JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

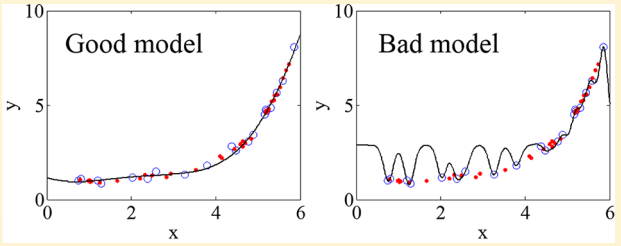
pubs.acs.org/jcim

Criterion for Evaluating the Predictive Ability of Nonlinear Regression Models without Cross-Validation

Hiromasa Kaneko and Kimito Funatsu*

Department of Chemical System Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

ABSTRACT: We propose predictive performance criteria for nonlinear regression models without cross-validation. The proposed criteria are the determination coefficient and the root-mean-square error for the midpoints between *k*-nearest-neighbor data points. These criteria can be used to evaluate predictive ability after the regression models are updated, whereas cross-validation cannot be performed in such a situation. The proposed method is effective and helpful in handling big data when cross-validation cannot be applied. By analyzing data from numerical simulations and quantitative structural relationships, we confirm that the proposed criteria enable the predictive ability of the nonlinear regression models to be appropriately quantified.



INTRODUCTION

Chemoinformatics[1] aims to solve chemistry problems using an informatics method. There is a great deal of previous research in this field, revolving around topics such as quantitative structure−activity relationships (QSARs), quantitative structural−property relationships (QSPRs), reaction design, and drug design.

Multivariate techniques, such as multiple linear regression, principal component regression,[2] and partial least-squares (PLS),[3] are powerful tools for handling several problems in chemoinformatics. In PLS modeling, it is possible to construct an accurate model by introducing the latent variables whose covariance with an objective variable $\mathbf{y}$ is maximized. However, even PLS cannot deal with a nonlinear relationship between $\mathbf{y}$ and its explanatory variables $\mathbf{X}$.

Many kinds of nonlinear regression methods have been developed, such as back-propagation neural network,[4] counter-propagation neural network,[5] kernel PLS,[6] Gaussian process,[7] support vector regression (SVR),[8] and relevance vector machine.[9] The aim of these techniques is to construct predictive models even in the existence of the nonlinear relationship between $\mathbf{y}$ and $\mathbf{X}$. On the one hand, such nonlinear regression methods can construct complex models, but on the other hand, we often face the possibility of overfitting, that is, the constructed models try to overfit the training data. The determination coefficient r^2 and root-mean-square error (RMSE) of the $\mathbf{y}$ values calculated using nonlinear regression models can be almost 1 and 0, respectively. However, the constructed models exhibit poor predictive performance for new data. Therefore, these regression models must be validated to quantify their predictive ability and allow the appropriate model and hyperparameters to be selected.

There are two main types of validation: external validation and cross-validation (CV).[10−12] Using adequate data for external validation, the predictive performance can be evaluated, but the model may overfit the validation data during variable selection. Additionally, it is difficult to divide data for training and validation, and the amount of data is not always large enough to prepare separate validation data.

Leave-one-out CV (LOOCV) is the simplest CV method. A single datum from the original data is used to validate the model constructed by the remaining data. This is repeated such that each datum is used once for the validation. However, there is no correlation between the predictive performance estimated by LOOCV and predictive ability for the test data.[13] In addition, as an objective function in variable selection, LOOCV has a strong tendency for overfitting.[14] The problems of LOOCV can be overcome by employing leave-multiple-out CV and *N*-fold CV, in which multiple data are used for validation. For example, in *N*-fold CV, the original data are first randomly divided into *N* groups containing a similar number of data. One group is then used to validate the model constructed by the data of the other $N - 1$ groups. This procedure is repeated *N* times, so that each of the *N* groups is used once as the validation data. The predictive ability of the regression models can be evaluated adequately, and the appropriate model can be selected using leave-multiple-out or *N*-fold cross-validated r^2.

As in other criteria based on CV, r_{cv}^2 is combined with another index,[15] Monte Carlo CV,[16] or time-split CV[17] and used to improve the estimation accuracy of the predictive performance. In addition, *y*-randomization[18] or *y*-scrambling[19] is used to check the change correlation and verify the statistical significance of the regression models. During variable selection, cross-model validation will reduce the tendency for overfitting compared with normal CV.[20]

However, CV methods require multiple model construction. For instance, the model construction and the prediction are repeated *N* times in *N*-fold CV. When we handle big data,[21,22]

Received: June 28, 2013
Published: August 23, 2013

   dx.doi.org/10.1021/ci4003766 | *J. Chem. Inf. Model.* 2013, 53, 2341−2348

that is, very large and complex data sets, the computational time of CV renders it impractical.

In addition, when new data are obtained and the regression models that were already constructed with big data require reconstruction, it is inefficient and impractical to construct new regression models with both the original big data and the new data. Therefore, the regression models that have already been built are efficiently reconstructed with new data using model updating methods[23] or online learning methods.[24] In process monitoring with soft sensors,[25] regression models are also reconstructed with new data to reduce the degradation of their predictive performance,[26] and nonlinear regression models such as SVR are updated online.[27] It is impossible to cross-validate regression models after they have been updated. Although the r^2 and RMSE can be calculated, these criteria tend to overestimate the predictive ability, as mentioned above, and cannot be used as criteria for validating regression models. There are other criteria such as Mallows' Cp,[28] Akaike's information criterion,[29] and Bayesian information criterion,[30] which can consider model complexity. Although these criteria can be used as the objective function for variable selection, their values will be the same with the same number of **X**-variables and training data, and so these criteria cannot be used as indexes for the model or hyperparameter selection.

Therefore, in this study, we propose criteria for evaluating the predictive ability of nonlinear regression models without CV. The criteria are based on the midpoints between training data. For each training datum, the *k*-nearest-neighbor data points are selected and the midpoints are calculated. These midpoints are assumed to be real data, and the proposed criteria, $r_{\mathrm{midknn}}{}^2$ and $\mathrm{RMSE}_{\mathrm{midknn}}$, are the r^2 and RMSE for these midpoint data.

To verify the effectiveness of the proposed method, we analyze the numerical simulation data where the relationship between **X** and **y** is nonlinear. The performance of the proposed criteria is compared with that of r^2 (RMSE) and 5-fold cross-validated r^2 and $r_{\mathrm{cv}}{}^2$ ($\mathrm{RMSE}_{\mathrm{cv}}$). The proposed method is then applied to QSAR and QSPR data.

