http://dx.doi.org/10.5012/bkcs.2012.33.5.1527

A New Variable Selection Method Based on Mutual Information Maximization by Replacing Collinear Variables for Nonlinear Quantitative Structure-Property Relationship Models

Jahan B. Ghasemi* **and Ehsan Zolfonoun**

Chemistry Department, Faculty of Sciences, K.N. Toosi University of Technology, Tehran, Iran
**E-mail: Jahan.ghasemi@gmail.com*
Received November 21, 2011, Accepted February 3, 2012

Selection of the most informative molecular descriptors from the original data set is a key step for development of quantitative structure activity/property relationship models. Recently, mutual information (MI) has gained increasing attention in feature selection problems. This paper presents an effective mutual information-based feature selection approach, named mutual information maximization by replacing collinear variables (MIMRCV), for nonlinear quantitative structure-property relationship models. The proposed variable selection method was applied to three different QSPR datasets, soil degradation half-life of 47 organophosphorus pesticides, GC-MS retention times of 85 volatile organic compounds, and water-to-micellar cetyltrimethylammonium bromide partition coefficients of 62 organic compounds.The obtained results revealed that using MIMRCV as feature selection method improves the predictive quality of the developed models compared to conventional MI based variable selection algorithms.

Key Words : Mutual information, Variable selection, Quantitative structure-property relationship

Introduction

Quantitative structure activity/property relationships (QSPR/QSAR) are of the most important methods in chemometrics, which attempts to correlate molecular descriptors with functions (*i.e.* physicochemical properties, biological activities, toxicity *etc.*) for a set of similar compounds, by means of statistical methods.[1-3]

For building a quantitative structure activity/property-relationship (QSAR/QSPR) model, molecular structural descriptors must be calculated for the compounds involved. Since, at the present time, thousands of molecular descriptors (including topological, geometric, electronic and quantum-chemical features) are available for QSAR/QSPR analyses and only a particular subset of the mare statistically significant in terms of correlation with activity/property for a particular analysis, a variable selection method is necessary for producing a useful predictive model.[4,5] Meanwhile, using techniques that allow the selection of a reduced set of variables containing the most informative features enables a better interpretation and comprehension of the model.[4]

Variable selection algorithms can be classified into filters, wrappers and embedded methods. Filter methods select subset of variables as a preprocessing step, independently of the model that eventually makes use of them. Filter methods use a given criterion like the value of correlation coefficient between predictors and response in order to select some variables and/or eliminate others.[6] Wrapper methods choose those features with high prediction performance estimated by specified learning algorithms. In these methods the optimum selection of variables is achieved by combining stochastic search algorithms such as simulated annealing and genetic algorithm with multivariate calibration methods such as multiple linear regressions (MLR) and partial least squares (PLS).[7,8] In the embedded methods, feature selection is integrated into the process of calibration for a given modeling algorithm. The subset of selected variables can be constructed by successive additions (forward), elimination (backward) or a combination of both approaches. Stepwise multiple linear regression[9] is the most commonly employed embedded variable selection method.

In QSPR studies, a linear regression model of the form $\mathbf{y} = \mathbf{Xb} + e$ is usually used to describe a set of predictor variables or descriptors (**X**) with a predicted variable or property (**y**) by means of a regression vector (**b**). Multiple linear regression (MLR) and partial least squares (PLS) are among mostly used linear methods in QSPR studies.[9-11] Because of the complexity of the relationships existing between the activity/property of the molecules and the structures, nonlinear modeling methods are often used to model the structure-activity/property relationships. Nonlinear multivariate maps use a nonlinear transformation of the input variable space to project inputs onto the designated attribute values in output space.[12] Artificial neural networks (ANN) and support vector machines (SVM) are nonlinear modeling techniques that have attracted increasing interest in recent years.[12,13] The ANN and SVM can incorporate nonlinear relationships between molecular descriptors and activity/property and often produce superior QSPR/QSAR models compared to models developed by the linear approaches MLR and PLS.[14,15]

Mutual Information (MI) is an alternative for the selection

of the most important variables.[16-18] Basically, the mutual information measures the amount of information contained in a variable or a group of variables, in order to predict the dependent one.[19] Unlike other parametric estimators, such as the correlation coefficient, the MI does not make any assumption about what type of relation could exist between the variables (linear or nonlinear) and estimation of MI is carried out independently from a regression model.[20]

For the selection of the most important variables based on the mutual information, two procedures have been proposed. The first one, named ranking procedure, is the selection of the variables that individually exhibit the largest MI with the dependent one.[20,21] However, this method may lead to the selection of highly collinear variables.[20,22] The second option consists of a forward procedure in which variables are sequentially added into the subset of selected variables. In this procedure, the variables that maximize the mutual information value between the set of selected variables and dependent one are selected.[20,21] The selected variables in the forward procedure are strongly dependent on the previously selected features, and thus, some important descriptors could be lost during the variable selection process.[20-22]

This paper proposes a new mutual information-based variable selection procedure, called mutual information maximization by replacing collinear variables (MIMRCV), for nonlinear quantitative structure-property relationship models. The proposed method was applied to develop QSPR models for three datasets by nonlinear approaches radial basis function neural networks (RBFNN) and support vector machines (SVM).

