CHEMICAL ENGINEERING TRANSACTIONS

VOL. 59, 2017

A publication of
AIDIC
The Italian Association
of Chemical Engineering
Online at www.aidic.it/cet

Guest Editors: Zhuo Yang, Junjie Ba, Jing Pan
Copyright © 2017, AIDIC Servizi S.r.l.
ISBN 978-88-95608- 49-5; **ISSN** 2283-9216

DOI: 10.3303/CET1759157

Application of Radial Basis Function Neural Network in Environmental Chemistry

Bing Chen

Xi'an International University, Business school, Xi'an 710077, China
cb3124@126.com

The retention behaviour of 66 organic pollutants in biocompatible micelles was studied by QSPR method. The linear and nonlinear models between the structures of these compounds and their chromatographic retention values were established by using the heuristic method and the RBFNN method, respectively. The correlation coefficients of the two methods are 0.8400 and 0.8642, respectively, and the corresponding root mean square error (RMS) is 0.1577 and 0.1562. The results show that there is a significant linear relationship between the selected descriptor and the retention value by comparing the stability and prediction ability of the two models. The nonlinear method cannot improve the prediction ability of the model.

1. Introduction

In the past 20 years, environmental pollution has brought great harm to people's lives. Many of the literature reports and studies the adverse effects of these substances on humans, and even some are highly lethal (Halali et al., 2015). However, it is very difficult to predict the toxicity of these contaminants in vivo. If a substance causes a biological reaction, it must go through a series of processes, including absorption, transfer and distribution. First, this substance must enter the organism and then dissolve in body fluids. Although there are many experimental methods to determine the retention data, the quantitative structure-retention correlation (QsRR) provides a valuable method for predicting retention values based on molecular descriptors (Daachi et al., 2015). As a result of the expansion of linear free energy (LFERs), QSRR has successfully predicted chromatographic behavior. The main steps of QSRR include: data collection, calculation and selection of molecular descriptors, establishment of related models and evaluation of model (Hassanzadeh et al., 2015).

In this study, QSRR is better than other methods in that the descriptors used for modeling are only calculated from the molecular structure and do not depend on other experimental properties (Alexandridis et al., 2014). The heuristic method (HM) and RBFNN were calculated using the CODESSA software and the selected descriptors were used to predict the retention factors for 66 organic pollutants in the BMC. The purpose is to establish a robust QSRR model to predict the log k of more contaminants while identifying structural factors associated with the mechanism of pollutant retention.

2. Brief introduction to radial basis function neural network (RBFNN)

Artificial neural network (ANN) is a kind of information processing system established by imitating the human brain neural network functions. The researches on it started from 1940s, but the development was always auite slow. After 1980s, J. Hopfield put forward a fully connected feedback network model used for associative memory and optimization computation. The proposal of “energy function” made the realization of neural network have a certain technical guarantee, which greatly promoted the study and application of artificial neural network (Golbabai et al., 2015). Since 1986, the artificial neural network theories have been applied in chemical field, and its study has been paid much attention to. At present, in the artificial neural network application, one of the most active fields is QSPR/QSAR studies, involving chemistry, biology, pharmacology, environmental science, material science and so on disciplines.

Please cite this article as: Bing Chen, 2017, Application of radial basis function neural network in environmental chemistry, Chemical Engineering Transactions, 59, 937-942 DOI:10.3303/CET1759157

ANN is the complex non-liear network connected by functional units (neurons) with a large number and simple functions. In that it absorbs many characteristics of biological neural network, it has the following advantages: firstly, although the function of each processing unit of ANN is rather simple, the parallel activities of massive simple processing units make the information processing ability and effect astonishing. Secondly, each neuron accepts the input of other neurons, produces the output through network, and affects other neurons. The restriction and impact between networks achieve the non-linear mapping from the input state to the output state space (Luo and Billings, 2015). At last, ANN can obtain the network weight and structure through the training and learning, and display strong self-learning ability and adaptability to the environment.

