Quantitative Structure-Activity Relationship Modeling Based on Improving Penalized Linear Regression Model

Rehad Emad Shamany and Zakariya Yahya Algamal
Department of Statistics and Informatics, University of Mosul, Mosul, Iraq
E-mail: zakariya.algamal@uomosul.edu.iq

Abstract. One of the powerful and a promising model which is used to understand the structural relationship between the chemical activity and the chemical compounds is the quantitative structure-activity relationship (QSAR). However, the huge in dimensionality is one of the major problems which affect the quality of the QSAR modeling. Penalized methods are an attractive framework that have been adapted and gained popularity among researchers as the key for performing descriptor selection and QSAR model estimation simultaneously. The choice of the tuning parameter of the penalized methods is critical. Our aim of this paper is to efficiently estimate such a tuning parameter by using bat algorithm (BA), which is a king of nature-inspired algorithms. Experimental results, obtained by running on two datasets, show that our proposed method performs better than other methods, in terms of prediction, number of selected descriptors, and running time. Further, the Y-randomization test and applicability domain confirm that the constructed QSAR model by BA method is reliable and robust.

Keywords: Bat optimization algorithm; penalized method; tuning parameter; QSAR.

1. Introduction

One of the powerful and a promising model which is used to understand the structural relationship between the chemical activity and the chemical compounds is the quantitative structure-activity relationship (QSAR) by explicitly considering the mathematical, statistical, and informatical methods [1-4]. “A common task in these models is the selection of relevant descriptors (variables), where researchers try to determine the smallest possible set of descriptors that can still achieve good predictive performance [4-17]. A typical data in QSAR modeling consist of a small sample size of compounds (molecules) and a very large number of descriptors. Consequently, QSAR modeling is challenged by the high dimensionality of the descriptors.

Descriptor selection plays a vital role in regression modeling for QSAR. It aims to find a subset of descriptor s to improve prediction accuracy and to make the interpretability of the QSAR model easy. Traditional subset selection methods like backward elimination, forward selection, and stepwise selection become more computationally expensive in such cases. Penalization methods are an attractive framework that have been adapted and gained popularity among researchers as the key for performing variable selection and model estimation simultaneously. These methods impose a penalty term to the loss function of the regression model. The advantage behind this term is to control the tradeoff between the variance and bias of the selected model.

Several penalties have been proposed and developed by researchers, among which is least absolute shrinkage and selection operator (LASSO) [18]. The efficiency of the LASSO penalty fully depends on the choice of the tuning parameter. The cross-validation approach (CV), a data-driven procedure, is a practically useful approach for handling the tuning parameter selection problem in the LASSO penalty. However, CV tends to have a high variability and a long computational time [18-20].

Nature-inspired algorithms, which they developed by drawing inspiration from nature, have attracted considerable interest and achieved competitive results when solving optimization problems including hyper parameters tuning problem [21-23].
In this paper, a bat algorithm, which is a nature-inspired continuous algorithm, is proposed to choose the tuning parameter in LASSO penalty. The proposed method will efficiently help to find the most important descriptors in the QSAR model with a high prediction. The superiority of the proposed method in establishing a reliable QSAR model for two different applications in the computational chemistry is proved.

2. QSAR model with LASSO penalty

In QSAR studies, the linear regression model has been commonly used to link the biological activities as a response variable to the molecular descriptors as predictor variables for data analysis. The resulting ordinary least square method (OLS) has a closed form, which is easy to compute. However, OLS fails when the number of molecular descriptors, p, greater than the number of compounds, n, because the design data matrix $X$ has more columns than rows and has multicollinearity between molecular descriptors, therefore $X^TX$ is singular [24, 25].

Consider that we have a data set \{(y_i, x_i)\}_{i=1}^n where $y_i \in \mathbb{R}$ is a response variable (chemical activity) and $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in \mathbb{R}^p$ is a $p \times 1$ known predictor vector (descriptors). Without loss of generality, we assume that the response variable is centered and the predictors are standardized. Consider a classical linear regression model for QSAR study,

$$y_i = x_i^T \beta + \varepsilon_i,$$  \hspace{1cm} (1)

where $\beta = (\beta_1, \ldots, \beta_p)$ is a $p \times 1$ vector of unknown regression coefficients, and $\varepsilon_i$ is an error variable with mean 0 and variance $\sigma^2$.

