ARTICLES

Modified Particle Swarm Optimization Algorithm for Adaptively Configuring Globally Optimal Classification and Regression Trees

Yan-Ping Zhou,*,†,‡ Li-Juan Tang,‡ Jian Jiao,† Dan-Dan Song,† Jian-Hui Jiang,*,‡ and Ru-Qin Yu*,‡

Key Laboratory of Pesticide and Chemical Biology of Ministry of Education, College of Chemistry, Central China Normal University, Wuhan 430079, P. R. China, and State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, P. R. China

Received October 10, 2008

The configuration of classification and regression trees (CART) used to include tree-growing by greedy recursive partitioning, which selects the splitting parameters (i.e., splitting variables and values) involved in tree, and tree-pruning, which aims to obtain a final tree of right size. This method is successful for most applications; however, it presents some well-known limitations and drawbacks, such as, less comprehensibility, inclination to overfitting, and suboptima. In the present study, the modified discrete particle swarm optimization method was invoked to adaptively configure the globally optimal CART (MPSOCART) via simultaneously selecting the optimal splitting parameters in CART and the appropriate structure of CART. A new objective function was formulated to decide the appropriate CART architecture and the optimum splitting parameters. The proposed MPSOCART was applied to predict the bioactivities of flavonoid derivatives and inhibitory activities of inhibitors of epidermal growth factor receptor tyrosine kinase, compared with partial least-squares and CART induced by greedy recursive partitioning. The comparison revealed that MPSO was a useful tool for inducing a globally optimal CART, which converges fast to the optimal solution and avoid overfitting in great extent.

1. INTRODUCTION

Classification and regression trees (CART),[1] as a nonparametric method for classification and regression tasks, aims to search for exclusive regions of the data space, each region containing homogeneous subset of the whole data. Compared to some traditional methods, such as, partial least-squares (PLS) or artificial neural network (ANN), CART offers several predominant strengths, including simplicity, interpretability, high capacity in handling large data sets and immunities to outliers, collinearities, and heteroscedasticity. Recently, CART has greatly increased in popularity for QSAR studies.[2−4]

Traditionally, the induction of CART is performed by first growing the largest tree applying greedy recursive partitioning for picking out the optimal splitting parameters (i.e., the splitting variables and the corresponding splitting values) involved in the tree and then pruning the grown tree to yield the final appropriately fit CART.[1] The recursive partitioning method is subjected to horizon effects (i.e., it selects the splitting parameters based on local rather than global measures), and so risks are trapped by local minima. CART induced by recursive partitioning approach tends to fit idiosyncrasies and noise in training cases, which are unlikely to occur with the same pattern in unseen ones, resulting in overfitting. Moreover, CART gets into suboptima with a higher frequency, since the tree-growing and tree-pruning phases are two sequential and completely independent procedures. Consequently, CART by conventional generator is exposed to a high risk of overfitting and local optima. In addition, the complexity of CART, measured as the number of the leaf nodes, is another most critical factor under consideration, when effectively solving practical problems using CART. Overly large trees can be fragmented, holding many leaves with only a few cases per leaf. Such leaf nodes are more error-prone predictors than leaf nodes containing many cases, and so are affected by noise and idiosyncrasies with higher probabilities, and then, resulting in overfitting and suboptima. While oversimplified tree might lead to poor modeling capacity. Besides the better generalization ability, smaller trees are preferred, because they are more comprehensible and usually faster and cheaper to be built. Typically, appropriately complex CART is identified by applying the minimal cost-complexity pruning (MCCP) criterion[1] either based on cross-validation or pruning set techniques. Via cross-validation way, a sequence of auxiliary subtrees is constructed. It will be computational burdensome. In addition, a large variance can be caused by cross validation procedure, especially for small training samples.[5] Gelfand et al.[6] claimed that this procedure is both inefficient and possibly ineffective in locating the optimal tree, suggesting an efficient iterative tree growing and pruning algorithm that is guaranteed to converge. When using the pruning set technique, a large number of training samples is needed to effectively identify the right sized CART.

* To whom correspondence should be addressed. Tel: +86-731-8822577; +86-15872406428. E-mail: hgzyp2005@yahoo.com.cn (Y.-P.Z.); jianhuijiang@hnu.cn (J.-H.J.); rqyu@hnu.cn (R.-Q.Y.).
† Central China Normal University.
‡ Hunan University.

10.1021/ci800374h CCC: $40.75
Published on Web 04/14/2009

In addition to the aforementioned typical approach, some optimizing techniques, such as genetic programming and simulated annealing, were introduced for inducing CART.[7−9] Such methods searched the space of all CARTs using random perturbations, additions, and deletions of the splits. Trees induced by these techniques tend to be overly large, with less comprehensibility and low generalization ability. In addition, artificial ant colony algorithm was used to induce CART.[10] As for this method, attention was paid to optimizing the splitting parameters, but no effort was placed on the tree complexity. Although many CART induction algorithms are shown to produce simpler and more comprehensible trees with good accuracy, tree simplification is always the secondary concern relative to accuracy. Therefore, it is highly desirable to develop an approach to configure a globally optimal CART model with a well trade-off between the accuracy and the complexity.

The particle swarm optimization (PSO) method,[11−14] a relatively new optimizing technique, can also be employed for CART configuration. This method is originated as simulating a simplified social system. Most PSO algorithms operate in a continuous or real-number space.[15−17] Recently, Yu et al.[16−19] developed a modified version of particle swarm optimization (MPSO) fast converging to the optima for discrete optimization issues. In the present study, MPSO was invoked for adaptively constructing the globally optimal CART (MPSOCART) by simultaneously searching the appropriate tree architecture and the optimal splitting parameters. On the basis of joint evaluation of the model complexity and error, an objective function was also formulated as a performance measure for the CART induction.

The newly proposed MPSOCART was applied to model the bioactivities of flavonoid derivatives as p56lck tyrosine kinase inhibitors and the inhibitory activities of inhibitors of epidermal growth factor (EGFR) tyrosine kinase, as compared with CART induced by recursive partitioning (RPCART) and PLS. The results displayed that MPSOCART compared favorably with the conventional algorithms, enabled a rapid convergence to the globally optimum, and had high capacity of overfitting avoidance, demonstrating that the newly designed algorithm holds great promise in adaptively configuring the globally optimal CART. In addition, the proposed objective function can be well used to fine-tune the balance between comprehensibility and accuracy to further substantially guarantee a globally optimal CART.

2. THEORY

2.1. Classification and Regression Trees. CART, as a relatively new method for classification and regression tasks, was developed by Bremain et al.[1] The dependent variable can be either numerical or categorical, respectively, resulting in regression or classification trees. CART, as a binary tree representation, is able to describe the relationships between dependent and independent variables with high flexibility and sufficient accuracy. An extended depiction on CART can be referred to Breiman et al.'s excellent book *Classification and Regression Trees*.[1] Here, since CART was used for regression tasks, only a concise description of regression tree is presented.

