Radial Basis Function Network-Based Transform for a Nonlinear Support Vector Machine as Optimized by a Particle Swarm Optimization Algorithm with Application to QSAR Studies

Li-Juan Tang, Yan-Ping Zhou, Jian-Hui Jiang,* Hong-Yan Zou, Hai-Long Wu, Guo-Li Shen, and Ru-Qin Yu*

State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, P. R. China

Received February 2, 2007

The support vector machine (SVM) has been receiving increasing interest in an area of QSAR study for its ability in function approximation and remarkable generalization performance. However, selection of support vectors and intensive optimization of kernel width of a nonlinear SVM are inclined to get trapped into local optima, leading to an increased risk of underfitting or overfitting. To overcome these problems, a new nonlinear SVM algorithm is proposed using adaptive kernel transform based on a radial basis function network (RBFN) as optimized by particle swarm optimization (PSO). The new algorithm incorporates a nonlinear transform of the original variables to feature space via a RBFN with one input and one hidden layer. Such a transform intrinsically yields a kernel transform of the original variables. A synergetic optimization of all parameters including kernel centers and kernel widths as well as SVM model coefficients using PSO enables the determination of a flexible kernel transform according to the performance of the total model. The implementation of PSO demonstrates a relatively high efficiency in convergence to a desired optimum. Applications of the proposed algorithm to QSAR studies of binding affinity of HIV-1 reverse transcriptase inhibitors and activity of 1-phenylbenzimidazoles reveal that the new algorithm provides superior performance to the backpropagation neural network and a conventional nonlinear SVM, indicating that this algorithm holds great promise in nonlinear SVM learning.

1. INTRODUCTION

A support vector machine (SVM) is a relatively novel machine learning technique based on a statistical learning theory (SLT) principle proposed by Vapnik and co-workers in 1995.[1,2] Compared with other learning machines, such as artificial neural networks (ANN) and a fuzzy learning machine (FLM), a SVM boasts its structural risk minimization (RSM) principle and has a desirable generalization performance. A SVM has demonstrated a good performance in model estimation problems by numerous successful applications.[3−7] It has also shown great promise in quantitative structure−activity relation (QSAR) studies due to its ability to interpret the nonlinear relationships between molecular structure and bioactivities.[8−11]

To solve the problem of nonlinearity, generally SVM uses a device called kernel mapping to transform the data from the original variables to a feature space in which the model becomes linear, followed by a linear SVM technique to get the solution. Gaussian radial basis function transform with consistent kernel width is frequently utilized when the exact nonlinear model is unknown. Unfortunately, selection of support vectors and intensive optimization of kernel width in conventional SVM are inclined to get trapped into local optima. Moreover, a consistent kernel width also limits the flexibility of the kernel transform to adaptively approximate the unknown nonlinear model. The kernel width reflects the interaction between support vectors. If it is too small, then the interaction between support vectors would be weak, resulting in an inferior generalization performance, while too large a kernel width cannot ensure the accuracy of a model due to too strong an interaction between support vectors. Therefore, proper selection of support vectors and kernel width plays a crucial role in determining the approximation accuracy and generalization performance. In cases where insufficient or excessive support vectors are utilized in kernel transform, conventional SVM is still exposed to a substantial risk of underfitting or overfitting. On the other hand, the number of free parameters in the SVM model is equal to the number of support vectors. Training a SVM becomes computationally intensive, even prohibitive, when the training set is too large. Development of a new effective approach to select the parameters in kernel transform is of considerable significance for combating these problems and improving the learning and generalization performance of SVM.

In the present study, radial basis function network-based transform for a nonlinear support vector machine (RBFN-SVM) is proposed. Radial basis function network (RBFN) is a kind of feedforward neural network[12] that is introduced in the solution of the real nonlinear interpolation problem. In the proposed algorithm RBFN is employed to implement the nonlinear kernel transform from the original variables to the feature ones. The original variables are used as the input of RBFN with one input and one hidden layer, and

* Corresponding author phone: +86-731-8822577; fax: +86-731-8822782; e-mail: jianhuijiang@hnu.cn (J.-H.J.), rqyu@hnu.cn (R.-Q.Y.).

10.1021/ci700047x CCC: $37.00
Published on Web 06/08/2007

the outputs of hidden nodes are obtained as the feature variables. Then, a linear model in the feature space is established using the SVM learning criterion. For effectively searching the optimal parameters in a RBFN and a linear model, a population stochastic optimization technique, particle swarm algorithm (PSO),[13−16] is invoked. Previous studies have shown that PSO works well in different optimization problems and has a relatively high efficiency in convergence to desirable optima.[17−20] With optimization of the RBFN parameters and the SVM model relating the feature variables to the dependent ones using PSO, a very flexible kernel transform with adaptive selection of kernel centers and widths in terms of the SVM model error can be accomplished. The proposed RBFN-SVM algorithm has been applied to QSAR studies of two data sets, 2-amino-6-arylsulfonylbenzonitriles and their thio and sulfinyl congeners as HIV-1 reverse transcriptase (RT) inhibitors[21] and 1-phenylbenzimidazoles as ATP-site inhibitors of the platelet derived growth factor receptor[22,23] (PDGFR). The results have demonstrated that the proposed method can converge quickly toward the optimum and avoid the overfitting or underfitting as compared to a conventional nonlinear SVM algorithm that exhibits a deteriorated performance due to a greedy search of optimized kernel centers and widths, indicating that the RBFN-SVM algorithm holds great promise in nonlinear SVM learning.