With a tuning parameter, $\alpha$, the penalized linear regression model using LASSO penalty is

$$\hat{\beta}_{\text{LASSO}} = \arg \min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \alpha \sum_j | \beta_j | \right\}$$  \hspace{1cm} (2)

where $\alpha \geq 0$. The main advantage of the penalty function is to perform the variable selection by shrinking the parameters to zero. For $\alpha = 0$, we obtain the ordinary least squares (OLS) estimation. In contrast, for large values of $\alpha$, the influence of the penalty function on the coefficient estimates increases. Choosing the tuning parameter is an important part of the model fitting, because the tuning parameter should find the right balance between bias and variance to minimize prediction error.

3. Selection of the tuning parameter

A crucial part of variable selection and model estimation using the penalized methods is the selection of the tuning parameter. In practice the tuning parameter, $\alpha$, has to be chosen by a data driven procedure [10, 19, 26]. This can be achieved by using cross-validation or generalized cross-validation. In addition, the information criteria, such as Akaike information criterion (AIC), and Bayesian information criterion (BIC), are considered. These criteria can be defined as

$$CV_\alpha = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \hat{y}_{i(\alpha)}}{1 - S_i} \right)^2,$$  \hspace{1cm} (3)

$$GCV_\alpha = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \hat{y}_{i(\alpha)}}{1 - trS / n} \right)^2,$$  \hspace{1cm} (4)

$$\text{AIC}_\alpha = -2 \log \left( \sum_{i=1}^{n} (y_i - \hat{y}_{i(\alpha)})^2 \right) + 2 trS.$$  \hspace{1cm} (5)
FISCAS 2021
Journal of Physics: Conference Series 1897 (2021) 012016 doi:10.1088/1742-6596/1897/1/012016

\[ \text{BIC}_{(\alpha)} = -2\log \left[ \sum_{i=1}^{n}(y_i - \hat{y}_{i(i(\alpha))})^2 \right] + \log(n) \times \text{tr} S, \]  
(6)

where \( \hat{y}_{i(i(\alpha))} \) is the fitted values and \( s_{ii} \) is the \( i^{th} \) diagonal element of the hat matrix, \( S \), of the selected predictors, where \( S = X [X'X + \sum_{i=1}^{m}(\hat{b}_{\text{LASSO}})^2]^{-1}X' \).

It is worth mentioning that CV and GCV method is greatly dependent on the fold assignment process which leads to large variability in selecting the shrinkage parameter value and, consequently, will negatively affect the prediction performance of the penalized linear regression using LASSO. This is happen because repeating the observations assignment to folds might result in significantly different values of \( \alpha \) [27, 28]. On the other hand, the AIC and BIC can be time consuming.  

4. The proposed method

In recent years, meta-heuristic algorithms have a great attraction and proved their efficiency for solving complex optimization problems [29]. Among them, bat algorithm (BA) is considered to be a powerful algorithm because it is easy to implement with few parameters. BA proposed by Yang [30] is based on the echolocation ability of the microbats that guides them on their foraging behavior. This algorithm starts with the random initial population of bats in a \( n \)-dimensional search space where the position of the bat \( i \) denoted by \( x_i \) and its velocity denoted by \( v_i \) at time \( t \). Therefore, the new positions \( x_{i+1} \) and new velocities \( v_{i+1} \) at time step \( t+1 \) can be determined by

\[ v_{i+1} = v_i + (x_i - G_{\text{best}})\delta_i, \]  
(7)

\[ x_{i+1} = x_i + v_{i+1}, \]  
(8)

\[ \delta_i = \delta_{\text{min}} + (\delta_{\text{max}} - \delta_{\text{min}})\theta, \]  
(9)

where \( \theta \) is a random number in \([0, 1]\), \( G_{\text{best}} \) represents the current global optimal solution, and \( \delta_i \) represents the pulse frequency emitted by bat \( i \) at the current moment, where \( \delta_{\text{min}} \) and \( \delta_{\text{max}} \) represent the minimum and maximum values of pulse frequency, respectively. Initially, \( \delta_i \) is assigned randomly for each bat which is elected uniformly from \([\delta_{\text{min}}, \delta_{\text{max}}]\) [30-33]. The corresponding position of the randomly bat is updated as