Theory

Mutual Information. In the information theory, the mutual information (MI) can be applied for evaluating any arbitrary dependency between random variables. In this theory, the uncertainty of a random variable is measured by entropy.[23,24] The MI between two random variables X and Y is a measure of the amount of knowledge on Y supplied by X. The MI of two random variables X and Y is defined as:

$$MI(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) \quad (1)$$

where $H(X)$ and $H(Y)$ are the entropy of variables X and Y, is the conditional entropy of X in the case of Y is known, and $H(X,Y)$ is the joint entropy of X and Y, which are defined as:

$$H(X) = -\int_x p(x)\log(p(x))dx \quad (2)$$

$$H(Y) = -\int_x p(y)\log(p(y))dx \quad (3)$$

$$H(X, Y) = -\int_x \int_y p(x, y)\log(p(x, y))dxdy \quad (4)$$

where $p(x, y)$ is the joint probability density function and $p(x)$ and $p(y)$ are marginal density functions of X and Y, respectively. The marginal density functions are

$$p(x) = -\int_y p(x, y)dx \quad (5)$$

$$p(y) = -\int_x p(x, y)dy \quad (6)$$

By substituting Eqs. (2)-(4) into Eq. (1), the MI equation will be

$$MI(X, Y) = --\int_x \int_y p(x, y)\log\frac{p(x, y)}{p(x)p(y)}dxdy \quad (7)$$

The estimation of the mutual information is generally carried out on histograms and kernel probability density functions,[25] which are derived from the available data sets. However, these estimators suffer from the curse of dimensionality[26] and their use is usually restricted to one- or two-dimensional variables. Kraskov *et al.*[27] proposed to use *k*-nearest neighbor statistics to estimate the entropies and compute mutual information. They estimated the MI between two random variables of any multi-dimensional space. The basic idea is to estimate the entropy from based on an average distance to the *k*-nearest neighbors.[27] Usually a mid-range value for number of neighbors (*k*), *i.e.*, $k = 6$, is selected for the estimation of MI.[21,28]

Proposed MI-Based Variable Selection Procedure. The proposed mutual information-based variable selection procedure can be described as follows. Let N be the maximum number of variables that can be included in the subset of selected descriptors. The main purpose of MIMRCV is to selectan optimal subset {X1, X2,…, XN} of descriptors with maximum mutual information MI({X1, X2, X3,…., XN}, Y). MIMRCV comprises two phases. The first phase consists of the selection of the variables that individually exhibit the largest MI value with the dependent variable and removing collinear variables with the selected ones from the matrix of descriptors, iteratively. The first variable to be chosen is the one that maximizes the mutual information with Y, MI(X1, Y). Then the correlation coefficient value (R) of the selected variable and the remaining ones is calculated. The variables that exhibit higher correlation coefficient value than a defined threshold (*C*) are eliminated from the original dataset. For the selection of the second variable among the remaining ones, the variable that has the largest mutual information with Y (MI(X2, Y)) is selected and all the variables that have higher correlation coefficient value than *C* with the selected one are removed from the remaining descriptors and saved for the second phase. This procedure is continued until the selection of variable XN.

The second phase of MIMRCV consists of replacing each one of selected variables (except variable X1) by corresponding collinear one that maximizes the mutual information of the selected subset with the dependent one (MI ({X1, X2, X3….XN}, Y)). Firstly, the variable with the lowest mutual information (XN) is replaced with itscollinear variables (the variables that have higher correlation coefficient value than *C* with variable XN) that removed from the data set in the previous phase and keeping the best one. If the replacement of the descriptor does not improve the value of MI, it remains unchanged. In the next step, variable XN–1 is

replaced by the same method. All the remaining variables in the initial set ({XN–2, XN–3,....,X2}) are replaced in the same way except variable X1. When finishing, above procedure starts again with the variable XN and the whole process is repeated. This process is repeated as many times as necessary until the set of descriptors remains unchanged.

Radial Basis Function Neural Networks. Artificial neural networks (ANN) are well known methods for solving nonlinear problems.[7,29,30] An ANN consists of a series of interconnected nodes (neurons) that receive and/or send number values to other nodes. The radial basis function-neural network (RBFNN) is a particular type of ANN applied to problems such as modeling and classification.[31] Recently, there is a growing interest in the use of RBFNN for its short training time and being guaranteed to reach the global minimum of error surface during training.[32] In RBFNN, the input layer does not process the information; it only distributes the input vectors to the hidden layer. The hidden layer consists of a number of RBF neurons and a bias (b_k). Each neuron in the hidden layer employs a radial basis function as the nonlinear transfer function to operate on the input data. A common RBF is the Gaussian function that is characterized by the center (c_j) and the width (r_j):

$$H_j(x) = \exp\left[-\left(\frac{\|x-c_j\|}{r_j}\right)^2\right] \quad (8)$$

where H represents the radial basis function, and $\|x-c_j\|$ is the Euclidean distance between x input vector and c_j. The outputs from the radial functions are fully connected to the neurons of the output layer by the strength of weight coefficients w_{jk}. The relation between the output value and the input variable can be represented by:

$$y_k(x) = \Sigma_{j=1}^{n_k} w_{kj} h_j(x) + b_k \quad (9)$$

Where y_k is the kth output unit for the input vector x, w_{kj} is the weight connection between the kth output unit and the jth hidden layer unit and b_k is the bias. When the error of network output reaches the pro-set error goal value in RBFNN, the procedure of adding hidden neurons will stop.