First, the largest tree T_{max} is grown. Greedy recursive binary partitioning in top-down fashion is the most commonly used method for tree-growing nowadays, starting from the root node containing the entire training compounds until each node reaches completely homogeneity or a user-specified minimal sample number (i.e., node size) and becomes a terminal or leaf node. The node split into two subordinate nodes is called parent node. Its two subordinate ones go without saying the child nodes. In CART, the nodes except leaf nodes are named internal nodes, too. The partition proceeds via choosing a certain descriptor and its certain value as the splitting variable and value, respectively. Thus, the crucial issue in growing T_{max} lies on determining the optimal splitting parameters (i.e., the splitting variable and value) in each internal node. Generally, such optimal splitting parameters are identified by a goodness-of-split criterion, i.e., minimizing the sum of squared deviations (SSD) of dependent variables (i.e., bioactivities). Via exhaustively exploring all the variables and all their possible values, respectively, as the splitting variables and values, the variable and its certain value that together offer the maximal change in SSD between the parent and prospectively child nodes are identified as the optimal split in the parent node. For example, as for an internal node t including n samples$\{(\vec{x}_n, y_n)\}$, the SSD associated with node t is computed by

$$\mathrm{SSD}(t) = \sum_{\vec{x}_n \in t} [y_n - \bar{y}(t)]^2 \quad (1)$$

where $\bar{y}(t)$ is the mean or the median of bioactivities of compounds in the tth node. Supposed that node t is divided into two prospectively subsidiary nodes t_L and t_R by an arbitrary split s, the split s^*, maxing $\phi(s,t) = \mathrm{SSD}(t) - \mathrm{SSD}(t_L) - \mathrm{SSD}(t_R)$, is the best splitting parameters. It can be expressed as $\phi(s^*,t) = \max_{s\in\Omega}\phi(s,t)$, in which Ω is the collection of all possible splits. This indicates that the higher is the value of $\phi(s,t)$, the better split is obtained because it increases the within-node homogeneity more significantly. In the above-mentioned way, a largest tree T_{max} is grown.

Second, T_{max} is pruned to obtain the final appropriately fit tree. The T_{max} induced by above-mentioned technique often yields overelaborate structure, and thus suffering from overfitting and incomprehensibility. To overcome these bottlenecks, tree-pruning procedure is inevitable, as long as the pruned smaller tree is comparable with the larger one in terms of performance. The minimal cost-complexity pruning (MCCP)[1] is the most popular tree-pruning method, expressed as follows:

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \quad (2)$$

where $R_\alpha(T)$ is a linear combination of the tree cost and complexity, representing the cost-complexity measure; $R(T)$ is the within-node sum of resubstitution error, representing the tree cost; $\alpha \geq 0$ refers to a complexity parameter, and $|\tilde{T}|$ measures the number of terminal nodes in the tree T as the tree complexity. The more complex is the tree, the lower is $R(T)$, but at the same time, the higher is the penalty $\alpha|\tilde{T}|$ and vice versa. As proven by Breiman,[1] α is a metric of additional accuracy introduced by a certain new bifurcation. The tree with α equaling to 0 refers to the unpruned largest tree T_{max} and α tends to increase from the leaf nodes to the root node. The MCCP tries to cut off the weakest branches of the tree in a bottom-up manner, gradually increasing α value. For a given α, among all subtrees of the same complexity, the optimal subtree $T \leq T_{max}$ minimizing $R_\alpha(T)$

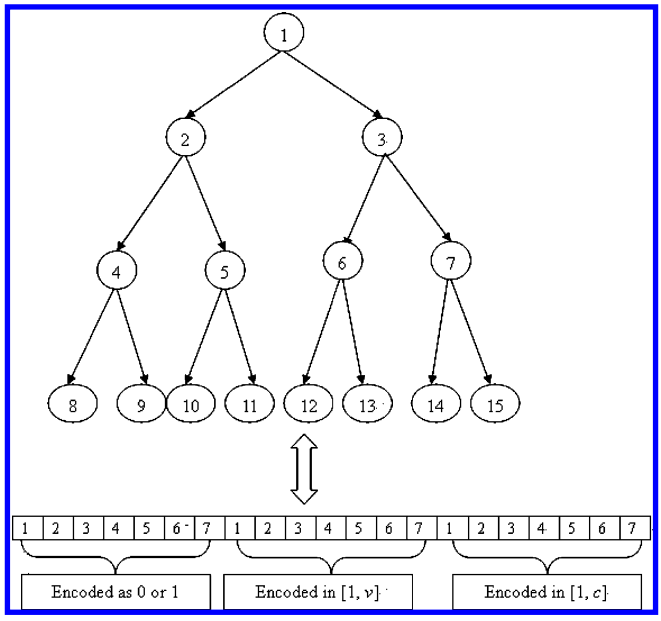


Figure 1. Mapping a full symmetric tree to an equivalent array.

uniquely exists, as proven by Bremain. In such way, a sequence of nested subtrees is gained, from which a final appropriately fit tree can be selected in terms of its best prediction accuracy either gained by cross validation method or pruning set technique.

Once the final tree is gained, some imminent node information is endowed. Each internal node is characterized by a splitting rule including the optimal parameters. Each node is assigned the mean or the median of the bioactivities of the involved compounds as the node output. In addition, the node size, that is, the number of compounds in the node, is provided in each node. A prediction of the bioactivity of an unseen compound from a given set of descriptors is made by traversing the tree until a leaf node is reached, and this leaf node output acts as the predicted bioactivity.

The principle of classification tree is similar to regression tree. The difference between them lies on the fact that the Gini index or Twoing rule is taken as the default goodness-of-split criterion for classification tree.[1]

From the above-mentioned delineation, one can obtain that the core of CART configuring includes the identification of splitting parameters and tree complexity. In the present study, a modified particle swarm optimization is invoked to adaptively configure the globally optimal CART by simultaneously searching the optimal splitting parameters and the tree architecture.

2.2. Modified Particle Swarm Optimization. The particle swarm optimization (PSO) algorithm,[11−14] as a stochastic global optimization method, simulates the social behavior of bird flocking. It explores the problem space by a population of particles (individuals), each standing for a single solution. In PSO, each particle flies over the problem space with a velocity directing the movement of the particle, keeping track of the best solutions encountered so far. PSO first randomly initializes the position and velocity of individuals in the swarm by dispersing them uniformly across the search space. The particle i is represented as $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{iD})$. Velocity, the rate of the position variation for the ith particle, is denoted as $\mathbf{v}_i = (v_{i1}, v_{i2}, ..., v_{iD})$. PSO postulates that particles should move to a location combining their personal best and the global best positions. The best previous position of the ith particle which yields the best fitness value is the personal best position expressed as $\mathbf{p}_i = (p_{i1}, p_{i2}, ..., p_{iD})$. The global best position $\mathbf{p}_g = (p_{g1}, p_{g2}, ..., p_{gD})$ refers to the best particle among all the particles in the population until the current iteration. During optimization, each particle is updated via following the above-mentioned two best values.

Most PSO versions operate in continuous real-number space. Recently, based on the information-sharing mechanism of the continuous PSO and the pattern of updating particle by following the two best positions, Yu et al.[16−19] proposed a modified PSO (MPSO) for a discrete problem expressed as an integer string varying from 0 to k. For such a discrete issue, each individual in MPSO moves in a search space confined to an integer from 0 to k on each bit. Renovating a particle indicates the variation of a bit that should be an integer. The velocity v_{id}, a random number in the range of (0, 1), represents the probability of a bit x_{id} taking the integer from 0 to k, respectively. The resulting change in position is then defined as follows:

$$\text{If } (0 < v_{id} \leq \beta), \text{ then } x_{id}(\text{new}) = x_{id}(\text{old}) \quad (3)$$

$$\text{If } (\beta < v_{id} \leq (1 + \beta)/2), \text{ then } x_{id}(\text{new}) = p_{id} \quad (4)$$

$$\text{If } ((1 + \beta)/2 < v_{id} \leq 1), \text{ then } x_{id}(\text{new}) = p_{gd} \quad (5)$$

where β, named the static probability, is a random number in the range of (0, 1), which is started with a value of 0.5 and decreases to 0.33 when the iteration ceases. A total of 10% of the particles are forced to fly randomly, not following the two best positions to improve the potential of the discrete PSO technique to get over local optima. Employing the gradually decreasing static probability and some percent of randomly flying particles to overleap the local optima, the modified PSO indicates satisfactory convergence performance for discrete problems.[16−19]