2. THEORY

2.1. Support Vector Machine (SVM) Regression. The basic theory of SVM will be briefly reviewed in the following. Consider the problem of approximating the set of data with a linear function

$$\mathbf{y} = \mathbf{w}^{\mathrm{T}}\mathbf{X} + b \tag{1}$$

where $\mathbf{w}$ is the weight vector to be identified in the function, and b is the threshold. The optimal regression function is given by the minimum of the cost function Φ

$$\Phi = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\frac{1}{I}\sum_{i=1}^{I} L_{\epsilon}(\mathrm{y}_i - y_{\mathrm{o}i}) \tag{2}$$

where

$$L_{\epsilon}(\mathrm{y}_i - \mathrm{y}_{\mathrm{o}i}) = \begin{cases} |\mathrm{y}_i - \mathrm{y}_{\mathrm{o}i}| - \epsilon & |\mathrm{y}_i - \mathrm{y}_{\mathrm{o}i}| \geq \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

is the ϵ-insensitive loss function measuring the error between the given observations ($\mathbf{y}_{\mathrm{o}}$) and the estimated ones ($\mathbf{y}$), ϵ is the tolerance zone, I is the number of the training compounds, and $^1/_2\mathbf{w}^{\mathrm{T}}\mathbf{w}$ is used as a measurement of the model complexity. The penalty constant C is introduced to determine the tradeoff between the empirical error and the model complexity. In this study, the penalty constant C is determined by sensitivity analysis and the tolerance zone ϵ as a desired precision band is estimated according to the mean absolute error (MAE) calculated by a backpropagation (BP) training neural network (BPNN). To solve the problem of nonlinearity, the radial basis function network-based transform for a nonlinear SVM is proposed.

2.2. Radial Basis Function Network-Based Transform for a Nonlinear SVM (RBFN-SVM). The construction of a RBFN-SVM involves two main parts. First, it comprises a nonlinear kernel transform from the original descriptors to the feature variables via a RBFN with one input layer and one hidden layer. The original descriptors are used as the input of the RBFN. The hidden layer applies a nonlinear mapping of the original variables to a feature space using Gaussian functions with corresponding parameters of centers and widths. The feature variables are then obtained as the output of the hidden layer. Second, a linear SVM relating the feature variables to the dependent variable is established. Given $\mathbf{x}$ as the input or the descriptor vector for a certain compound, the output of the kth hidden node ($k = 1,...,K$) can be represented as follows

$$\mathbf{o}_k = \exp\left\{-\frac{||\mathbf{x} - \mathbf{c}_k||^2}{\sigma_k^2}\right\} \tag{4}$$

where σ and $\mathbf{c}_k$ are the kernel width and center, respectively, of the kth hidden node. The hidden layer outputs $\mathbf{O} = (\mathbf{o}_1, ...,\mathbf{o}_K)$ can be regarded as a set of extracted feature variables obtained via the RBFN-based nonlinear kernel transform. Then, the SVM procedure is employed to model the following relationship between the extracted feature variables $\mathbf{O}$ and the bioactivity, $\mathbf{y}$, of the compound.

$$\mathbf{y} = \mathbf{w}^{\mathrm{T}} \times \mathbf{O} + b \tag{5}$$

Compared with a conventional kernel transform-based nonlinear SVM, a RBFN-based nonlinear transform allows varying kernel widths for different hidden nodes. Moreover, the kernel centers are not limited as the sample points or some cluster centers. Instead, kernel centers and widths are optimized together with the model parameters, $\mathbf{w}$ and b, in terms of the model cost function as defined by eq 2 using the PSO algorithm. Note that eq 5 actually gives a model exactly generated by the RBFN. Considering the fact that a RBFN can approximate arbitrarily well any nonlinear continuous function,[24] one can conclude that, with this flexible kernel transform, the model of eq 5 is capable of approximating any nonlinear function. This implies that such a RBFN-based kernel transform offers a very flexible nonlinear mapping for the original variables and is expected to have improved learning ability than a conventional nonlinear SVM. Moreover, due to this flexible kernel transform, one could reach a model with the desired cost using relatively few hidden nodes, thus reducing the complexity of kernel transform. This offers the possibility of avoiding the risk of overfitting caused by excessive support vectors utilized in kernel transform. On the other hand, a conventional nonlinear SVM algorithm involves two separate steps: one to select the kernel transform using a greedy search and the other to determine the linear model. Consequently, the kernel transform cannot be guaranteed to be optimal in terms of the structural risk minimization principle. In contrast, the formulation of the RBFN-SVM learning in terms of the structural risk minimization allows a synergetic optimization of all parameters including kernel centers $\mathbf{c}_k$ and kernel widths σ_k as well as linear model coefficients $\mathbf{w}$ and b using a global optimization technique, thereby enabling the determination of a kernel transform according to the performance of the total model. In this study, the PSO algorithm is invoked for the search of the optimal parameters

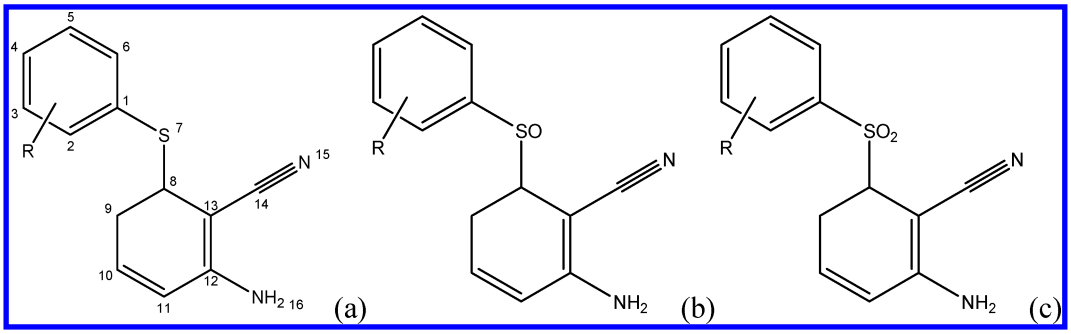

Figure 1. Molecular basic structure of 2-amino-6-arylsulfonylbenzonitriles and their thio and sulfinyl congeners as HIV-1 RT inhibitors. The atoms have been numbered 1−16. (a) The parent structure of compound nos. 1−19 in Table 1 of the Supporting Information. (b) The parent structure of compound nos. 20−28 in Table 1 of the Supporting Information. (c) The parent structure of compound nos. 29−51 in Table 1 of the Supporting Information.