\[ x_{\text{new}} = x_{\text{old}} + cL', \]  
(10)

where \( x_{\text{old}} \) represents a random solution chosen from the current best solutions, \( L' \) is the loudness and \( c \) is a random vector that is drawn from \([-1,1]\). The pulse emission rate \( r \) and the loudness \( L \) are updated by controlling the balance between these techniques as follows

\[ r_{i+1} = r_i \times [1 - \exp(-b_1 t)], \]  
(11)

\[ L_{i+1} = b_2 L_i', \]  
(12)

where \( b_1 \) and \( b_2 \) are constants.

For LASSO penalty, we have one parameter, \( \alpha \). This tuning parameter treated as a position in BA. Specifically, our improving of penalized regression model is depending on giving a wide search space for the \( \alpha \) values with less time. Consequently, the steps of our proposed improving are as:

Step 1: The number of bats is \( nb = 30 \), \( \delta_{\text{min}} = 0 \), \( \delta_{\text{max}} = 2 \), \( b_1 = b_2 = 0.9 \), and the maximum number of iterations is \( t_{\text{max}} = 100 \).
Step 2: The positions of each bat are randomly specified. For the $\alpha$, the position is randomly generated from uniform distribution $U(0,500)$.

Step 3: The fitness function is defined as

$$\text{fitness} = \min \left[ \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_{i,\text{test}} - \hat{y}_{i,\text{test}})^2 \right],$$

where the fitness is calculated for the testing dataset.

Step 4: The positions of the bats are updated using Eq. (10).

Step 5: Steps 3 and 4 are repeated until a $t_{\text{max}}$ is reached.

5. Datasets

To test the predicting performance of our proposed algorithm, BA, comprehensive comparative experiments with other tuning selecting methods. Two different sets of chemical datasets were used in this research: “The first data set concerning the QSAR study of the anticancer potency of imidazole[4,5-b] pyridine derivatives (ACIP) which were used as anticancer compounds in Aurora kinase. The biological activities of 65 imidazo[4,5-b]pyridine derivatives, which was represented by the half maximal inhibitory concentration (IC50) were collected from the literature [34-38].

The second data set is related to the QSAR study of the non-peptidic inhibitors of procollagen C-proteinase (PCP). The PCP is an enzyme which cleavages of the C-terminal propeptide at Ala-Asp in type I and II procollagens and at Arg-Asp in type III. This data set consists of a set of 54 molecules of sulfonamide derivatives as non-peptidic PCP inhibitors with their associated biological activities (IC50) which were collected from the literature [39, 40]. The chemical structures and their associated IC50 for both datasets are listed in the supplementary file.

In order to build a persuasive QSAR model, each data set was randomly separated into a training data set and a test data set according to the proportion of 70%:30%. The training data set was used to find the tuning parameter and to build the QSAR model, and the test data set was used to validate the constructed QSAR model.

The molecular structures of the used dataset compounds were sketched using Chem3D software. The molecular structures were optimized using the molecular mechanics (MM2) method and then by a molecular orbital package (MOPAC) module in Chem3D software. Dragon software (version 6.0) was used to generate 4885 molecular descriptors including all 29 blocks based on the optimized molecular structures [41]. To include consistent and useful molecular descriptors, preprocessing steps were carried out as follows: first, those that had constant value for all compounds were excluded from the QSAR study. Second, molecular descriptors in which 60% of their values were zeros were removed. Then, those that had zero values for all compounds were discarded. After that, molecular descriptors with a relative standard deviation of less than 0.001 were removed. In addition, correlation of remained molecular descriptors and inhibition activity was examined to omit the multicollinearity by disregarding those highly correlated ($r_{ij} \geq 0.90$). Finally, 2053 and 2479 descriptors for both ACIP and PCP datasets are kept.