2.3. Adaptive Configuring of CART by MPSO (MPSOCART). In the present study, a modified discrete PSO was invoked to adaptively configure the globally optimal CART for simultaneously searching the tree architecture and the splitting parameters. In MPSO, each particle was encoded as a string of integers including three equal length and mutually corresponding parts. The length of a particle is thrice as the number of internal nodes involved in a completely symmetrical tree shown in Figure 1. A fully symmetrical tree with L levels holds internal nodes of $\sum_{l=1}^{L-1} 2^{(l-1)}$. Moreover, the number of the child nodes is twice and twice plus 1 as that of their parent node, and the nodes in lth level are $2^{(l-1)}$. These attributes offer the convenience of the particle coding and decoding in MPSO. All of these are the reasons for the tree initially encoded as symmetrical one. The nodes in the last level of the full symmetric tree are not under consideration, since they either are leaf nodes or do not exist at all. No optimization is needed. The first part in a string encoded as a binary bit represents the tree structure, indicating whether the corresponding internal node exists or not. The second and third parts encoded as integers, respectively, refer to the indices of the splitting variables and values in training data matrix, since the splitting variable and value, respectively, are the certain descriptor and its certain value. Once the matrix indices of the descriptor and its value as splitting parameters are identified, the splitting

parameters can be easily drawn out. A bit of 0 in the first part implies that the corresponding node should be a leaf node or does not exist at all. Of course, its descendent nodes are automatically excluded in the final CART model. This also means that the splitting parameters associated with these nodes are useless, and vice versa, a bit of 1 indicates the existence of the associated node. The splitting variable and value for this node were obtained, respectively, according to the information provided by the corresponding bits in the second and third parts. In addition, for enhancing the optimization efficiency, the first three bits in the first part of a particle were always fixed to 1 to guarantee sufficient accuracy of the tree. Consequently, given a training data matrix **X** containing c compounds (rows) and v descriptors (columns), the last two parts in a particle were encoded as the integers, respectively, from 1 to v and 1 to c, as shown in Figure 1. Provided a bit in the second part is v and its corresponding bit of the third part is encoded as c, the splitting variable and value for the associated node are the vth descriptor and x_{cv}, respectively. What follows is the depiction of adaptive configuration of CART by MPSO.

Step 1: Generate randomly all of the initial strings **STR**, **SPVAR**, and **SPVAL** in MPSO with a population of proper size. **SPVAR** and **SPVAL** are encoded as integers in the range of [1, v] and [1, c], respectively. With the first three bits fixed at 1, **STR**s are binary strings corresponding to **SPVAR** and **SPVAL** encoded as delineated above.

Step 2: Decode a particle as a CART, whose structure identified by **STR**. The final splitting variables and values (i.e., splitting parameters **P**) are, respectively, the Hadamard product of **STR** and **SPVAR**, as well as **STR** and **SPVAL**. If **STR** = $(str_{ij})_{m \times n}$ and **SPVAR** = $(spvar_{ij})_{m \times n}$, then the Hadamard mulplication yields a product of **STR**·**SPVAR** = $(str_{ij} \times sp\ \mathrm{var}._{ij})_{m \times n}$. At the same time, during particle decoding, the minimal node size of 5 is considered. The node holding less than or equaling to 5 samples will be labeled as a leaf node. Of course, its associated bit is forced to be changed to 0 wherever this bit in the particle are encoded as 1 or 0 before, and its subsidiary nodes ought to disappear.

Step 3: Compute the fitness function (vide infra) of the CART associated with each particle of the population **P**. If the current iteration number reaches the predefined maximum iteration number or the user-specified minimal error criterion, the iteration is ceased with the results output; otherwise, go to the next step.

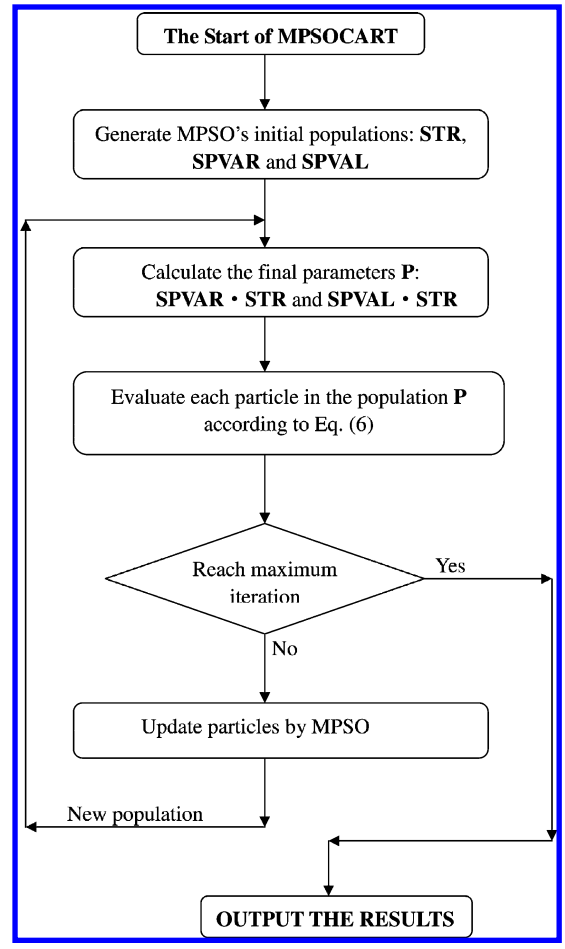
Step 4: Update the particles with the first three bits in **STR** fixed at 1. Return to step 2 to run the next iteration. The flowchart of MPSOCART is given in Scheme 1.

2.4. Fitness Function. In the current study, a predefined fitness function is employed to evaluate the performance of each particle in MPSOCART, whose minimization would yield a globally optimal CART. Constructing a CART by MPSO should inevitably search the optimal tree architecture and the optimal series of splitting parameters. In the light of these demands, an objective function is designed as follows:

$$f = \mathrm{RMSE}(1 + \lambda \times (m/\sum_{l=1}^{L} 2^{(l-1)})) \qquad (6)$$

where RMSE (root-mean-squared error) refers to the accuracy of CART model; the term $m/\sum_{l=1}^{L}2^{(l-1)}$ stands for the tree complexity, in which m is the number of nodes really involved in the CART computation; $\sum_{l=1}^{L}2^{(l-1)}$ is the number of nodes in a fully symmetrical tree, including the nodes in the last level which are not encoded in the particles; the term λ is a weighting coefficient controlling the trade-off between the tree accuracy and complexity. The larger is the value of λ, the simpler is the CART structure. The larger value of λ may prevent the convergence of CART, and the algorithm may converge to a poor solution. On the contrary, a smaller value of λ is in favor of the algorithm to converge more easily but may possibly result in overfitting. From experience, λ is set to 1 to keep the balance between the CART accuracy and complexity.

Scheme 1. Flow Chart of the MPSOCART Scheme



3. DATA SETS

3.1. Flavonoid Derivatives as p56lck Tyrosine Kinase Inhibitors Data. To evaluate the performance of the newly proposed MPSOCART algorithm, 104 flavonoid derivatives with their corresponding inhibitory activities to p56lck tyrosine kinase were used as a data set.[20] The inhibitory activity is expressed as IC_{50}, the molar concentration of the compound causing 50% inhibition of p56lck tyrosine kinase. These104 flavonoid derivatives were randomly divided into a training set of 80 compounds and a test set of 24 compounds. The parent structure of flavonoid derivatives is presented in Figure 2. A series of molecular descriptors representing the chemical structure were calculated by using Material Studio 4.0 software system, including structure, spatial, thermodynamic, electronic, topological descriptors, and *E*-state indices. In addition to the preceding descriptors, the seven variables used by Thakur[20] were also considered in the current study, including hydration energy (He) and

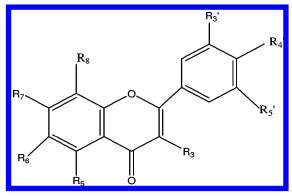


Figure 2. Parent structure of flavonoid derivatives used in the current study.