Table 1. Results of QSAR Analysis of HIV-1 RT Inhibitors Using a RBFN-SVM Compared with Those Obtained by PLS, BPNN, and Conventional Nonlinear SVM

	R (correlation coefficient)				RSS (sum of squared residual)			
data set	PLS	BPNN	SVM[a]	RBFN-SVM	PLS	BPNN	SVM[a]	RBFN-SVM
training set	0.8846	0.9639	0.9199	0.9540	6.3093	1.5517	7.4748	2.6270
test set	0.8998	0.8485	0.8680	0.9404	4.5633	6.0126	7.8195	2.7295

[a] SVM: conventional nonlinear SVM using Gaussian kernel transform and optimized by quadratic programming.

in the RBFN-SVM model. This circumvents the risk of getting trapped in local optima and improves the convergence efficiency.

2.3. Particle Swarm Optimization (PSO). PSO[13−16] is an evolutionary computation technique, derived from simulating the behavior of birds searching for food. As a popular optimization tool, PSO has been widely used in the training of artificial neural networks, function optimization, and other genetic algorithm areas. In PSO, the potential solutions, called particles, fly through the problem space by following the current optimum particles. Each particle keeps track of its coordinate in the problem space which is associated with the best solution (fitness) it has achieved so far. This value is called the personal best position (*p*Best) for particle *i* represented as $\mathbf{p}_i = (p_{i1},p_{i2},...,p_{iD})$. Another best value that is tracked by the particle swarm optimizer is the best value obtained so far by all particles in the solution space called global best position (*g*Best) which is represented as $\mathbf{p}_g = (p_{g1},p_{g2},...,p_{gD})$. Each particle updates its velocity $\mathbf{v}_i = (v_{i1},v_{i2},...,v_{iD})$ and position $\mathbf{x}_i = (x_{i1},x_{i2},...,x_{iD})$ by tracking these two best values according to the following equations

$$v_{id}(\text{new}) = w \times v_{id}(\text{old}) + c_1 \times r_1 \times (p_{id}\text{-}x_{id}) + c_2 \times r_2 \times (p_{gd}\text{-}x_{id}) \quad (6)$$

$$x_{id}(\text{new}) = x_{id}(\text{old}) + \mu \times v_{id}(\text{new}) \quad (7)$$

where *w* is an inertia weight which is brought into eq 6 to balance the global search and local search, and r_1 and r_2 are random numbers between 0 and 1. Two positive constants, c_1 and c_2, called learning factors are introduced and generally both take the integer value 2. In eq 7, μ is the time parameter determining the different flying times for each particle. The particle swarm optimization concept consists of, at each time step, changing the velocity of each particle toward its *p*Best and *g*Best locations. Acceleration is weighted by a random term, with separate random numbers being generated for acceleration toward *p*Best and *g*Best locations.

In this paper, PSO is employed to search the optimal solution of a RBFN-SVM by minimizing the cost function Φ as the fitness function. Each particle is encoded as a real string representing the kernel centers (**c**) and widths (σ) as well as linear model coefficients **w** and *b*. With the movement of the particles in the problem space, the optimal solution with a minimum value of the cost function Φ will be obtained. Optimizing the kernel centers and widths and the weights of the SVM model relating the feature variables synergistically keeps the model from getting trapped into local optima and improves the model performance.

3. DATA SETS

3.1. HIV-1 RT Binding Affinity Data Set. To demonstrate the performance of the proposed algorithm, the RBFN-SVM algorithm is applied to QSAR modeling of HIV-1 reverse transcriptase (RT) inhibitor 2-amino-6-arylsulfonylbenzonitriles and their thio and sulfinyl congeners including 51 compounds reported by Chan et al.,[21] which later were studied by Roy and Leonard[25] using molecular connectivity and E-state parameters. Molecular basic structures of 2-amino-6-arylsulfonylbenzonitriles and their congeners as HIV-1 RT inhibitors are presented in Figure 1(parts a, b, and c, respectively). The detailed structural formulas of the compounds are listed in Table 1 of the Supporting Information. The binding affinity (BA) is expressed as IC_{50} values, which are the concentrations of compounds that would produce a 50% decrease in the cytopathic effect. The binding affinity data taken from ref 21 were converted to the logarithmic scale [pC (mM)] and used as the response variable. Besides the indicator variable (***I***), molecular connectivity parameters ($[^1\chi^v]_{mb}$), and two E-state indices (S_4, S_{15}) used by Roy and Leonard,[25] we calculated a series of molecular descriptors representing the chemical structure using the Cerius[2] 3.5 software system on a Silicon Graphics R3000 workstation, including structural, spatial, thermodynamic, electronic, topological descriptors, and E-state indices. We randomly divided these into a training set of 35 compounds and a test

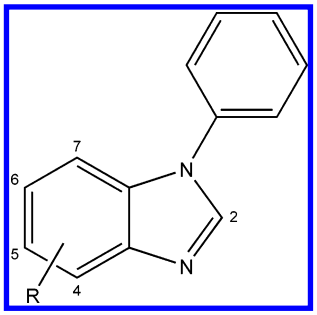

Figure 2. Molecular basic structure of 1-phenylbenzimidazoles as PDGFR.