METHOD

Midpoint. The *i*th data point $\mathbf{z}^{(i)}$ is represented as follows:

$$\mathbf{z}^{(i)} = [\mathbf{x}^{(i)} y^{(i)}] = [x_1{}^{(i)} x_2{}^{(i)} \cdots x_n{}^{(i)} y^{(i)}] \quad (1)$$

where $\mathbf{x}^{(i)}$ is the *i*th data point of **X**, $x_p{}^{(i)}$ is the value of the *p*th **X**-variable of the *i*th data point, $y^{(i)}$ is the *y* value of the *i*th data point, and *n* is the number of **X**-variables. The midpoint of $\mathbf{z}^{(i)}$ and $\mathbf{z}^{(j)}$, $\mathbf{z}^{\mathrm{mid}(i,j)}$, is as follows:

$$\mathbf{z}^{\mathrm{mid}(i,j)} = [\mathbf{x}^{\mathrm{mid}(i,j)} y^{\mathrm{mid}(i,j)}] = \left[\frac{x_1{}^{(i)} + x_1{}^{(j)}}{2} \frac{x_2{}^{(i)} + x_2{}^{(j)}}{2} \cdots \frac{x_n{}^{(i)} + x_n{}^{(j)}}{2} \frac{y^{(i)} + y^{(j)}}{2}\right] \quad (2)$$

where $\mathbf{x}^{\mathrm{mid}(i,j)}$ is the midpoint of $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ and $y^{\mathrm{mid}(i,j)}$ is the midpoint of $y^{(i)}$ and $y^{(j)}$. When the number of data points is *m*, we have $m - 1$ midpoints, based on the *i*th data point of $\mathbf{z}^{(i)}$. After the elimination of overlapping midpoints, the total number of midpoints is $m(m-1)/2$.

Midpoint between *k*-Nearest-Neighbor Data Points. In this study, we consider only the midpoints between the *k*-nearest-neighbor data points (MIDKNNs). In the calculation of the MIDKNNs based on $\mathbf{z}^{(i)}$, the *k* data points that are most similar to $\mathbf{z}^{(i)}$ are selected. For example, the Euclidean distance may be used as an index of the similarity. The Euclidean distance between $\mathbf{z}^{(i)}$ and $\mathbf{z}^{(j)}$, $\mathrm{ED}^{(i,j)}$, is calculated as follows:

$$\mathrm{ED}^{(i,j)} = \sqrt{\|\mathbf{z}^{(i)} - \mathbf{z}^{(j)}\|^2} = \sqrt{(x_1^{(i)} - x_1^{(j)})^2 + (x_2^{(i)} - x_2^{(j)})^2 + \cdots + (x_n^{(i)} - x_n^{(j)})^2 + (y^{(i)} - y^{(j)})^2} \quad (3)$$

The *k* midpoints between $\mathbf{z}^{(i)}$ and the *k* data points that have the lowest $\mathrm{ED}^{(i,j)}$ values are obtained. When the number of data points is *m*, the number of MIDKNNs is $m \times k$ and the overlapping midpoints are deleted from the MIDKNNs. The Mahalanobis distance,[31] correlation, kernel functions, and so on can be used as indexes of the similarity.

r^2 and RMSE of Midpoints between *k*-Nearest-Neighbor Data Points. We propose the r^2 and *RMSE* values of the MIDKNNs as new indexes for the predictive performance of regression models. We refer to these as $r_{\mathrm{midknn}}{}^2$ and $\mathrm{RMSE}_{\mathrm{midknn}}$, respectively. Figure 1 illustrates the concept of model validation using the MIDKNNs. Figure 1a and b show the training data, in which a nonlinear relationship exists between **x** and **y**, and each of the midpoints of the training data. Some midpoints seem to be far from the distribution of the training data because of the nonlinearity between **x** and **y**. These midpoints cannot be used to validate the regression model. For example, the point indicated by the arrow is the midpoint for the data points at the two ends of the curve in Figure 1b. The regression model fitting to this midpoint would

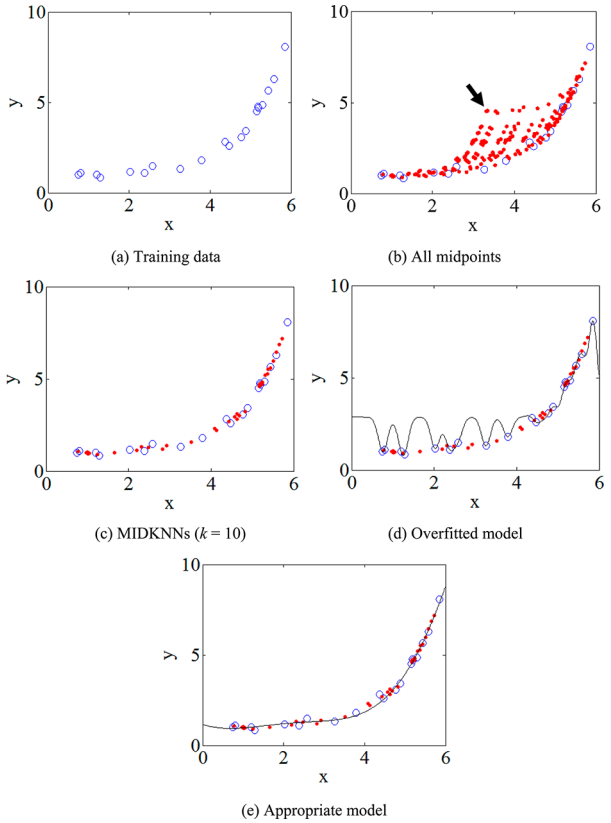


Figure 1. Concept of model validation using the MIDKNNs. The circles, points and lines represent the training data, midpoints, and regression models, respectively.

dx.doi.org/10.1021/ci4003766 | J. Chem. Inf. Model. 2013, 53, 2341−2348

be useless. Figure 1c shows the MIDKNNs when k equals 3. The MIDKNNs interpolate the training data and have a distribution that is similar to that of the training data. These MIDKNNs are the validation data. The regression model in Figure 1d overfits to the training data, and the prediction errors of the MIDKNNs are large. On the other hand, an appropriate regression model can be constructed, as shown in Figure 1e, leading to the MIDKNNs giving small prediction errors. The MIDKNNs enable the predictive ability of the regression models to be quantified.