6. Results and discussion

In order to study the performance of the proposed algorithm, BA, “for improving the penalized regression model using LASSO, comparisons with each of CV, GCV, AIC, and BIC were carried out with $k=10$ Several prediction assessment criteria were used to test the effectiveness of our proposed method. Depending on the training data sets, the mean-squared error ($MSE_{\text{train}}$), the number of selected descriptors, and the computational time. On the other hand, the mean-squared error of the testing data set ($MSE_{\text{test}}$) was used as external validation.
Table 1 summarizes the number of features selected by each of use approach in the training set and the averaged $\text{MSE}_{\text{train}}$. The number of descriptors selected by each approach is an important factor, in which methods with a small number of selected features are preferred. As can be seen from Table 1, the proposed method, BA, possesses a fewer descriptors than the other four methods. For instance, in PCP dataset, BA selected 4 descriptors compared to 7, 6, 6, and 8 descriptors for the CV, GCV, BIC, and AIC, respectively. In terms of predictive performance, as we can see from Table 1, the BA was superior to all the compared methods. Hence, the utilization of the BA yielded a lowest $\text{MSE}_{\text{train}}$.

Furthermore, in ACIP dataset, it can be seen that the reduction in $\text{MSE}_{\text{train}}$ of the BA was about 44.03%, 42.10%, 41.39%, and 43.02% lower than that of the CV, GCV, BIC, and the AIC, respectively. Moreover, with respect to the results of Table 1, CV method is ranked last.

Once again, based on the test set results in Table 1, the proposed method, BA, yields significantly better predictive ability as compared to CV, GCV, BIC, and AIC. The predictive performance achieved by the BA in ACIP dataset, for example, was 0.317, which was better than 0.726, 0.531, 0.628, and 0.691 obtained by CV, GCV, BIC, and AIC, respectively. Among the methods used, it is clearly seen that the CV is the worst predictive method.

To further highlight the computational efficiency, Table 2 shows the average CPU time of the proposed algorithm based on 20 repeating, BA, CV, GCV, BIC, and AIC. As can be seen, in terms of computational efficiency, the BA has less time than CV, GCV, BIC, and AIC. The p-values (*) from Wilcoxon’s rank sum test (nonparametric statistical test) with 5% significance level are adopted. The statistical test is needed to indicate that the BA provides a significant improvement compared to the other methods. It can be seen that there is a statistical difference between BA and all the others for all datasets. This is not surprising because the CV and GCV is computationally consuming time.

The obtained results in Tables 1 and 2 strongly prove that high exploration of BA is able to explore the search area extensively and give promising regions of the search area comparing to other used methods”.

| Table 1. Evaluation criteria based on training dataset |
|---------------------------------|-----------------|--------------|-----------------|-----------------|
|                                | $\alpha$ | No. of selected descriptors | Time in seconds | $\text{MSE}_{\text{train}}$ | $\text{MSE}_{\text{test}}$ |
| ACIP                           | BA      | 0.097 | 4   | 81   | 0.286 | 0.317 |
|                                | CV      | 0.119 | 8   | 116  | 0.511 | 0.726 |
|                                | GCV     | 0.108 | 7   | 101  | 0.494 | 0.531 |
|                                | BIC     | 0.111 | 7   | 97   | 0.488 | 0.628 |
|                                | AIC     | 0.122 | 8   | 110  | 0.502 | 0.691 |
| PCP                            | BA      | 0.171 | 4   | 93   | 0.513 | 0.652 |
|                                | CV      | 0.188 | 7   | 131  | 0.747 | 0.853 |
|                                | GCV     | 0.182 | 6   | 122  | 0.703 | 0.814 |
|                                | BIC     | 0.185 | 6   | 119  | 0.681 | 0.792 |
|                                | AIC     | 0.191 | 8   | 127  | 0.739 | 0.844 |
Table 2. Computing time criteria based on training dataset

|        | Time in seconds | P-value  |
|--------|-----------------|----------|
| ACIP   |                 |          |
| BA     | 81              | -        |
| CV     | 116             | 0.0001 (*)|
| GCV    | 101             | 0.0002 (*)|
| BIC    | 97              | 0.0071 (*)|
| AIC    | 110             | 0.0001 (*)|
| PCP    |                 |          |
| BA     | 93              | -        |
| CV     | 131             | 0.0002 (*)|
| GCV    | 122             | 0.0002 (*)|
| BIC    | 119             | 0.0034 (*)|
| AIC    | 127             | 0.0004 (*)|