set of 16 compounds. Each sample was described by the following 14 parameters including quadrupole and octupole polarizabilities (Quadrupole xz, Quadrupole xxx, Octupole xxz, Octupole yyz), total charge, subgraph count indices (SC-3 cluster), spatial descriptor (Shadow-Zlength), Jurs descriptor PPSA2 (total charge weighted partial positively charged molecular surface area), Connolly surface occupied volume, indicator variable ***I*** (having value 1 for sulfonyl compounds, value 0 otherwise), molecular connectivity parameters ($[^1\chi^v]_{mb}$), and three E-state indices (S-aaCH, S_4, S_{15}). $[^1\chi^v]_m$ is the first-order valence molecular connectivity index of *meta* substituents. $[^1\chi^v]_{mb}$ was defined as if fragmental $^1\chi^v$ values of the two *meta* substituents are different, then the lower value is called $[^1\chi^v]_{mb}$.[25] In the symbol S-aaCH, 'S' represents the electronic topological state of atom, and 'aaCH' stands for CH with two aromatic bonds. S_n was defined as an E-state value of atom numbered n.[25]

3.2. PDGFR Data Set. A set of 75 1-phenylbenzimidazoles reported by Brian et al.[22,23] was also studied using the proposed method. 1-Phenylbenzimidazoles are shown to be a new class of ATP-site inhibitors of the platelet derived growth factor receptor (PDGFR). The activity IC_{50} values were defined as the concentration of an inhibitor to reduce the level of ^{32}P (from added [^{32}P]-ATP) incorporated into the copolymer substrate. The molecular basic structure of 1-phenylbenzimidazoles as PDGFR and the substituent sites is presented in Figure 2. Table 2 of the Supporting Information lists the detailed structures of the compounds. The data set was randomly divided into two groups, with 55 compounds used as the training set for regression modeling and the rest of the 20 compounds used as the prediction set. More than 100 descriptors calculated by the Cerius^2 3.5 software system were utilized to describe these compounds. Among these descriptors, 16 variables were selected for describing each compound by the modified PSO for nonlinear modeling, including lowest unoccupied molecular orbital energy (LUMO), dipole moments (Dipole-Y), heat of formation, quadrupole and octupole polarizabilities (Quadrupole xx, Quadrupole yy, Octuploe xxx), shadow indices (Shadow-XYfrac, Shadow ratio), the number of rotatable bonds (Rotbands), bond information content (BIC), molecular connectivity parameters ($^3\chi_c^v$,$^3\chi_p^v$), the third-order kappa index (kappa-3), subgraph count indices (SC-3 path), E-state indices (S-dssC, S-aaaC) besides indicator variable $\boldsymbol{I}_5$ ($\boldsymbol{I}_5$ is an indicator parameter for the 7-substituted compounds, that is, $\boldsymbol{I}_5 = 1$ for 7-substituent present and $\boldsymbol{I}_5 = 0$ for 7-substituent absent). The symbol $^n\chi_t^v$ is the nth order valence molecular connectivity index, where t stands for the types of Chi indices: path (p) and cluster (c). In the symbol S-dssC and S-aasC, 'dssC' represents a carbon atom with two single bonds and one double bond attached to it, and 'aasC' represents a carbon atom with two aromatic and a single bond attached to it.

All the algorithms used in this study were written in Matlab 5.3 and run on a personal computer (Intel Pentium processor 4/2.80GHz 256MB RAM).

4. RESULTS AND DISCUSSION

4.1. HIV-1 RT Binding Affinity Data Set. The performance of the RBFN-SVM was first examined by modeling the binding affinity of HIV-1 RT inhibitor data. As a comparison, partial least-squares (PLS) method with six latent variables was invoked using the selected 14 variables. The calculated results are listed in Table 1. The correlation coefficient (R) for the training set and the test set are 0.8846 and 0.8998, respectively. The correlation between the calculated and observed values of binding affinity is shown in Figure 3a. To further examine nonlinearity correlation between the molecule structure and the binding affinity, ANN trained by BP with 20 hidden nodes was employed. To reduce the possibility of the BPNN to overfitting the training data, a monitoring set[26] including 12 samples was randomly picked from the training samples for training the BPNN. But the results are even poorer than those obtained by PLS because of serious overfitting, shown in Figure 3b. The correlation coefficient for the training set trained by the BPNN is 0.9639 but 0.8485 for the test set, and the sum of squared residual (RSS) is 1.5517 for the training set but 6.0126 for the test set compared with 6.3093 and 4.5633, respectively, which are obtained by PLS. A conventional nonlinear SVM gives a correlation coefficient of 0.9199 and a RSS of 7.4748 for the training set and a correlation coefficient of 0.8680 and a RSS of 7.8195 for the test set. It can be seen that the modeling error is rather high from the correlation between the calculated and observed values of binding affinity shown in Figure 3c.

To improve the QSAR model, a RBFN-SVM algorithm is applied to the HIV-1 RT binding affinity data. The hidden layer with 20 nodes was considered in this case. According to the MAE calculated by the BPNN, ϵ in the loss function is estimated as 0.0450. The sensitivity analysis of the penalty constant C is revealed in Figure 4. It can be observed that

Table 2. Results of QSAR Analysis of PDGFR Using a RBFN-SVM Compared with Those Obtained by PLS, BPNN, and Conventional Nonlinear SVM

	R (correlation coefficient)				RSS (sum of squared residual)			
data set	PLS	BPNN	SVM[a]	RBFN-SVM	PLS	BPNN	SVM[a]	RBFN-SVM
training set	0.8766	0.9677	0.9747	0.9110	8.9140	2.0244	2.0629	6.5349
test set	0.8614	0.8506	0.8641	0.9191	4.6372	7.5141	4.5288	3.1859

[a] SVM: a conventional nonlinear SVM using Gaussian kernel transform and optimized by quadratic programming.