r_{midknn}^2 and $\mathrm{RMSE}_{\mathrm{midknn}}$ are defined as follows:

$$r_{\mathrm{midknn}}^2 = 1 - \frac{\sum_{\mathrm{midknn}=1}^{N_{\mathrm{midknn}}} (y^{\mathrm{midknn}} - y_{\mathrm{pred}}^{\mathrm{midknn}})^2}{\sum_{\mathrm{midknn}=1}^{N_{\mathrm{midknn}}} (y^{\mathrm{midknn}} - \hat{y}^{\mathrm{midknn}})^2} \quad (4)$$

$$\mathrm{RMSE}_{\mathrm{midknn}} = \sqrt{\frac{\sum_{\mathrm{midknn}=1}^{N_{\mathrm{midknn}}} (y^{\mathrm{midknn}} - y_{\mathrm{pred}}^{\mathrm{midknn}})^2}{N_{\mathrm{midknn}}}} \quad (5)$$

where N_{midknn} is the number of MIDKNNs, y^{midknn} is the y value of the MIDKNN, and $y_{\mathrm{pred}}^{\mathrm{midknn}}$ is the predicted y value of the MIDKNN. The corrected $\mathrm{RMSE}_{\mathrm{midknn}}$ resulting from the reduction of the variance in the calculation of the midpoints is given as follows:

$$\mathrm{RMSE}_{\mathrm{midknn}} = \sqrt{\frac{2(m+1)\sum_{\mathrm{midknn}=1}^{N_{\mathrm{midknn}}} (y^{\mathrm{midknn}} - y_{\mathrm{pred}}^{\mathrm{midknn}})^2}{m(N_{\mathrm{midknn}} - 1)}} \quad (6)$$

The derivation of eq 6 is shown in Appendix A. The predictive performance can be estimated without CV from eqs 4 and 6.

In Figure 1d, the regression model accurately adapts to the training data, leading to a high r^2 value and low RMSE. However, the shape of the model shows that it obviously overfits to the training data. The prediction errors of the MIDKNN are large, and thus the r_{midknn}^2 value will be low and the $\mathrm{RMSE}_{\mathrm{midknn}}$ will be large. On the other hand, in Figure 1e, the shape of the model shows that it is more appropriate than that in Figure 1d. The prediction errors of the MIDKNN are small, giving a high value of r_{midknn}^2 and a low $\mathrm{RMSE}_{\mathrm{midknn}}$. As described above, the predictive ability of regression models can be evaluated with r_{midknn}^2 and $\mathrm{RMSE}_{\mathrm{midknn}}$.

The number of MIDKNNs increases exponentially with the number of training data. If the number of MIDKNNs is too large, it can be limited arbitrarily.

RESULTS AND DISCUSSION

To verify the effectiveness of the proposed method, we analyzed numerical simulation data and applied this method to QSAR and QSPR data. The relationship between $\mathbf{X}$ and $\mathbf{y}$ is nonlinear for the numerical simulation data. In this study, we used the SVR method,[32] a nonlinear regression approach, to model the relation between $\mathbf{X}$ and $\mathbf{y}$. Details of the SVR method are given in Appendix B. In the SVR model, the parameter C was varied from 2^{-2} to 2^{12} by multiplying by 2, ν was similarly varied from 2^{-8} to 2^{-4} and then from 0.1 to 0.9 in steps of 0.1, and γ was varied from 2^{-10} to 2^{5} by multiplying by 2. The k-values for the MIDKNN range from 1−30 in steps of 1.

The configurations of the computer used in this study are given as follows: OS, Windows 7 Professional (64 bit); CPU, Intel(R) Xeon(R) X5690 3.47 GHz; RAM, 48.0 GB. The version of MATLAB is R2012a.

Modeling of Numerical Simulation Data. The number of $\mathbf{X}$-variables was set as two. First, we prepared pseudorandom numbers $\mathbf{x}_1$ and $\mathbf{x}_2$ in the range $[-2, 2]$. The $\mathbf{y}$ values were generated using the following relationship:

$$\exp(\mathbf{y}) = \{1 + (\mathbf{x}_1 + \mathbf{x}_2 + 1)^2(19 - 14\mathbf{x}_1 + 3\mathbf{x}_1^2 - 14\mathbf{x}_2 + 6\mathbf{x}_1\mathbf{x}_2 + 3\mathbf{x}_2^2)\} \times \{30 + (2\mathbf{x}_1 - 3\mathbf{x}_2)^2 (18 - 32\mathbf{x}_1 + 12\mathbf{x}_1^2 + 48\mathbf{x}_2 - 36\mathbf{x}_1\mathbf{x}_2 + 27\mathbf{x}_2^2)\} \quad (7)$$

Equation 7 is described[33] as a test problem. Figure 2 shows the shape of eq 7. The training data and test data both contain 500 data points.

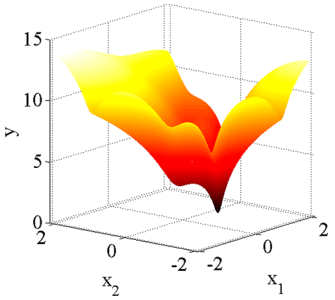


Figure 2. Shape of eq 7

The SVR model was constructed for each combination of C, ν, and γ, and values of r^2, r_{cv}^2, and r_{midknn}^2 were calculated. A 5-fold CV was used for r_{cv}^2. The optimal SVR model for r^2, r_{cv}^2, and r_{midknn}^2 was that which gave the maximum value of r^2, r_{cv}^2, and r_{midknn}^2, respectively. In other words, the values of RMSE, $\mathrm{RMSE}_{\mathrm{cv}}$, and $\mathrm{RMSE}_{\mathrm{midknn}}$ were minimized in the optimal model. Figure 3 shows the r_{pred}^2 and $\mathrm{RMSE}_{\mathrm{p}}$ values for each of the optimal SVR models. The r_{pred}^2 value is an r^2 value calculated with the test data. The higher the r_{pred}^2-value of a model, the greater that model's predictive accuracy. The $\mathrm{RMSE}_{\mathrm{P}}$ value is an RMSE value calculated with the test data. Lower $\mathrm{RMSE}_{\mathrm{P}}$ values indicate higher prediction accuracy. The r_{pred}^2 and $\mathrm{RMSE}_{\mathrm{p}}$ values for r_{midknn}^2 and $\mathrm{RMSE}_{\mathrm{midknn}}$ depend on k, whereas those for r^2, r_{cv}^2, RMSE, and $\mathrm{RMSE}_{\mathrm{cv}}$ are constant for each k value. The r_{pred}^2 values of the models selected using r_{cv}^2 and r_{midknn}^2 were higher than those given by the models selected with r^2. The trend of the $\mathrm{RMSE}_{\mathrm{p}}$ values for RMSE, $\mathrm{RMSE}_{\mathrm{cv}}$, and $\mathrm{RMSE}_{\mathrm{midknn}}$ was the opposite to that of the r_{pred}^2 values. Regardless of the k values, the difference in the r_{pred}^2 values given by r_{cv}^2 and r_{midknn}^2 is small. The r_{pred}^2 values were slightly lower, and the $\mathrm{RMSE}_{\mathrm{p}}$ values were slightly higher, for larger k, which reflects the distribution of the MIDKNNs separating from that of the training data, as in Figure 1b for large k. Nevertheless, we can say that r_{midknn}^2 and $RMSE_{\mathrm{midknn}}$