Support Vector Machines. Support vector machine (SVM) introduced by Vapnik[33,8] is a valuable tool for solving pattern recognition and classification problem. SVM can be applied to regression problems by the introduction of an alternative loss function. Due to its advantages and remarkable generalization performance over other methods, SVM has attracted attention and gained extensive application.[8] In support vector machine, the input data is first mapped into high dimensional feature space by the use of kernel function and then linear regression is performed in the feature space. The nonlinear feature mapping will allow the treatment of nonlinear problems in a linear space. After training on set data SVM can be used to predict the objects whose values are unknown. The prediction or approximation function used by SVM is:

$$y = \Sigma_{i=1}^{N} a_i k(x_i, x) + b \quad (10)$$

Where k (x_i, x) is the kernel function, x_i is the input vector, a_i is Lagrange multipliers called support value, b is bias term. Training points with nonzero weight a_i are called support vectors. The elegance of using a kernel function lies in the fact that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\Phi(x)$ explicitly, and it may be useful to think of the kernel, $K(x_i, x)$ as comparing patterns or evaluating the proximity of objects in their feature space. Thus, a test point is evaluated by comparing it with all training points. Many functions can be used as the kernel function. However, the kernel function more used is the radial basis function (RBF), $\exp(-(\|x_i - x\|^2)/2\sigma^2)$, a simple Gaussian function, where σ^2 is the width of the Gaussian function, which should be optimized by the user, to obtain the support vector.

Experimental

Datasets.

Degradation Half-life of Organophosphoruspesticides in Soil: Organophosphorus pesticides (OPPs) are located among the most commonly employed pesticides because of their high activity, ease of use and rapid degradation under natural conditions.[34,35] They have been widely applied as insecticides, herbicides, acaricides, fungicides, and plant growth regulators for controlling disease and growth.[36,37] Organophosphorus pesticides are generally depredated by the reactions including oxidation, reduction, hydrolysis, hydroxylation, dehydrochlorination, dealkylation, methylation, isomerization, and conjugate formation.[38] Their ability to degrade made them an attractive alternative to the persistent organochlorines pesticides, such as DDT, aldrin and dieldrin. Although the degradation process of pesticides in soils is complicated, main factors may be soil constituents, soil microflora, and chemical structures of pesticides. Chemical structures are especially important for soil metabolism of organophosphorus pesticides, because the priority of the reactions mentioned above is decided.[38] The soil degradation half-life (DT50) is a measure of the persistence of a pesticide in soil. Pesticides can be categorized on the basis of their half life as non-persistent, degrading to half the original concentration in less than 30 days; moderately persistent, degrading to half the original concentration in 30 to 100 days; or persistent, taking longer than 100 days to degrade to half the original concentration.[39]

The experimental values for the degradation half-life (DT_{50}) of 47 organophosphoruspesticides taken from the literature[39] are presented in Tables 1. The dataset was split into a training set and a test set. The training set of 37 compounds was used to adjust the parameters of the models, and the test set of 10 compounds was used to evaluate their prediction ability.

Gas Chromatography-Mass Spectrometry Retention Times of Volatile Organic Compounds: Volatile organic compounds (VOCs) are a large group of carbon-based chemicals that have a highvapor pressure at ordinary, room-temperature conditions. VOCs are introduced into the

Table 1. Experimental and predicted values of DT_{50} of 47 organophosphorus pesticides for training and test sets

No.	Pesticide	DT_{50} (Exp.)	DT_{50} (Pred.)		No.	Pesticide	DT_{50} (Exp.)	DT_{50} (Pred.)	
			SVM	RBFNN				SVM	RBFNN
1	Acephate	3	5.3	6.9	25	Methyl parathion	5	7.8	7.3
2	Azinphos-methyl	10	13.9	13.7	26	Mevinphos	3	8.2	10.8
3[a]	Bensulide	120	110	94.2	27	Monocrotophos	30	11.3	8.6
4[a]	Carbophenothion	30	41.8	47.8	28	Naled	1	4.9	0.46
5	Chlorpyrifos	30	33.9	26.0	29	Oxydemeton methyl	10	13.4	5.9
6	Chlorpyrifos-methyl	7	9.1	10.9	30	Parathion	14	10.0	10.3
7	Demeton	15	16.97	18.98	31	Phenthoate	35	32.5	31.0
8	Diazinon	40	43.9	38.0	32[a]	Phorate	60	45.0	75.0
9	Dichlorvos	0.5	3.49	4.48	33	Phosalone	21	24.6	25.1
10	Dicrotophos	20	16.0	15.9	34	Phosmet	19	16.7	15.0
11[a]	Dimethoate	7	7.5	20.3	35	Phosphamidon	17	17.1	20.9
12	Disulfoton	30	33.9	33.7	36[a]	Profenofos	8	32.6	6.5
13	Ethephon	10	13.7	13.9	37	Sulprofos	140	125.3	126.0
14	Ethion	150	164.0	146.0	38	Temephos	30	34.0	33.4
15[a]	Ethoprop	25	17.8	21.3	39	Terbufos	5	8.9	28.1
16[a]	Fenamiphos	50	56.9	30.3	40	Tetrachlorvinphos	2	5.9	4.6
17[a]	Fenitrothion	4	10.8	1.6	41	Tribufos	10	11.9	13.9
18	Fensulfothion	30	33.9	43.4	42[a]	Trichlorfon	10	3.9	9.3
19[a]	Fenthion	34	29.9	48.1	43	Trichloronate	139	118.2	112.6
20	Fonofos	40	46.9	43.9	44	Glyphosate	16	18.1	12.0
21	Isazofos	34	32.7	30.0	45	Pirimiphos-methyl	10	13.9	22.5
22	Isofenphos	150	126.7	155.0	46	Pirimiphos-ethyl	45	55.9	48.9
23	Methamidophos	6	9.9	2.0	47	Tolclofos-methyl	30	26.0	27.3
24	Methidathion	7	10.9	10.9					

[a]Test set

atmosphere *via* a wide range of anthropogenic, biogenic and photochemical sources. These compounds can pose a serious hazard to human health and the environment due to the well-known toxicity of several compounds (*e.g.*, benzene and 1,3-butadiene).[40] They also play an important role in physico-chemical processes of the troposphere, as they contribute significantly to the formation of ozone and other photochemical oxidants.[41] Gas chromatography-mass spectrometry retention times (t_R) of 85 volatile organic compounds were taken from theEPA Method 8260C.[42] The dataset was divided into two subsets: 65 in the training setand 20 in the test set.