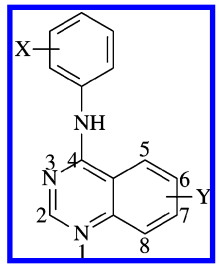


Figure 3. Basic molecuar structure of 4-(X-phenylamino)-Y-quinazoline.

six indicatory parameters (i.e., I_1, I_3, I_{OH}, I_{NH}, I_{NO_2}, I_{OMe}). I_3 is an indication of the presence of substituent at R_3 position by $I_3 = 1$ otherwise 0. I_{NH}, I_1, I_{NO_2}, and I_{OMe}, respectively, represent the presence of amino, hydroxyl, nitro, and methoxy groups at any position by $I_{NH} = 1$, $I_1 = 1$, $I_{NO_2} = 1$, and $I_{OMe} = 1$, otherwise 0. I_{OH} equals to 1, when hydroxyl is present at phenyl ring, or else, I_{OH} equals to 0.

3.2. Inhibitors of the Epidermal Growth Factor Receptor Tyrosine Kinase Data. Sixty-one 4-(X-phenylamino-)-Y-quinazolines as inhibitors of epidermal growth factor receptor (EGFR) tyrosine kinase[21] were employed as another data set to further check the validity of MPSOCART in QSAR studies, coupled with their corresponding bioactivities reported by Bridges et al.[22] The inhibitory activity is expressed as the molar concentration IC_{50}, that is, the 50% inhibition concentration. Figure 3 shows the molecular structure of EGFR inhibitors. The data set was stochastically split into two subsets: including a training set of 49 compounds and a test set of 12 compounds. For each compound, over 70 descriptors were calculated using Cerius[2], 3.5, software system on a Silicon Graphic R3000 workstation, including structure, spatial, thermodynamic, electronic, topological descriptors, and *E*-state indices. In addition to the descriptors calculated, another five descriptors reported in ref 21 were also under consideration for QSAR modeling, including the hydrophobic character (ClogP); indicator variable I ($I = 1$ indicates the presence of 6,7-di-OMe derivatives), steric parameter $B1_{Y,7}$ and $B1_{X,3}$, the electronic descriptor σ^-_Y.

The algorithms used in the present study were written in Matlab 5.3 and run on a personal computer (Intel Pentium processor 4/2.66 GHz 256 MB RAM).

4. RESULTS AND DISCUSSION

4.1. Flavonoid Derivatives as p56lck Tyrosine Kinase Inhibitors Data. Generally, the performance of CART is sensitive to the total levels of a tree. Since the nodes of no use are automatically set to be vanished during CART architecture optimization, increasing the levels of CART pays little effect on the tree performance. In the present study, as for the flavonoid derivatives data, even the number of levels is set as 20, the symptom of overfitting does not easily occur.

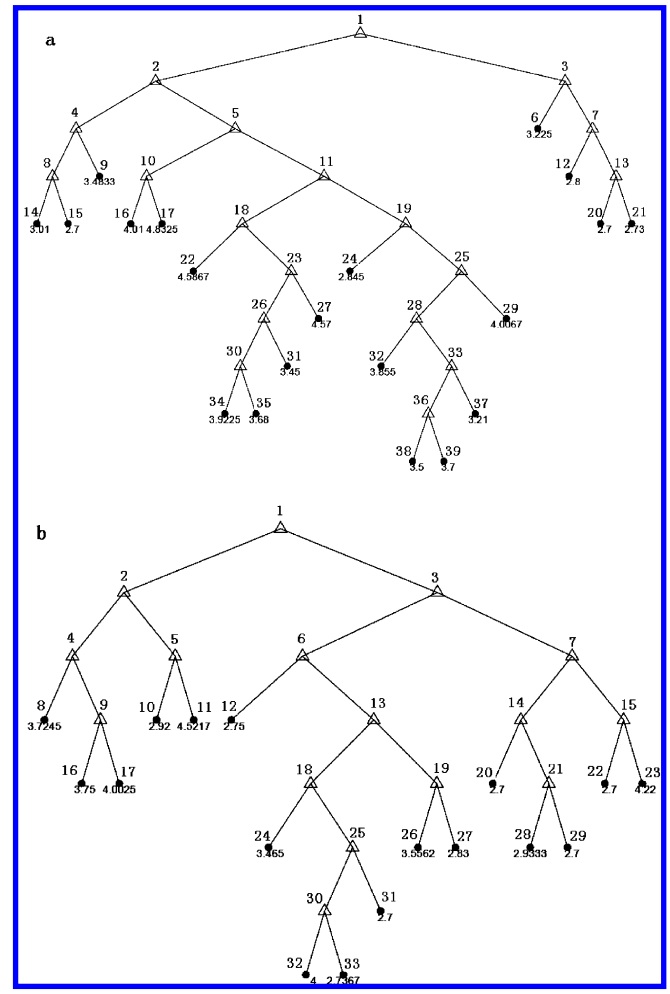


Figure 4. CARTs configured by recursive partition (a) and MPSO (b) for flavonoid derivatives, where Δ represents the internal node and ● refers to the leaf node.

Table 1. Results of QSAR Analysis of Flavonoid Derivatives Using MPSOCART Compared with Those Obtained by PLS and CART Configured by Greedy Recursive Partition

	R (correlation coefficient)			RSS (sum of squared residual)		
data set	method 1[a]	method 2[b]	method 3[c]	method 1[a]	method 2[b]	method 3[c]
training set	0.78	0.99	0.90	0.43	0.12	0.30
test set	0.79	0.74	0.92	0.48	0.55	0.32

[a] QSAR study by PLS. [b] QSAR study by RPCART. [c] QSAR study using MPSOCART.

Consequently, the fully symmetric tree is set as 10 levels, based on the immediate two considerations. One is the size of data set, and the other refers to the fact that CART with less complexity can reduce the enormous learning search space and is usually faster and cheaper to be constructed.

To check the validity of newly proposed MPSOCART approach in QSAR studies, as a comparison, the CART induced by greedy recursive partitioning (RPCART) was first used to model the flavonoid derivatives. Tree by RPCART for this data set is shown in Figure 4a. Table 1 of Supporting Information lists its associated node information, that is, the splitting variable and value in splitable node, node output as the calculated bioactivity of unseen sample to be allocated in this node, and node size in each node. Here, the mean

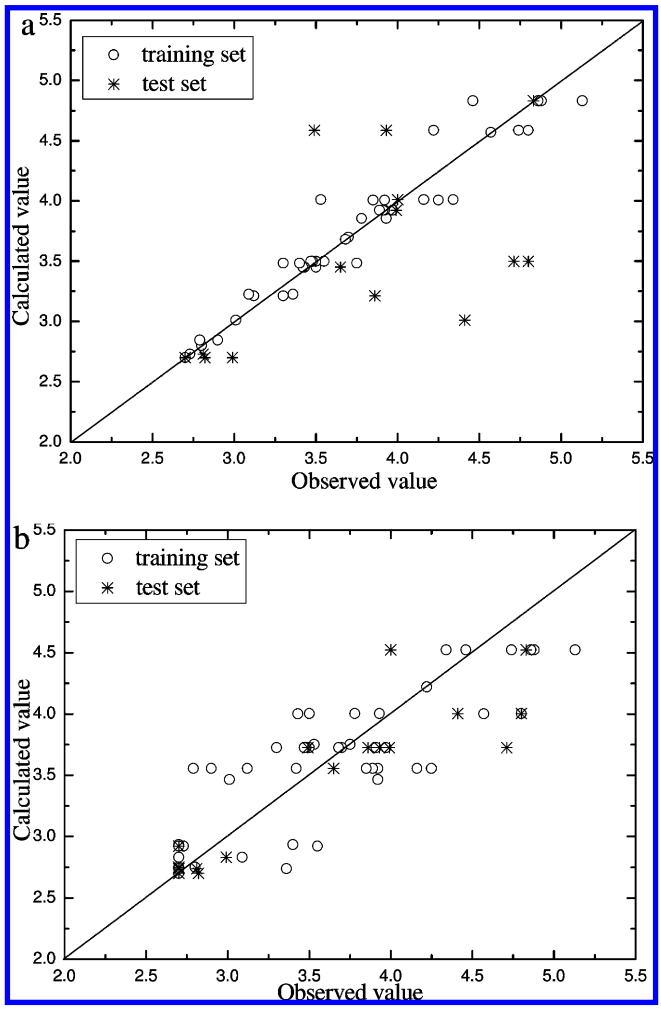


Figure 5. Calculated versus observed inhibitory activities of flavonoid derivatives by RPCART (a) and MPSOCART (b).

