TRAINED NEURAL NETWORK CHARACTERIZING VARIABLES FOR PREDICTING ORGANIC RETENTION BY NANOFILTRATION MEMBRANES

Arcadio Sotto, Ana Martinez, Angel Castellanos

Abstract: Many organic compounds cause an irreversible damage to human health and the ecosystem and are present in water resources. Among these hazard substances, phenolic compounds play an important role on the actual contamination. Utilization of membrane technology is increasing exponentially in drinking water production and waste water treatment. The removal of organic compounds by nanofiltration membranes is characterized not only by molecular sieving effects but also by membrane-solute interactions. Influence of the sieving parameters (molecular weight and molecular diameter) and the physicochemical interactions (dissociation constant and molecular hydrophobicity) on the membrane rejection of the organic solutes were studied. The molecular hydrophobicity is expressed as logarithm of octanol-water partition coefficient. This paper proposes a method used that can be used for symbolic knowledge extraction from a trained neural network, once they have been trained with the desired performance and is based on detect the more important variables in problems where exist multicolineality among the input variables.

Keywords: Neural Networks, Radial Basis Functions, Nanofiltration; Membranes; Retention.

ACM Classification Keywords: K.3.2 Learning (Knowledge acquisition)

Introduction

Phenolic compounds are commonly used as raw materials in the manufacture of polymers, plasticizers, hydraulic fluids and various industrial chemicals. Therefore, there are many wastewater effluents contain discharges amounts of these recalcitrant organic compounds.

Nanofiltration (NF) is a viable treatment for the removal of dissolved organic pollutants for production of drinking water and as combined method with advanced and traditional water treatment process [Van der Bruggen et al 2003] [Hellebrand et al 1997]. Many reported studies indicate that several physical phenomena can play a role in the solute transport through nanofiltration membranes: solution-diffusion, convection (sieving), electrostatic (charge) repulsion and dielectric exclusion. In addition, NF strongly depends on the feed water composition, membrane and solute properties, and operational conditions [Bellona et al 2004]. Therefore retention of organic compounds is influenced either by pore size and charge of membrane, or by the molecular size, hydrophobicity and ionization constant of solutes [Boussu K. et al 2008] [Arsuaga et al 2008].

Neural network is proposed as suitable tool to prediction the membrane performance on solute retention and detect the more important variables when the input variables exist high correlation. Artificial Neural networks perform adaptative learning. This advantage can be used to improve the knowledge acquisition in knowledge engineering .This paper proposes extracting knowledge from a neural network that has learned using sensitivity analysis used to determinate which are the most important variables for the prediction. These will guide the process of create one model for prediction with a few variables, at least the most important variables.

Characteristics about the Forecast Method

Neural networks [Anderson, James A. and Edward Rosenfield., 1988] are non-linear systems whose structure is based on principles observed in biological neuronal systems [Hanson, Stephen J. and David J. Burr. 1990]. A neural network could be seen as a system that can be able to answer a query or give an output as answer to a

specific input. The in/out combination, i.e. the transfer function of the network is not programmed, but obtained through a training process on empiric datasets. In practice the network learns the function that links input together with output by processing correct input/output couples. Actually, for each given input, within the learning process, the network gives a certain output that is not exactly the desired output, so the training algorithm modifies some parameters of the network in the desired direction. Hence, every time an example is input, the algorithm adjusts its network parameters to the optimal values for the given solution: in this way the algorithm tries to reach the best solution for all the examples. These parameters we are speaking about are essentially the weights or linking factors between each neuron that forms our network.

There is a great number of Neural Networks [Anderson, James A. 1995] which are substantially distinguished by: type of use, learning model (supervised/non-supervised), learning algorithm, architecture, etc. Multilayer perceptrons (MLPs) are layered feed forward networks typically trained with static backpropagation. These networks have found their way into countless applications requiring static pattern classification. Their main advantage is that they are easy to use, and that they can approximate any input-output map. In principle, backpropagation provides a way to train networks with any number of hidden units arranged in any number of layers.

The research community has developed several different neural network models, such as, radial basis function, growing cell structures and self-organizing feature maps. A common characteristic of the mentioned models is that they distinguish between learning and a performance phase. Neural networks with radial basis functions have proven to be an excellent tool in approximation with few patterns. Most relevant research in theory, design and applications of radial basis function neural networks is due to Moody and Darken [Moody and Darken, 1989]. Radial basis function (RBF) neural networks provide a powerful alternative to multilayer perceptron (MLP) neural networks to approximate or to classify a pattern set. RBFs differ from MLPs in that the overall input-output map is constructed from local contributions of Gaussian axons, require fewer training samples and train faster than MLP. The most widely used method to estimate centers and widths consist on using an unsupervised technique called the k-nearest neighbour rule. The centers of the clusters give the centers of the RBFs and the distance between the clusters provides the width of the Gaussians.