Water-To-Micellar CTAB Partition Coefficients of Some of Organic Compounds: Surfactants are known to play a vital role in many processes of interest in both fundamental and applied science. One important property of surfactants is the formation of colloidal-sized clusters in solutions, known as micelles. Dissolved solutes in micellar solutions are distributed between the micelles and the bulk aqueous solvent medium. The enhanced solubility that results from solute partitioning into the micellar aggregates has been used successfully in practical applications such as tertiary oil recovery, design of controlled drug delivery systems, remediation of contaminated waste sites, removal of hazardous materials from industrial waste effluents, and chemical separations by micellarelectrokinetic chromatography.[43] The experimental data of the logarithm of water-to-micellar cetyltrimethylammonium bromide partition coefficients (Log $P_{CTAB/water}$) of 62 organic compoundswere taken from Sprunger *et al*.[43] The total 62 samples are split into a training set with 50 samples and a test set with 12 samples.

Computer Hardware and Software. All calculations were run on a Toshiba computer with Pentium IV as central processing unit (4Gb RAM) with windows XP operating system. The ChemDrawUltra version 11.01 (ChemOffice 2008, CambridgeSoft Corporation) software was used for drawing the molecular structures. The optimizations of molecular structures were done by the HyperChem version 8.05 using molecular mechanics and semi empirical AM1 tools. For the calculation of molecular descriptors, Dragon (Milano Chemometrics group, version 3.0) software's was used. All data analyses were performed using MATLAB software, version 7.7 (Mathworks, Inc.). SVM regression (PLS Toolbox, version 6, Eigenvector Company) and RBFNN analysis (Neural Network Toolbox) were performed in the MATLAB. The mutual information was calculated with a MATLAB/C implementation provided by Astakhov *et al*.[44]

Molecular Descriptors. The main step in every QSPR study is calculating the structural descriptors as numerical encoded parameters representing the chemical structures. In the present work the molecular descriptors were generated using Dragon software. Descriptors with constant or near constant values inside each group were discarded. In addition, pairs of variables with a correlation coefficient greater

than 0.95 were classified as intercorrelated and only one of them were considered in developing the models. The final number of descriptors was 505, 497 and 477 for the DT_{50}, Log $P_{CTAB/water}$ and t_R datasets, respectively.

Model Development and Validation. The quality of each model was assessed by applying the *k*-fold cross-validation procedure. In *k*-fold cross-validation, the data is first partitioned into *k* equally (or nearly equally) sized segments or folds. The regression model will then be trained and tested *k* times. Each time the model is built using (*k*–1) folds as the training sample and the remaining single fold is retained for testing.

For the evaluation of the performance of multivariate calibration models, the root mean square error (RMSE) can be used:

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]^{0.5} \quad (11)$$

The square of the correlation coefficient (R^2), which indicates the quality of fit of all the data to a straight line is calculated for the checking of each calibration, and is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (12)$$

where $\hat{y}_i$ is the estimated value of the ith object and y_i is the corresponding reference value of this object, $\bar{y}$, is the mean of reference values and *n* is the total number of objectsin the corresponding set.

Results and Discussion

Variable Selection with Mutual Information. A major step in QSPR studies is the selection of minimum number of variables which tightly describe dependencies between the chemical structures of compounds and their property. In most of reported QSPR studies, the selection of variables for nonlinear models has been performed using linear models such as MLR and PLS.[4,5] However, these methods may not be appropriate if there was some kind of nonlinearity in the data. In this context, mutual information (MI), a measure that captures linear and nonlinear relationships between variables, is preferable.

In this work, MIMRCV as a nonlinear variable selection method was applied in order to select the most suitable molecular descriptors to describe the degradation half-life of organophosphorus pesticides in soil, water-to-micellar CTAB partition coefficients of organic compounds and GC-MS retention times of volatile organic compounds. The mutual information estimation is conducted with $k = 6$ nearest neighbors. For the selection of the most important descriptors by the proposed algorithm, the parameter *C* (threshold for correlation coefficient) needs to be optimized.

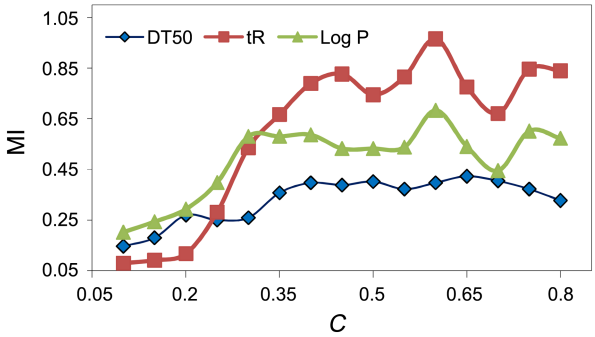

Figure 1. Influences of the parameter *C* on the MI value of the selected descriptors.

For this purpose, this value was varied from 0.1 to 0.8, and after the selection of variables for the each dataset, the mutual information of the variables X2, X3….XN with the dependent one (MI ({X2, X3….XN},Y)) was measured.The obtained results are exhibited in Fig. 1. As can be seen in Figure 1, the maximum mutual information value of the selected subsets was obtained at $C = 0.65$, 0.60 and 0.60 for the DT_{50}, Log $P_{CTAB/water}$ and t_R datasets, respectively. At lower *C* values there was a decrease in the MI value of the selected variables, probably due to losing some informative descriptors. The mutual information value decreases at higher *C* values, probably due to the interrelations between the selected descriptors.