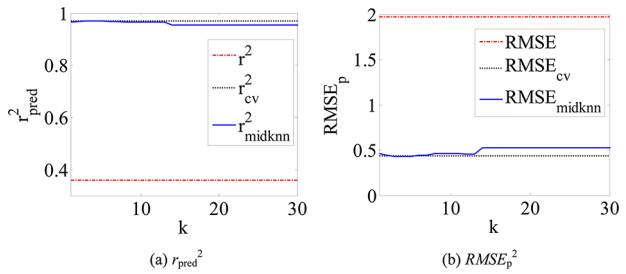


Figure 3. r_{pred}^2 and $\mathrm{RMSE}_{\mathrm{p}}$ values of the optimal SVR model for r^2, r_{cv}^2, and r_{midknn}^2 when the numerical simulation data were used.

dx.doi.org/10.1021/ci4003766 | J. Chem. Inf. Model. 2013, 53, 2341−2348

with a suitable k value enable the selection of an appropriate regression model, as is the case for r_{cv}^2 and $RMSE_{cv}$.

Figure 4 shows the relationships between r_{pred}^2 and r^2, r_{cv}^2, and r_{midknn}^2 (k = 10) for different combinations of C, ν, and γ.

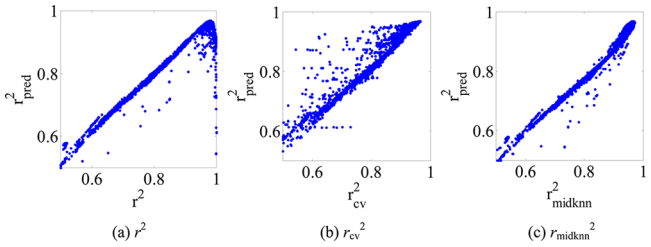


Figure 4. Relationships between r_{pred}^2 and each criterion when k = 10 for the numerical simulation data. The range of both axes was limited from 0.5 to 1.

As shown in Figure 4a, there are many combinations where the r_{pred}^2 values are low in spite of the high r^2 values. This means that the models constructed with these hyperparameters overfit to the training data. The r^2 could not select appropriate combinations of the SVR hyperparameters. From the r_{cv}^2 and r_{midknn}^2 (k = 10) results, there appears to be a strong correlation between r_{cv}^2 and r_{pred}^2, and between r_{midknn}^2 and r_{pred}^2 (Figure 4b and c). This reflects the fact that the predictive ability could be quantified from the validation results. In Figure 4b, the r_{cv}^2 values are low and the r_{pred}^2 values are high for some points. This is because the regression model is constructed with only part of the training data in the repeated CV, meaning that the predictive ability of the model constructed with all the training data is underestimated. Although the r_{midknn}^2 values were much higher than the r_{pred}^2 values for some combinations of the SVR hyperparameters, the absolute r_{midknn}^2 values were not high and did not affect the model selection.

To compare the performance of each criterion quantitatively, the numbers of the models whose r_{pred}^2-values exceed the thresholds in the highest 500 models of each criterion is shown in Table 1. The larger number of the models means that the criterion can evaluate the prediction performance of the SVR model more appropriately. As shown in Table 1, the numbers of the models for r^2 were much less than those for r_{cv}^2 and r_{midknn}^2. Although the number of the models for r_{midknn}^2 was lower than that for r_{cv}^2 when the threshold of r_{pred}^2 was 0.95, the difference was small, and in addition, not only the number of the models for r_{cv}^2 but also that for r_{midknn}^2 achieved 500 when the threshold was 0.90. It was therefore confirmed that the predictive ability of the nonlinear regression models can be evaluated appropriately using the proposed r_{midknn}^2 and $RMSE_{midknn}$ criteria.

Table 1. Numbers of the Models Whose r_{pred}^2-Values Exceed the Thresholds in the Highest 500 Models of Each Criterion for the Numerical Simulation Data

threshold of r_{pred}^2	no. of the models for r^2	no. of the models for r_{cv}^2	no. of the models for r_{midknn}^2
0.95	233	412	390
0.90	371	500	500

QSAR Study Using pIGC$_{50}$. We analyzed data downloaded from the Environmental Toxicity Prediction Challenge 2009 Web site.[34] This is an online challenge that invites researchers to predict the toxicity of molecules against *Tetrahymena pyriformis*, expressed as the logarithm of 50% growth inhibitory concentration (pIGC$_{50}$). The data set includes 1,093 compounds and has been analyzed for the visualization of molecular fingerprints.[35]

Some 2232 molecular descriptors were calculated on this data set by the Dragon 6.0 software.[36] Descriptors with the same values for more than 50% of molecules were removed. When the correlation coefficient of a pair of descriptors was greater than 0.9, one was removed. We thus reduced the number of descriptors, leaving 333 for the construction of predictive models. This data set was randomly divided into a training set of 700 molecules and a test set of 393 molecules. The 700 molecules were used to construct the prediction model, and the 393 molecules were used to test the predictive accuracy of the obtained model.