The QSAR model by using BA was further validated by applying the Y-randomization test [42]. This was in order to ensure that the predictive power of the BA model was not based on chance. This test randomly shuffled the retention indices values several times and applied BA each time. In each time, the $Q^2_{\text{int}} = 1 - \left[ \frac{\sum_{i=1}^{n_{\text{sub}}} (y_{i,\text{train}} - \bar{y}_{\text{train}})^2}{\sum_{i=1}^{n_{\text{sub}}} (y_{i,\text{train}} - \bar{y})^2} \right]$ was calculated. If all the obtained values were less than the $Q^2_{\text{int}}$ of the constructed QSAR by BA, then the constructed QSAR was not due to chance correlation, indicating that the BA method could lead to an acceptable method using the training data. Figure 1 shows the results for the Y-randomization test for 500 times of $Q^2_{\text{int}}$ values.

For ACIP, it can be clearly seen from Figure 1 that the $Q^2_{\text{int}}$ values were in the range of 0.0101 to 0.3643. In comparison to true $Q^2_{\text{int}}$ values of BA ($Q^2_{\text{int}} = 0.963$), these values indicate that the QSAR model of imidazo[4,5-b]pyridine derivatives by BA method was not due to chance correlation or structural dependence of the training data. On the other hand, for non-peptidic inhibitors of procollagen C-proteinase (PCP), it is clearly seen from Figure 2 that true $Q^2_{\text{int}}$ values of BA ($Q^2_{\text{int}} = 0.974$). This indicating that the QSAR model of non-peptidic inhibitors of procollagen C-proteinase by BA method was not due to chance correlation or structural dependence of the training data.
Figure 1. Y-randomization test for BA method for ACIP dataset.
To further evaluate the ability of BA method to construct a robust QSAR model, the leverage approach was used as an applicability domain (AD) assessment [43]. Figures 3 and 4 display the Williams plot of the leverage values against the standardized residuals for each compound for the BA model (the dotted line indicates the leverage threshold, while the dashed line represents the standardized residual limits). The influential compound can be detected when its leverage value is greater than the leverage threshold \( h^* = \frac{3(p + 1)}{n} \) where \( p \) is the number of the selected descriptors in the final QSAR model, and \( n \) represents the number of compounds.

It is obvious from Figures 2 and 3 that no compounds have a standardized residual higher than the limit ±3, which can be considered as outliers, or with a high leverage value. Thus, it is clearly demonstrated from these two figures that all the results confirm that the constructed QSAR model using the BA method is reliable and robust for both datasets.
Figure 3. AD plot for the training and testing data for ACIP dataset.
7. Conclusion
In this study, a new tuning parameter selection method, which is based on nature-inspired algorithm, was proposed. Bat algorithm was utilized for this paper. The results of the experimental study and the statistical analysis on two datasets have demonstrated that the performance of the proposed approach compared with the other methods leads to a better performance in terms of prediction, computational time, and number of selected features. In addition, the results which were obtained by the Y-randomization test and applicability domain confirm that the constructed QSAR model by BA algorithm is robust and reliable. In conclusion, the current study proposes a useful approach to be appropriately used in other QSAR studies.
REFERENCES

[1] Gupta S, Basant N and Singh K P 2015 Qualitative and quantitative structure–activity relationship modelling for predicting blood-brain barrier permeability of structurally diverse chemicals SAR QSAR Environ. Res. 26 95-124

[2] Nantasenamat C, Monnor T, Worachartcheewan A, Mandi P, Isarakura-Na-Ayudhya C and Prachayasittikul V 2014 Predictive QSAR modeling of aldose reductase inhibitors using Monte Carlo feature selection Eur. J. Med. Chem. 76 352-9

[3] Pourbasheer E, Aalizadeh R, Shokouhi T S, Ganjali M R, Norouzi P and Shadmanesh J 2014 2D and 3D quantitative structure–activity relationship study of hepatitis C Virus NS5B polymerase inhibitors by comparative molecular field analysis and comparative molecular similarity indices analysis methods J. Chem. Inf. Model. 54 2902-14

[4] Al-Fakih A M, Algamal Z Y, Lee M H, Abdallah H H, Maarof H and Aziz M 2016 Quantitative structure-activity relationship model for prediction study of corrosion inhibition efficiency using two-stage sparse multiple linear regression J. Chemom. 30 361-8