The optimal number of variables was determined by the maximum value of mutual information of the variables X2, X3….XN with the dependent one (MI ({X2, X3….XN}, Y)). Figure 2 shows MI value as a function of the number of selected variables. As can be seen, MI value increases with the number of variables in the range of 3-10, 3-6 and 3-5 for the DT_{50}, Log $P_{CTAB/water}$ and t_R datasets, respectively. Subsequent addition of variables does not improve the MI value. Hence 10, 6 and 5 were selected as the optimum number of variables. Table 2 presents the notation and a short description of the selected molecular descriptors.

Figure 3 shows MI value as a function of the number of steps in the second phase (replacement of the selected variables with the collinear ones). The graph reveals that the MI value increases with replacing variables until step 8, 4

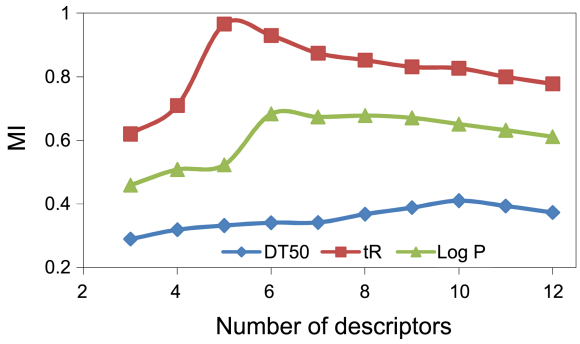

Figure 2. MI as a function of the number of descriptors ($C = 0.65$, 0.60 and 0.60 for the DT_{50}, Log $P_{CTAB/water}$ and t_R datasets, respectively).

Table 2. Variables selected by the MIMRCV procedure

Data set	Name	Description
DT_{50}	BELm3	Lowest eigenvalue n. 3 of Burden matrix/weighted by atomic masses
	MAXDP	Maximal electrotopological positive variation
	Mor31u	3D-MoRSE-signal 31/unweighted
	GATS3e	Geary autocorrelation-log 3/weighted by atomic Sanderson electronegativities
	RDF075p	Radial Distribution Function-7.5/weighted by atomic polarizabilities
	GATS3v	Geary autocorrelation-log 3/weighted by atomic van der waals volume
	RDF055u	Radial Distribution Function-5.5/unweighted
	Mv	Mean atomic van der waals volume (scaled on Carbon atom)
	G3u	3st component symmetry directional WHIM index/unweighted
	BEHv6	Highest eigenvalue n. 6 of Burden matrix/weighted by atomic van der waals volume
t_R	X1sol	Solvation connectivity index chi-1
	Ss	Sum of Kier-Hall electrotopological states
	H4m	H autocorrelation of lag 4/weighted by atomic masses
	BEHm2	Highest eigenvalue n. 2 of Burden matrix/weighted by atomic masses
	TI1	First Mohar index TI1
Log $P_{CTAB/water}$	MLogP	Moriguchioctanol-water partition coefficient
	Mv	Mean atomic van der waals volume (scaled on Carbon atom)
	Mor18m	3D-MoRSE-signal 18/weighted by atomic masses
	H2m	H autocorrelation of lag 2/weighted by atomic masses
	H5u	H autocorrelation of lag 5/unweighted
	CIC0	Complementary information content

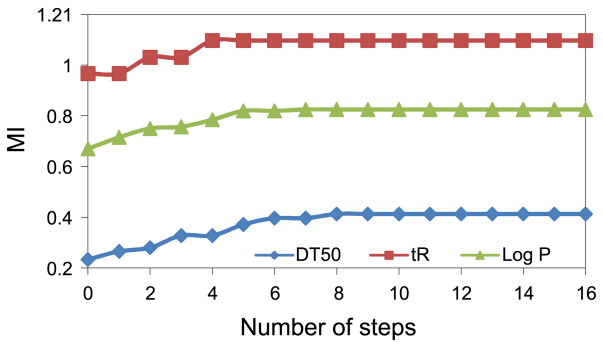

Figure 3. MI *vs*. number of steps in the second phase.

and 7 for the DT_{50}, Log $P_{CTAB/water}$ and t_R datasets, respectively. Afterward, the replacement of descriptors does not improve the value of MI and therefore set of descriptors remains unchanged.

For comparison purposes, both ranking and forward mutual information-based variable selection procedures were also employed to select variables for modeling of above datasets. In the ranking procedure, 10, 6 and 5 variables with the highest MIwere selectedfor the DT_{50}, Log $P_{CTAB/water}$ and t_R datasets, respectively, while in forward procedure, the optimal number of variables was determined by the maximum value of MI. In addition, a comparison is performed with genetic algorithm (GA), which is a commonly used technique for the selection of variables in the QSPR studies.[14]

SVM Modeling. In this work, SVM was performed with radial basis function (RBF) as a kernel function. To determine the optimum values of the kernel width and penalty constant a grid search was performed based on leave-one-out cross-validation on the original training set.