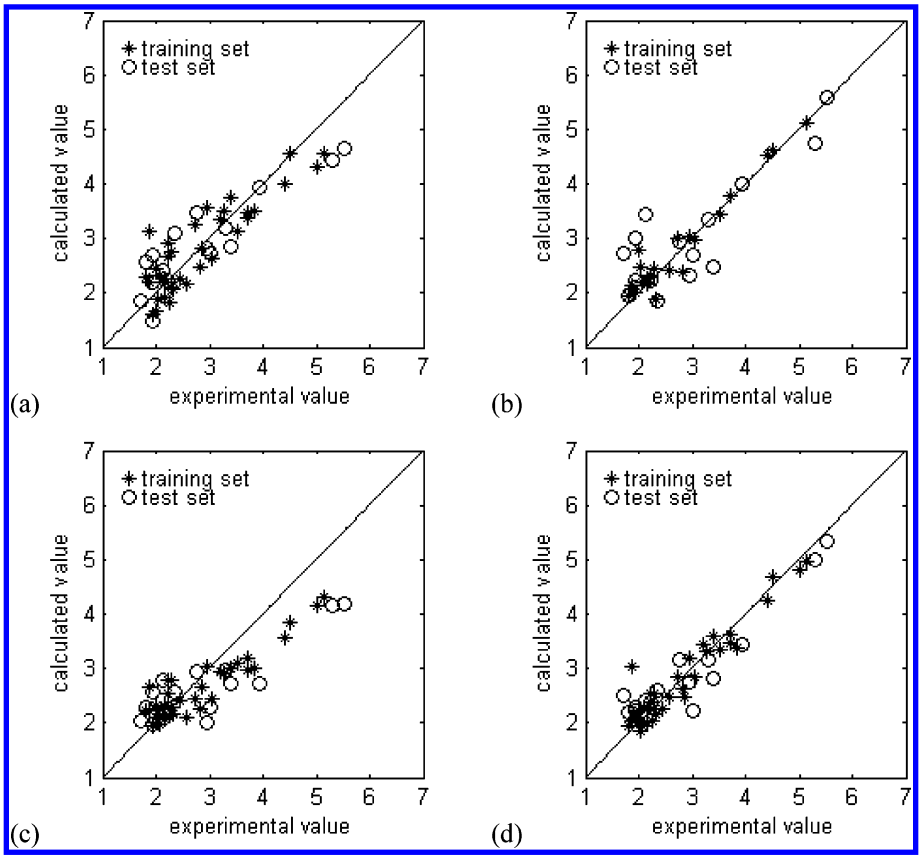

Figure 3. (a) Calculated versus experimental values of the binding affinity using PLS modeling with six latent variables for HIV-1 RT binding affinity data set. (b) Calculated versus experimental values of the binding affinity using BPNN modeling. (c) Calculated versus experimental values of the binding affinity using conventional nonlinear SVM modeling. (d) Calculated versus experimental values of the binding affinity using RBFN-SVM modeling.

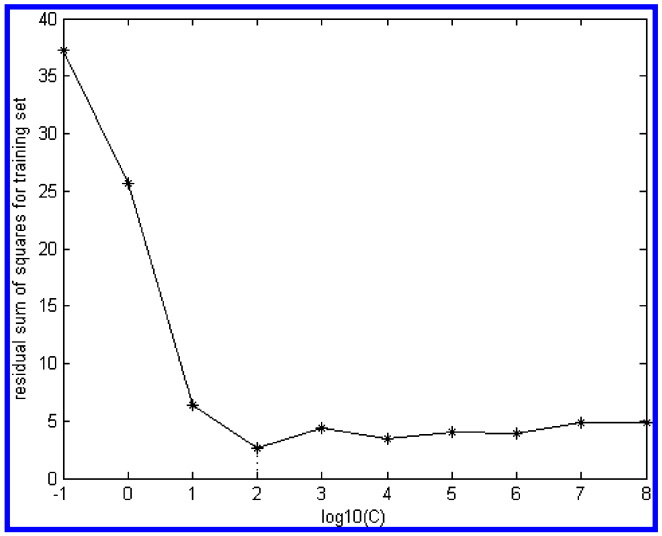

Figure 4. The sensitivity analysis of penalty constant *C* for HIV-1 RT binding affinity data set.

the RBFN-SVM model reaches the best performance when *C* takes the value of 100. Taking advantage of both the RBFN and PSO, the RBFN-SVM shows a good performance for modeling and prediction. A correlation coefficient for a training set of 0.9540 was obtained by a RBFN-SVM and that for the test set was improved to 0.9404, and the RSS was also improved greatly, down to 2.6270 and 2.7295 for training and test sets, respectively, indicating that the RBFN-SVM yielded a QSAR model with desirable generalization ability. The correlation between the calculated and experimental values of binding affinity is shown in Figure 3d. It shows that a RBFN-SVM is very effective in conquering overfitting which is a serious problem in the BPNN as shown in Figure 1b. Compared with a conventional nonlinear SVM, a RBFN-SVM provides an enhanced performance for both the training set and the test set, exhibiting that the proposed algorithm has better precision in modeling and superior generalization in prediction. Useful information was sufficiently extracted via a nonlinear RBFN transform. Flexible kernel widths bring a higher accuracy of model, especially for the HIV-1 RT binding affinity data set in which samples differ from each other obviously with respect that they are actually three kinds of compounds in the data set. Given equal width to different kernels would result in poor model precision which can be seen from the results obtained by a conventional nonlinear SVM in which the parameter σ is uniform 0.72 (selected by a grid search and cross-validation). In a RBFN-SVM, each kernel has the best appropriate width with it as optimized by PSO. More flexible nonlinear mapping, with kernel centers, widths, and linear model coefficients optimized synergistically using the PSO algorithm guaranteed that a RBFN-SVM could generate an excellent model.