[5] Eklund M, Norinder U, Boyer S and Carlsson L 2012 Benchmarking Variable Selection in QSAR Mol Inform 31 173-9

[6] Al-Fakih A M, Algamal Z Y, Lee M H, Aziz M and Ali H T M 2019 QSAR classification model for diverse series of antifungal agents based on improved binary differential search algorithm SAR QSAR Environ Res 30 131-43

[7] Al-Fakih A M, Aziz M, Abdallah H H, Algamal Z Y, Lee M H and Maarof H 2015 High dimensional QSAR study of mild steel corrosion inhibition in acidic medium by furan derivatives International Journal of Electrochemical Science 10 3568 - 83

[8] Algamal Z 2017 An efficient gene selection method for high-dimensional microarray data based on sparse logistic regression Electronic Journal of Applied Statistical Analysis 10 242-56

[9] Algamal Z Y and Lee M H 2015 Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification Expert Systems with Applications 42 9326–32

[10] Algamal Z Y and Lee M H 2015 Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification Comput Biol Med 67 136-45

[11] Algamal Z Y and Lee M H 2017 A new adaptive L1-norm for optimal descriptor selection of high-dimensional QSAR classification model for anti-hepatitis C virus activity of thiourea derivatives SAR QSAR Environ Res 28 75-90

[12] Algamal Z Y and Lee M H 2018 A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification Advances in Data Analysis and Classification

[13] Algamal Z Y, Qasim M K and Ali H T M 2017 A QSAR classification model for neuraminidase inhibitors of influenza A viruses (H1N1) based on weighted penalized support vector machine SAR QSAR Environ Res 28 415-26

[14] Qasim M K, Algamal Z Y and Ali H T M 2018 A binary QSAR model for classifying neuraminidase inhibitors of influenza A viruses (H1N1) using the combined minimum redundancy maximum relevancy criterion with the sparse support vector machine SAR QSAR Environ Res 29 517-27

