2022 19th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) | 978-1-6654-9389-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCWAMTIP56608.2022.10016572

OPTIMIZING DRUG SCREENING WITH MACHINE LEARNING

CHEN LIN[1], ZHOU XIAOXIAO[1]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China,
Chengdu 611731, China
E-MAIL: chicochin@163.com, 202021080938@std.uestc.edu.cn

Abstract:

Drug screening is the process by which potential drugs are identified and optimized before the selection of a candidate drug to progress to clinical trials. To find drug candidates with good pharmacokinetic properties and adequate safety in the human body, pharmaceutical researchers need to comprehensively consider the biological activity of compounds and their influence on the human body. More specifically, only when the compound has good biological activity and ADMET (i.e., absorption, distribution, metabolism, excretion, and toxicity) properties can it qualify as a drug candidate.

To improve the efficiency of drug screening, we propose a drug candidate screening approach based on machine learning methods, which not only discovers appropriate compounds but also reveals the potential effects of molecular descriptor (i.e., features) values on the properties of compounds. First, an accurate prediction model is trained based on independent variables (i.e., feature values) and dependent variables (i.e., bioactivity values or ADMET properties). Second, we use a feature interpretation algorithm to pick out features with a significant impact on the dependent variables. Third, we search for the approximate optimal values of these important features and analyze their numerical ranges that are beneficial to obtaining better bioactivity and ADMET properties. Experimental results demonstrate that our scheme is accurate, efficient, and reliable.

Keywords:

Drug screening; Biological activity; ADMET; Feature selection and analysis

1. Introduction

Drug screening is an important process in the pharmaceutical industry for discovering and optimizing new candidate medications. For screening potentially active medical compounds, researchers used to establish a prediction model of compound activity, which significantly reduces development time and cost [1][2]. A typical drug modeling approach based on machine learning consists of the following steps: 1) for a disease-related target, collecting a series of compounds that act on the target and their biological activity data; 2) constructing the QSAR (Quantitative Structure-Activity Relationship) model of compounds based on molecular structure descriptors and bioactivity values; 3) using the model to predict whether new compounds have good bioactivity, or to guide the structural optimization of existing compounds.

Besides biological activity, pharmacokinetic properties and drug safety of the compound are also important evaluation indicators for screening drug candidates, which contain five specific properties of the compound, e.g., absorption, distribution, metabolism, excretion, and toxicity, collectively called *ADMET*. Among them, *ADME* mainly refers to the pharmacokinetic properties that describe the change of the concentration of the compound in vivo with time, while *T* refers to the possible toxic and side effects of the compound on the human body. Only with both good biological activity and ADMET properties, a compound is eligible for drug candidate selection. Moreover, the specific values of some molecular descriptors have a greater impact on the corresponding dependent variables. Therefore, finding out these numerical ranges will also help pharmaceutical researchers discover effective drugs more quickly or develop new drugs more efficiently.

In this paper, we propose a drug candidate screening approach based on machine learning techniques. This approach can not only discover compounds with good bioactivity and ADMET properties but also analyze the values of important features. The workflow of our scheme is displayed in Fig.1. Our main work is as follows:

(1) We establish the prediction model that can accurately determine whether a new compound has the potential to be a drug candidate.

(2) Important features are visually selected by a feature interpretation algorithm, i.e., SHAP. We vividly show the effect of these important features on the dependent variables.

(3) The feature values that can improve the prediction result (i.e., making the compound have better bioactivity and ADMET properties) are searched by the PSO algorithm, which constitutes the effective numerical ranges of these important features.

978-1-6654-9389-5/22/$31.00 ©2022 IEEE

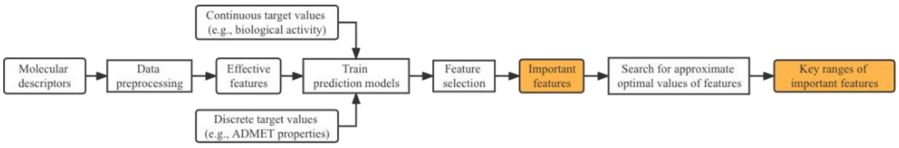

Fig.1 Workflow of our approach.