4.2. PDGFR Data Set. For further testing the performance of a RBFN-SVM, this new algorithm was applied to model the activity of 1-phenylbenzimidazoles, as comparisons, PLS, BPNN, and a conventional nonlinear SVM were also employed. The 17 selected variables were used in these four models. Table 2 summarizes the statistical results. PLS with six latent variables gives a correlation coefficient of 0.8766

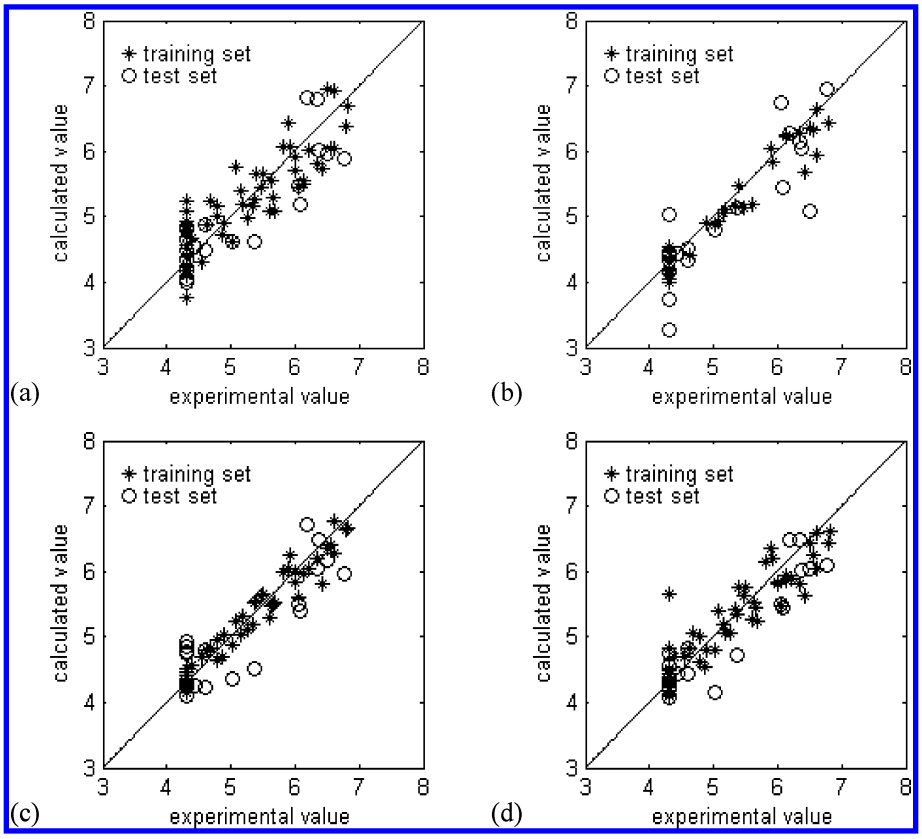

Figure 5. (a) Calculated versus experimental values of the binding affinity using PLS modeling with six latent variables for PDGFR data set. (b) Calculated versus experimental values of the binding affinity using BPNN modeling. (c) Calculated versus experimental values of the binding affinity using conventional nonlinear SVM modeling. (d) Calculated versus experimental values of the binding affinity using RBFN-SVM modeling.

and a RSS of 8.9140 for the training set and a correlation coefficient of 0.8614 and a RSS of 4.6372 for the test set. The poor results shown in Figure 5a reveal the nonlinearity between the molecular structures of 1-phenylbenzimidazoles and their biological activities failed to be represented in a linear plot. The correlation between the results calculated by the BPNN with 20 hidden nodes and observed values of bioactivities is shown in Figure 5b. A monitoring set[26] was used for training the BPNN to mitigate the probability of overfitting the training data. The BPNN brings good results of model training, with a correlation coefficient of 0.9677 and a RSS of 2.0244, but poor prediction, with a correlation coefficient of 0.8506 and a RSS of 7.5141. A conventional nonlinear SVM failed to improve the model. The correlation coefficients for the training set and the test set are 0.9747 and 0.8641, respectively. The high correlation coefficient of the training set and the low correlation coefficient for prediction imply a serious problem of overfitting due to excessive support vectors selected in the process of modeling.

To improve the QSAR model, a RBFN-SVM algorithm was applied to predict the inhibitory activity of 1-phenylbenzimidazoles. The same number of hidden nodes with the BPNN was given to the RBFN-SVM. According to the MAE calculated by the BPNN, ϵ in the loss function is estimated as 0.0644. The sensitivity analysis of the penalty constant C is demonstrated in Figure 6. It is clear that when C is greater than or equal to 100, a large change in C results in a relatively small change in the outcomes, so C takes 100 in this model. The correlation coefficient is 0.9110 for the training set and 0.9191 for the test set. In contrast to PLS, BPNN, and a conventional nonlinear SVM, the newly proposed method presents a better performance. At a certain extent, a RBFN-SVM enhanced the precision of the model, that is, the RSS of 8.9140 for the training set of PLS is down to 6.5349. A RBFN-SVM also exhibits good generalization ability. By the comparison of Figure 5b−d, it can be seen that a RBFN-SVM improved the model prediction performance effectively, and the RSS for the test set is down to 3.1859 from 7.5141 (BPNN) and 4.5288 (a conventional nonlinear SVM) as shown in Table 2, respectively. As a whole, the results show