Table 2. Node Information Associated with CART Configured by MPSO for Flavonoid Derivatives[a]

node	splitting variable	splitting value	node size	node output
1	Shadow_ZX	48.3466	80	3.24
2	PMI_Y	5760	26	3.83
3	Jurs-WPSA-1	205.424	54	2.97
4	quadrupole zz	−13.757	16	3.80
5	FPSA-3	0.1068	10	3.88
6	RadOfGyration	4.17885	36	3.04
7	ellipsoidal volume	1370	18	2.82
8	–	–	11	3.72
9	S_aasC	0.5379	5	3.95
10	–	–	4	2.92
11	–	–	6	4.52
12	–	–	2	2.75
13	FPSA-3	0.9553	34	3.05
14	quadrupole zz	−17.552	8	2.79
15	dipole moment Y	3.7862	10	2.85
16	–	–	1	3.75
17	–	–	4	4.00
18	S_dO	12.082	23	2.91
19	Jurs-RPCG	0.1273	11	3.36
20	–	–	2	2.70
21	Jurs-FNSA-3	−0.1294	6	2.82
22	–	–	9	2.70
23	–	–	1	4.22
24	–	–	2	3.47
25	octupole xzz	57.258	21	2.86
26	–	–	8	3.56
27	–	–	3	2.83
28	–	–	3	2.93
29	–	–	3	2.70
30	Hf	259.15	20	2.86
31	–	–	1	2.70
32	–	–	2	4.00
33	–	–	18	2.74

[a] Symbol "–" represents the leaf nodes without splitting variables and values.

value of the bioactivities of training compounds assigned in the *t*th node is taken as the output of node *t*. The splitting rule means that the compounds with their values of splitting variable larger than or equaling to are placed into the right subsidiary node, or else into the left subsidiary node.

Table 1 indicates the statistical results by RPCART for this data set, from which one can obtain that RPCART yielded the correlation coefficients of 0.99 and 0.74, respectively, for the training and test sets. The correlation coefficient for the training set is significantly higher than that for the test set, indicating the occurrence of overfitting and suboptima. Table 1 also lists the RMSEs of 0.12 and 0.55 by RPCART, respectively, for the training and test sets. The correlation between the calculated and observed values of $\log 1/IC_{50}$ of flavonoid derivatives is presented in Figure 5a. As shown in Figure 5a and Table 1, CART configured by conventional inducer is home to poor generalization ability, limiting its application in practical issues. This may be due to the fact that greedy splitting heuristics are per se suboptimal and overfitting, and the tree-growing and tree-pruning are two separately steps.

To improve the QSAR model, the MPSOCART algorithm was designed to predict the bioactivities of flavonoid derivatives. In MPSOCART, MPSO was invoked to adaptively configure a globally optimal CART by simultaneously searching the optimal splitting parameters and the appropriate tree size. Taking the size of this data set into account, a fully symmetrical CART with 10 levels was used and the population size of PSO was identified as 50 in MPSOCART. The statistical results by MPSOCART are also enumerated in Table 1. MPSOCART yielded the correlation coefficients of 0.90 and 0.92, respectively, for the training and test sets. The correlation coefficients of the two subsets are comparable, indicating that CART configured by MPSO shows enough accuracy with no symptom of overfitting. The same is true for the RMSEs for the training and test sets. As shown in Table 1, compared with RPCART, MPSOCART yielded a slightly larger RMSE for the training set while yielding a much smaller RMSE for the test set, indicating that this newly designed MPSOCART algorithm is capable of circumventing the issues of overfitting and suboptima encountered in conventional CART generator. The correlation of calculated and observed bioactivities is depicted in Figure 5b. The comparison between Figure 5a and b reveals that MPSOCART offers a relatively smaller deviation between the calculated and observed activities than RPCART. These outcomes further testify the superior generalization ability of MPSOCART to RPCART. Tree by MPSOCART is demonstrated in Figure 4b, and its corresponding node information is itemized in Table 2. A comparison of Figure 4a with b indicates that the tree obtained by MPSO is more parsimonious and thus holds better generalization ability than that by greedy recursive partitioning. All of these reveals that the introduction of MPSO to simultaneously carry out

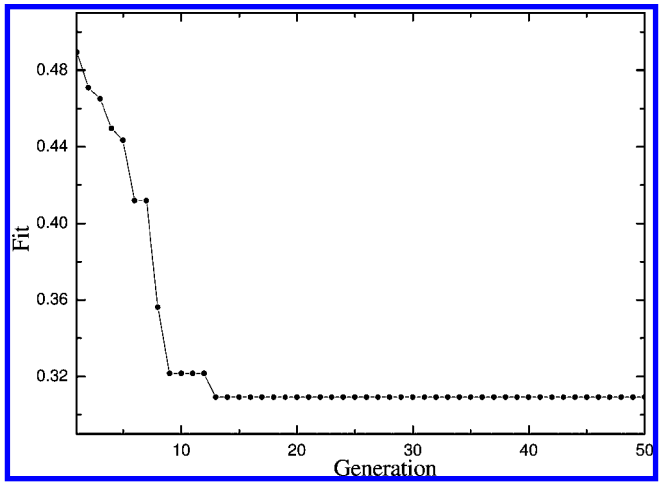


Figure 6. Convergence curve of MPSOCART for flavonoid derivatives.

the parameter and structure optimization tasks is very beneficial for overcoming the overfitting and suboptima issues and enhancing the characteristic performance of the configured CART. This may be due to the fact that MPSO can both holds population multiformity and maintain the best particle among all the particles in the population. The way used to retain the diversity of populations well prevents from the issue of trapping into suboptima. In addition, when growing a tree, the joint action between the tree complexity and the splitting parameters is taken into account, that is, aggregating the procedures of tree growing and pruning as one step while these procedures in RPCART are operated step by step. In addition, greedy recursive partitioning is per se suboptima and overfitting. This may be the essential factor resulting in serious overfitting and suboptima of RPCART. Figure 6 depicts the convergence process for MPSOCART. By visual observation of Figure 6, one can see that MPSO converges to the best solution quickly. Only 13 iterations are needed for obtaining a best solution. The time required to perform the proposed algorithm is only several minutes.

It is well-known that CART holds great promise in approximating strong nonlinearity relationship among data set and selecting automatically the optimal variable from a large number of variables submitted for analysis. As to this data set, the important descriptors chosen by MPSOCART, which acted as the splitting variables, are as follows: two *E*-state indices (S_dO and S_aasC), dipole moment Y, four Jurs descriptors (Jurs-FPSA-3, Jurs-WPSA-1, Jurs-RPCG, and Jurs-FNSA-3), VAMP electrostatics descriptors (quadrupole zz and octupole xzz), Hf, RadOfGyration, PMI_Y, ellipsoidal volume, and Shadow_ZX, shown in Table 2. In the present study, to reveal the strong competence of CART to fit nonlinearities, PLS was also used to model the flavonoid derivatives data using the descriptors selected by MPSOCART as the original input variables. Its results, together with those of MPSOCART and RPCART, are summarized in Table 1. From Table 1, one can obtain that the model by PLS yielded a correlation coefficient of 0.78 and a RMSE of 0.43 for the training set, and a correlation coefficient of 0.79 and a RMSE of 0.48 for the test set, indicating the correlation is rather poor and the modeling error is quite high. The results by PLS is obviously inferior to those by MPSOCART. It seems that PLS is inadequate for modeling this data set because of the presence of nonlinearity among this data set. Consequently, it is inevitable to introduce CART to well-perform QSAR studies of flavonoid derivatives.

