. Dubois, J. Anal. Chem. 49 (1971) 747-751.
[9] E. Dubois, R. Chretien, L. Sojrka. J. Chromatogr. A 194 (1980) 121-134.
[10] S.H. Hilal, L.A. Carreira, S.W. Karickholl, C.M. Melton. J. Chromatogr. A 662 (1994) 269-280.
[11] D.M. Haaland, E.V. Thomas, Anal. Chem. 60 (1988) 1193-1202.
[12] J.A.K. Suykens, T. van Gestel, J. de Brabanter, B. de Moor, J. Vandewalle, Least squares support vector machines, World Scientifics, Singapore, 2002.