The quality of each model was assessed by applying the 5-fold cross-validation procedure. The statistical parameters for four different variable selection methods are presented in Table 3 to compare the performance of the models. It can be seen from this table that for the DT_{50} dataset, the statistical parameters of MIMRCV-SVM model are superior to that of other models. The best model is obtained with MIMRCV-SVM with the highest R^2 value of 0.891 for the test set and the lowest RMSEP value (11.293). The value of the determination coefficient of the 5-fold cross-validation for the model obtained with the MIMRCV-SVM method (R^2_{CV} = 0.801) is higher than GA-SVM model (R^2_{CV} = 0.729). Two other mutual information based-variables election methods (ranking and forward selection) do not have satisfactory results for the prediction of soil half-life of OPPs. Ranking yielded the worst model for the OPPs degradation half-life data set with a very low R^2 value of 0.313 and a very high RMSEP value of 35.420 for the prediction set. Forward MI based variable selection results are slightly better than those of ranking variable selection. In the case of the Log $P_{CTAB/water}$ dataset, GA gives better results than other variables selection methods. The coefficient of determination for the test set by MIMRCV-SVM is 0.962 and the root mean square error (RMSE) is 0.324. Ranking and forward procedures are also not very satisfactory with poor test results. MIMRCV-SVM method yielded the best model for the GC-

Table 3. Statistical parameters for SVM models

Data set	Parameter	Forward	Ranking	GA	MIMRCV
DT_{50}	Number of descriptors	4	10	11	10
	$R^2_{Training\ set}$	0.681	0.782	0.903	0.964
	R^2_{CV}	0.401	0.313	0.729	0.801
	RMSECV	29.731	35.420	22.122	20.051
	$R^2_{Test\ set}$	0.368	0.310	0.836	0.891
	RMSEP	29.584	29.566	14.248	11.293
t_R	Number of descriptors	5	5	7	5
	$R^2_{Training\ set}$	0.965	0.964	0.972	0.975
	R^2_{CV}	0.915	0.950	0.943	0.964
	RMSECV	2.994	2.330	2.461	1.981
	$R^2_{Test\ set}$	0.920	0.958	0.965	0.972
	RMSEP	3.365	2.755	2.308	2.090
Log $P_{CTAB/water}$	Number of descriptors	3	6	10	6
	$R^2_{Training\ set}$	0.948	0.943	0.984	0.971
	R^2_{CV}	0.928	0.912	0.976	0.956
	RMSECV	0.289	0.319	0.170	0.190
	$R^2_{Test\ set}$	0.922	0.938	0.979	0.962
	RMSEP	0.482	0.416	0.238	0.324

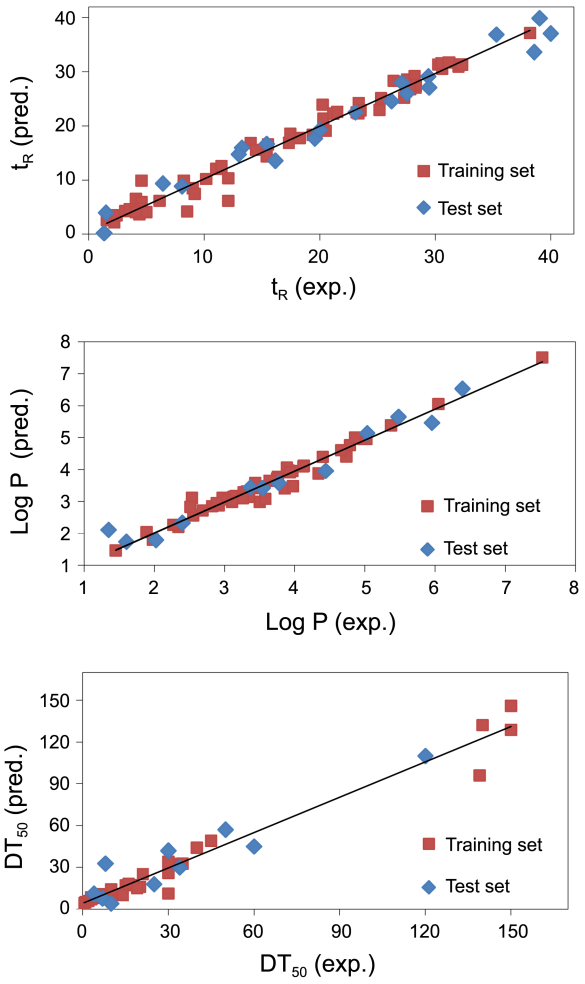

Figure 4. Estimated *versus* experimental values of DT_{50}, Log $P_{CTAB/water}$ and t_R using SVM modeling for the training and prediction sets.

MS retention time data with R^2 value of 0.972 and RMSE value of 0.192 for the test set. The value of the determination coefficient of the 5-fold cross-validation for the model obtained with the MIMRCV-SVM method (R^2_{CV} = 0.964) is slightly higher than GA-SVM model (R^2_{CV} = 0.943). The forward and ranking methods give R^2 values of 0.920 and 0.958 for the test set, respectively. In Figure 4, the plots of the predicted DT_{50}, Log $P_{CTAB/water}$ and t_R by the regression models *versus* experimental DT_{50}, Log $P_{CTAB/water}$ and t_R are represented. The agreement between observed (experimental) and predicted values, high correlation coefficient, low RMSE confirms the good predictive ability of MIMRCV-SVM modeling.

RBFNN Modeling. In RBFNN, the number of neurons on the hidden layer and the width of the radial basis function (spread) are the two important parameters affecting the performance of RBFNN. In order to find the optimum values of these two parameters and prohibit the over fitting of the model, leave-one-out cross-validation of the whole training set was performed.