As in the analysis of the numerical simulation data, the SVR model was constructed for each combination of C, ν, and γ, and the combination that maximized the value of r^2, r_{cv}^2, or r_{midknn}^2 (thus minimizing RMSE, $RMSE_{cv}$, or $RMSE_{midknn}$) was selected. Figure 5 shows the r_{pred}^2 and $RMSE_p$ values of the SVR model selected by each criterion. The r_{pred}^2 and $RMSE_p$ values depend on the k values for r_{midknn}^2 and $RMSE_{midknn}$, respectively. In this case study, the difference between the r_{pred}^2 values of the SVR models selected by r^2, r_{cv}^2, and r_{midknn}^2 was small. For the $RMSE_p$ values, the model selected with r_{midknn}^2 was the same as the model selected with r_{cv}^2. When k = 1, the value of r_{midknn}^2 was similar to that of r^2. This is because the MIDKNNs are close to the original data for a small value of k, and r_{midknn}^2 tends to overestimate the predictive ability of the models. By increasing the value of k, the r_{pred}^2 values became high and stable. Even when real data were used, r_{midknn}^2 and $RMSE_{midknn}$ could select a model whose predictive ability was almost the same as that selected using CV.

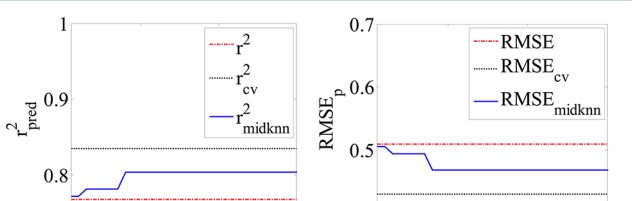


Figure 5. r_{pred}^2 and $RMSE_p$ values of the optimal SVR model for r^2, r_{cv}^2, and r_{midknn}^2 using the QSAR data.

The relationships between r_{pred}^2 and r^2, r_{cv}^2, and r_{midknn}^2 (k = 10) for different combinations of C, ν, and γ are shown in Figure 6. There are some models with both high r^2 values and

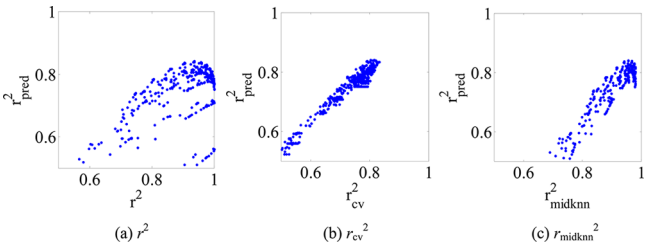


Figure 6. Relationships between r_{pred}^2 and each criterion when k = 10 with the QSAR data. The range of both axes was limited from 0.5 to 1.

dx.doi.org/10.1021/ci4003766 | J. Chem. Inf. Model. 2013, 53, 2341−2348

low r_{pred}^2 values in Figure 6a, which indicates that the success of the model selection with r^2 and RMSE (Figure 5) was coincidental, and r^2 could not in fact quantify the predictive ability of the models. From the r_{cv}^2 and r_{midknn}^2 results (Figure 6b and c), there appears to be some collinearity between r_{cv}^2 and r_{pred}^2 and between r_{midknn}^2 and r_{pred}^2. The r_{midknn}^2 values were consistently higher than r_{pred}^2, which posed no problem for the model selection. None of the models had both high r_{midknn}^2 values and low r_{midknn}^2 values.

Table 2 shows the numbers of the models whose r_{pred}^2-values exceed the thresholds in the highest 500 models of each criterion. When the threshold of r_{pred}^2 was 0.80, the number of the models for r_{midknn}^2 was the same as that for r_{cv}^2 whereas that of r^2 was 0. For the threshold of 0.75, although the number of the models for r^2 increased, this is much lower than those for r_{midknn}^2 and r_{cv}^2, which had almost the same value. We thus confirmed that r_{midknn}^2 and $RMSE_{midknn}$ enable adequate quantification of the predictive ability of the nonlinear regression models.

Table 2. Numbers of the Models Whose r_{pred}^2-Values Exceed the Thresholds in the Highest 500 Models of Each Criterion with the QSAR Data

threshold of r_{pred}^2	no. of the models for r^2	no. of the models for r_{cv}^2	no. of the models for r_{midknn}^2
0.80	0	128	128
0.75	220	492	474

QSPR Study on Aqueous Solubility. We constructed QSPR models on aqueous solubility using the SVR method, and compared the predictive ability using r^2 (RMSE), r_{cv}^2 ($RMSE_{cv}$), and r_{midknn}^2 ($RMSE_{midknn}$). In this paper, we analyzed the aqueous solubility data investigated by Hou.[37] Aqueous solubility is expressed as log S, where S is the solubility at a temperature of 20−25 °C in moles per liter. The Hou data set includes 1290 diverse compounds and has been analyzed by several groups.[37−44]

For this data set, 1935 molecular descriptors were calculated by the Dragon 6.0 software.[35] Descriptors with the same values for more than 50% of molecules were removed. When the correlation coefficient of a pair of descriptors was greater than 0.9, one was removed. This process reduced the number of descriptors to 225. The data set was divided into a training set of 878 molecules and a test set of 412 molecules, as in the original study.[37] The 878 molecules were used to construct the prediction model, and the 412 molecules were used to test the predictive accuracy of the obtained model.

As for the numerical simulation data and the QSAR data, we determined the optimal parameters for r^2, r_{cv}^2, and r_{midknn}^2 ($RMSE$, $RMSE_{cv}$, and $RMSE_{midknn}$) by varying the values of C, ν, and γ. Figure 7 shows the r_{pred}^2 and $RMSE_p$ values of the model selected with each criterion. The r_{pred}^2 values of the models selected with r_{cv}^2 and r_{midknn}^2 were higher than those obtained with r^2, and the opposite is true for the $RMSE_p$ values. The r_{pred}^2 and $RMSE_p$ values for r_{midknn}^2 and r_{cv}^2 are almost the same, irrespective of the value of k. In addition, the r_{pred}^2 values stabilized as k increased. We have therefore confirmed that the appropriate model can be selected using r_{midknn}^2 and $RMSE_{midknn}$.

Figure 8 shows the relationship between r_{pred}^2 and r^2, r_{cv}^2, and r_{midknn}^2 ($k = 10$) for different combinations of C, ν, and γ. When r^2 was used (Figure 8a), there are some combinations of the SVR hyperparameters, that is, the SVR models, where the r^2 values are high but the r_{pred}^2 values are low. This means that the SVR models were overfitted to the training data and had low predictive ability. That is, r^2 failed to select the appropriate SVR parameters. On the other hand, r_{cv}^2 and r_{midknn}^2 were strongly correlated with r_{pred}^2 (Figure 8b and c). The r_{midknn}^2 values were consistently higher than r_{pred}^2, as was the case for the QSAR data, but this does not affect the model selection. Although the relationship between r_{midknn}^2 and r_{pred}^2 was a little vague for small values of r_{midknn}^2, the region of high r_{midknn}^2 is more important, and the relationship between r_{midknn}^2 and r_{pred}^2 was clear in this region.