Generally, the descriptor, as a splitting variable in the root node, is the most important variable for describing the compounds. In addition, the descriptor, which is closer to the root node or emerges in a CART with higher frequency, is the more important variable for describing the compound. The emerging order and frequencies of these descriptors in the tree by MPSOCART can be observed in Table 2. From Figure 4b and Table 2, one can see that Shadow_ZX acts as the splitting variable in the root node to divide the entire training data into two subsidiary nodes, indicating that Shadow_ZX pays the most important effect on the inhibitory activities of flavonoid derivatives to p56lck tyrosine kinase. This set of geometric descriptors helps to characterize the shape of the molecules, calculated by projecting the molecular surface on three mutually perpendicular planes XY, YZ, and XZ. These descriptors depend not only on the conformation but also on the orientation of the molecule. Shadow_ZX represents the area of the molecular shadow in XZ plane. Moreover, as shown in Figure 4b and Table 2, the compound with its value of Shadow_ZX less than 48.3466 is distributed into the left subsidiary node of the root node, whose output is 3.83, or else, is assigned in the right subsidiary one with an output of 2.97. The node output of the left node (i.e., node 2) is larger than that of the right one (i.e., node 3), revealing that the increase of Shadow_ZX makes against the bioactivity, that is, the smaller shadow area of the molecule on the XZ plane favors the inhibitory action. From Figure 4b and Table 2, one can easily gain that PMI_Y as the splitting variable in node 2 plays a positive role in compound bioactivity. PMI_Y is a spatial descriptor to compute the principal moments of inertia about the principal axes of a molecule. The spatial descriptor Jurs-WPSA-1 in node 3 is well correlated with inhibitory activity, as demonstrated in Figure 4b and Table 2. The compounds with the Jurs-WPSA-1 values larger than or equaling to 205.424 are placed into the right child node (i.e., node 6); otherwise, these compounds are into the left child one (i.e., node 7). As shown in Table 2, the outputs of the nodes 6 and 7 are 3.04 and 2.82, respectively, indicating that the decrease of the value of Jurs-WPSA-1 enhances the bioactivities of flavonoid derivatives. From Figure 4b and Table 2, one can also easily see the information that another three Jurs descriptors including Jurs-FPSA-3, Jurs-RPCG and Jurs-FNSA-3 also present their ponderance in the inhibitory action. The Jurs-like descriptors combine shape and electronic information to characterize the molecules, computed by mapping the atomic partial charges on solvent accessible surface areas of individual atoms. Jurs-FPSA-3 (in node 5) favors the inhibitory action, while Jurs-RPCG ((in node 19) and Jurs-FNSA-3 ((in node 21) show the negative role in bioactivity. Moreover, as shown in Table 2, the Jurs-FPSA-3 is twice taken as the splitting variable in nodes 5 and 13, further indicating the importance of this descriptor for inhibitory activities. All of these revealed that Jurs descriptors play an important role in predicting the compound bioactivities. Another two spatial descriptors, radius of gyration (RadOfGyration) and ellipsoidal volume, respectively, as the splitting variables in nodes 6 and 7, play more or less active role in the compound bioactivities, as demonstrated in Figure 4b and Table 2. Except for the spatial descriptors, the two

Table 3. Results of QSAR Analysis of Inhibitors of Epidermal Growth Factor Receptor (EGFR) Tyrosine Kinase Using MPSOCART Compared with Those Obtained by CART Configured by Recursive Partition

	R (correlation coefficient)		RSS (sum of squared residual)	
data set	method 2[a]	method 3[b]	method 2[a]	method 3[b]
training set	0.96	0.86	0.42	0.71
test set	0.60	0.87	1.21	0.76

[a] QSAR study by RPCART. [b] QSAR study using MPSOCART.

E-state indices (S_aasC (in node 9) and S_dO (in node 18) are well associated with the compound bioactivity, representing the electron accessibility associated with each atom type. Via a visual inspection of the node information of nodes 9, 16, 17, 18, 24, and 25, one can gain that S_aasC is beneficial for the compound inhibitory activity, while S_dO goes against the inhibitory activity. They indicate the presence/absence of a given atom type. In addition to the above-mentioned descriptors, dipole moment Y (in node 15) makes for the inhibitory activities of flavonoid derivatives. As to descriptors quadrupole zz (in nodes 4 and 14), octupole xzz (in node 25), and Hf (in node 30), the increase of their values leads to the drop of the compound bioactivities. In addition, quadrupole zz acts as the splitting variables twice, indicating the compact relationship between quadrupole zz and the bioactivity. From the above-mentioned analysis, we can acquire some information that the inhibitive action of flavonoid derivatives to p56lck tyrosine kinase is complex, including spatial, thermodynamic, and electronic effects, especially for spatial aspect.

4.2. Inhibitors of the Epidermal Growth Factor Receptor Tyrosine Kinase Data. For the sake of further verifying the validity of the proposed algorithm in QSAR studies, we applied MPSOCART for predicting the inhibitory activities of EGFR inhibitors. The RPCART, as a comparison with MPSOCART, was also investigated. The statistical results by these two CART generators are demonstrated in Table 3. For the training set, the RMSE of 0.42 and the correlation coefficient of 0.96 were obtained by RPCART, as shown in Table 3. Generalization of the model by RPCART to the test set produced a RMSE of 1.21 and a correlation coefficient of 0.60. From Table 3, one can acquire that the correlation coefficient for the training set is obviously higher than that for the test one, and meantime, RMSE for the training set is significantly less than that for the test one. All of these can easily lead to a conclusion that the phenomenon of overfitting occurs. Figure 7a describes the tree model obtained by RPCART. Its associated node information is listed in Table 2 of Supporting Information. The correlation of the calculated and observed bioactivities of EGFR inhibitors is presented in Figure 8a. Via a visual inspection of Figure 8a, one can obtain the information that the deviation between the calculated and observed bioactivities for the test set is rather large, further indicating that RPCART is unsuitable for modeling the EGFR inhibitor data for its poor generalization ability.

For this data set, the levels of the fully symmetrical CART and the population size of PSO in MPSOCART were, respectively, set as 8 and 50, with a view to the size of EGFR inhibitors. The correlation coefficients for the training and test sets by MPSOCART were, respectively, 0.86 and 0.87. The RMSEs of 0.71 and 0.76 were obtained by MPSOCART for these two subsets. The above-listed statistical results reveal that RMSEs of the two subsets are comparable for MPSOCART, so are the correlation coefficients of the two subsets. This indicates that no sign of overfitting occurs. When refers to the comparison with RPCART algorithm, MPSOCART yields a relatively lower correlation coefficient for the training set while much higher correlation coefficient for the test set, indicating that MPSOCART holds much more ameliorated generalization capabilities. The calculated versus observed bioactivity values of EGFR inhibitors by MPSOCART is shown in Figure 8b. The comparison between Figure 8a and b obviously demonstrates that MPSOCART shows much smaller deviations in the bioactivity estimation for test set than RPCART. These results further confirm that MPSOCART is provided with the strength of compensating for overfitting always encountered in conventional CART inducing algorithm. All of these well suggest that MPSO is well suitable for configuring globally optimal CART. The CART by MPSO and its associated node information for this data set are represented in Figure 7b and Table 4, respectively. When compared with Figure 7a and b, one can gain that a more parsimonious CART with excellent generalization potential can be obtained by MPSOCART. Figure 9 shows the rapid convergence of MPSO to the optimal solution. The obvious plane section of the curve suggests the optimal iteration number is 17.