Artificial neural network has many kinds, and the category results are different because of various angles. According to the network structure, it can be divided into feedforward network and backward network; in accordance with the learning means, it is divided into learning network with supervision and learning network without supervision. Among a variety of algorithms, the most frequently used are BP (Back Propagation) neural network and RBF (Radial Basis Function) neural network.

Radial basis function (RBF) neural network is also a commonly used neural network model. It has the characteristics of optimal approximation, simple optimization process and fast training speed so taht it has been widely used in many fields. Radial basis function neural network is feedforward layered neural network. With the input of training sample and the Euclidean distance of nodes weight vector of the hidden layer as the input, the Gauss function that reflects the probability density is adopted as the role function of hidden layer. It is characterized by: the smaller the distance from the training set sample to the node center, namely the Gauss peak center, the greater the output value with the interaction of the Gauss function (close to 1), and the farther the distance to the Gauss peak, the smaller the output value (close to 0). The optimization of network is just to adjust the weights near the input vector, so as to adjust the output. And the corresponding weights with larger distance can be ignored, which belongs to the local regulatory network. The RBF netwrok is used for function approximation, and compared with using the BP network, the former is easier to approximate the local characteristics of the function, and the latter is to approximate by using a global function.

3. Data and methods

3.1 The source of the data

The data for the compounds that used in this study were taken from the second reference. The 66 organic pollutants belong to different classes, and most of them are phenol derivatives containing chlorine atoms and sulfur atoms. We randomly selected 44 compounds as a training set to model, and the remaining 22 compounds were used as test sets to test the predictive power of the model.

3.2 The calculation of descriptor

In the HyperChem 4.0 software, the molecular structure of the organic matter is plotted, and then the structure is optimized by the PM3 method in the semi-empirical quantum chemistry software MOPAC 6.0, and the output file is imported into the CODESSA software to calculate the descriptor. Five types of descriptors are obtained: composition, topology, geometry, electrostatic, quantization.

In this study, the RBFNN algorithm is based on the RBFNN function in Matlab. All the program running environment is Pentium IV 2.5G processor, 256M memory compatible machine.

3.3 Radial basis function neural network

Radial basis function neural network is a commonly used neural network model (Molani et al., 2014). It has the characteristics of the best approximation, the simple optimization process and the fast training speed, and it has been widely used in many fields. RBFNN is also a forward layered neural network. Its structure can be described by a three-layer network: the first layer is the input layer, the second layer is the hidden layer, that is, the radial basis function layer, and the third layer is the output layer (Zou et al., 2015). The input layer only enters information without any other processing, and the implicit layer is usually composed of a series of RBF functions. The most commonly used RBF function is the Gaussian function. The RBF function that used in this paper adopts the Gaussian function. The expression is as follows:

$$h_j(x) = esp(-\left\|x - c_j\right\|^2 / r_j^{\,2}) \tag{1}$$

In the formula, x is the input vector, $h_j(x)$ is the output of the first RBF function, c_j is the center of the jth hidden layer node, and r_j is the radius of the jth RBF. Each RBF node represents an RBF function, and each RBF has its corresponding center and radius. RBFNN output layer is usually a linear function.

RBFNN training and optimization process is to determine the following network parameters: (1) The number of RBF functions n_h. (2) The center c_j and radius r_j of BRBF. (3) The connection weight w_{kj} between the jth hidden layer node and the kth output node. When all the network parameters are determined, a network with generalization capability is formed, so that the network can be used for prediction. The number of RBFs affects RBFNN's prediction ability to a large extent. If the number is too small, the network is too simple to reflect the complexity of the research system. On the other hand, if the number of RBF is too large, it may cause over-fitting, resulting in deterioration of the generalization ability of the model. At present, many methods have been developed for selecting the number of RBFs in the RBFNN, such as PLS-RBF, KNN, and genetic algorithms (Zhu, et al., 2014). In this study, we chose the ORR proposed forward subset selection method to select the RBF center and the number of RBF. The characteristic of the ORR method is that all the RBFNN centers are selected by the training set. A RBF is introduced at each time until the introduction of the new RBF no longer changes the performance of the network. The advantage of this approach is that it does not need to pre-determine the number of RBFs, which can simultaneously select the center and determine the number of RBFs. The selected centers are from the training set of samples. The selected criteria are based on the one-out (LOO) cross-test method:

$$\sigma_{LOO}{}^{2} = \frac{\hat{y}P(diagP))^{-2}P\hat{y}}{p} \qquad (2)$$

In the formula, $\hat{y}$ is the output of the network, P is the projection matrix, $p=I_p$-ZZ', Z is the output of the hidden layer, and I is the unit matrix of dimension P. P is the number of samples in the training set. The leave one out (LOO) interaction test is mainly used to prevent the network from over fitting.

After selecting the center and hidden layer nodes of RBF, the linear least squares method can be used to calculate the connection weight between the hidden layer and the output layer:

$$w = yZ'(ZZ')^{-1} \qquad (3)$$

In the formula, Y is the target value of the training set sample, Z is the output of the hidden layer node, Z 'is the transpose of Z, and w is the connection weight between the hidden layer and the output layer. The optimal radius value is determined using the interactive test method.

4. Results and discussion

4.1 The results of HM

A large number of non-empirical molecular descriptors were calculated by CODESSA, and the descriptors were analyzed by regression analysis. In order to obtain the optimal descriptor number to establish the model prediction log k, the optimal linear regression equation containing 1-8 different descriptors is calculated by the training set data.

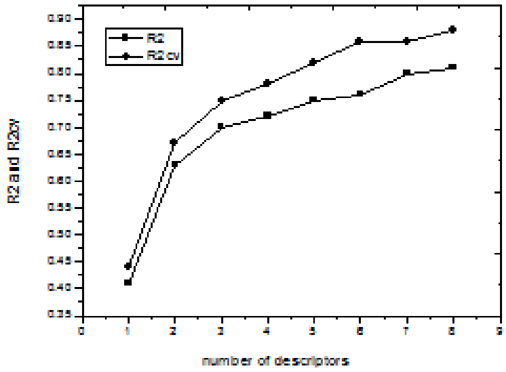

Figure 1: Influence of the number of descriptors on R^2 and R^2_{cv} *of the regression model.*

Figure 1 shows the relationship between R^2 and R^2_{cv} and the number of descriptors. R^2 and R^2_{cv} increase with the increase of the number of descriptors. In order to avoid the over parameterization of the model, the value of R^2 is less than 0.02 as the selection criteria of the number of parameters, that is, when a descriptor is added, the value of R^2 is less than 0.02. At this point, there is no need to add new parameters in the model. The final six parameters are the best number of parameters. The statistical parameters of the training set model are listed in Table 1.

Table 1: The selected descriptors, regression coefficients and t-values for the linear model.

Descriptor	Chemical meaning	Coefficient	DX	t-test
constant	Intercept	-0.0032	0.1883	-0.0170
KHI	Kier Hall index (order 0)	0.1441	0.0131	11.0200
MNACc	max net atomic for a C atom	-1.9554	0.2897	-6.7500
HDCA-1	HA dependent HDCA-1 [Zefirov's PC]	-0.1348	0.0460	-2.9310
HDSA-2/TMSA	HA dependent HDSA-2/TMSA [quantum-chemical PC]	5.3329	3.5634	1.4970
RNS	relative number of S atoms	-6.9348	2.4311	-2.8530
RPCG	Relative positive charge (QMPOS/QTPLUS) [Zefirov's PC]	2.8931	0.6260	4.6210

As can be seen from Table 1, at the confidence level P = 95%, the t-test value of each descriptor is compared with the t-test threshold (t_c = 2.0262), and it can be seen that the selected descriptor and log k have significant linear statistical relationships (except HDSA-2 / TMSA). The heuristic method is modeled as follows:

$$\begin{aligned}\log k &= 0.0032(\pm 0.188) + 0.1441(\pm 0.0131)\times KHI - 1.9554(\pm 0.2897)\times MNACc \\ &\quad - 0.1348(\pm 0.0460)\times (HDCA-1) + 5.3329(\pm 3.5634)\times (HDSA-2/TMSA) \\ &\quad - 6.9348(\pm 2.4311)\times RNS + 2.8931(\pm 0.6260)\times RPCG \\ R^2 &= 0.8550; s^2 = 0.0319; F = 35.3800; R^2_{\mathrm{cv}} = 0.8024; n = 44\end{aligned} \tag{4}$$