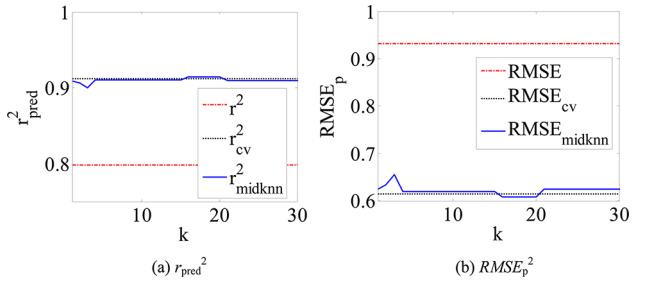


Figure 7. r_{pred}^2 and $RMSE_p$ values of the optimal SVR model for the r^2, r_{cv}^2, and r_{midknn}^2 when the QSPR data were used.

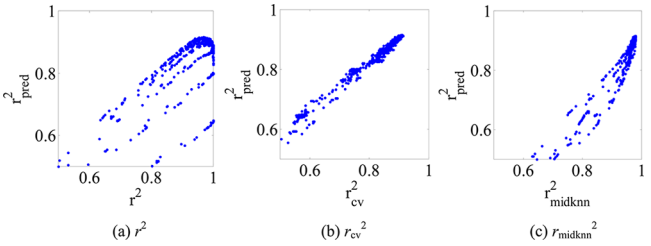


Figure 8. Relationships between r_{pred}^2 and each criteria when $k = 10$ with the QSPR data. The range of both axes was limited from 0.5 to 1.

The numbers of the models whose r_{pred}^2-values exceed the thresholds in the highest 500 models of each criterion with the QSPR data were shown in Table 3. As were the cases in the numerical simulation data analysis and the QSAR analysis, the numbers of the models for r_{cv}^2 and r_{midknn}^2 were almost the same while that of r^2 was low. Thus, r_{midknn}^2 and $RMSE_{midknn}$ enable the predictive ability of the SVR models to be evaluated appropriately.

Table 3. Numbers of the Models Whose r_{pred}^2-Values Exceed the Thresholds in the Highest 500 Models of Each Criterion with the QSPR Data

threshold of r_{pred}^2	no. of the models for r^2	no. of the models for r_{cv}^2	no. of the models for r_{midknn}^2
0.85	116	499	479
0.80	116	500	500

Through these analyses of numerical simulation data, QSAR data, and QSPR data, it was confirmed that $k = 10$ is an appropriate value for r_{midknn}^2 and $RMSE_{midknn}$.

Computation Time. Table 4 shows the computation time for each validation method with the numerical simulation data, the QSAR data and the QSPR data. For each case, the proposed method could perform the model validation in less than half the computation time it took the 10-fold and 5-fold cross-validation methods. The computation time for the

dx.doi.org/10.1021/ci4003766 | J. Chem. Inf. Model. 2013, 53, 2341−2348

Table 4. Computation Time for Each Validation Method with the Numerical Simulation Data, the QSAR Data, and the QSPR Data

	computation time [min]		
data set	10-fold cross-validation method	5-fold cross-validation method	proposed method
simulation data	1832	671	250
QSAR data	206	87	37
QSPR data	1562	562	222

proposed method was more than one-tenth of that of the 10-fold cross-validation and more than one-fifth of that of the 5-fold cross-validation because the number of data used for the model construction in cross-validation is less than that of all training data (nine-tenths for the 10-fold cross-validation and four-fifths for the 5-fold cross-validation).

CONCLUSION

In this study, we developed the r_{midknn}^2 and $\mathrm{RMSE}_{\mathrm{midknn}}$ criteria, based on the MIDKNNs, for evaluating the predictive ability of nonlinear regression models. Through case studies on simulation data with a nonlinear relationship between **X** and **y**, QSAR data, and QSPR data, it was confirmed that r_{midknn}^2 and $\mathrm{RMSE}_{\mathrm{midknn}}$ could evaluate the predictive ability of the nonlinear regression models and select the hyperparameters with which the predictive model could be constructed. The resulting models performed similarly to those derived from r_{cv}^2 and $\mathrm{RMSE}_{\mathrm{cv}}$ using CV with less computation time. The proposed criteria are simple but work effectively. Additionally, the case studies demonstrated that $k = 10$ is suitable for r_{midknn}^2 and $RMSE_{\mathrm{midknn}}$, regardless of differences in the number of training data and **X**-variables.

We used ED as a similarity index in this paper, but the similarity cannot be appropriately quantified when there is a degree of collinearity between variables. The Mahalanobis distance,[31] correlation, and so on can overcome this problem. Indeed, adequate kernel functions should be used when there is some nonlinearity between variables. The proposed r_{midknn}^2 and $\mathrm{RMSE}_{\mathrm{midknn}}$ are effective and helpful criteria for when online learning methods with big data become mainstream.

APPENDIX A

A variable $\mathbf{x}_i$ is autoscaled, and we first calculate the variance of all midpoints of $\mathbf{x}_i$. The average of $\mathbf{x}_i$ is zero and the variance of $\mathbf{x}_i$ is one

$$\frac{\sum_{j=1}^{m} x_i^{(j)}}{m} = 0 \tag{A.1}$$

$$\frac{\sum_{j=1}^{m} \{x_i^{(j)}\}^2}{m-1} = 1 \tag{A.2}$$

where m is the number of training data. To calculate the average of the midpoints of $\mathbf{x}_i$, set A as follows:

$$A = \sum_{j=1}^{m} \sum_{k=j+1}^{m} \{x^{(j)} + x^{(k)}\} \tag{A.3}$$

The following equation holds owing to the symmetry property:

$$\sum_{j=1}^{m} \sum_{k=1}^{m} \{x^{(j)} + x^{(k)}\} = 2A + \sum_{j=1}^{m} x_i^{(j)} \tag{A.4}$$