The comparative data for the statistical parameters of ranking and forward MI based variable selection methods and also genetic algorithm followed by radial basis function neural network and those for the proposed method are summarized in Table 4. For the DT_{50} dataset, a comparison between the results obtained by the MIMRCV, GA, forward and ranking variable selection methods clearly indicates the superiority of MIMRCV-RBFNN over that of the other models. As can be seen from Table 4 the statistical parameters of MIMRCV-RBFNN model are superior to GA-RBFNN model. Two other mutual information based-variable selection methods are also not very satisfactory with poor results. Compared to ranking, forward method offers

Table 4. Statistical parameters for RBFNN models

Data set	Parameter	Forward	Ranking	GA	MIMRCV
DT_{50}	Number of descriptors	4	10	11	10
	$R^2_{Training\ set}$	0.857	0.815	0.934	0.956
	R^2_{CV}	0.561	0.490	0.701	0.763
	RMSECV	28.937	31.552	23.916	20.59
	$R^2_{Test\ set}$	0.611	0.528	0.725	0.826
	RMSEP	21.633	23.144	17.638	14.134
t_R	Number of descriptors	5	5	7	5
	$R^2_{Training\ set}$	0.970	0.968	0.979	0.974
	R^2_{CV}	0.90	0.942	0.940	0.957
	RMSECV	3.408	2.537	2.613	2.221
	$R^2_{Test\ set}$	0.901	0.915	0.952	0.960
	RMSEP	3.844	3.571	2.657	2.384
Log $P_{CTAB/water}$	Number of descriptors	3	6	10	6
	$R^2_{Training\ set}$	0.902	0.912	0.986	0.985
	R^2_{CV}	0.877	0.862	0.942	0.936
	RMSECV	0.391	0.401	0.257	0.264
	$R^2_{Test\ set}$	0.883	0.881	0.975	0.967
	RMSEP	0.604	0.638	0.273	0.349

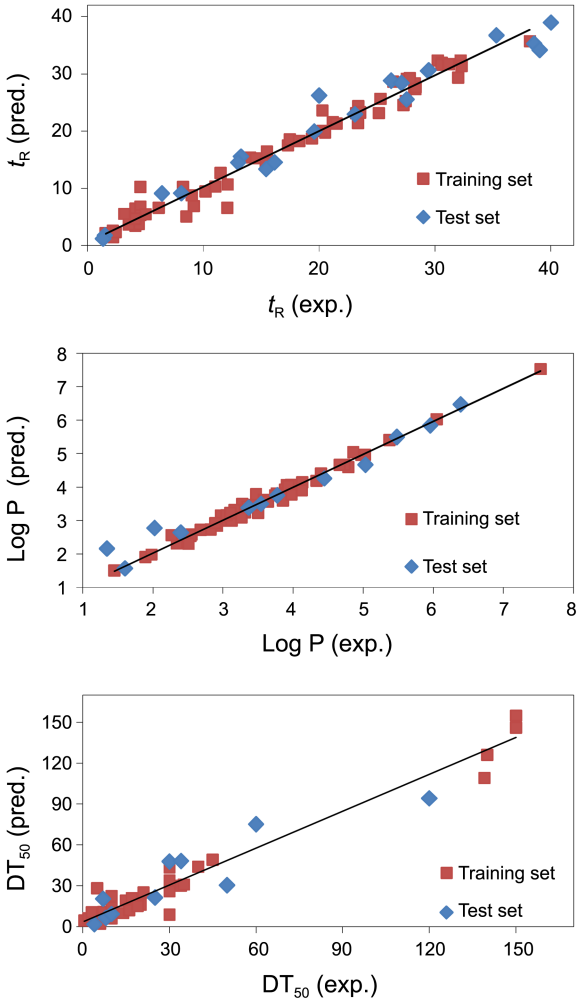

Figure 5. Estimated *versus* experimental values of DT_{50}, Log $P_{CTAB/water}$ and t_R using RBFNN modeling for the training and prediction sets.

better results. In the case of the Log $P_{CTAB/water}$ data set, the coefficient of determination (R^2) values of forward and ranking methods for the test set and also 5-fold cross validation were lower than those of MIMRCV method. The coefficient of determination obtained by GA method was higher than of other variable selection methods. For the GC-MS retention time data, MIMRCV-RBFNN gives highest R^2 and lowest RMSE values, so this model gives the most satisfactory results, compared with there sults obtained from forward, ranking and GA methods. In Figure 5, the plots of the predicted values of the DT_{50}, Log $P_{CTAB/water}$ and t_R by the MIMRCV-RBFNN models *versus* experimental DT_{50} Log $P_{CTAB/water}$ and t_R are represented.

Conclusion

In this paper, an effective feature selection based on mutual information for nonlinear QSPR models is proposed. In comparison with the parametric estimators such as correlation coefficient, mutual information has the unique advantage to be model independent and nonlinear at the same time. The proposed MI based variable selection algorithm was applied for modeling of degradation half-life of organo phosphorus pesticides in soil, GC-MS retention times of volatile organic compounds, and water-to-micellar CTAB partition coefficient of organic compounds by using SVM and RBFNN regression techniques. The QSPR models obtained by MIMRCV showed the better statistical parameters than the other MI based variable selection methods and also genetics algorithm. The results show the strong potential of MIMRCV, as a nonlinear feature selection method, to be applied to solve descriptor selection problem in QSAR/QSPR studies.