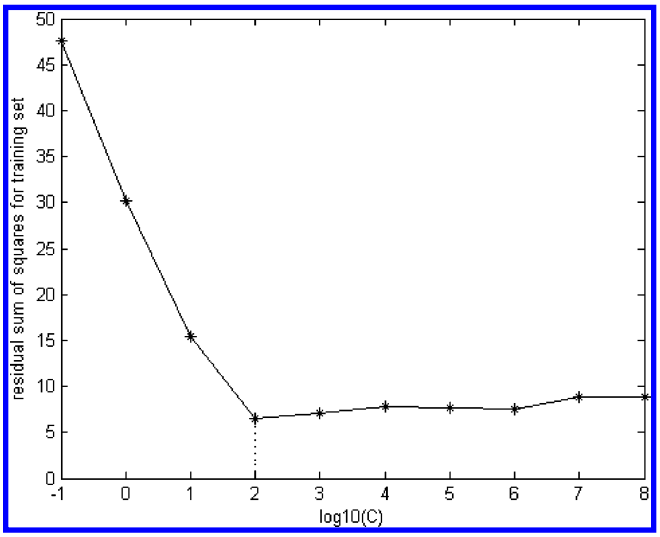

Figure 6. The sensitivity analysis of penalty constant C for PDGFR data set.

the RBFN-SVM has the best performance rather than PLS, BPNN, and a conventional nonlinear SVM on modeling the activity of 1-phenylbenzimidazoles. Using a small quantity of hidden nodes in the RBFN-SVM based on the desired accuracy, one reduced the complexity of the model and avoided the problem of overfitting. Moreover, a synergetic optimization of all parameters using PSO substantially improved the performance of the total model of the RBFN-SVM, which also guaranteed the model from getting trapped in local optima.

To further demonstrate the behavior of the algorithms, five different random combinations of training and test sets were investigated. The model obtained by the RBFN-SVM training was used to predict the activity of 1-phenylbenzimidazoles. The procedure was repeated five times for the five combinations. The mean correlation coefficients in the five computations were 0.9126 and 0.9104 for the training and test sets, respectively. The mean RSS of 6.6503 for the training sets and 3.2655 for the test sets were obtained. In contrast, the mean correlation coefficients for the training and test sets over the five computations calculated by the BPNN were 0.9683 and 0.8510, respectively. A conventional nonlinear SVM gives the mean correlation coefficient of 0.9801 and the mean RSS of 1.6640 for the training sets and the mean correlation coefficient of 0.8601 and the mean RSS of 4.5058 for the test sets. These results showed that a RBFN-SVM performed better than a BPNN and a conventional nonlinear SVM and that the good performace was not due to a fortuitous choice of the training and test sets.

5. CONCLUSION

In this paper, a radial basis function network-based transform for a nonlinear SVM as optimized by PSO has been proposed. The introduction of a RBFN makes a more flexible nonlinear kernel transform, which greatly enhanced the ability of overcoming underfitting and overfitting. The use of PSO makes it possible for synergetic optimization of all parameters including kernel centers and kernel widths as well as linear model coefficients which are also essential for a quick convergence. The performance of a RBFN-SVM was evaluated by using two QSAR data sets which revealed the proposed method is of great promise in nonlinear QSAR modeling.

ACKNOWLEDGMENT

This work was financially supported by National Natural Science Foundation of China (grant nos. 20375012, 20205005, 20105007, 20435010, and 20675028).

Supporting Information Available: Detailed structural formulas of the compounds. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Vapnik, V. N. In *The Nature of Statistical Learning Theory*; Springer: New York, U.S.A., 1995.

(2) Vapnik, V. N. In *Statistical Learning Theory*; Wiley: New York, 1998.

(3) Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. Drug Discovery Using Support Vector Machines. The Case Studies of Drug-likeness, Agrochemical-likeness, and Enzyme Inhibition Predictions. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 2048−2056.

(4) Thissen, U.; Ustun, B.; Melssen, W. J.; Buydens, L. M. C. Multivariate Calibration with Least-Squares Support Vector Machines. *Anal. Chem*. **2004**, *76*, 3099−3105.

(5) Lee, Y.; Lee, C. K. Classification of Multiple Cancer Types by Multicategory Support Vector Machines Using Gene Expression Data. *Bioinformatics* **2003**, *19*, 1132−1139.

(6) Zhang, S. W.; Pan, Q.; Zhang, H. C.; Zhang, Y. L.; Wang, H. Y. Classification of Protein Quaternary Structure with Support Vector Machine. *Bioinformatics* **2003**, *19*, 2390−2396.

(7) Mukherjee, S.; Osuna, E.; Girosi, F. Nonlinear Prediction of Chaotic Time Series Using a Support Vector Machine. In *Proceeding of IEEE NNSP*; 1997; pp 24−26.

(8) Zhou, Y. P.; Jiang, J. H.; Lin, W. H.; Zou, H. Y.; Wu, H. L.; Shen, G. L.; Yu, R. Q. Boosting Support Vector Regression in QSAR Studies of Bioactivities of Chemical Compounds. *Eur. Pharm. Sci*. **2006**, *28*, 344−353.

(9) Yao, X. J.; Panaye, A.; Doucet, J. P.; Zhang, R. S.; Chen, H. F.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Comparative Study of QSAR/QSPR Correlations Using Support Vector Machines, Radial Basis Function Neural Networks, and Multiple Linear Regression. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1257−1266.