The left-hand side and the second term of the right-hand side of eq A.4 are zero, from eq A.1, and hence A equals zero. Therefore, the average of the midpoints is zero:

$$\frac{\sum_{j=1}^{m} \sum_{k=j+1}^{m} \left\{\frac{x^{(j)} + x^{(k)}}{2}\right\}}{m(m-1)/2} = \frac{A/2}{m(m-1)/2} = 0 \tag{A.5}$$

Then, to calculate the variance of the midpoints of $\mathbf{x}_i$, set B as follows:

$$B = \sum_{j=1}^{m} \sum_{k=j+1}^{m} \{x^{(j)} + x^{(k)}\}^2 \tag{A.6}$$

The following equation holds because of the symmetry property:

$$\sum_{j=1}^{m} \sum_{k=1}^{m} \{x^{(j)} + x^{(k)}\}^2 = 2B + \sum_{j=1}^{m} \{2x^{(j)}\}^2 \tag{A.7}$$

The left-hand side of eq A.7 can be expanded as follows:

$$\sum_{j=1}^{m} \sum_{k=1}^{m} \{x^{(j)} + x^{(k)}\}^2 = \sum_{j=1}^{m} \sum_{k=1}^{m} [\{x^{(j)}\}^2 + \{x^{(k)}\}^2 + 2x^{(j)}x^{(k)}] \tag{A.8}$$

From eq A.2, the following equation holds:

$$\sum_{j=1}^{m} \sum_{k=1}^{m} \{x^{(j)}\}^2 = \sum_{k=1}^{m} (m-1) = m(m-1) \tag{A.9}$$

Using eq A.1, the third term of eq A.8 can be transformed as follows:

$$\sum_{j=1}^{m} \sum_{k=1}^{m} x^{(j)}x^{(k)} = \sum_{j=1}^{m} x^{(j)} \sum_{k=1}^{m} x^{(k)} = 0 \tag{A.10}$$

From eqs A.2, A.8, A.9, and A.10, eq A.7 can be written as

$$2m(m-1) = 2B + 4(m-1) \tag{A.11}$$

Thus, the following equation holds:

$$B = (m-1)(m-2) \tag{A.12}$$

Therefore, from eq A.5, the variance of the midpoints is given as follows:

$$\sum_{j=1}^{m} \sum_{k=j+1}^{m} \left\{\frac{x^{(j)} + x^{(k)}}{2}\right\}^2 \Big/ \left\{\frac{m(m-1)}{2} - 1\right\} = \frac{1}{2}\frac{(m-1)(m-2)}{(m+1)(m-2)} = \frac{m-1}{2(m+1)} \tag{A.13}$$

When m is sufficiently large, the variance of the midpoints can be approximated as 1/2. $RMSE_{\mathrm{midknn}}$ (eq 5) is divided by the standard deviation of the midpoints, that is, the square root of eq A.13, and is corrected because the denominator in the square root of $\mathrm{RMSE}_{\mathrm{midknn}}$ (eq 5) is N_{midknn} as follows:

dx.doi.org/10.1021/ci4003766 | J. Chem. Inf. Model. 2013, 53, 2341−2348

$$\mathrm{RMSE}_{\mathrm{midknn}} = \sqrt{\frac{\sum_{\mathrm{midknn}=1}^{N_{\mathrm{midknn}}} (y^{\mathrm{midknn}} - y_{\mathrm{pred}}^{\mathrm{midknn}})^2}{N_{\mathrm{midknn}}} \times \frac{2(m+1)}{m-1} \times \frac{N_{\mathrm{midknn}}}{N_{\mathrm{midknn}}-1} \times \frac{m-1}{m}}$$
$$= \sqrt{\frac{2(m+1)\sum_{\mathrm{midknn}=1}^{N_{\mathrm{midknn}}} (y^{\mathrm{midknn}} - y_{\mathrm{pred}}^{\mathrm{midknn}})^2}{m(N_{\mathrm{midknn}}-1)}} \quad (A.14)$$

APPENDIX B

The SVR method applies a support vector machine (SVM) to regression analysis and can be used to construct nonlinear models by applying a kernel trick, as well as the SVM. The primal form of SVR can be expressed as the following optimization problem.

Minimize

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i |y_i - f(\mathbf{x}_i)|_e \quad (B.1)$$

where y_i, and $\mathbf{x}_i$ are training data, $\mathbf{w}$ is a weight vector, e is a threshold, and C is a penalizing factor that controls the trade-off between model complexity and training errors. The second term of eq 1 is the e-insensitive loss function, which is written as follows:

$$|y_i - f(\mathbf{x}_i)|_e = \max(0, |y_i - f(\mathbf{x}_i)| - e) \quad (B.2)$$

where e is a threshold. Through the minimization of eq B.1, we can construct a regression model that has a good balance between generalization capabilities and the ability to adapt to the training data. The kernel function in our application is a radial basis function:

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma|\mathbf{x} - \mathbf{x}'|^2) \quad (B.3)$$

where γ is a tuning parameter controlling the width of the kernel function. The ν-SVR method[45] that assigns an upper threshold ν of the ratio of data whose error exceeds e instead of e itself was used in this study. LIBSVM[46] was used as the machine learning software.

AUTHOR INFORMATION

Corresponding Author

*E-mail: funatsu@chemsys.t.u-tokyo.ac.jp.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge the financial support of the Japan Society for the Promotion of Science (JSPS) through a Grant-in-Aid for Young Scientists (B) (Grant 24760629).

ABBREVIATIONS

QSAR, quantitative structure−activity relationship; QSPR, quantitative structural-property relationship; PLS, partial least-squares; SVR, support vector regression; RMSE, root-mean-square error; CV, cross-validation; LOOCV, leave-one-out CV; MIDKNN, midpoints between k-nearest-neighbor data points; pIGC_{50}, logarithm of 50% growth inhibitory concentration

REFERENCES

(1) Gasteiger, J.; Engel, T. *Chemoinformatics—A Textbook*; Wiley-VCH: Weinheim, Germany, 2003.

(2) Wold, S. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37−52.

(3) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(4) Marini, F.; Roncaglioni, A.; Novič, M. Variable selection and interpretation in structure−affinity correlation modeling of estrogen receptor binders. *J. Chem. Inf. Model.* **2005**, *45*, 1507−1519.

(5) Mazzatorta, M.; Smiesko, M.; Piparo, E. L.; Benfenati, E. QSAR model for predicting pesticide aquatic toxicity. *J. Chem. Inf. Model.* **2005**, *45*, 1767−1774.