References

1. Livingstone, D. J. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195.
2. Ghasemi, J.; Saaidpour, S. *Anal. Chim. Acta* **2007**, *604*, 99.
3. Chen, K. X.; Li, Z. G.; Xie, H. Y.; Gao, J. R.; Zou, J. W. *Eur. J. Med. Chem.* **2009**, *44*, 4367.
4. Mercader, A. G.; Duchowicz, P. R.; Fernández, F. M.; Castro, E. A. *J. Chem. Inf. Model* **2010**, *50*, 1542.
5. Shamsipur, M.; Zare-Shahabadi, V.; Hemmateenejad, B.; Akhond, M. *Anal. Chim. Acta* **2009**, *646*, 39.
6. Jouan-Rimbaud, D.; Walczack, B.; Massart, D.; Last, I.; Prebble, K. *Anal. Chim. Acta* **1995**, *304*, 285.
7. Ghasemi, J.; Abdolmaleki, A.; Mandoumi, N. *J. Hazard. Mater.* **2009**, *161*, 74.
8. Guptaa, V. K.; Khanic, H.; Ahmadi-Roudid, B.; Mirakhorlic, S.; Fereyduni, E.; Agarwale, S. *Talanta* **2011**, *83*, 1014.
9. Ghasemi, J.; Asadpour, S.; Abdolmaleki, A. *Anal. Chim. Acta* **2007**, *588*, 200.
10. Deswal, S.; Roy, N. *Eur. J. Med. Chem.* **2006**, *41*, 1339.
11. Xia, B.; Ma, W.; Zheng, B.; Zhang, X.; Fan, B. *Eur. J. Med. Chem.* **2008**, *43*, 1489.
12. Blank, T. B.; Brown, S. D. *Anal. Chem.* **1993**, *65*, 3081.
13. Vapnik, V. *Statistical Learning Theory*; John Wiley: New York, 1998.
14. Pourbasheer, E.; Riahi, S.; Ganjali, M. R.; Norouzi, P. *Eur. J. Med. Chem.* **2010**, *45*, 1087.
15. Hemmateenejad, B.; Shamsipur, M.; Miri, R.; Elyasi, M.; Foroghini, F.; Sharghi, H. *Anal. Chim. Acta* **2008**, *610*, 25.
16. Benoudjita, N.; François, D.; Meurensc, M.; Verleysen, M. *Chemom. Intell. Lab. Syst.* **2004**, *74*, 243.
17. Amiri, F.; Rezaei Yousefi, M.; Lucas, C.; Shakery, A.; Yazdani, N. *J. Netw. Comput. Appl.* **2011**, *34*, 1184.
18. Liu, H.; Sun, J.; Liu, L.; Zhang, H. *Pattern Recogn.* **2009**, *42*, 1330.
19. Huang, D.; Chow, T. W. S. *Neurocomputing* **2005**, *63*, 325.
20. Rossi, F.; Lendasse, A.; François, D.; Wertz, V.; Verleysen, M. *Chemom. Intell. Lab. Syst.* **2006**, *80*, 215.
21. Durand, A.; Devos, O.; Ruckebusch, C.; Huvenne, J. P. *Anal. Chim. Acta* **2007**, *595*, 72.
22. Caetano, S.; Krier, C.; Verleysen, M.; Vander Heyden, Y. *Anal. Chim. Acta* **2007**, *602*, 37.
23. Eckschlager, K.; Danzer, K. *Information Theory in Analytical Chemistry*; John Wiley and Sons: Wiley Interscience, 1994.
24. Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; Wiley: New Jersey, 2005.
25. Kojadinovic, I. *Comput. Stat. Data Anal.* **2005**, *49*, 1205.
26. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer-Verlag: New York, 2001.
27. Kraskov, A.; Stogbauer, H.; Grassberger, P. *Phys. Rev. E* **2004**, *69*, 066138.
28. Harald, S.; Alexander, K.; Sergey, A. A.; Peter, G. *Phys. Rev. E* **2004**, *70*, 066123.
29. Despagne, F.; Massart, D. L. *Analyst* **1998**, *123*, 157.
30. Perez-Marin, D.; Garrido-Varo, A.; Guerrero, J. E. *Talanta* **2007**, *72*, 28.
31. Park, J.; Sandberg, I. W. *Neural Comput.* **1993**, *5*, 305.
32. Akhlaghi, Y.; Kompany-Zareh, M. *J. Chemom.* **2006**, *20*, 1.
33. Cortes, C.; Vapnik, V. *Mach. Learn.* **1995**, *20*, 273.
34. Zvinavashe, E.; Du, T.; Griff, T.; Van den berg, H. H. J.; Soffers, J. Vervoort, A. E. M. F.; Murk, A. J.; Rietjens, I. M. C. M. *Chemosphere* **2009**, *75*, 1531.
35. FAO, Agriculture Towards 2010; C 93/24 Document of 27th Session of FAO Conference: Rome, 1993.
36. Cai, C. P.; Liang, M.; Wen, R. R. *Chromatographia* **1995**, *40*, 417.
37. Yan, D.; Jiang, X.; Xu, S.; Wang, L.; Bian, Y.; Yu, G. *Chemosphere* **2008**, *71*, 1809.
38. Tomizawa, L. *Environ. Qual. Saf.* **1975**, *4*, 117.
39. Vogue, P. A.; Kerle, E. A.; Jenkins, J. J. *National Pesticide Information Center*; OSU Extension Pesticide Properties Database, 1994.
40. Forst, L.; Conroy, L. M. In Rafson, H. J., Ed.; *Odor and VOC Control Handbook*; McGraw-Hill: New York, 1998; p 3.1.
41. Calvert, J. G. Chemistry for the 21st Century. *The Chemistry of the Atmosphere: Its Impact on Global Change*; Blackwell Scientific Publications: Oxford, 1994.
42. EPA Method 8260C: Volatile organic compounds by Gas chromatography-mass/spectrometry (GC/MS), 2006.
43. Sprunger, L. M.; Gibbs, J.; Acree, W. E.; Abraham, M. H. *QSAR Comb. Sci.* **2009**, *28*, 72.
44. Astakhov, S. A.; Grassberger, P.; Kraskov, A.; Stögbauer, H. MILCA algorithm, available at http://www.klab.caltech.edu/kraskov/MILCA/ index.html.