2. Methodology

Let $y = \{y_1, y_2, \dots, y_n\}$ denote the true target values and $\hat{y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n\}$ denote the predicted values (i.e., model outputs). As shown in Fig.1, our scheme consists of four main steps: 1) training the prediction model f based on y, e.g., bioactivity values and ADMET properties; 2) selecting a subset $\mathcal{V} = \{v_1, v_2, \dots, v_m\}$ from all features $\mathcal{U}$ with the greatest impact on $\hat{y}$; 3) searching for the approximate optimal values $\mathcal{W}_i = \{w_{i,1}, w_{i,2}, \dots, w_{i,m}\}$ of $\mathcal{V}$ for each compound C_i, then using f to obtain the new predicted value $\hat{y}_i'$ of C_i with $\mathcal{W}_i$; 4) determining whether $\hat{y}'$ have improved compared to $\hat{y}$, and if so, calculating the main ranges of the approximate optimal values of $\mathcal{V}$. It should be noted that the numerical ranges of $\mathcal{V}$ in the fourth step are credible only when the accuracy of f exceeds a certain threshold. The main steps are detailed below.

2.1. Training Prediction Models

The prediction model f ought to be trained on the preprocessed compound dataset. The dataset is divided into a training set and a test set according to the ratio of 7:3. The preprocessing procedure usually includes missing value padding, normalization, feature selection, etc. If the target value is continuous (e.g., biological activity values), then f is a regression model. Similarly, if the target value is discrete (e.g., five ADMET properties), then f is a classification model. The key to predicting the target value of compounds is to find the relationship between features and the dependent variable. Some nonlinear models can be used for training, such as random forest, SVM, GBDT, etc.

2.2. Important Feature Selection

In the actual drug development process, each compound contains a large number of features, but most of them may have little or no effect on the bioactivity values and ADMET properties. Therefore, we need to remove invalid features before feature selection, which has already been done in the data preprocessing procedure in Section 2.1. Traditional feature selection methods (e.g., filter, embedded, and wrapper methods) usually have poor interpretability as the model complexity increases. To solve this problem, we use the SHAP (**Sh**apley **A**dditive ex**P**lanations) algorithm [3] to interpret the model and select important features. To be more specific, it generates an explainer h based on f, then use h to calculate the SHAP values of the original features. The SHAP value is a pivotal indicator for measuring the importance of features. Finally, we select m features with the largest SHAP value, denoted as $\mathcal{V} = \{v_1, v_2, \dots, v_m\}$.

2.3. Analysis of Feature Values

In addition to predicting the properties of compounds, another goal of drug screening is to find which molecular descriptors, and at what value or in what range of values, can make compounds have better bioactivity and ADMET properties. Therefore, we need to analyze the effective values of the important features $\mathcal{V}$.

Even if traditional QSAR models achieve good accuracy, they often fail to reveal the potential impact of descriptor values. To address this issue, we use the PSO (Particle Swarm Optimization) algorithm [4] to search for approximate optimal values of $\mathcal{V}$ and evaluate the impact of these values on the corresponding variables. Given a measure of quality, PSO can iteratively improve the candidate solution to approach the optimization objective, which fits well with the optimization goal in this study. In other words, PSO is suitable for searching for feature values that can improve the bioactivity and ADMET properties of a compound.

We assume that f is accurate enough and that the larger the $\hat{y}$, the better the medical effect of the compound. Next, we set f as the function to be optimized (i.e., the measure of quality), and set $\mathcal{V}$ as the variables to be searched (i.e., candidate solution). More specifically, for each compound C_i in the test set, we fix the values of $\mathcal{U} - \mathcal{V}$, then only change the values of $\mathcal{V}$, and search for the approximate optimal values $\mathcal{W}_i = \{w_{i,1}, w_{i,2}, \dots, w_{i,m}\}$ through PSO to improve $\hat{y}_i'$. Let $w_{min,j}$ and $w_{max,j}$ denote the minimum and maximum of the j-th feature in $\mathcal{V}$:

$$w_{min,j} = min\{w_{n_1+1,j}, w_{n_1+2,j}, \dots, w_{n,j}\} \quad (1)$$

$$w_{max,j} = max\{w_{n_1+1,j}, w_{n_1+2,j}, \dots, w_{n,j}\} \quad (2)$$

At last, we extract the numerical range with the highest data coverage from $[w_{min,j}, w_{max,j}]$, as the key range of the *j*-th feature. In general, the greater the data coverage, the more representative the numerical range is. Data coverage is defined as follows:

$$\text{Data coverage} = \frac{\text{number of samples in the range}}{\text{total number of samples}} \quad (3)$$