The prediction results of HM are shown in Table 1, and the experimental values of log k and the scatter plot of the predicted values are shown in Figure 2. The RMSE of the training set, test set and the whole data set are 0.2032, 0.1745 and 0.1577 respectively. The corresponding correlation coefficients (R2) are 0.8550, 0.8549 and 0.8400 respectively. As can be seen from Table 1 and Figure 2, the prediction results are not very accurate. In other words, the relationship between the descriptor and log k is not a simple linear relationship, there may be a more complex relationship between them. Therefore, it is necessary to use the nonlinear RBFNN method to further study the quantitative relationship between the molecular structure and log k.

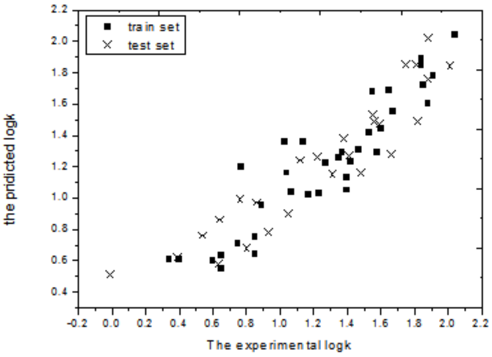

Figure 2: Scatter plot of the predicted log k vs. the experimental log k (HM model).

4.2 The results of RBFNN

After establishing the linear model, the same descriptor establishes the nonlinear model as the input of RBFNN. In order to get good results, the parameters that affect RBFNN are optimized. The best radius selection is achieved by changing the value of the system during the training process. We should select the value that can get the best results of the leave one out method. As can be clearly seen from Figure 3, the best radius for this data set is 1.6. The corresponding hidden layer nodes are 17. The experimental and logarithmic values of log k are shown in Figure 4. The statistical results of the obtained model are: training set: RMSE=0.1695, R^2 = 0.8761. Test set: RMSE = 0.1925, R^2 = 0.8464.

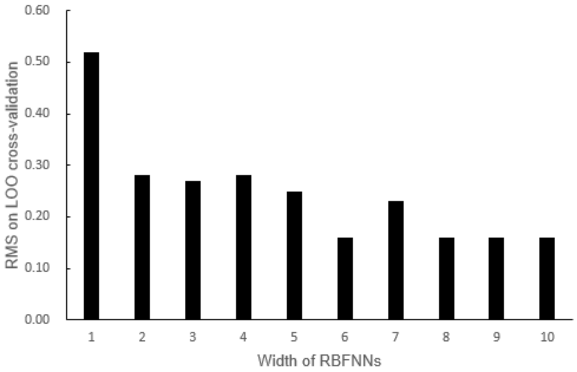

Figure 3: The width of RBFNN vs. RMS error on LOO cross-validation.

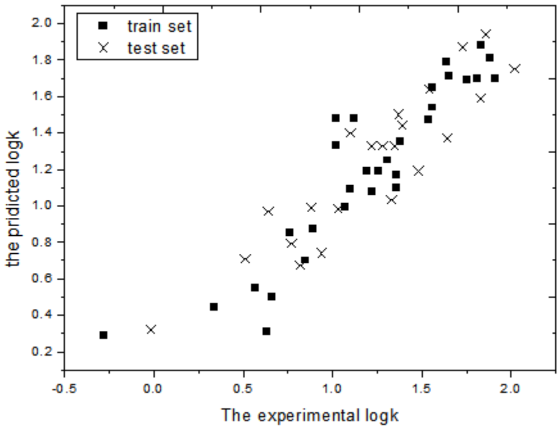

Figure 4: Scatter plot of the predicted log k vs. the experimental log k (RBFNN model).

By comparing the results of HM and RBFNN, we can see that the results of the nonlinear model are superior to the heuristic linear method. This indicates that there is a certain non-linear relationship between the selected descriptor and log k. Furthermore, the 1/3 compound is used as a test set, which also demonstrates the applicability and reliability of the method.

4.3 Discussion of the parameters

By explaining the descriptors in the model, we can identify the microscopic factors that affect the retention factor of the organic compound in the BMC. In the selected six descriptors, it contains one topology, three electrons, one composition, and one quantum chemical descriptor. These descriptors reflect the different aspects of the molecular structure of the information.

The Kier Hall index (order 0) is a topological descriptor that reflects the connectivity of atoms in a molecule. Level 0 indicates the number of atoms and the branching of the molecule, which is not related to the chemical bond. The more the number of atoms in the molecule, the higher the value of KHI is. If the molecule contains the same number of atoms, the smaller the branch, the higher the KHI value. The positive coefficients in the model indicate that increasing the value of KHI is advantageous for increasing the retention. The reason for this is that the intermolecular interactions of the stationary phase in the organic compound molecules and the biocompatible micelles are affected by the size of the molecule and the molecular branching.

Under certain experimental conditions, the above descriptors reflect the structural characteristics associated with the pollutant retention factor. By analyzing the meaning of descriptors in detail, it can be seen that molecular size, hydrogen bonding and charge action play a key role in the retention of compounds. The smaller retention values can also indicate that the compounds are easy to penetrate the biofilm, leading to higher toxicity.

5. Conclusions

In this study, a QSRR model for predicting the retention of organic compounds in BMC was established. The proposed linear model clearly reflects which descriptors play a role in the retention of these substances. In addition, the nonlinear RBFNN model established with the same molecular descriptor embodies the strong prediction ability, which improves predictive accuracy. At the same time, it shows the existence of nonlinear relations. Therefore, the model that proposed in this study can be used to predict the retention of more organic pollutants in BMC.

Reference

Alexandridis A., Chondrodima E., Efthimiou E., Papadakis G., 2014, Large earthquake occurrence estimation based on radial basis function neural networks, Geoscience & Remote Sensing IEEE Transactions on, 52(9), 5443-5453.

Daachi M.E., Madani T., Daachi B., Djouani K., 2015, A radial basis function neural network adaptive controller to drive a powered lower limb knee joint orthosis, Applied Soft Computing, 34(C), 324-336.

Golbabai A., Mohebianfar E., Rabiei H., 2015, On the new variable shape parameter strategies for radial basis functions, Computational and Applied Mathematics, 34(2), 691-704.

Halali M.A., Azari V., Arabloo M., Mohammadi A.H., Bahadori A., 2015, Application of a radial basis function neural network to estimate pressure gradient in water–oil pipelines, Journal of the Taiwan Institute of Chemical Engineers, 58, 189-202.

Hassanzadeh Z., Kompany-Zareh M., Ghavami R., Gholami S., Malek-Khatabi A., 2015, Combining radial basis function neural network with genetic algorithm to qspr modeling of adsorption on multi-walled carbonnanotubes surface, Journal of Molecular Structure, 1098, 191-198.

Luo W., Billings S.A., 2015, Structure selective updating for nonlinear models and radial basis function neural networks, International Journal of Adaptive Control & Signal Processing, 12(4), 325-345.

Molani M., Ghaffari A., Jafarian A., 2014, A new approach to software project cost estimation using a hybrid model of radial basis function neural network and genetic algorithm, Indian Journal of Science & Technology, 7(6), 838-843.

Zhu J. Z., Cao J.X., Zhu Y., Zhu Y., 2014, Traffic volume forecasting based on radial basis function neural network with the consideration of traffic flows at the adjacent intersections, Transportation Research Part C Emerging Technologies, 47(2), 139-154.

Zou B., Wang M., Wan N., Wilson J.G., Fang X., Tang Y., 2015, Spatial modeling of pm 2.5, concentrations with a multifactoral radial basis function neural network, Environmental Science and Pollution Research, 22(14), 10395.