The object of the present study is to ascertain whether the membrane type NF90 has a quantitative effect on the values of the retention for different components analyzed, and affects the relationships between the different variables considered as input in the model propose for prediction retention. Retention behavior of the phenolic compounds by NF90 membrane was investigated in order to clarify the influence of the molecular weight (MW), size (diameter), acid dissociation constant (pk_a) and molecular hydrophobicity (log*P*) of selected compounds on membrane performance. This paper proposes a method in order to detect the importance of the input variables. In multivariate analysis problems, when there exists correlation among different variables of forecasting, the importance and the sequence when adding variables in the model, can be detected from the knowledge stored in NN, and must be taken into account when the study of the correlations detect relationships among a set of variables

Neural networks can predict any continuous relationship between inputs and the target; artificial neural networks develop a gain term that allows prediction of target variables for a given set of input variables., we use neural networks models with analysis of sensibility, this model predict more accurately the relationship existing between variables, and is a suitable way to find the individual effects of forecasting variables over the variable to forecast , and the way to find a set of forecasting variables to include in the new model.

The addition of a given variable into a forecasting model does not implies that this variable will have an important effect over the response of the model, that is, if a researcher identifies a set of forecasting variables, he must check if they really affect the response. A frequent problem is that some of the forecasting variables are

correlated. If the correlation is small, then consequences will be less important. However, if there is a high correlation between two or more forecasting variables, then the model results will be ambiguous but not for obtain a bad prediction, the problem is the high correlation between variables (high lineal association) decrease in a drastic way the individual effect over the response for each correlation variable and sometimes is difficult to detect and is not possible measure the real effect for each variable over the output.

The process of finding relevant data components is based on the concept of sensitivity analysis applied to trained neural networks. Two ANN models predict changes for certain combinations of input variables, detecting the most important influence in the output variable. We have studied different analysis for detecting relationships between molecular diameter, molecular weight, log*P* and pk_a in the two membranes during the process of nanofiltration. Retention organics compounds by correlated with characteristics of membrane and also with phisico-chemical properties of organic solutes. In order to study the relationships between different variables it has been used neural networks models with a single hidden layer and Tanh as transfer function in both cases. One ANN model uses MLP (multilayer perceptron) and the other ANN model uses a normal radial basis function (RBF) for model development.

Two ANNs models have been implemented with four input neurons: molecular weight, molecular diameter, pka and logP to estimate the membrane solute retention. The MLP network uses a sigmoid activation function with a single hidden layer with four neurons. The general form of a feed- forward neural network expresses a transformation of the expected target as a linear combination of no-linear functions of linear combinations of the inputs. A normalized radial basis function (RBF) network is a feed-forward network with a single hidden layer using in this case, the same function sigmoid (Tanh), in the hidden layer with 15 clusters and one output layer. In contrast to MLP, each basis function is the ratio of a bell-shaped Gaussian surface. For all the learning process has been performed with the momentum algorithm. Unsupervised learning stage is based on 100 epochs and the supervised learning control uses as maximum epoch 10000, and threshold 0.001. We have performed an initial study using 17 patterns in training set.

Materials and methods

Seventeen phenolic compounds were selected to carry out membrane retention experiments. Table 1 summarizes the most important properties of selected compounds.

Thin-film composite polyamide membrane, NF90 supplied by Dow/Filmtec was evaluated in this study. It's classified as nanofiltration membrane. According to the manufacturers, NF90 membrane is polyamide thin-film composite with a microporous polysulfone supporting layer. A cross flow system (SEPA CF II, Osmonics) was used for membrane retention measurements. Organic solution concentrations were fixed at 100 mg L^{-1} and system temperature was maintained constant in all experiments at 25ºC. It was controlled by circulating feed water through a stainless-still coil immersed in the thermostatic bath. Quantitative analysis of the organic compounds was carried out by means of their respective absorptions in the ultraviolet region, using a Varian Cary 500 Scan UV-VIS-NIR spectrophotometer. Concentration of PEGs and saccarides were measured with a Total Organic Carbon (TOC) analyzer (model TOC-V CSN Shimadzu). Regression factor (R2) obtained for calibrations within the range of experimental concentration used was greater than 0.99.

Retention *R* (%) of a solute was calculated using the expression:

$$R = 1 - \frac{C_p}{C_r} \times 100\% \quad (1)$$

where Cp and Cr are the concentrations for the permeate and retentate, respectively.

Compound	Formula	Molecular diameter (nm)	Molecular Weight ($gmol^{-1}$)	pK_a	log$P$
Phenol	C_6H_6O	0.1945	94.11	9.86	1.48
Resorcinol	$C_6H_6O_2$	0.1948	110.11	9.45	0.76
Hydroquinone	$C_6H_6O_2$	0.1908	110.11	10.33	0.66
Cathecol	$C_6H_6O_2$	0.2160	110.11	9.5	0.88
3-Nitrophenol	$C_6H_5NO_3$	0.2142	139.11	8.33	1.93
3-Chlorophenol	C_6H_5ClO	0.2134	128.56	9.00	2.40
2-Chlorophenol	C_6H_5ClO	0.2157	128.56	8.5	2.04
2-Nitrophenol	$C_6H_5NO_3$	0.2112	139.11	7.14	1.71
4-Chlorophenol	C_6H_5ClO	0.1915	128.56	9.47	2.43
4-Nitrophenol	$C_6H_5NO_3$	0.1849	139.11	7.23	1.57
Pirogallol	$C_6H_6O_3$	0.2154	126.11	9.12	0.29
Phloroglucinol	$C_6H_6O_3$	0.2331	126.11	7.97	0.06
Oxalic acid	$C_2H_2O_4$	0.1148	90.04	1.38	-0.24
Maleic acid	$C_4H_4O_4$	0.1291	116.07	3.15	0.04
Malonic acid	$C_3H_4O_4$	0.1378	104.06	2.92	-0.31
Acetic acid	$C_2H_4O_2$	0.1218	60.05	4.79	-0.17
Formic acid	CH_2O_2	0.1335	46.03	3.74	-0.37
Ribose	$C_5H_{10}O_5$	0.20856	150.13	12.46	-2.39
Glucose	$C_6H_{12}O_6$	0.28356	180.16	12.45	-3.169
Sucrose	$C_{12}H_{22}O_{11}$	0.38956	342.3	12.81	-3.484
Raffinose	$C_{18}H_{32}O_{16}$	0.50256	504.42	12.81	-6.76

The solute permeation (B) was calculated from retention values and defined as follows:

$$B = \frac{1-R}{R}$$

Results and Conclusions: determining the important inputs for the model

This example is based on detect the more important variables when exist multicolineality.

Multilayer feedforward networks are often used for modeling complex relationships between the data sets. Deleting unimportant data components in the training sets could lead to smaller networks and reduced-size data vectors. The process of finding relevant data components is based on the concept of sensitivity analysis applied to a trained neural network. ANN models predict changes for certain combinations of input variables, detecting the most important influence in the output variables.

After work with both neural network MLP and RBNF, in both case the variable Mw is the less signification above the model which propose for prediction of the retention B, and is consequence of high correlation between Diameter and MW .If we are looking for a model for prediction the retention of the membrane, the most important is the variable diameter as the first to include in the model forward the variable log$P$.

Analysis of the results obtained about the weight importance in percent is listed in the tables. . MLP results are in Table 2, and in table 3 have been shown the RBF results.

Table 2 Multilayer Perceptron results (MLP)

Sensitivity of criterion %			
Variables			
MW	Diameter	pK_a	log$P$
10.182	42.634	18.170	29.014
16.357	42.664	–	38.979
–	50.497	17.565	31.938
Diameter		log$P$	
51.689		48.311	

Table 3 Radial Basis Function results (RBF)

Sensitivity of criterion %			
Variables			
MW	Diameter	pK_a	log$P$
11.709	34.042	13.720	40.529
18.939	56.182	–	24.879
–	36.545	18.174	45.281
Diameter		log$P$	
65.599		34.401	

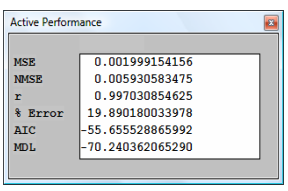
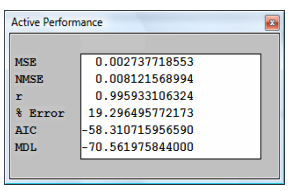
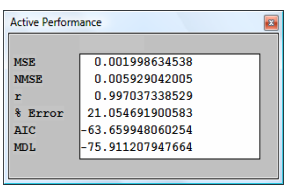
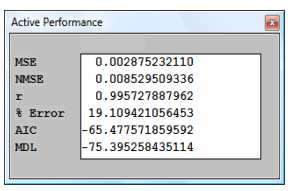
Tables 2 and 3 show how there is correspondence between the importance of the variables in percentage and the membrane retention for both variable and trained neural network.