(10) Golbraikh, A. P.; Oloff, S.; Xiao, Y.; Tropsha, A. Combinatorial QSAR Modeling of P-Glycoprotein Substrates de Cerqueira Lima. *J. Chem. Inf. Model*. **2006**, *46*, 1245−1254.

(11) Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. QSAR Models for the Prediction of Binding Affinities to Human Serum Albumin Using the Heuristic Method and a Support Vector Machine Xue. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 1693−1700.

(12) Powell, M. J. D.; Mason, J. C.; Cox, M. G. Radial Basis Functions for Multivariable Interpolation: A review. In *Algorithm for Approximation*; Clarendon Press: 1987; pp 143−167.

(13) Kennedy, J.; Eberhart, R. Particle Swarm Optimization. In *Proceedings of IEEE International Conference on Neural Networks*; Perth, Australian, 1995; Institute of Electrical and Electronics Engineers: Piscataway, NJ, 1995; Vol. 4, pp 1942−1948.

(14) Shi, Y.; Eberhart, R. A Modified Particle Swarm Optimizer. In *Proceedings of IEEE World Congress on Computational Intelligence*; Piscataway, NJ, 1998; Institute of Electrical and Electronics Engineers: Piscataway, NJ, 1998; pp 69−73.

(15) Clerc, M.; Kennedy, J. The Particle Swarms Explosion, Stability, and Convergence in a Multidimensional Complex Space. In *IEEE Transactions on Evolutionary Computation*; Institute of Electrical and Electronics Engineers: Piscataway, NJ, 2002; Vol. 6, pp 58−73.

(16) Shi, Y.; Eberhart, R. Fuzzy Adaptive Particle Swarm Optimization. In *Proceedings of the 2001 Congress on Evolutionary Computation*; Seoul, South Korea, 2001; Institute of Electrical and Electronics Engineers: Piscataway, NJ, 2001; Vol. 1, pp 101−106.

(17) Shen, Q.; Jiang, J. H.; Jiao, C. X.; Lin, W. Q.; Shen, G. L.; Yu, R. Q. Hybridized Particle Swarm Algorithm for Adaptive Structure Training of Multilayer Feed-Forward Neural Network: QSAR Studies of Bioactivity of Organic Compounds. *J. Comput. Chem*. **2004**, *25*, 1726−1735.

(18) Lin, W. Q.; Jiang, J. H.; Wu, H .L.; Shen, G. L.; Yu, R. Q. Piecewise Hypersphere Modeling by Particle Swarm Optimization in QSAR Studies of Bioactivities of Chemical Compounds. *J. Chem. Inf. Model*. **2005**, *45*, 535−541.

(19) Lin, W. Q.; Jiang, J. H.; Shen, G. L.; Yu, R. Q. Optimized Block-Wise Variable Combination by Particle Swarm Optimization for Partial Least Squares Modeling in Quantitative Structure-Activity Relationship Studies. *J. Chem. Inf. Model*. **2005**, *45*, 486−493.

(20) Shinzawa, H.; Jiang, J. H.; Iwahashi, M.; Noda, I.; Ozaki, Y. Self-modeling Curve Resolution (SMCR) by Particle Swarm Optimization (PSO). *Anal. Chim. Acta* In press.

(21) Chan, J. H.; Hong, J. S.; Hunter, R. N., III; Orr, G. F.; Cowan, J. L.; Sherman, D. L.; Sparks, S. M.; Reitter, B. E.; Andrews, C. W., III; Hazen, R. J.; St Clair, M.; Boone, L. R.; Ferris, R. G.; Creech, K. L.; Roberts, G. B.; Short, S. A.; Weaver, K.; Ott, R. J.; Ren, J.; Hopkins, A.; Stuart, D. I.; Stammers, D. K. 2-Amino-6-arylsulfonylbenzonitriles as Non-nucleoside Reverse Transcriptase Inhibitors of HIV-1. *J. Med. Chem*. **2001**, *44*, 1866−1882.

(22) Palmer, B. D.; Smaill, J. B.; Boyd, M.; Boschelli, D. H.; Doherty, A. M.; Hamby, J. M.; Khatana, S. S.; Kramer, J. B.; Kraker, A. J.; Panek, R. L.; Lu, G. H.; Dahring, T. K.; Winters, R. T.; Showalter, H. D. H.; Denny, W. A. Structure-Activity Relationships for 1-Phenylbenzimidazoles as Selective ATP Site Inhibitors of the Platelet-Derived Growth Factor Receptor. *J. Med. Chem*. **1998**, *41*, 5457−5465.

(23) Palmer, B. D.; Kraker, A. J.; Hartl, B. G.; Panopoulos, A. D.; Panek, R. L.; Batley, B .L.; Lu, G. H.; Trumpp-Kallmeyer, S.; Showalter, H. D. H.; Denny, W. A. Structure-Activity Relationships for 5-Substi-

tuted 1-Phenylbenzimidazoles as Selective Inhibitors of the Platelet-Derived Growth Factor Receptor. *J. Med. Chem.* **1999**, *42*, 2373−2382.
(24) Gurumoorthy, A.; Kosanovich, K. A. Improving the Prediction Capability of Radial Basis Function Networks. *Ind. Eng. Chem. Res.* **1998**, *37*, 3956−3970.
(25) Roy, K.; Leonard, J. T. QSAR Modeling of HIV_1 Reverse Transcriptase Inhibitor 2-Amino-6-arylsulfonylbenzontriles and Congeners Using Molecular Connectivity and E-state Parameters. *Bioorg. Med. Chem.* **2004**, *12*, 745−754.
(26) Despagne, F.; Massart, D. L. Neural Networks in Multivariate Calibration. *Analyst (Cambridge, U. K.)* **1998**, *123*, 157−178.

CI700047X