[15] Qasim O S and Algamal Z Y 2018 Feature selection using particle swarm optimization-based logistic regression model Chemometrics and Intelligent Laboratory Systems 182 41-6
[16] Eklund M, Norinder U, Boyer S and Carlsson L 2014 Choosing feature selection and learning algorithms in QSAR J Chem Inf Model 54 837-43
[17] Goodarzi M, Dejaegher B and Heyden Y V 2012 Feature selection methods in QSAR studies Journal of AOAC International 95 636-51
[18] Tibshirani R 1996 Regression shrinkage and selection via the lasso J. Roy. Statist. Soc. Ser. B 58 267-88
[19] Kawano S 2014 Selection of tuning parameters in bridge regression models via Bayesian information criterion Statistical Papers 55 1207-23
[20] Park H, Sakaori F and Konishi S 2014 Robust sparse regression and tuning parameter selection via the efficient bootstrap information criteria Journal of Statistical Computation and Simulation 84 1596-607
[21] Chou J-S and Pham A-D 2017 Nature-inspired metaheuristic optimization in least squares support vector regression for obtaining bridge scour information Information Sciences 399 64-80
[22] Laref R, Losson E, Sava A and Siadat M 2019 On the optimization of the support vector machine regression hyperparameters setting for gas sensors array applications Chemometrics and Intelligent Laboratory Systems 184 22-7
[23] Li S, Fang H and Liu X 2018 Parameter optimization of support vector regression based on sine cosine algorithm Expert Systems with Applications 91 63-77
[24] Filzmoser P, Gschwandtner M and Todorov V 2012 Review of sparse methods in regression and classification with application to chemometrics J. Chemom. 26 42-51
[25] Ross Kunz M and She Y 2013 Multivariate calibration maintenance and transfer through robust fused LASSO Journal of Chemometrics 27 233-42
[26] Algamal Z Y and Lee M H 2015 Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification Expert Systems with Applications 42 9326-32
[27] Algamal Z Y 2018 Shrinkage parameter selection via modified cross-validation approach for ridge regression model Communications in Statistics - Simulation and Computation 1-9
[28] Algamal Z Y 2018 A new method for choosing the biasing parameter in ridge estimator for generalized linear model Chemom. Intell. Lab. Syst. 183 96-101
[29] Lin S-W, Ying K-C, Chen S-C and Lee Z-J 2008 Particle swarm optimization for parameter determination and feature selection of support vector machines Expert Systems with Applications 35 1817-24
[30] Yang X-S 2010 Nature inspired cooperative strategies for optimization (NICSO 2010): Springer pp 65-74
[31] Alomari O A, Khader A T, Al-Betar M A and Awadallah M A 2018 A novel gene selection method using modified MRMR and hybrid bat-inspired algorithm with β-hill climbing Applied Intelligence 48 4429-47
[32] Rizk-Allah R M and Hassanien A E 2017 New binary bat algorithm for solving 0–1 knapsack problem Complex & Intelligent Systems 4 31-53
[33] Zhao D and He Y 2015 Chaotic binary bat algorithm for analog test point selection Analog Integrated Circuits and Signal Processing 84 201-14
[34] Bababdani B M and Mousavi M 2013 Gravitational search algorithm: A new feature selection method for QSAR study of anticancer potency of imidazo[4,5-b]pyridine derivatives Chemom. Intell. Lab. Syst. 122 1-11
[35] Bavetsias V, Crumpler S, Sun C, Avery S, Atrash B, Faisal A, Moore A S, Kosmopoulos M, Brown N, Sheildrake P W, Bush K, Henley A, Box G, Valenti M, de Haven Brandon A, Raynaud F I, Workman P, Eccles S A, Bulyiss R, Linardopoulos S and Blagg J 2012 Optimization of imidazo[4,5-b]pyridine-based kinase inhibitors: Identification of a dual FLT3/Aurora kinase inhibitor as an orally bioavailable preclinical development candidate for the treatment of acute myeloid leukemia J. Med. Chem. 55 8721-34
Bavetsias V, Faisal A, Crumpler S, Brown N, Kosmopoulou M, Joshi A, Atrash B, Pérez-Fuertes Y, Schmitt J A, Boxall K J, Burke R, Sun C, Avery S, Bush K, Henley A, Raynaud F I, Workman P, Bayliss R, Linardopoulos S and Blagg J 2013 Aurora isoform selectivity: Design and synthesis of imidazo[4,5-b]pyridine derivatives as highly selective inhibitors of Aurora-A kinase in cells J. Med. Chem. 56 9122-35

Bavetsias V, Large J M, Sun C, Bouloc N, Kosmopoulou M, Matteucci M, Wilsher N E, Martins V, Reynisson J h, Atrash B, Faisal A, Urban F, Valenti M, de Haven Brandon A, Box G, Raynaud F I, Workman P, Eccles S A, Bayliss R, Blagg J, Linardopoulos S and McDonald E 2010 Imidazo[4,5-b]pyridine derivatives as inhibitors of Aurora kinases: Lead optimization studies toward the identification of an orally bioavailable preclinical development candidate J. Med. Chem. 53 5213-28

Lan P, Chen W-N and Chen W-M 2011 Molecular modeling studies on imidazo[4,5-b]pyridine derivatives as Aurora A kinase inhibitors using 3D-QSAR and docking approaches Eur. J. Med. Chem. 46 77-94

Khazaei A, Sarmasti N, Seyf J Y, Rostami Z and Zolfigol M A 2015 QSAR study of the non-peptidic inhibitors of procollagen C-proteinase based on Multiple linear regression, principle component regression, and partial least squares Arabian J. Chem.

Turtle E, Chow N, Yang C, Sosa S, Bauer U, Brenner M, Solow-Cordero D and Ho W B 2012 Design and synthesis of procollagen C-proteinase inhibitors Bioorg. Med. Chem. Lett. 22 7397-401

Todeschini R, Consonni V, Mauri A and Pavan M 2012 Dragon 6.

Rücker C, Rücker G and Meringer M 2007 y-randomization and its variants in QSPR/QSAR J. Chem. Inf. Model. 47 2345-57

Algamal Z Y, Lee M H and Al-Fakih A M 2016 High-dimensional quantitative structure–activity relationship modeling of influenza neuraminidase a/PR/8/34 (H1N1) inhibitors based on a two-stage adaptive penalized rank regression J. Chemom. 30 50-7