Prediction of Solubility of Organic Compound for High-temperature Water by Machine Learning

Mitsumasa Osada*, and Kotaro Tamura

Department of Chemistry and Materials, Faculty of Textile Science and Technology,
Shinshu University
3-15-1, Tokida, Ueda, Nagano 386-8567, Japan

E-mail : osadam@shinshu-u.ac.jp Tel : +81-268-21-5458 FAX : +81-268-21-5391

Abstract

The objective of this study is to predict the solubility (log$S$ (mol/kg)) of the organic compounds in high-temperature water by machine learning. The chemical structure of the organic compound was converted into 196 descriptors (parameters). We added the temperature (T) and water density (ρ) to the above 196 descriptors. The observed solubilities (log$S$ (mol/kg)) experimentally were regressed by the 196 parameters, temperature, and water density. We compared the regression methods of multiple regression (MR), least absolute shrinkage and selection operator (Lasso), and support vector regression using Gaussian kernel (SVR). In this work, we suggested a new regression method that combines Lasso and SVR (Lasso + SVR). As a result, it was found that the model obtained by the combination of Lasso and SVR can predict the solubility of organic compound in high-temperature water with good accuracy. By inputting chemical structure, temperature, and water density into this model, the solubility of any organic compound can be predicted.

Keywords

Solubility, Machine learning, Subcritical water

1. Introduction

In recent years, high-temperature water has attracted attention for processes such as organic reactions and extractions due to its low environmental load and low waste [1-3]. The experimental data of the solubility (log$S$ (mol/kg)) of organic compound for high-temperature water is important for designing the processes.

In previous studies, some models for estimating the solubility based on the chemical structure of organic compounds have been suggested. Typical examples are the UNIFAC

method [4] and the COSMO-RS method [5]. However, there are some problems for these methods. For the UNIFAC method, it can only be applied to molecules consisting of specific functional groups because it uses the number of functional groups in the solute to predict the solubility. For the COSMO-RS method, the computational cost is high in exchange for the high prediction accuracy because the charge density of the molecular surface is used by obtaining from quantum chemical calculations.

In this study, we used Quantitative Structure-Property Relationships (QSPR), which is effective for estimating the solubility of organic substances in water under ambient temperature and pressure [6]. The QSPR is a method for statistically modeling the relationship between chemical structures and physical properties [7]. When predicting the solubility of a new compound, only the chemical structure is used, and additional experiments are not required. Once the model has been developed, the estimation can be done quickly. The QSPR is an effective method for estimating the solubility of organic compound in water under ambient temperature and pressure because the data on the solubility is abundant.

On the other hand, the data of the solubility of organic compound for high-temperature water is limited due to the difficulty of the experiments. In addition, for the case of water at high-temperature and high-pressure, the estimation of the solubility under arbitrary temperature and water density is required. In this study, we propose a new method of QSPR, which uses not only chemical structure but also the parameters such as temperature and water density. We developed a prediction model of the solubility by machine learning by combining the experimental data at ambient temperature with those at high temperatures.

2. Method

The solubility data (log$S$ (mol/kg)) of 1,290 organic compounds at ambient temperature and pressure [8], and 12 organic compounds (included in the above 1,290 compounds) at high temperatures of 100-250°C and above saturated pressures, with 55 experimental points [9-11], were used to develop the model. At first, the chemical structure of the organic compound was converted into descriptors using Python RDKit [12], which is an open source of chemical informatics and machine learning. The chemical structure was converted into 196 descriptors (parameters) by the RDKit. A data set was created by combining 196 descriptors and the temperature (T) and water density (ρ) corresponding to the conditions for measuring the solubility of each organic compound in high-temperature water. In advance, we aligned the mean of all descriptors (parameters) to 0 and the standard deviation of those to 1. The 80% of data was selected randomly as training data for developing a model, and the remaining 20% was used as test data. Multiple regression (MR), least absolute shrinkage and selection operator (Lasso), and support vector regression using Gaussian kernel (SVR) were employed for developing models. In addition, a new method that combines Lasso and SVR

(Lasso + SVR) was also considered, which is our original method. In this method, the parameters are selected by Lasso firstly and then the selected parameters are used for SVR. The regression was calculated using the scikit-learn [13] library. 5-fold cross-validation was used to determine the hyperparameters for each regression. The test data was used to validate the model. Finally, the validity of the model was confirmed using data from 41 experimental points in high-temperature water for 7 organic compounds [9-11] that were not used in the model development. The validity of the model was checked from the R-squared (R^2), root mean square error (RMSE), and mean absolute error (MAE).

3. Results and Discussion

The 196 descriptors (parameters) obtained from the chemical structure of the organic compound were molecular weight, substituents, partition coefficient, surface area, partial charge, etc. Among these descriptors, the molecular weight of the organic compound and the number of substituents are primary information. The partition coefficient and the partial charge are secondary information and are calculated from the primary information. Both the primary and secondary information is used for the regression of the solubility (log$S$ (mol/kg)).

Table 1 shows the comparison of R^2, RMSE, and MAE for each the regression method. The parity plot of the traing, test, and validation data obtained by MR, Lasso, SVR, and Lasso + SVR are shown in Fig. 1, 2, 3, and 4, respectively. Both MR and Lasso, which are linear regressions, and SVR and Lasso + SVR, which are nonlinear regressions, can be used for the estimation, but the R^2 values of validation were higher for Lasso + SVR. The RMSE and MAE of validation are also lower for Lasso + SVR, indicating that Lasso + SVR estimation models are better than MR, Lasso, and SVR. For the MR and Lasso, which are linear regressions, the comparison of the coefficients allows us to consider the parameters those have a large contribution to the solubility. 35 parameters were selected from 198 parameters by Lasso. The order of the coefficients in Lasso was the partition coefficient > water density > maximum absolute partial charge > molar refractivity > other parameters. However, the coefficients of these 4 parameters were not particularly large among the 35 parameters, indicating that dozens of parameters are needed for accurate estimation. The partition coefficient of organic compounds to water and octanol is practically the same as the solubility of organic compound in water, so it is natural to show a strong correlation.

For the Lasso + SVR, the accuracy for the vailidation data is very high compared with other methods. In general, if there is collinearity between parameters, the regression of validation data becomes more difficult. In this method, we reduced the paramers from 198 to 35 by Lasso, which may be the same as avoiding collinearity. After that, we regressed 35 parameters by SVR, which may allow SVR to develop a more valid model.

Table 1

Comparison of the regression methods.

Method	Data	R^2	RMSE	MAE
MR	Training	0.922	0.568	0.424
	Test	0.890	0.706	0.522
	Validation	0.753	0.937	0.707
Lasso	Training	0.856	0.766	0.593
	Test	0.851	0.820	0.622
	Validation	0.765	0.886	0.718
SVR	Training	0.974	0.327	0.177
	Test	0.930	0.561	0.399
	Validation	0.837	0.785	0.512
Lasso+SVR	Training	0.966	0.376	0.287
	Test	0.923	0.589	0.437
	Validation	0.963	0.365	0.286

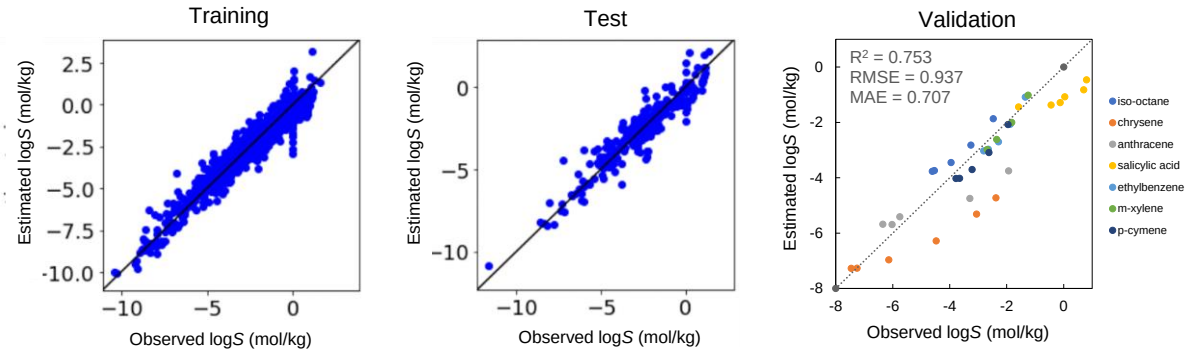

Fig. 1. The observed versus estimated aqueous solubilities by the multiple regression (MR).

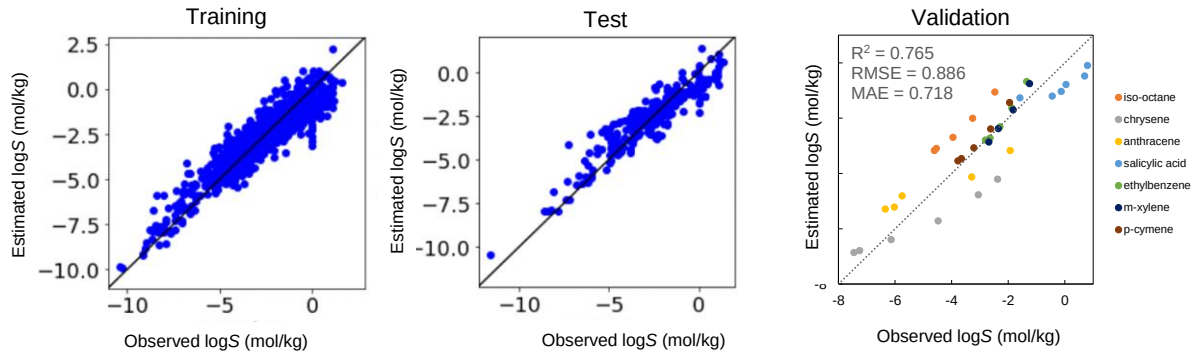

Fig. 2. The observed versus estimated aqueous solubilities by the Lasso.

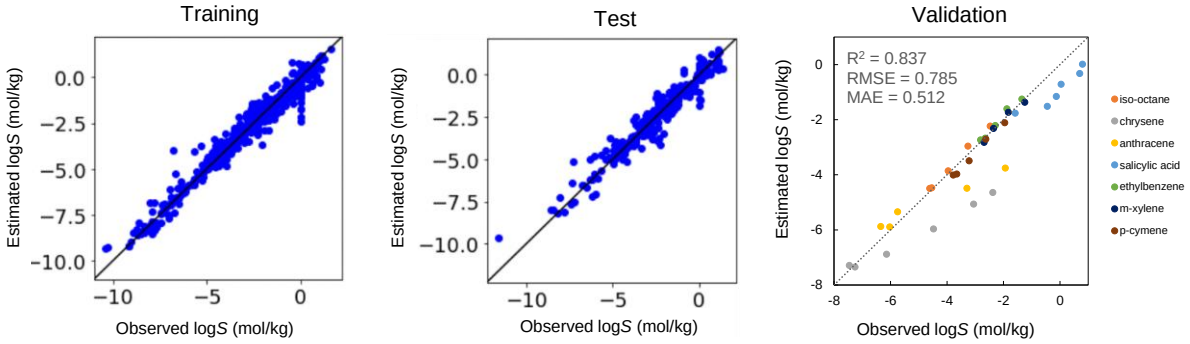

Fig. 3. The observed versus estimated aqueous solubilities by the SVR.

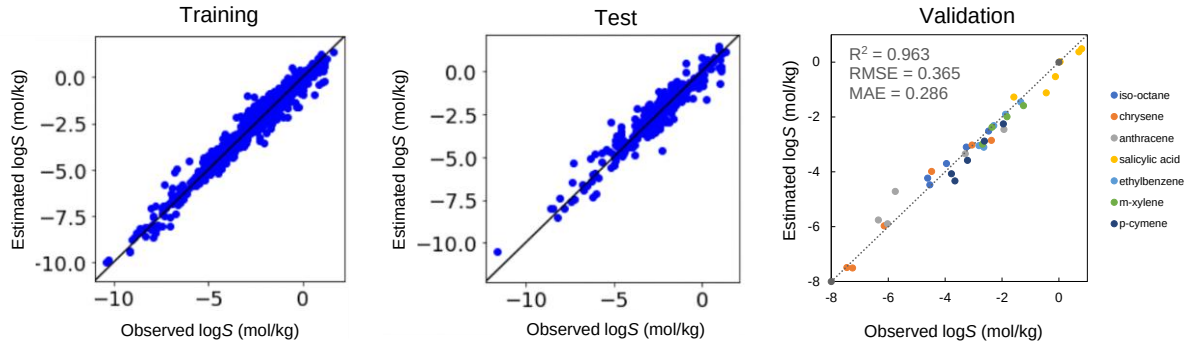

Fig. 4. The observed versus estimated aqueous solubilities by the combination of Lasso + SVR.

4. Conclusion

The solubility (logS (mol/kg)) of organic compounds was predicted by the regression of the reported experimental data of 1,290 organic compounds at ambient temperature and pressure, and 12 organic compounds at high temperatures of 100-250ºC. The experimental solubility data were represented by the models obtained by the combination of Lasso and SVR with good accuracy. The machine learning is a promising method to predict the solubility of the organic compound for high-temperature water.

References

[1] M. Osada, S. Shoji, S. Suenaga, M. Ogata, Fuel Process. Technol. 195 (2019) 106154.

[2] M. Osada, H. Kobayashi, T. Miyazawa, S. Suenaga, M. Ogata, Int. J. Biol. Macromol. 136 (2019) 994-999.

[3] S. Suenaga, M. Osada, ACS Appl. Polym. Mater. 1 (5) (2019) 1045–1053

[4] A. Fredenslund, R. L. Jones, J. M. Prausnitz, AIChE J. 21 (1975) 1086-1099.

[5] A. Klamt, V. Jonas, T. Bürger, and J. C. W. Lohrenz et al., J. Phys. Chem. A 102 (1998) 5074-5085.
[6] H. Kaneko, K. Funatsu, Chemom. Intell. Lab. Syst. 142 (2015) 64-69.
[7] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, V. Prachayasittikul, EXCLI J. 8 (2009) 74-88.
[8] T. J. Hou, K. Xia, W. Zhang, X. J. Xu, J. Chem. Inf. Comput. Sci. 44 (2004) 266-275.
[9] D. J. Miller, S. B. Hawthorne, Anal. Chem. 70 (1998) 1618-1621.
[10] D. J. Miller, S. B. Hawthorne, A. M. Gizir, A. A. Clifford, J. Chem. Eng. Data 43 (1998) 1043-1047.
[11] P. Khuwijitjaru, S. Adachi, R. Matsuno, Biosci. Biotechnol. Biochem. 66 (8) (2002) 1723-1726.
[12] https://www.rdkit.org/.
[13] https://scikit-learn.org/stable/.