3. Experimental Evaluation

In this study, we evaluate our approach on the dataset of Problem D in the 18th China Postgraduate Mathematical Contest in Modeling. The dataset contains information on 729 molecular descriptors (i.e., features), biological activity values (pIC50), and five ADMET properties of 1974 compounds. Among them, molecular descriptors are independent variables while biological activity values and five ADMET properties are dependent variables. In general, the larger the pIC50 value, the better the bioactivity. In addition, each ADMET property only has two values, i.e., 0 or 1, which represent the corresponding property is bad or good, respectively. According to Section 2.1, 1381 samples (70%) are used for training and 593 samples (30%) are used for testing. Among the 729 features, 225 features with a single value are eliminated, and the remaining 504 features are retained.

3.1. Results of Prediction Models

Several regression models are trained only with the bioactivity value pIC50 as the target value, including LightGBM [5], RandomForest, LinearRegression, etc. We use mean absolute error (MAE) and mean squared error (MSE) as evaluation metrics for regression models. The experimental results of the above models are reported in Table 1, which shows that LightGBM is the optimal model for predicting the biological activity.

Table 1 Loss of regression models on biological activity values

Model	Test MAE	Test MSE
LightGBM	**0.5068**	**0.4938**
RandomForest	0.5090	0.4953
DecisionTree	0.6865	0.9492
LinearRegression	0.7691	0.9522

In addition, for each ADMET property, several classification models are trained and the corresponding evaluation metric is accuracy. The experimental results of these classification models are reported in Table 2, which shows that LightGBM is the optimal model for four properties. Through a large number of comparative experiments, LightGBM is selected as the final model.

To account for the two dependent variables together, we relabel the dataset and convert the learning task into a binary classification task. For a compound, if its pIC50 value is greater than a certain threshold τ and it has three or more good ADMET properties, then it qualifies as a drug candidate and its label is set to 1, otherwise its label is set to 0. Here, we take the median of the pIC50 values as the threshold, i.e., $\tau = 6.58$. Further, we train a binary LightGBM model f_{lgb} based on the 504 features and the new labels. The output value $\hat{y}$ of f_{lgb} represents the probability that a compound is selected as a drug candidate. The accuracy of f_{lgb} is 0.8725 and the AUC score is 0.9335, which demonstrates that the prediction results of f_{lgb} have high confidence.

3.2. Results of Important Feature Selection

With f_{lgb} as the kernel model, we use the SHAP algorithm to build an explainer h. Next, we use h to calculate the SHAP value $s_{i,j}$ of the j-th feature of the i-th training sample ($1 \leq i \leq 1381$, $1 \leq j \leq 504$), which is vividly displayed in the form of a heat map, i.e., Fig.2. Take maxsssN in Fig.2 as an example, when maxsssN takes a larger value (red dots), its SHAP value increases accordingly, that is, $\hat{y}$ increases. Similarly, the smaller value (blue dots) of maxsssN, the smaller $\hat{y}$. To speed up the subsequent PSO process, we only select the six most important features that need to be searched, i.e., $\mathcal{V}$ = {LipoaffinityIndex, maxsssN, FMF, SP-5, SHBd, minHBa}.

Table 3 The impact of PSO on some prediction results

ID	y	$\hat{y}$	$\hat{y}_{pso}$	$\Delta\hat{y}$
691	0	0.0315	0.9609	0.9258
889	0	0.1123	0.9485	0.8362
971	0	0.2017	0.8403	0.6386
1377	0	0.0003	0.6564	0.6561
1512	1	0.7107	0.9993	0.2976

Table 2 Accuracy of classification models on ADMET properties

Model	Absorption	Distribution	Metabolism	Excretion	Toxicity
LightGBM	**0.9140**	**0.9444**	0.9073	**0.8887**	**0.9646**
RandomForest	0.9073	**0.9444**	**0.9123**	0.8836	0.9511
DecisionTree	0.8668	0.9174	0.8634	0.8229	0.9224
LogisticRegression	0.8550	0.8988	0.8516	0.7960	0.8752

Table 4 Important features and their numerical ranges that are beneficial for converting compounds into drug candidates