It can be seen that, from tables 2 and 3, how the most important variable in percent % is the diameter followed by the log$P$. The pK_a is not very important and finally the MW has no influence, but this last variable is correlated with the diameter and in some type of membrane it is possible obtain confuse measure about the importance over the output. It can be also seeing how the diameter is the most important variable through the different possible combinations of models, and error is softly decreasing.

The General performance probe displays the Mean Squared Error (MSE), the Normalized Mean Squared Error (NMSE), the Correlation Coefficient (r), and the Percent Error.

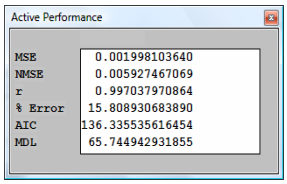
Table 3 General performance probe

MLP

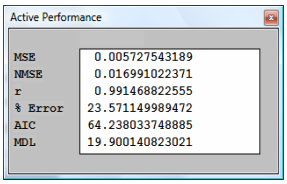
All variables	without MW	without pk_a	without MW and pk_a
			

RBF

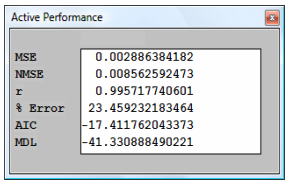
All variables



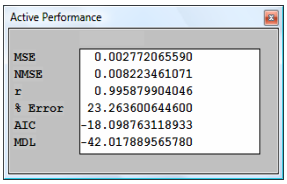
without MW



without pk_a



without MW and pk_a



Once the most important variables for the model have been determined, we can train again the neural network with three or two variables, in this case with diameter and log*P* we obtained a very good results Squared Error SME less 0.001 for prediction solute retention.

This paper presents a method for prediction. In this method, firstly the global problem is obtain the most important variables, extracted and finally the solution is globalize with a model or prediction. Two stages have been judiciously combined, which allow selected to be a more efficient, effective and easy to control process. The obtained results show that this mixed system could be applied to different situations other than the one considered in this paper, due to the general nature of the proposed solution.

Bibliography

[Anderson 1995] Anderson, James A, 1995. An Introduction to Neural Networks Cambridge, MA: MIT Press.

[Arsuaga et al 2008] Arsuaga J.M., Lopez-Muñoz M. J., Aguado J., Sotto A., 2008. Temperature pH and concentration effects on retention and transport of organic pollutants across thin-film composite nanofiltration membranes, Desalination 221. 253-258.

[Bellona C. et al 2004] Bellona, Drewes J., P. Xu and G. Amy, 2004. Factors affecting the rejection of organic solutes during NF/RO treatment. Water Res. 38. 2795-2809.

[Boussu K. et al 2008] K. Boussu C. Vandecasteele and B. Van der Bruggen, 2008. Relation between membrane characteristics and performance in nanofiltration, J. Membr. Sci. 310. 51-65.

[Hellebrand R. et al 1997] Hellebrand R., D. Mantzavinos, I. S. Metcalfe y A.G. Livingston, 1997. Integration of Wet Oxidation and Nanofiltration for Treatment of Recalcitrant Organics in wastewater, Ind. Eng. Chem. Res. 36. 5054-5062.

[Moody, J. and Darken C. (1989)]. Moody, J. and Darken C., 1989. Fast learning in networks of locally-tuned processing units. Neural Computation, 1:281-294 .

[Van der Bruggen et al 2003] Van der Bruggen B., Vandescasteele C. 2003. Removal of pollutants from surface water and groundwater by nanofiltration: overview of possible applications in the drinking water industry, Environ. Pollut. 122. 435.

Authors' Information

Sotto A. - Department of Chemical and Environmental Technology, University Rey Juan Carlos. C/ Tulipán s/n, 28933-Móstoles, Madrid, Spain. arcadio.sotto@urjc.es

Martinez A. - Natural computing group. Universidad Politécnica de Madrid, Spain. ana.martinez@upm.es

Castellanos A. - Departamento de Ciencias Básicas aplicadas a la Ingeniería Forestal. Escuela de Ingeniería Técnica Forestal. Universidad Politécnica de Madrid, Avda. de Ramiro de Maeztu s/n 28040 Madrid, Spain. angel.castellanos@upm.es