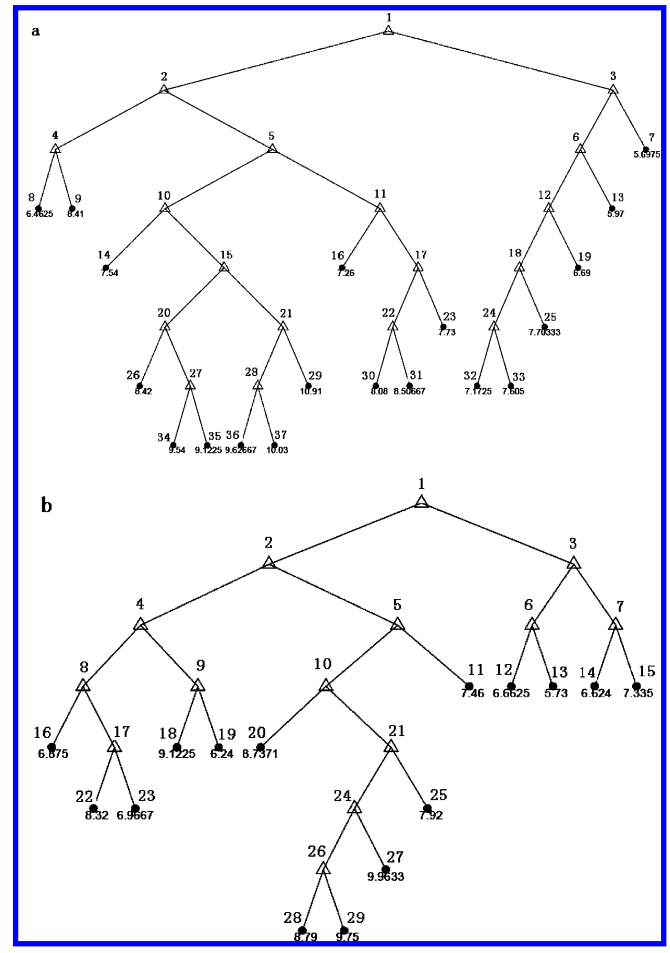


Figure 7. CARTs configured by recursive partition (a) and MPSO (b) for EGFR tyrosine kinase inhibitors, where Δ represents the internal node and • refers to the leaf node.

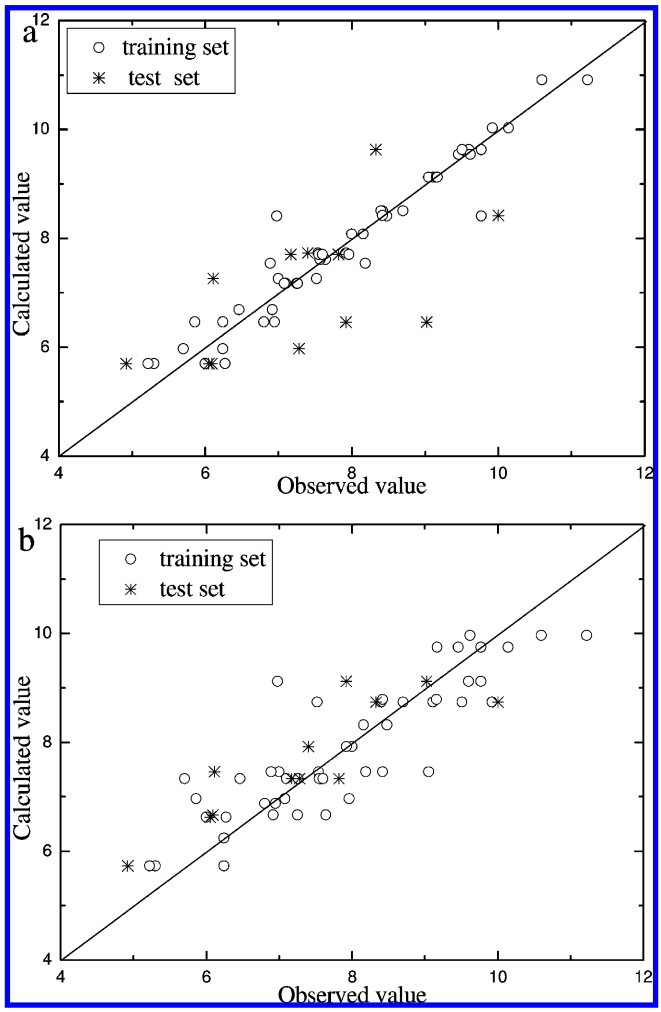


Figure 8. Calculated versus observed bioactivities of EGFR tyrosine kinase inhibitors by RPCART (a) and MPSOCART (b).

The descriptors selected as splitting variables in MPSOCART consist of two desolvation energy (Fh2o and Foct), LogP, two dipole moments (Dipole_X and Dipole_Y), highest occupied molecular orbit energy (HOMO), three shadow indices (Shadow_nu, Shadow_Zlength, and Shadow_XZ), Sr, and PMI_Y. As shown in Table 4, Fh2o twice acts as the splitting variables, respectively, in nodes 1 and 24, indicating that Fh2o is the foremost variable for the inhibitory activity. The compound in root node, whose Fh2o value is smaller than the splitting value of −20.133, is assigned into the left child node (i.e., node 2) with a node output of 8.38, or else, it is split into the right child one (i.e., node 3) with a node output of 6.61. Such a phenomenon means that Fh2o goes against the bioactivity of the EGFR inhibitor. Although Fh2o, as the splitting variable in node 24, seems to be favorable for the compound activity. However, it is worth to be notified that the splitting variable in root node is the most important factor in effecting the inhibitive action. Another descriptor associated with dissolvation energy, Foct in node 21, pays an effect on the inhibitory operation in some extent, as shown in Table 4. The compounds, with Foct value not less than the critical value of −26.607, are divided into the right subsidiary node (i.e., node 25) of its corresponding parent node. Otherwise, they are assigned into the left one (i. e., node 24). Fh2o and Foct are physiochemical properties associated with linear free energy model of a molecule, proven useful as molecular

Table 4. Node Information Associated with CART Configured by MPSO for EGFR Tyrosine Kinase Inhibitors[a]

node	splitting variable	splitting value	node size	node output
1	Fh2o	−20.133	49	7.91
2	Dipole_Y	−2.319	36	8.38
3	HOMO	−10.4191	13	6.61
4	Shadow_nu	2.08	12	7.84
5	Dipole_Y	1.068	24	8.66
6	Shadow_Zlength	6.279	6	6.35
7	Sr	0.7046	7	6.83
8	Shadow_XZ	0.573	7	7.33
9	Dipole_X	10.125	5	8.55
10	PMI_Y	788.133	17	9.15
11	–	–	7	7.46
12	–	–	4	6.66
13	–	–	2	5.73
14	–	–	5	6.62
15	–	–	2	7.34
16	–	–	2	6.88
17	LogP	1.12	5	7.51
18	–	–	4	9.12
19	–	–	1	6.24
20	–	–	7	8.74
21	Foct	−26.607	10	9.44
22	–	–	2	8.32
23	–	–	3	6.97
24	Fh2o	25.494	9	9.61
25	–	–	1	7.92
26	LogP	0.4976	6	9.43
27	–	–	3	9.96
28	–	–	2	8.79
29	–	–	4	9.75

[a] Symbol "–" represents the leaf nodes without splitting variables and values.