Molecular descriptor	Range	Data coverage
LipoaffinityIndex	[11.22, 14.58], [16.10, 23.00]	94.94%
maxsssN	[2.68, 2.73]	98.14%
FMF	[0.040, 0.10], [0.29, 0.52]	94.78%
SP-5	[1.85, 5.7], [8.011, 27.22]	99.83%
SHBd	[0.48, 1.54], [13.8, 18.16]	92.59%
minHBa	[-2.32, 0.88]	91.24%

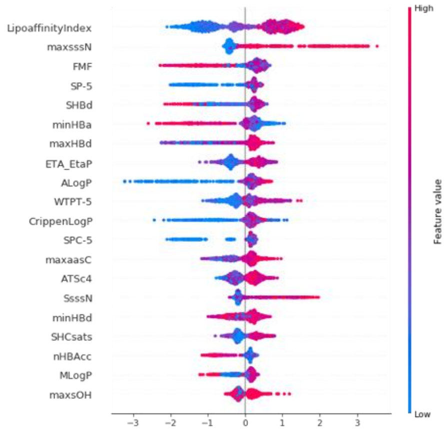

Fig.2 SHAP value of features (impact on model output).

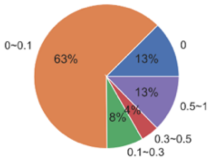

Fig.3 The distribution of $\hat{y}$.

3.3. Results of Feature Value Analysis

Regarding the settings of PSO, we set $\mathcal{V}$ as the independent variables to be searched and f_{lgb} as the function to be optimized. Only the six values of $\mathcal{V}$ can be changed by PSO, and the remaining 498 features $\mathcal{U}-\mathcal{V}$ are fixed to the original values, so $\hat{y}$ changes with $\mathcal{V}$, that is, from $\hat{y}$ to $\hat{y}_{pso}$. Let $\Delta\hat{y}=\hat{y}_{pso}-\hat{y}$ denotes the probability improvement. The PSO algorithm needs to be executed once on each test set sample, and thereby we obtain 593 sets of approximate optimal values (i.e., candidate solutions) of $\mathcal{V}$. Among them, we only keep 519 sets of candidate solutions with correct predictions, and the distribution of their $\Delta\hat{y}$ is shown in Fig.3. The mean of $\Delta\hat{y}$ is 0.1361 and about 25% of $\Delta\hat{y}$ is greater than 0.1, which demonstrates that the improvement is significant. Besides, we sample five sets of experimental data of PSO optimization, which are recorded in Table 3. Next, we focus on analyzing the above candidate solutions, then compute the main numerical ranges of these feature values and obtain the key ranges in Table 4. It should be noted that the numerical ranges in Table 4 are for reference only, and do not refer to any specific values, because the feature values of each compound are slightly different.

4. Conclusions

With the increase in the number of compounds and the complexity of screening purposes, traditional drug screening methods have become inefficient and unable to meet complex drug development needs. To address this issue, we propose a drug screening scheme based on multiple machine learning methods to screen drugs that have both good biological activity and ADMET properties. With this approach, we can efficiently and accurately determine whether a new compound can be selected as a drug candidate. Furthermore, we can find important features and their effective numerical ranges, so as to assist pharmaceutical researchers in the subsequent drug development work, which is of great significance for improving the efficiency of drug screening.

References

[1] B. J. Neves, R. C. Braga, C. C. Melo-Filho, J. T. Moreira-Filho, E. N. Muratov, and C. H. Andrade, “QSAR-based virtual screening: Advances and applications in Drug Discovery,” Frontiers in Pharmacology, vol. 9, 2018.

[2] X. Lin, X. Li, and X. Lin, “A Review on Applications of Computational Methods in Drug Screening and Design,” Molecules, vol. 25, no. 6, p. 1375, Mar. 2020.

[3] S. M. Lundberg and S.-I. Lee., “A unified approach to interpreting model predictions.” Advances in neural information processing systems, vol. 30, 2017.

[4] F. Marini and B. Walczak, “Particle Swarm Optimization (PSO). A tutorial,” Chemometrics and Intelligent Laboratory Systems, vol. 149, pp. 153–165, 2015.

[5] G. Ke et al., “Lightgbm: A highly efficient gradient boosting decision tree.” Advances in neural information processing systems, vol. 30, 2017.