(6) Kim, K.; Lee, J. M.; Lee, I. B. A novel multivariate regression approach based on kernel partial least squares with orthogonal signal correction. *Chemom. Intell. Lab. Syst.* **2005**, *79*, 22−30.

(7) Obrezanova, O.; Csanyi, G.; Gola, J. M. R.; Segall, M. D. Gaussian processes: A method for automatic QSAR modeling of ADME properties. *J. Chem. Inf. Model.* **2007**, *47*, 1847−1857.

(8) Song, M. H.; Clark, M. Development and evaluation of an in silico model for hERG binding. *J. Chem. Inf. Model.* **2006**, *46*, 392−400.

(9) Lowe, R.; Mussa, H. Y.; Mitchell, J. B. O.; Glen, R. C. Classifying molecules using a sparse probabilistic kernel binary classifier. *J. Chem. Inf. Model.* **2011**, *51*, 1539−1544.

(10) Schuurmann, G.; Ebert, R. U.; Chen, J. W.; Wang, B.; Kuhne, R. External validation and prediction employing the predictive squared correlation coefficient—Test set activity mean vs training set activity mean. *J. Chem. Inf. Model.* **2008**, *48*, 2140−2145.

(11) Chirico, N.; Gramatica, P. Real external predictivity of QSAR models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *J. Chem. Inf. Model.* **2011**, *51*, 2320−2335.

(12) Chirico, N.; Gramatica, P. Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. *J. Chem. Inf. Model.* **2012**, *51*, 2044−2058.

(13) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graph Model.* **2002**, *20*, 269−276.

(14) Baumann, K. Cross-validation as the objective function for variable-selection techniques. *TrAC, Trends Anal. Chem.* **2003**, *22*, 395−406.

(15) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. Detecting "bad" regression models: multicriteria fitness functions in regression analysis. *Anal. Chim. Acta* **2004**, *515*, 199−208.

(16) Konovalov, D. A.; Sim, N.; Deconinck, E.; Heyden, Y. V.; Coomans, D. Statistical confidence for variable selection in QSAR models via Monte Carlo cross-validation. *J. Chem. Inf. Model.* **2008**, *48*, 370−383.

(17) Sheridan, R. P. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J. Chem. Inf. Model.* **2013**, *53*, 783−790.

(18) Rucker, C.; Rucker, G.; Meringer, M. y-Randomization and its variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345−2357.

(19) Papa, E.; Villa, F.; Gramatica, P. Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in *Pimephales promelas* (fathead minnow). *J. Chem. Inf. Model.* **2005**, *45*, 1256−1266.

(20) Anderssen, E.; Dyrstad, K.; Westad, F.; Martens, H. Reducing over-optimism in variable selection by cross-model validation. *Chemom. Intell. Lab. Syst.* **2006**, *84*, 69−74.

(21) Howe, D.; Costanzo, M.; Fey, P.; Gojobori, T.; Hannick, L.; Hide, W.; Hill, D. P.; Kania, R.; Schaeffer, M.; Pierre, S., St; Twigger, S.; White, O.; Rhee, S. Y. Big data: The future of biocuration. *Nature.* **2008**, *455*, 47−50.

(22) Li, J. Z.; Gramatica, P. Classification and virtual screening of androgen receptor antagonists. *J. Chem. Inf. Model.* **2010**, *50*, 861−874.

(23) Rodgers, S. L.; Davis, A. M.; Tomkinson, N. P.; Waterbeemd, H. V. D. Predictivity of simulated ADME autoQSAR models over time. *Mol. Inf.* **2011**, *30*, 256−266.

(24) D'Souza, A.; Schaal, S. Incremental online learning in high dimensions. *Neural Comput.* **2005**, *17*, 2602−2634.

(25) Kadlec, P.; Gabrys, B.; Strandt, S. Data-driven soft sensors in the process industry. *Comput. Chem. Eng.* **2009**, *33*, 795−814.

(26) Kaneko, H.; Funatsu, K. Classification of the degradation of soft sensor models and discussion on adaptive models. *AIChE J.* **2013**, *59*, 2339−2347.
(27) Kaneko, H.; Funatsu, K. Adaptive soft sensor model using online support vector regression with time variable and discussion of appropriate hyperparameter settings and window size. *Comput. Chem. Eng.* **2013**, *58*, 288−297.
(28) Mallow, C. L. Some comments on Cp. *Technometrics* **1973**, *15*, 661−675.
(29) Akaike, H. Factor analysis and AIC. *Psychometrika* **1987**, *52*, 317−332.
(30) Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461−464.
(31) Maesschalck, R. D.; Jouan-Rimbaud, D.; Massart, D. L. The Mahalanobis distance. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 1−18.
(32) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer: New York, 1999.
(33) Li, G.; Aute, V.; Azarm, S. An accumulative error based adaptive design of experiments for offline metamodeling. *Struct. Multidiscip. Optim.* **2010**, *40*, 137−155.
(34) http://www.cadaster.eu/node/65 (accessed June 12, 2013).
(35) Owen, J. R.; Nabney, I. T.; Medina-Franco, J. L.; Lopez-Vallejo, F. Visualization of molecular fingerprints. *J. Chem. Inf. Model.* **2011**, *21*, 1552−1563.
(36) http://www.talete.mi.it/products/dragon_description.htm (accessed June 12, 2013).
(37) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 266−275.
(38) Baurin, N. Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 643−651.
(39) Sun, H. A Universal molecular descriptor system for prediction of logP, logS, logBB, and absorption. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 748−757.
(40) Wegner, J. K.; Fröhlich, H.; Zell, A. Feature selection for descriptor based classification models. 1. Theory and GA-SEC algorithm. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 921−930.
(41) Bergström, C. A. S.; Wassvik, C. M.; Norinder, U.; Luthman, K.; Artursson, P. Global and local computational models for aqueous solubility prediction of drug-like molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1477−1488.
(42) Clark, M. Generalized fragment-substructure based property prediction method. *J. Chem. Inf. Model.* **2005**, *45*, 30−38.
(43) Vidal, D.; Thormann, M.; Pons, M. LINGO, An efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *J. Chem. Inf. Model.* **2005**, *45*, 386−393.
(44) Kaneko, H.; Funatsu, K. Development of a new regression analysis method using independent component analysis. *J. Chem. Inf. Model.* **2008**, *48*, 534−541.
(45) Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer: New York, 2006.
(46) Chang, C. C.; Lin, C. J. LIBSVM: A library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm.