JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

ARTICLE

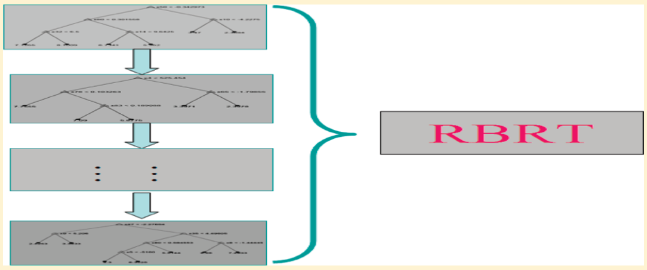
pubs.acs.org/jcim

A Robust Boosting Regression Tree with Applications in Quantitative Structure–Activity Relationship Studies of Organic Compounds

Jian Jiao, Shi-Miao Tan, Rui-Ming Luo, and Yan-Ping Zhou*

Key Laboratory of Pesticide and Chemical Biology of Ministry of Education, College of Chemistry, Central China Normal University, Wuhan 430079, P. R. China

S Supporting Information



ABSTRACT: A regression tree (RT) was extensively utilized in quantitative structure–activity relationship studies (QSAR), due to its inherently promising attributes. The issues of instability and inclination to overfitting and suboptima, however, often occur in RT. In the present study, a robust version of boosting was invoked to simultaneously improve the stability and generalization ability of RT, forming a new method called robust boosting regression tree (RBRT). RBRT works by sequentially employing the RT method to model the robustly reweighted versions of the original training set and then aggregating these resultant predictors via weighted median. The designed RBRT was applied to predict the bioactivities of flavoniod derivatives and the anti-HIV activities of HIV-1 inhibitors, compared with boosting RT (BRT) and RT. The results of these two data sets demonstrated that the introduction of robust boosting drastically enhances the stability and generalization ability of RT, and RBRT is superior to BRT in QSAR studies.

1. INTRODUCTION

Regression tree (RT)[1] as an important modeling approach has been extensively employed in quantitative structure–activity relationship (QSAR) studies.[2−5] The superiority of RT to other methods, such as partial least-squares (PLS) regression and artificial neural network (ANN), are attributed to its inherent potentials, including simplicity, interpretability, high capacity in handling large data sets, and immunities to heteroscedasticity, collinearity, and outliers, just to name a few.

Traditionally, the configuration of RT is performed by first growing the largest tree via greedy recursive partitioning and then pruning the grown tree to yield the final appropriately fit RT.[1] For the greedy recursive partitioning method, the splitting parameters are selected on the basis of the local rather than global measures, thus yielding local minima. RT induced by the recursive partitioning approach tends to fit noises and idiosyncrasies in training cases, which are unlikely to occur with the same pattern in unseen ones, therefore resulting in overfitting. Since the tree-growing and tree-pruning phases are two sequential and completely independent procedures, RT gets into suboptima with high frequency. In addition, it is well-known that RT is highly instable, that is, a small variation in training data leads to a large change in the computed results.[6] In brief, RT is exposed to high risks of poor generalization ability (i.e., overfitting and local optima) and high instability. Therefore, it is highly demanding to develop an approach capable of improving simultaneously the generalization ability and stability of RT.

Because of the great appeal of RT, many efforts have been made to improve the performance of RT. To improve the generalization ability of RT, some global optimization techniques have been invoked for inducing the globally optimal RTs.[4,7−9] Although the generalization ability of RT is indeed enhanced to a great extent, the problem of high instability still exists. It has recently been discovered that an ensemble of trees is one of the best ways to improve both the generalization ability and stability of a single tree.[10] The essence of the ensemble learning lies in the fact that the combination of multiple learners offers an improved performance over the individual one. The most prominent ensemble learning strategies are bagging, boosting, and random forest. In addition, a series of "forest" methods, e.g., random FIRM,[11] decision forest,[12] and recursive forest,[13] has attracted much attention in improving the performance of a single tree. Until now, it has been reported that boosting usually outperforms bagging, offering a comparable performance to random forest.[14] These methods are always used to improve the performance of trees for classification tasks,[6,15−17] while only a few studies dealing with ensemble of regression trees have been reported nowadays.[17,18]

As stated above, boosting is one of the most prominent ensemble learning procedures. Boosting,[19] which originated from the realm of machine learning, has recently attracted much increasing interest in cheminformatics.[6,15,20−22] Nowadays, only a few investigations have focused on the introduction of boosting to improve the performance of RT.[6,18] Boosting deals with the generic problem of formulating an accurate prediction rule by combining a series of rough and inaccurate rules of thumb. Such a

Received: October 26, 2010
Published: March 18, 2011

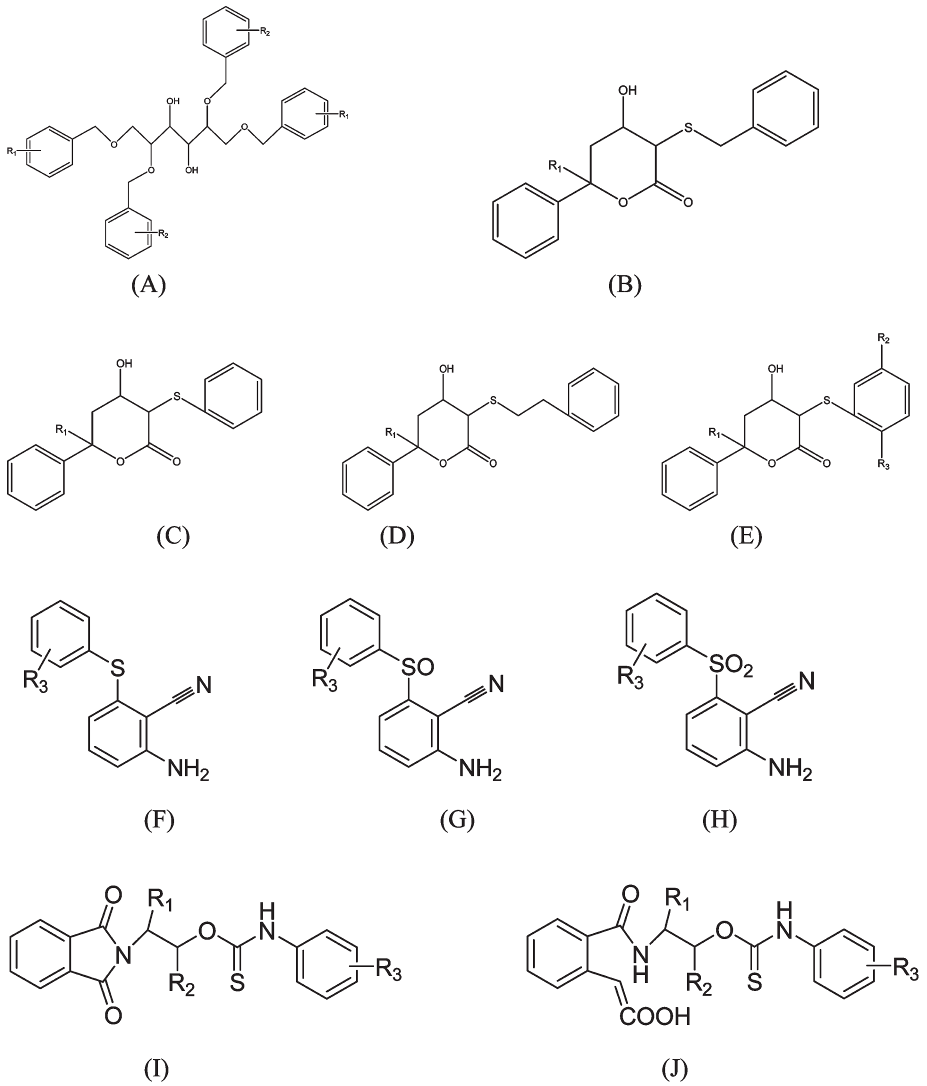


Figure 1. Parent structures of HIV-1 inhibitors used in the current study.

series of rules is iteratively constructed on the various weighted versions of the original training set. For the first rule, the weights of all the samples in the original training set are given unit. Then, the weight distributions for the subsequent rule are sequentially altered according to the performance of the previous rules constructed. The samples in the original training set worse predicted by the preceding rule are allocated higher weights for the next rule, implying that those instances most in error are more likely to be picked as the member of the training set for the subsequent rule. Such a weight-renovating strategy makes the instances of large errors be more and more overemphasized during iteration and then might lead boosting to be sensitive to outliers and noise, even overfitting them greatly.[22,23] Grove et al.,[24] Friedman et al.,[25] and Freund et al.[26] have noted that boosting actually exhibits noticeable overfitting on some practical data sets. To overcome such an issue of boosting, recently we have designed a robust version of boosting by introducing an error-trimmed technique before renovating the weights and used it to effectively improve the performance of PLS.[22] This error-trimmed technique is carried out by trimming those sample errors not smaller than a certain critical value as the median of all the original training sample errors obtained by the previous rule. Such an error-trimming procedure can effectively prevent the latter constructed rules from being excessively dependent on the samples of large errors so as to improve the robustness of boosting.

Inspired by the appealing properties of robust boosting and the drawbacks of RT, robust boosting has been invoked for improving the performance of RT (RBRT) in the current study. Two QSAR data sets have been employed to support the viewpoint that robust boosting is an efficacious way to improve the generalization ability and stability of RT while retain most of the appealing properties of RT. Thereinto, data analysis has been performed for each data set before QSAR modeling, including nonlinearity, clustering tendency, and outlier detections. Of course, we also keep in mind Wolpert's no free lunch theorem;[27] i.e., there is no one best algorithm for all the issues. Our goal is to show that RBRT is among the topmost algorithms, rather than the top one according to the prediction accuracy.

2. THEORY

2.1. Regression Tree (RT). RT, as a nonparametric method for regression, originated from classification and regression trees

(CART) by Bremain et al.[1] Here only a concise description of RT is presented.

Generally, the configuration of RT consists of three basic steps. First, the largest tree is grown by applying recursive partitioning. Recursive partitioning is conducted in a top-down fashion, starting from the root node containing all the training compounds until each node reaches complete homogeneity or a user-specified minimal sample number (i.e., minimal node size) and becomes a terminal or leaf node. Then, based on the minimal cost-complexity pruning (MCCP) criterion,[1] the largest tree is pruned to yield a sequence of nested subtrees, each holding an immanent complex parameter. Ultimately, from such nested subtrees, the final appropriately fit RT is selected in terms of its best prediction accuracy gained either by cross-validation or pruning set methods.[1] This is virtually performed by using cross-validation or pruning set methods to select the optimal complex parameter of the tree so as to identify the final tree of right size.

Once the final appropriately fit tree is gained, some immanent node information is endowed. Each splittable node is characterized by a splitting rule including the splitting variable and value. Each node is assigned the mean or median of the bioactivities of the involved compounds as the node output. The node size, i.e., 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.

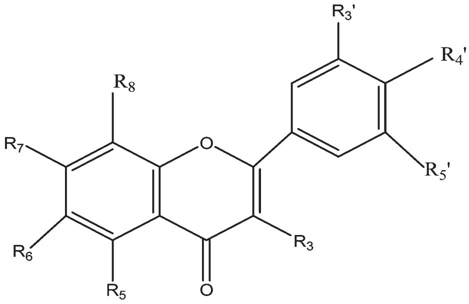


Figure 2. Parent structure of flavonoid derivatives used in the current study.

Table 1. Hopkins statistic for the HIV-1 inhibitors (n = 222)

number of iterations	population size (%)	$HS_{average}$	HS_{max}	HS_{min}	HS_{range}
5	20	0.9600	0.9729	0.9444	0.0285
5	100	0.9220	0.9492	0.8845	0.0647
10	10	0.9579	0.9786	0.9110	0.0675
20	5	0.9277	0.9932	0.5929	0.4003
50	5	0.9478	0.9936	0.5929	0.4008

2.2. Robust Boosting Regression Tree (RBRT).

To improve the performance of RT, a robust version of boosting (i.e., Adaboost.R)[22] was combined with RT to form a new algorithm, i.e., robust boosting regression tree (RBRT). RBRT is carried out by constructing a series of robust RT models followed by combining the predictions from the resultant RT models. An error-trimming technique was utilized to make the algorithm insensitive to outliers and noisy samples and then be resistant to overfitting and suboptima. The procedure is described as follows.

First, initialize the identical weights of all the original training compounds, $w_{i,1} = 1/I$ ($i = 1, ..., I$, where I is the size of the original training set). Then, for t = 1 to T (T is the ensemble size), perform the following steps.

(1) According to the weights $\mathbf{w}_t$, I samples, called the boosting set, are picked up with replacement from the original training set.

(2) Construct a RT model on the boosting set and use this RT model to estimate the bioactivities of the original training samples $y_{i,t}$ ($i = 1, ..., I$).

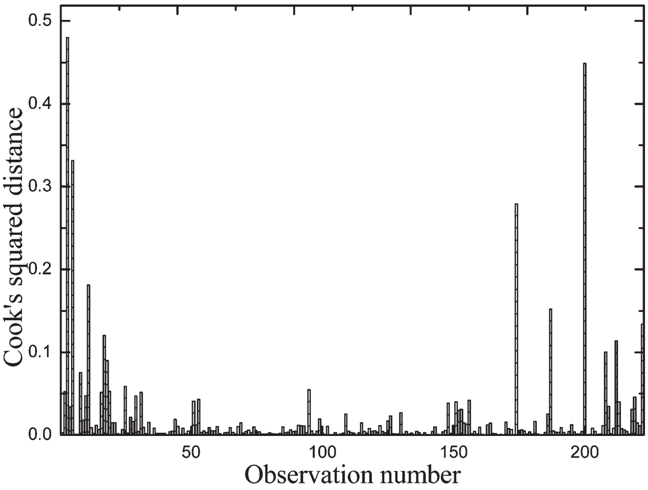


Figure 3. $CD^2_{(i)}$ values gained by Cook's squared distance method for HIV-1 inhibitors.

Table 2. Mean Performance over 10 Random Training/Test Splits of the HIV-1 Data Using RT, BRT, and RBRT[a]

		correlation coefficient (R)		root mean square error (RMSE)	
method	data set	MR	RR	MRMSE	RRMSE
RT	training set	0.9192	0.0683	0.7944	0.3252
	test set	0.9067	0.0901	0.8668	0.2690
BRT	training set	0.9783	0.0236	0.4254	0.2347
	test set	0.9353	0.0554	0.7352	0.1863
RBRT	training set	0.9754	0.0204	0.4564	0.1871
	test set	0.9428	0.0492	0.6914	0.1864

[a] MRMSE represents the mean of RMSEs for 10 random training/test divisions. RRMSE refers to the range of RMSEs for 10 random training/test divisions. MR and RR indicate the mean and range of correlation coefficients for 10 random training/test divisions, respectively.

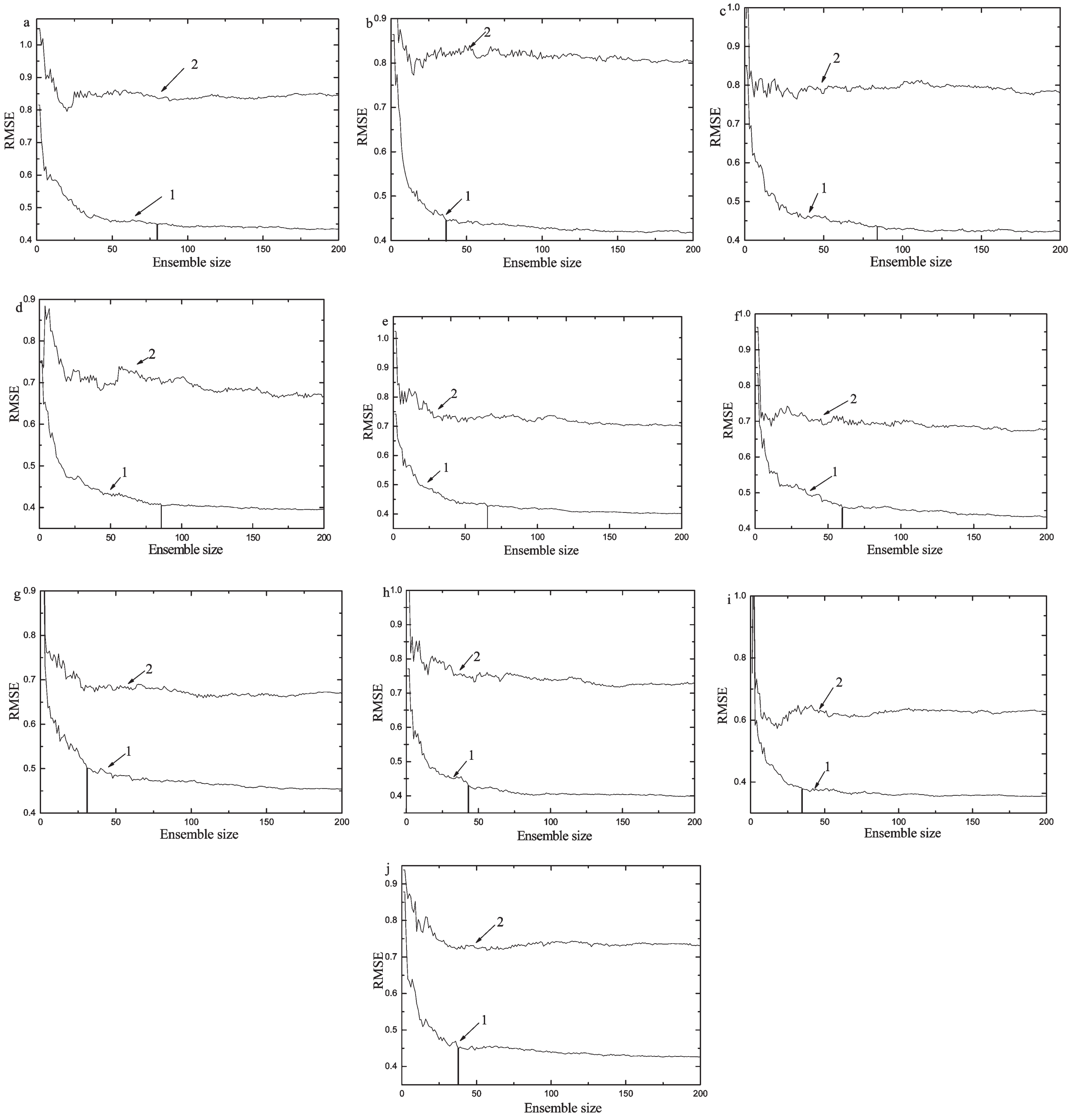


Figure 4. Convergence curves of BRT for HIV-1 inhibitor data for 10 random divisions of the whole data set. Curve 1 refers to the influence of ensemble size on RMSE of the original training set for each division. Curve 2 is the curve of RMSE for the test set versus ensemble size for each division. The vertical real lines indicate the optimal ensemble sizes selected for BRT in 10 computations.

(3) Let $e_{i,t} = |y_{i,t} - d_i|$ (where $e_{i,t}$ represents the absolute error of the ith sample in the tth cycle and d_i is the observed bioactivity of the ith training compound) and trim the large sample errors by the following formulation:

$$e_{\mathrm{inew},t} = \begin{cases} \mathrm{median}(\mathbf{e}_t) & e_{i,t} \geq \gamma \times \mathrm{median}(\mathbf{e}_t) \\ e_{i,t} & e_{i,t} < \gamma \times \mathrm{median}(\mathbf{e}_t) \end{cases} \qquad (1)$$

The parameter γ is a coefficient to weight the robustness of RBRT, generally determined by experience.

(4) Compute a loss value for every sample in the original training set:

$$L_{i,t} = 1 - \exp\left[- \frac{(e_{\mathrm{inew},t})}{\max(\mathbf{e}_{\mathrm{new},t})} \right] \qquad (2)$$

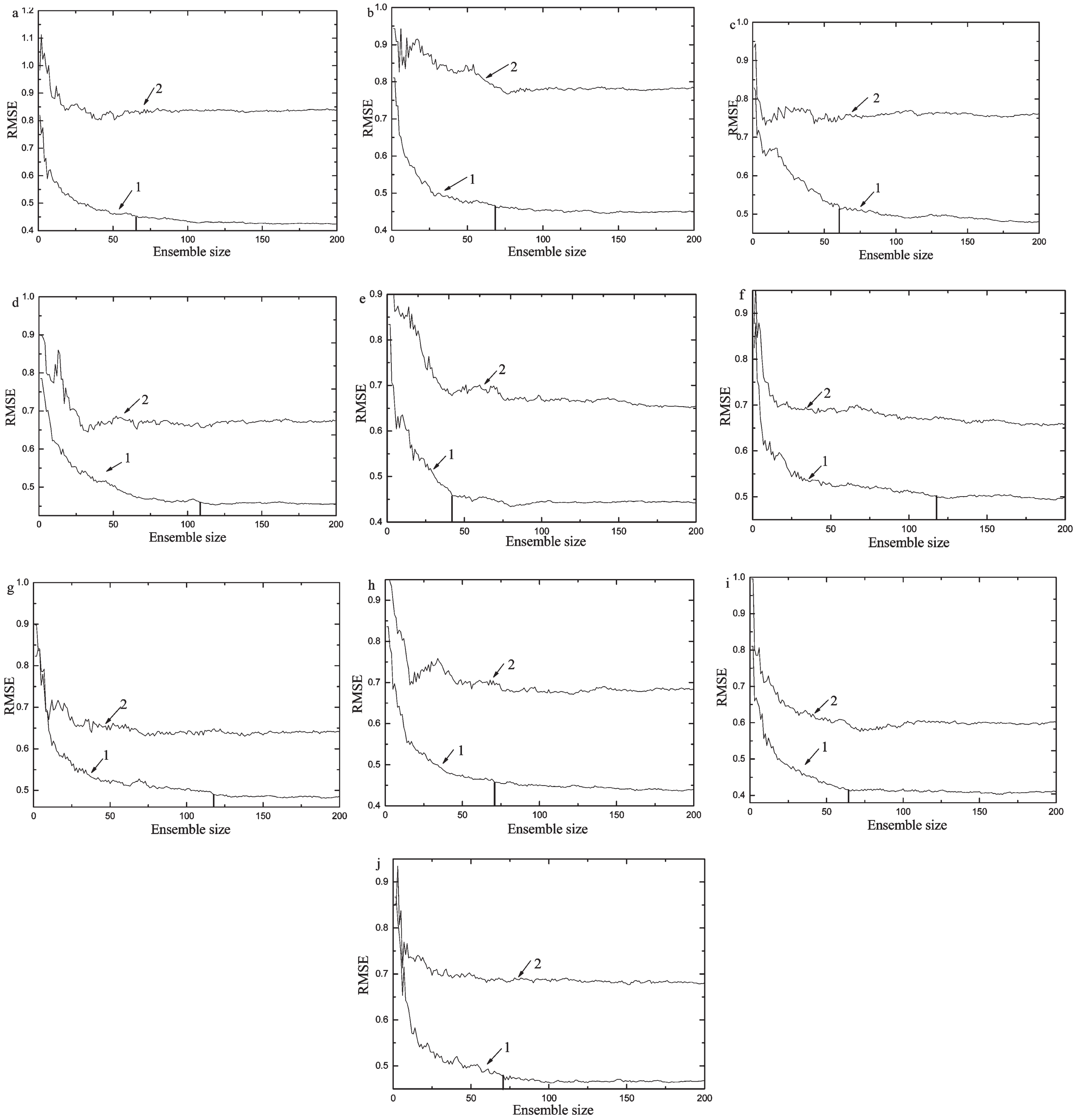


Figure 5. Convergence curves of RBRT for HIV-1 inhibitor data for 10 random divisions of the whole data set. Curve 1 refers to the influence of ensemble size on RMSE of the original training set for each division. Curve 2 is the curve of RMSE for the test set versus ensemble size for each division. The vertical real lines indicate the optimal ensemble sizes selected for RBRT in 10 computations.

(5) Calculate the mean loss:

$$\bar{L}_t = \sum_{i=1}^{I} L_{i,t} w_{i,t} \quad (3)$$

(6) Let $\beta_t = \bar{L}_t/(1 - \bar{L}_t)$ and update the weight of each compound in the original training set by using

$$w_{i,t+1} = w_{i,t}\beta_t^{(1 - L_{i,t})} \quad (4)$$

The new weights should be normalized, so that $\sum_{i=1}^{I} w_{i,t+1} = 1$. In eq 4, β_t, ranging from 0 to 1, acts as a confidence measure of the predictor. Higher β_t indicates lower confidence of the RT model. Such a weight-updating strategy suggests that sample weight would be increased with the increase of the sample error gained in step 3. The ensemble size T is vital to decrease the variability of the ensemble prediction and offer reliable estimation of the test samples. It

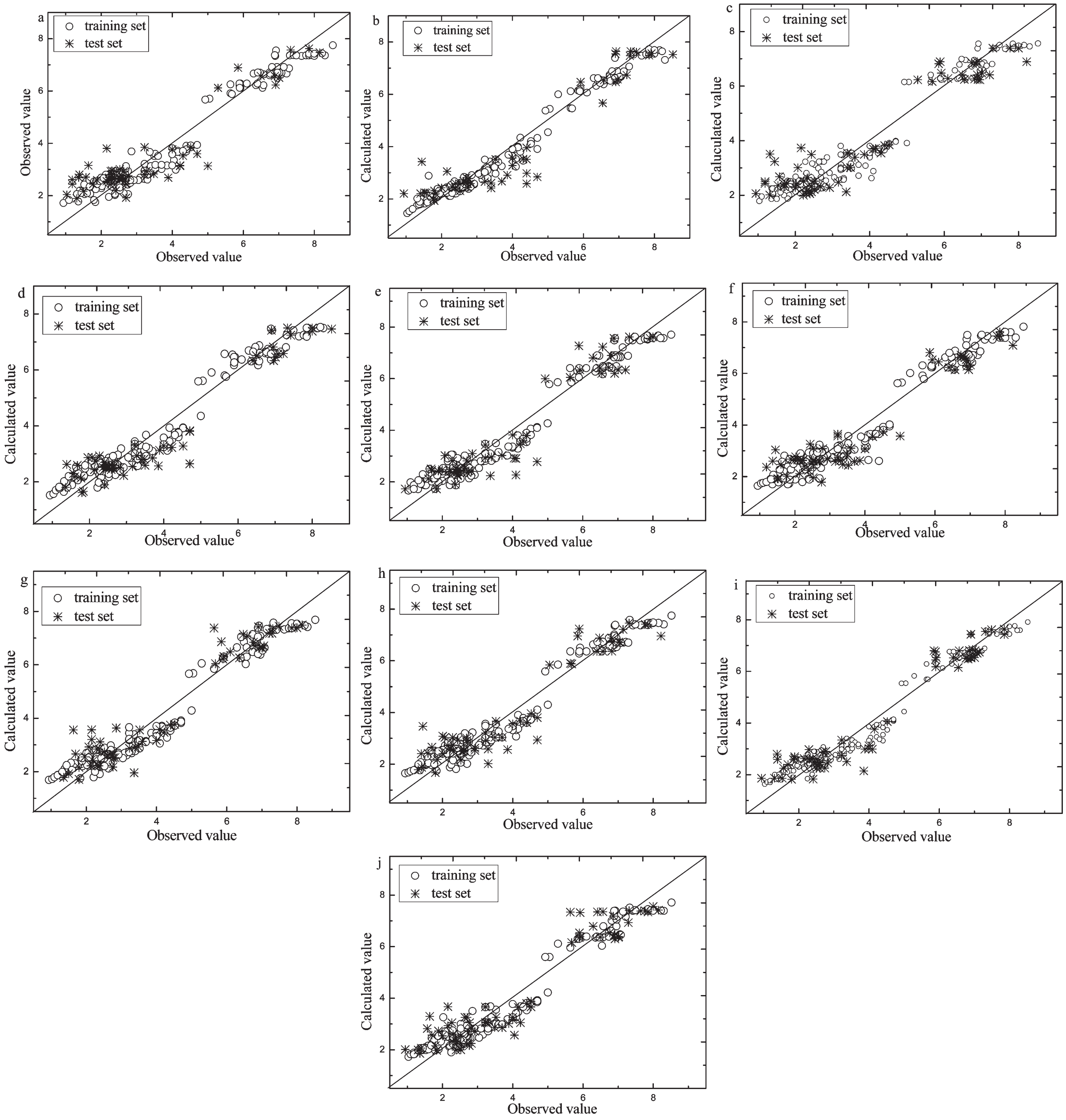


Figure 6. Observed versus calculated values of the bioactivities by RBRT for HIV-1 inhibitors in 10 computations.

used to be determined via investigating the impact of the ensemble size on the root mean squared error (RMSE) of the original training set.

After T cycles are finished, T RT models are induced. As far as the prediction is concerned, each RT model gives a prediction for the ith unknown sample (i.e., $y_{i,t}$) and a corresponding β_t. A final prediction is obtained by aggregating these T predictions via weighted median,[22,28] as delineated as follows.

Sort the T predictions $y_{i,t}$ (t = 1, ..., T) for the ith sample in ascending order and simultaneously retain the association of β_t with $y_{i,t}$:

$$\begin{array}{cccc} y_{i,n_1} \leq & y_{i,n_2} \leq & \cdots & \leq y_{i,n_T} \\ \uparrow & \uparrow & \cdots & \uparrow \\ \beta_{n_1} & \beta_{n_2} & & \beta_{n_T} \end{array} \qquad (5)$$

where n_t (t = 1, 2, ..., T) is a permutation of 1, 2, ..., T. For example, supposing that the largest estimated bioactivity of the ith

Table 3. Performance Comparison among RT, BRT, and RBRT (leave-one-out cross-validation on the HIV inhibitor data)

	RT	BRT	RBRT
R (correlation coefficient)	0.8890	0.9379	0.9438
RMSE	0.9455	0.7129	0.6789
optimal complex parameter	9.2771×10^{-4}	9.2771×10^{-4}	9.2771×10^{-4}

Table 4. Hopkins Statistic for the Flavonoid Derivatives (n = 104)

number of iterations	population size (%)	$HS_{average}$	HS_{max}	HS_{min}	HS_{range}
5	20	0.8239	0.8939	0.6523	0.2416
5	100	0.8204	0.8602	0.7612	0.0990
10	10	0.7806	0.8967	0.4264	0.4703
20	5	0.8095	0.9479	0.4285	0.5193
50	5	0.8172	0.9568	0.2074	0.7494

compound was the one gained by the RT model built in the third calculation cycle, then y_{i,n_T} should be $y_{i,3}$. If one accumulatively sums $\log(1/\beta_{n_t})$ over t, until one gets the smallest t (supposed as r) satisfying the following inequality

$$\sum_{t=1}^{r} \log(1/\beta_{n_t}) \geq \frac{1}{2} \sum_{t=1}^{T} \log(1/\beta_{n_t}) \qquad (6)$$

then y_{i,n_r} is identified as the ensemble prediction for the ith compound.[28] If all β_t are equal, the median of the T prediction values would be the ensemble prediction.

The difference between RBRT and BRT lies in the fact that, in each cycle, the former trims the large sample errors before the weight updating to avoid the models established latter being excessively dependent on the samples with very large errors.

3. DATA SETS

3.1. HIV-1 Inhibitor Data. To evaluate the performance of the newly proposed RBRT algorithm, 222 chemicals as HIV-1 inhibitors,[29−32] coupled with their associated bioactivities, were used as a data set. This large data set is home to a series of structurally heterogeneous compounds, with the parent structures presented in Figure 1. Their detailed structures and corresponding anti-HIV activities are enumerated in Table 1 of Supporting Information. For each compound, over 100 variables were calculated by Material Studio 4.0 software system, such as structure, spatial, thermodynamic, topological descriptors and E-state indices, as listed in Table 2 of Supporting Information. We eliminated the variables whose values were strictly correlated with that of another descriptor (i.e., with the correlation coefficient between them larger than or equaling to 0.95) to avoid useless redundancy. This data set was randomly divided into a training set (about three-fourths of the whole data) and a test set (about one-fourth of the whole data). The model constructed on the training set was used to predict the bioactivities of compounds in the test set. Ten different such random splits of the data set were performed, and the average over these 10 training/test performances was taken as the overall measure of performance.

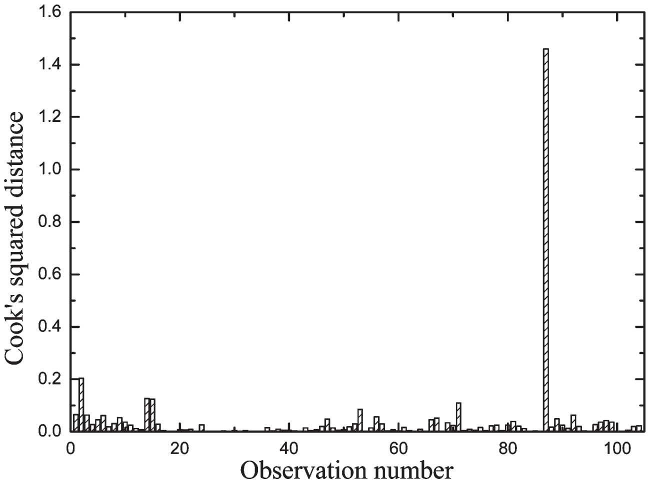


Figure 7. CD$^2_{(i)}$ values obtained by Cook's squared distance method for flavonoid derivatives.

3.2. Flavonoid Derivatives as p56lck Tyrosine Kinase Inhibitor Data. A set of 104 flavonoid derivatives with their corresponding inhibitory activities to p56lck tyrosine kinase[33] was used as another data set to further check the validity of the newly proposed algorithm. Figure 2 depicts the parent structure of flavonoid derivatives. The detailed structures and the associated bioactivities are listed in Table 3 of Supporting Information. Similar to HIV-1 inhibitor data, more than 100 descriptors were calculated as the original variables. In addition, seven variables used by Thakur[33] were considered in the current study, including hydration energy (He) and 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, the value is 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, the value is 0. I_{OH} equals to 1, when hydroxyl is present on the phenyl ring; otherwise I_{OH} equals to 0. Eliminating the variables whose values were strictly correlated with that of another descriptor was also carried out. We also performed 10 random partitions of the data set into the training and test sets, setting the ratio of the training set to the test one as about 3/1.

The algorithms used in the current study were written in Matlab environment and run on a personal computer with the processor being an Intel Pentium Dual-Core CPU E6300 @ 2.80 GHz, 2.79 GHz and the RAM as 2GB.

4. RESULTS AND DISCUSSION

4.1. HIV-1 Inhibitor Data. In the present study, to more effectively execute the QSAR modeling, data quality analysis was performed before QSAR modeling, including nonlinearity, clustering tendency, and outlier detections. It will be of great advantage for further selecting the suitable multivariate modeling methods. Runs test method was applied to test whether nonlinearity exists or not and to quantify the extent if nonlinearity is present. For this data set, the presence of serious nonlinearity was proven by the runs test that yielded a statistical value of −10.2920, whose absolute value is much larger than the critical value of 1.96.[34] To diagnose the clustering tendency, Hopkins statistic[35] was invoked. The results by Hopkins statistic are listed

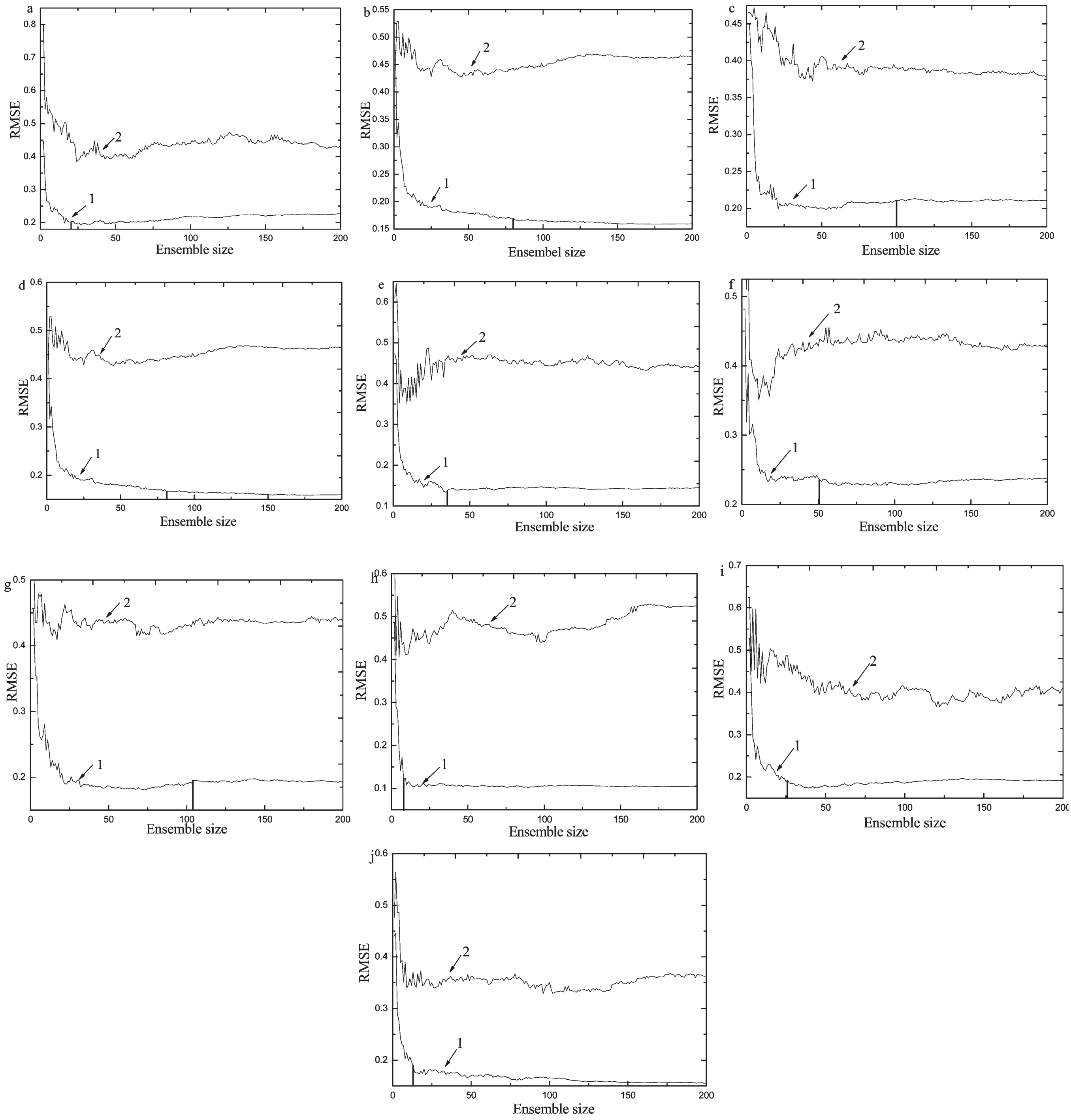


Figure 8. Convergence curves of BRT for flavonoid derivatives for 10 random divisions of the whole data set. Curve 1 refers to the influence of ensemble size on RMSE of the original training set for each division. Curve 2 is the curve of RMSE for the test set versus ensemble size for each division. The vertical real lines indicate the optimal ensemble sizes selected for BRT in 10 computations.

in Table 1, including those from different combinations of the population size and the iteration number. From Table 1, one can obtain that the $HS_{average}$ of each execution was larger than 0.75,[34] indicating the occurrence of clustering tendency in this data set. As to the outlier detection, Cook's squared distance $(CD^2_{(i)})$ method[36] was employed. The large value of $CD^2_{(i)}$ indicates that the *i*th compound has a considerable influence on the regression estimator. Generally, $CD^2_{(i)} = 1$ is considered to be large.[35] The results of the outlier investigation are depicted in Figure 3, indicating the absence of an outlier. All of these may be the very reasons that regression tree and its modified versions were considered for modeling this data set.

As a comparison, RT was first used to model this data set. In RT, the minimal node size was specified as five compounds, that is, the nodes covering less than five compounds cannot be split further and are specified as the leaf nodes. The final appropriately fit RT was identified in terms of the optimal complex parameter gained by cross-validation method. For 10 random splits of the

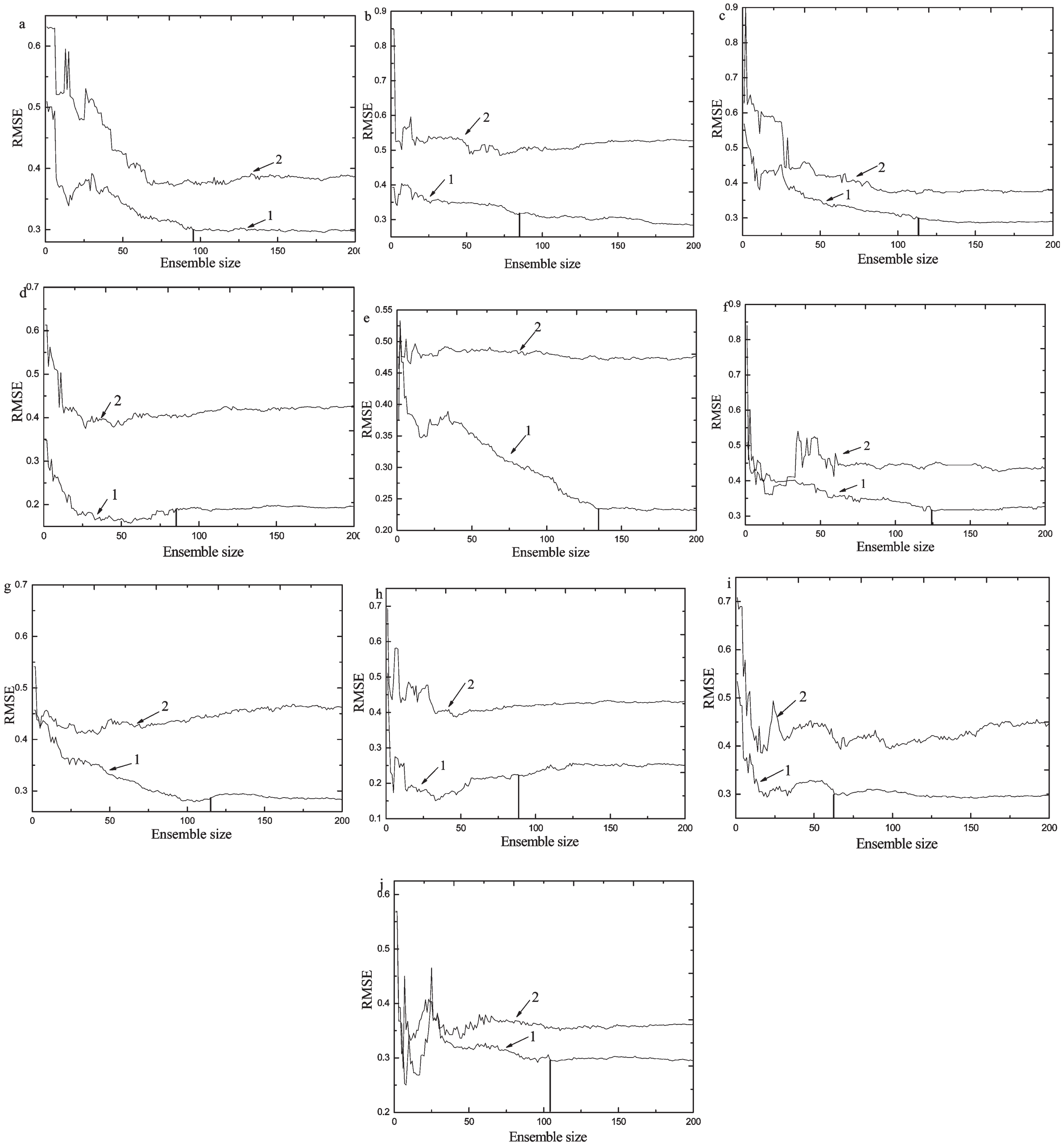


Figure 9. Convergence curves of RBRT for flavonoid derivatives for 10 random divisions of the whole data set. Curve 1 refers to the influence of ensemble size on RMSE of the original training set for each division. Curve 2 is the curve of RMSE for the test set versus ensemble size for each division. The vertical real lines indicate the optimal ensemble sizes selected for RBRT in 10 computations.

HIV-1 inhibitor data, the optimal complex parameters are as follows: 1.2×10^{-3}, 2.8×10^{-4}, 1.3×10^{-3}, 7.2×10^{-4}, 9.6×10^{-4}, 1.1×10^{-3}, 1.3×10^{-3}, 9.0×10^{-4}, 6.1×10^{-4}, and 1.3×10^{-3}. Table 2 lists the statistical results by RT for this data set, from which one can obtain that RT yielded the mean RMSEs of 0.7944 and 0.8668, respectively, for the training and test sets in 10 random splits of the whole data set into the training and test sets. The modeling error is quite high for RT. Moreover, as shown in Table 2, the ranges of the correlation coefficients (R) and RMSEs among 10 test sets are respectively 0.0901 and 0.2690, indicating the high instability of RT. It seems that RT is not accurate enough for modeling the HIV-1 inhibitor data. These may be due to the fact that the greedy splitting heuristics are per se suboptimal and overfitting, and the tree-growing and tree-pruning are two separate steps.

Table 5. Mean Performance over 10 Random Training/Test Splits of the Flavonoid Derivatives Using RT, BRT, and RBRT

		correlation coefficient (R)		root mean square error (RMSE)	
method	data set	MR	RR	MRMSE	RRMSE
RT	training set	0.8549	0.2302	0.3513	0.2856
	test set	0.7699	0.1971	0.4757	0.2908
BRT	training set	0.9816	0.0248	0.1738	0.1530
	test set	0.7973	0.1749	0.4465	0.1750
RBRT	training set	0.9431	0.0958	0.2838	0.1581
	test set	0.8422	0.1506	0.4183	0.1480[a]

[a] MRMSE represents the mean of RMSEs for 10 random training/test divisions. RRMSE refers to the range of RMSEs for 10 random training/test divisions. MR and RR indicate the mean and range of correlation coefficients for 10 random training/test divisions, respectively.

To compare with RBRT, BRT was also carried out to model the HIV-1 inhibitors. In BRT, three parameters were to be identified, i.e., the minimal node size, the optimal complex parameter for the tree of right size in each cycle, and the ensemble size. The minimal node size for BRT is also specified as five compounds. For each division of the whole data set, the optimal complex parameter value for RT was used to identify the appropriately fit tree based on the boosting set in each cycle for BRT. For each training/test division, since all the boosting sets were constructed by drawing samples from the original training set according to the sample weights, the optimal complex parameter optimized on the original training set can be a good approximation to that for RT in every cycle. Actually, it is computationally intolerable to optimize the complex parameters of trees for all cycles. The curves of the ensemble size versus RMSE for HIV-1 inhibitor data in 10 computations are delineated in Figure 4. The RMSE of the original training set became relatively stable when the ensemble size was not smaller than the optimal value. Here, the optimal ensemble sizes for 10 computations were ascertained as 80, 37, 80, 80, 65, 60, 30, 43, 35, and 40, respectively. When modeling the HIV-1 inhibitor data by BRT, these 10 ensemble sizes keep one-to-one correspondence with the above-mentioned 10 optimal complex parameters. BRT offered a mean R of 0.9783 and a mean RMSE of 0.4254 for the training set and a mean R of 0.9353 and a mean RMSE of 0.7352 for the test set, as shown in Table 2. The ranges of the RMSEs and Rs obtained by BRT among 10 computations are also listed in Table 2. A comparison of RT with BRT indicates that better results are obtained by BRT, which should benefit from the introduction of boosting, which has the promising ability to improve the stability of RT and reduce simultaneously bias as well as variance via combining multiple models.

To further improve the QSAR model, RBRT was employed to model the bioactivities of a series of HIV-1 inhibitors. For each training/test division, RBRT shared the minimal node size and optimal complex parameter with BRT. The curves of ensemble size versus RMSE in 10 computations are shown in Figure 5, from which, the optimal ensemble sizes for 10 random divisions of the HIV-1 inhibitor data can be determined as 65, 68, 61, 115, 42, 120, 120, 71, 65, and 70, respectively. The correlations between the observed and calculated bioactivities by RBRT for HIV-1 inhibitors in 10 computations are presented in Figure 6, indicating well correlation between the observed and calculated bioactivities. The statistical results by RBRT, together with those by BRT and RT, are also documented in Table 2. The mean Rs of 0.9754 and 0.9428 were obtained by RBRT, respectively, for the training and test sets. The mean R for the test sets is comparable to that for the training sets by RBRT, indicating no sign of overfitting in RBRT. By using RBRT, the mean RMSE for the test sets was reduced from 0.8668 by RT and 0.7352 by BRT to 0.6914, demonstrating the superior performance of RBRT compared to those of RT and BRT. Actually, RBRT significantly and uniformly outperformed RT and BRT for 10 random training/test splits (results not shown). In addition, in 10 computations, via utilizing RBRT, the ranges of Rs and RMSEs for 10 test sets dropped respectively from 0.0901 and 0.2690 by RT to 0.0492 and 0.1864, demonstrating the promising potential of RBRT in improving the stability of RT. To further show the performance of RT, BRT, and RBRT, we carried out the leave-one-out cross-validation (LOOCV) on the whole data set for these three modeling methods. The results of the LOOCV obtained by RBRT, BRT, and RT are summarized in Table 3. It is very clear from Table 3 that RBRT compared favorably with RT and BRT, and the good results are not due to fortuitous choices of the training and test sets for RBRT. In addition, the time required to perform the proposed algorithm is only within 1 min.

The convergence processes for BRT and RBRT in 10 computations can be examined in Figures 4 and 5 by plotting RMSE versus ensemble size. The two algorithms were both ceased after 200 cycles. Curve 1 is the convergence curve drawn with RMSE of the original training set versus ensemble size. Curve 2 represents the RMSE curve for the test set. From Figures 4 and 5, one can obtain that both BRT and RBRT converge quickly. By visual inspection of Figures 4 and 5, one can discern that the variation trends of curves 2 and 1 appear to be parallel to each other for both BRT and RBRT for each execution, indicating no sign of overfitting of the training data, as often occurred in RT configuring. Such a situation was also indicated in Table 2, that is, the mean R of training sets in 10 computations is comparable to that of the test sets for both BRT and RBRT.

4.2. Flavonoid Derivatives as p56lck Tyrosine Kinase Inhibitor Data.

For further checking the performance of the proposed RBRT, a series of flavonoid derivatives with their corresponding inhibitory activities to p56lck tyrosine kinase were used as another data set. Before QSAR modeling, data quality analysis was also implemented. The presence of nonlinearity was proven by runs test yielding a statistical value of −3.4262. The results of clustering tendency diagnosis are listed

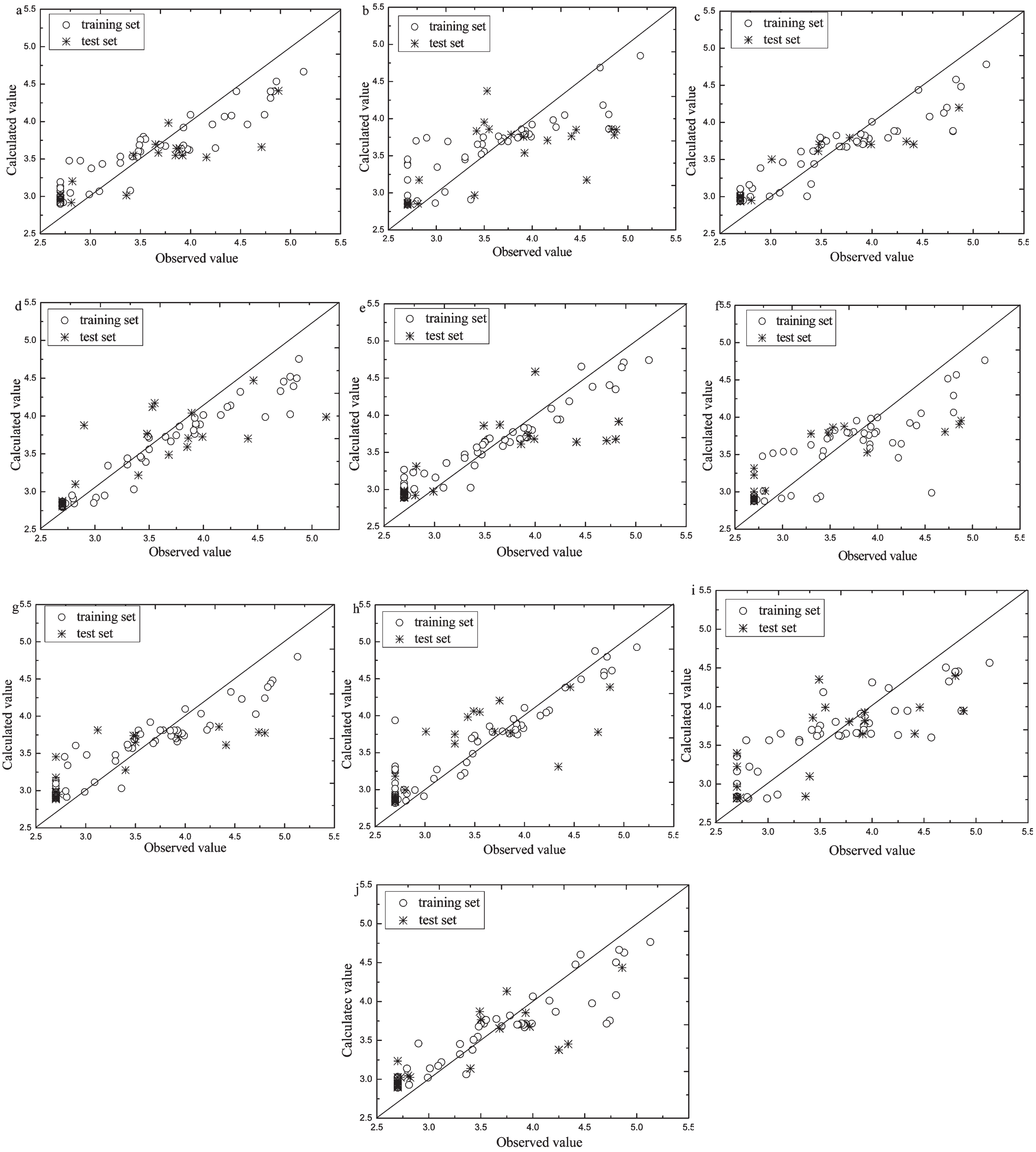


Figure 10. Observed versus calculated values of the inhibitory activities by RBRT for flavonoid derivatives in 10 computations.

in Table 4, indicating the presence of obvious clustering tendency in this data set. Figure 7 depicts the results of the outlier investigation, indicating the presence of one outlier.

For investigating the QSAR modeling of flavonoid derivatives, the newly developed RBRT, compared with RT and BRT, was applied. For this data set, the involved minimal node size is also specified as five compounds. The optimal complex parameters for the three methods in 10 computations are 6.5 × 10^{-3}, 4.1 × 10^{-3}, 3.7 × 10^{-3}, 1.4 × 10^{-3}, 1.9 × 10^{-3}, 5.8 × 10^{-3}, 4.2 × 10^{-3}, 1.4 × 10^{-3}, 5.7 × 10^{-3}, and 3.0 × 10^{-3}, respectively. For this data set, the ensemble sizes for BRT are ascertained as 20, 80, 100, 81, 35, 51, 104, 8, 26, and 13, respectively, for 10 training/test partitions, indicated in Figure 8. As shown in Figure 9, the ensemble sizes for RBRT in 10 computations can be ascertained

Table 6. Performance Comparison among RT, BRT, and RBRT (LOO cross-validation on the flavonoid derivatives)

	RT	BRT	RBRT
R (correlation coefficient)	0.6193	0.7064	0.7254
RMSE	0.6052	0.5085	0.4904
The optimal complex parameter	0.0019	0.0019	0.0019

as 95, 85, 113, 85, 134, 124, 115, 89, 62, and 104, respectively. When applying BRT and RBRT for modeling this data set, these 10 ensemble sizes keep the one-to-one correspondence with the above-mentioned 10 optimal complex parameters. To compare the QSAR modeling performance of the above-mentioned three procedures, the means of *R*s and RMSEs as well as the ranges of *R*s and RMSEs among 10 computations were also used as the criteria. The values of these statistics for the three methods were summarized in Table 5. As shown in Table 5, the correlation was rather poor and the modeling error was quite high for RT. The mean *R*s of 0.9816 and 0.7973 were offered by BRT, respectively, for the training and test sets. The mean *R* by BRT for the test sets is much lower than that for the training sets in 10 computations. This seems to be a sign that overfitting occurred in BRT for this data set. As shown in Table 5, RBRT offered the mean *R*s of 0.9431 and 0.8422, respectively, for the training and test sets in 10 computations. By using RBRT, although the mean of RMSEs for the training sets in 10 computations was increased from 0.1738 by BRT to 0.2838, the one for the test sets was decreased from 0.4465 by BRT to 0.4183. These statistical results demonstrated that a sign of overfitting was mitigated by using RBRT. From Table 5, one can obtain that RBRT and BRT show higher stability than RT. Figure 10 indicates the correlation between the observed and calculated bioactivities of flavonoid derivatives by RBRT for 10 random training/test splits. In addition, the LOOCV was also performed on the whole data set to test the stability of the three methods, with the results shown Table 6. It is very clear from Table 6 that the best results are obtained by RBRT as compared to BRT and RT, confirming that the performance of RBRT is very stable and that fortuitous choices of the training and test sets for RBRT are not responsible for the good results.

Figures 8 and 9 indicate the convergence processes for BRT and RBRT, respectively. From Figures 8 and 9, one can see that the RMSE drops quickly in BRT and RBRT, confirming that BRT and RBRT converge quickly. By visual inspection of Figure 8, one can discern that, in eight of 10 computations (i.e., Figures 8b,d−j), with the decrease of the RMSE for the training set (i.e., curve 1), the RMSE for the test set (i.e., curve 2) turned to increase fluctuantly with a low speed, and no further improvement was observed. This means that overfitting occurs in BRT under this condition. For RBRT, curves 2 and 1 appear to be parallel to each other in eight out of 10 computations, as shown in Figure 9. Even now, a conclusion can be still drawn that it seems that RBRT shows considerable antioverfitting capability under this situation. This may be the very reason that the mean *R* by BRT for the test sets is much lower than that for the training sets in 10 computations and the difference between the mean *R*s by RBRT for the training and tests was lessened, as shown in Table 5. Such a phenomenon may be due to the fact that the training set for a small data set may not be representative enough to contain all of the noises and idiosyncrasies occurred in the test set, resulting in overfitting of the training data. The large amounts of variability among the small data set (via comparing Tables 5 and 6) and the presence of one outlier may be the other factor for the occurrence of overfitting for BRT for this data set. In brief, the results of this data set revealed again that the introduction of robust boosting can substantially improve the performance of RT.

From the above-mentioned description, RBRT provides superior performance to BRT and RT. It shows that the introduction of robust boosting is very beneficial for overcoming the issue of poor generalization ability and improving the stability of RT. The method may benefit from the fact that robust boosting shares the promising ensemble learning with boosting for improving the performance via combining multiple learners. Besides this, RBRT takes an error-trimming technique before weight updating for RT modeling in the next cycle to avoid the models established latter being excessively dependent on the samples with large errors so as to be robust.

5. CONCLUSION

The ensemble technique can be the most efficient procedure for improving the performance of RT. In the present study, robust boosting was invoked for simultaneously enhancing the generalization ability and stability of RT, forming a new method named robust boosting regression tree (RBRT). The RBRT attempts to establish a sequence of robust RT models by introducing an error-trimming technique before the weight renovation for the next cycle and then integrate the outputs of all these resultant RT models to obtain a final prediction. The performance of the designed algorithm was assessed using two QSAR data sets. Experimental results revealed that the combination of robust boosting with RT offered the possibility of improving the generalization ability and stability of RT, and RBRT compared favorably with BRT in QSAR studies.

■ ASSOCIATED CONTENT

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

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +86-15872406428. Fax: 86-27 67867141. E-mail: hgzyp2005@yahoo.com.cn.

■ ACKNOWLEDGMENT

The authors are grateful to the self-determined research funds of CCNU from the colleges' basic research and operation of MOE (No. CCNU09A01012), the Fundamental Research Funds for the Central Universities (No. 09ZZKY077), and the open funds of State Key Laboratory of Chemo/Biosensing and Chemometrics of Hunan University (No. 200910) for supporting the research work.

■ REFERENCES

(1) Breiman, L.; Friedman, J. H.; Olshen, R. J.; Stone, C. J. In *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) Gleeson, M. P.; Waters, N. J.; Paine, S. W.; Davis, A. M. In Silico Human and Rat Vss Quantitative Structure—Activity Relationship Models. *J. Med. Chem.* **2006**, *49*, 1953–1963.

(4) Zhou, Y. P.; Tang., L. J.; Jiao., J.; Song, D. D.; Jiang, J. H.; Yu, R. Q. Modified Particle Swarm Optimization Algorithm for Adaptively Configuring Globally Optimal Classification and Regression Trees. *J. Chem. Inf. Model.* **2009**, *49*, 1144–1153.

(5) Tan, S. M.; Jiao, J.; Zhu, X. L.; Zhou, Y. P.; Song, D. D.; Gong, H.; Yu, R. Q. QSAR Studies of a Diverse Series of Antimicrobial Agents against *Candida albicans* by Classification and Regression Trees. *Chemom. Intell. Lab. Syst.* **2010**, *103*, 184–190.

(6) Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R. P.; Song, Q. H. Boosting: An Ensemble Learning Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Model.* **2005**, *45*, 786–799.

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

(8) 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.

(9) 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.

(10) Dietterich, T. G. Ensemble Learning. In *The Handbook of Brain Theory and Neural Networks*, 2nd ed.; Arbib, M. A., Ed.; The MIT Press: Cambridge, 2002.

(11) Hawkins, D. M.; Musser, B. J. One Tree or a Forest? Alternative Dendrographic Models. *Comput. Sci. Stat.* **1999**, *30*, 534–542.

(12) Tong, W.; Hong, H.; Fang, H.; Xie, Q.; Perkins, R. Decision Forest: Combining the Predictions of Multiple Independent Decision Tree Models. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 525–531.

(13) van Rhee, A. M. Use of Recursion Forests in the Sequential Screening Process: Consensus Selection by Multiple Recursion Trees. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 941–948.

(14) Meyer, D.; Leisch, F.; Hornik, K. The Support Vector Machine under Test. *Neurocomputing.* **2003**, *55*, 169–186.

(15) He, P.; Xu, C. J.; Liang, Y. Z.; Fang, K. T. Improving the Classification Accuracy in Chemistry via Boosting Technique. *Chemom. Intell. Lab. Syst.* **2004**, *70*, 39–46.

(16) Culp, M.; Johnson, K.; Michailidis, G. The Ensemble Bridge Algorithm: A New Modeling Tool for Drug Discovery Problems. *J. Chem. Inf. Model.* **2010**, *50*, 309–316.

(17) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.

(18) Deconinck, E.; Zhang, M. H.; Coomans, D.; Heyden, Y. V. Evaluation of Boosted Regression Trees (BRTs) and Two-Step BRT Procedures to Model and Predict Blood—Brain Barrier Passage. *J. Chemom.* **2007**, *21*, 280–291.

(19) Schapire, R. E. The Strength of Weak Learnability. *Mach. Learn.* **1990**, *5*, 197–227.

(20) Zhou, Y. P.; Jiang, J. H.; Lin, W. Q.; 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. J. Pharm. Sci.* **2006**, *28*, 344–353.

(21) Zhang, M. H.; Xu, Q. S.; Massart, D. L. Boosting Partial Least Square. *Anal. Chem.* **2005**, *77*, 1423–1431.

(22) Zhou, Y. P.; Cai, C. B.; Huan, S.; Jiang, J. H.; Wu, H. L.; Shen, G. L.; Yu, R. Q. QSAR Study of Angiotensin II Antagonists Using Robust Boosting Partial Least Squares Regression. *Anal. Chim. Acta* **2007**, *593*, 68–74.

(23) Dietterich, T. G. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Mach. Learn.* **2000**, *40*, 139–157.

(24) Grove, A. J.; Schuurmans, D. Boosting in the Limit: Maximizing the Margin of Learned Ensembles. *Proceedings of the 15th National Conference on Artificial Intelligence*; Madison, WI, July 1998; pp 692—299.

(25) Friedman, J.; Hastie, T.; Tibshirani, R. Additive Logistic Regression: A Statistical View of Boosting. *Ann. Statist.* **2000**, *28*, 337–407.

(26) Freund, Y.Schapire, R.Experiments with a New Boosting Algorithm. *Proceeding of the 13th International Conference on Machine Learning*; 1996; pp 148156

(27) *The Mathematics of Generalization*; Wolpert, D. H., Ed.; Addison-Wesley: Reading, 1995.

(28) Drucker, H. Improving Regressors Using Boosting Techniques. *Proceedings of the 14th National Conference on Machine Learning*; Nashville, TN, July 8—12, 1997; pp 107—115.

(29) Bhhatarai, B.; Garg, R. From SAR to Comparative QSAR: Role of Hydrophobicity in the Design of 4-Hydroxy-5,6-dihydropyran-2-ones HIV-1 Protease Inhibitors. *Bioorg. Med. Chem.* **2005**, *13*, 4078–4084.

(30) Leonard, J. T.; Roy, K. QSAR by LFER Model of HIV Protease Inhibitor Mannitol Derivatives Using FA-MLR, PCRA, and RT Techniques. *Bioorg. Med. Chem.* **2006**, *14*, 1039–1046.

(31) Roy, P. P.; Leonard, J. T.; Roy, K. Exploring the Impact of Size of Training Sets for the Development of Predictive QSAR Models. *Chemom. Intell. Lab. Syst.* **2008**, *90*, 31–42.

(32) Freitas, M. P. MIA-QSAR Modelling of Anti-HIV-1 Activities of Some 2-Amino-6-Arylsulfonylbenzonitriles and Their Thio and Sulfinyl Congeners. *Org. Biomol. Chem.* **2006**, *4*, 1154–1159.

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

(34) Centner, V.; de Noord, O. E.; Massart, D. L. Detection of Nonlinearity in Multivariate Calibration. *Anal. Chim. Acta* **1998**, *376*, 153–168.

(35) Centnera, V.; Massart, D. L.; de Noord, O. X. Detection of Inhomogeneities in Sets of NIR Spectra. *Anal. Chim. Acta* **1996**, *330*, 1–17.

(36) Cook, R. D. Detection of Influential Observations in Linear Regression. *Technometrics* **1977**, *19*, 15–18.