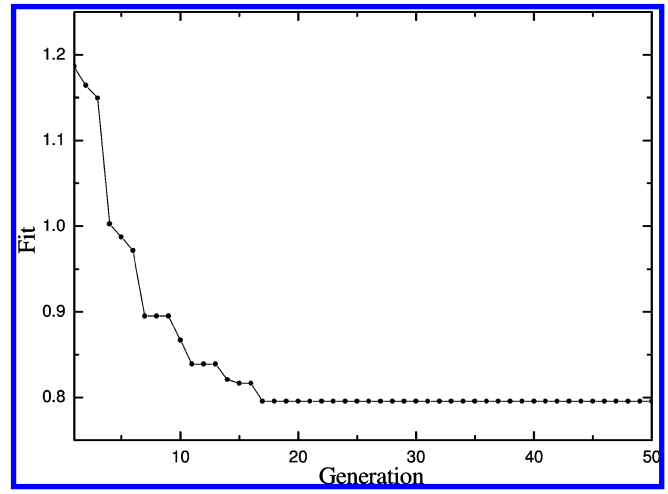


Figure 9. Convergence curve for MPSOCART for EGFR tyrosine kinase inhibitors.

descriptors in structure− activity analysis. For this data set, both Fh2o and Foct go against the compound bioactivity. In addition, electronic descriptors Dipole_Y (in nodes 2 and 5) and Dipole_X (in node 9) correlate well with inhibitory action, shown in Figure 7b and Table 4. The dipole moment descriptors are 3D electronic ones that refer to the strength and orientation behavior of a molecule in an electrostatic field, computed by using partial atomic charges and coordinates. Another electronic descriptor Sr (in node 7) plays an important role in predicting the compound activity. A visual inspection of node information of nodes 7, 14 and 15, one can discover that increasing Sr helps to improve the compound bioactivity. HOMO (in node 3) is crucial in

governing molecular reactivity and properties. Molecules with high HOMOs are more likely to be able to donate their electrons and are hence relatively reactive compared to molecules with low-lying HOMOs. HOMO is a measure of the nucleophilicity of a molecule. From the CART configured by MPSOCART, the same conclusion is obtained, i.e., HOMO partially accounts for the enhancement of activity. Shadow_nu (in node 4), Shadow_Zlength (in node 6) and Shadow_XZ (in node 8) also make contribution on inhibitory action with the positive ones by Shadow_nu and Shadow_XZ and the negative one by Shadow_Zlength, as indicated in Figure 7b and Table 4. Except the aforementioned descriptors, PMI_Y (in node 10) and LogP (in nodes 17 and 26) are the effective variables in the sense of EGFR inhibitors, too. Consequently, the interaction of EGFR inhibitors is a complex one, including thermodynamic, electronic and spatial effects.

5. CONCLUSION

In the present paper, the modified particle swarm optimization (MPSO) method was introduced to configure a globally optimal CART, simultaneously searching the optimal splitting parameters and the right architecture of CART. During optimization, a fitness function was formulated to identify the proper tree structure and optimal splitting parameters. Coupled with PLS and RPCART, MPSOCART was assessed in terms of the performance in a case study of two QSAR data sets. The results demonstrated that MPSOCART yielded consistent improvements in comprehensibility, accuracy and generalization abilities over the conventional CART inducer and substantially superior ability in modeling nonlinearity to PLS. In addition, MPSO was effective for configuring globally optimal CART, converging quickly to the optimal solution and improving the generalization ability of CART in a great extent.

ACKNOWLEDGMENT

This work was supported by "973" National Key Basic Research Program (2007CB310500) and National Nature Science Foundation (Grant No. 20805017, 20775023, 20675028) of China and Scientific and Technical Key Problem Program of Wuhan (200860423220).

Supporting Information Available: Detailed node information for the Trees by RPCART for two data sets. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Breiman, L.; Friedman, J. H.; Olshen, R. J.; Stone, C. J. *Classification and Regression Trees*; Bickel, P. J., Cleveland W. S., Dudley R. M., Eds.; Wadsworth Internal Group: Belmont, CA, 1984.

(2) Daszykowski, M.; Walczak, B.; Xu, Q. S.; Daeyaert, F.; de Jonge, M. R.; Heeres, J.; Koymans, L. M. H.; Lewi, P. J.; Vinkers, H. M.; Janssen, P. A.; Massart, D. L. Classification and Regression Trees-Studies of HIV Reverse Transcriptase Inhibitors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 716–726.

(3) Chohan, K. K.; Paine, S. W.; Mistry, J.; Barton, P.; Davis, A. M. A Rapid Computational Filter for Cytochrome P450 1A2 Inhibition Potential of Compound Libraries. *J. Med. Chem.* **2005**, *48*, 5154–5161.

(4) Gleeson, M. P.; Waters, N. J.; Paine, S. W.; Davis, A. M. In Silico Human and Rat V_{ss} Quantitative Structure-Activity Relationship Models. *J. Med. Chem.* **2006**, *49*, 1953–1963.

(5) Crawford, S. L. Extensions to the CART Algorithm. *Int. J. of Man-Machine Studies* **1989**, *31*, 197–217.

(6) Gelfand, S. B.; Ravishankar, C. S.; Delp, E. J. An Iterative Growing and Pruning Algorithm for Classification Tree Design. *IEEE Trans. on Pattern Anal. and Mach. Intell.* **1991**, *13*, 163–174.

(7) Lutsko, J. F.; Kuijpers, B. Simulated Annealing in the Construction of Near-Optimal Decision Trees. In *AI and Statistics*; Cheesman P., Oldford R. W., Eds.; Springer-Verlag: New York, 1994; pp 453−462.

(8) Delisle, R. K.; Dioxon, S. L. Induction of Decision Trees via Evolutionary Programming. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 862–870.

(9) Buontempo, F. V.; Wang, X. Z.; Mwense, M.; Horan, N.; Young, A.; Osborn, D. Genetic Programming for the Induction of Decision Trees to Model Ecotoxicity Data. *J. Chem. Inf. Model.* **2005**, *45*, 904–912.

(10) Izrailev, S.; Agrafiotis, D. A Novel Method for Building Regression Tree Models for QSAR Based on Artificial Ant Colony Systems. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 176–180.

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

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

(13) Clerc, M.; Kennedy, J. The Particle Swarms Explosion, Stability, and Convergence in a Multidimensional Complex Space. *IEEE Trans Evol. Comput.* **2002**, *6*, 58–73.

(14) Shi, Y.; Eberhart, R. *Fuzzy Adaptive Particle Swarm Optimization*; Proceeding 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.

(15) Lin, W. Q.; Jiang, J. H.; Zhou, Y. P.; Wu, H. L.; Shen, G. L.; Yu, R. Q. Support Vector Machine Based Training of Multilayer Perceptron as Optimized by PSO: Application to QSAR Studies of Bioactivities of Chemical Compounds. *J. Comput. Chem.* **2007**, *28*, 519–527.

(16) Zhou, Y. P.; Jiang, J. H.; Lin, W. Q.; Zou, H. Y.; Wu, H. L.; Shen, G. L.; Yu, R. Q. Adaptive Configuring of Radial Basis Function Network by Hybrid Particle Swarm Algorithm for QSAR Studies of Organic Compounds. *J. Chem. Inf. Model.* **2006**, *46*, 2494–2501.

(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) Shen, Q.; Jiang, J. H.; Jiao, C. X.; Shen, G. L.; Yu, R. Q. Modified Particle Swarm Optimization Algorithm for Variable Selection in MLR and PLS Modeling: QSAR Studies of Antagonism of Angiotensin II Antagonists. *Eur. Pharm. Sci.* **2004**, *22*, 145–152.

(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) Thakur, A.; Vishwakarma, S.; Thakur, M. QSAR Study of Flavonoid Derivatives as P56lck Tyrosinkinase Inhibitors. *Bioorg. Med. Chem.* **2004**, *12*, 1209–1214.

(21) Kurup, A.; Garg, R.; Hansch, C. Comparative QSAR Study of Tyrosine Kinase Inhibitors. *Chem. Rev.* **2001**, *101*, 2573–2600.

(22) Bridges, A. J.; Zhous, H.; Cody, D. R.; Rewcastle, G. W.; Mcmichael, A.; Showalter, H. D. H.; Fry, D. W.; Kraker, A. J.; Demmy, W. A. Tyrosine Kinase Inhibitors. 8. An Unusually Steep Structure-Activity Relationship for Analogues of 4-(3-Bromoanilino)-6,7-dimethoxyquinazoline (PD153035), A Potent Inhibitor of the Epidermal Growth Factor Receptor. *J. Med. Chem.* **1996**, *39*, 267–276.

CI800374H