Research Article

Research on Bioengineering and Biomedical science

Volume 1 | Issue 1

Accelerating the screening and development of drugs for the treatment of breast cancer

Bin Zhao[1*], Xia Jiang[2]

[1]*School of Science, Hubei University of Technology, Wuhan, Hubei, China.*

[2]*Hospital, Hubei University of Technology, Wuhan, Hubei, China.*

[*]**Corresponding author:** Dr. Bin Zhao, School of Science, Hubei University of Technology, Wuhan, Hubei, China Tel./Fax: +86 130 2851 7572. E-mail address: zhaobin835@nwsuaf.edu.cn

Submitted: 23 Jan 2022 **Accepted:** 10 Feb 2022 **Published:** 26 Feb 2022

Citation: *Bin Zhao, Xia Jiang. (2022). Accelerating the screening and development of drugs for the treatment of breast cancer. Research on Bioengineering and Biomedical science. Vol: 1 | Issue: 1 | Pg: 01-06.*

Abstract

Breast cancer has become one of the most common malignant tumors, which seriously affects the physical and mental health of women. However, the current targeted drugs for breast cancer are not effective. Therefore, the screening of potential active compounds has become an important part of drug design and development. Recently, screening active ingredients from numerous compounds has become a hot topic in cancer treatment worldwide. In this process, it is necessary to maintain high biological activity and satisfy certain kinetic properties as much as possible. Starting from the idea of machine learning, this work perform reasonable feature reduction and importance ranking of molecular descriptors by judging the weight of a single variable and the degree of correlation between variables. And based on the perspective of data mining, we further analyze the association rules of the characteristic variables. Moreover, through the PSO-XGBoost optimization method, a quantitative prediction model of the biological activity were established, providing a new plan for accelerating the screening and development of drugs for the treatment of breast cancer.

Keywords: Compound biological activity, ADMET properties prediction, PSO - XGBoost model, Association rules algorithm

1. Introduction

1.1 Research Background and Significance

Breast cancer has become one of the most common malignant tumors. The number of new breast cancer cases worldwide were 2.26 million in 2020 year, and the number of new cases in China will reach 420,000. Therefore, there is an urgent need to find drugs that can effectively inhibit breast cancer cell pathology. In recent years, screening active ingredients with targeted cancer inhibitory effects from numerous compounds has become a hot topic in cancer prevention and treatment worldwide. But in this process, often faced with factors such as large original sample base, strong correlation between variables, and difficulty in establishing model relationships, some researchers have tried to use radioactivity assays to use targets such as enzymes and receptors in the body as drugs for screening drugs. However, due to the complex relationship between the variable factors in the body and the measurement error, the long screening cycle, and environmental pollution, the effect do not work well.

1.2 Explore Importance of the Problem

At present, in order to save time and cost in drug development, methods of establishing compound activity prediction models are usually used to screen potential active compounds. The specific method is to collect a series of compounds and their biological activity data for a target related to the disease (ERα), and then use a series of molecular structure descriptors as independent variables. The biological activity value of the compound is used as the dependent variable to construct the Quantitative Structure-Activity Relationship model of the compound, and then use the model to predict new compound molecules with better biological activity, or to guide the development of existing active compounds' structural optimization. Now for the existing data set, it is required to give the value range of the molecular descriptor when it has better biological activity and at least three of the given five ADMET properties are positive.

1.3 Describe Relevant Scholarship

With the advent of the era of information technology, the application of modern computer methods to assist drug screening has moved from theory to practical application. This method can not only perform targeted screening for drug activity, but can even complete the design of new drug active compounds. Chen Hengwei et al.[1] constructed a molecular-level collaborative anti-tumor prediction model based on deep learning to solve the optimization and evaluation of multi-drug synergy. Luo Yao et al.[2] constructed a phosphatidylinositol 3-kinase family inhibitor classification model based on naive Bayes in machine learning. Chen et al.[3] used the interaction between drug targets to establish a random forest model, which promoted the high-throughput screening and prediction of the mechanism of action of drug combinations.

Furthermore, association rules are also used in the field of medical parameter optimization. For example, the intelligent classification of heart valve diseases in the field of disease treatment and diagnosis[4], the diagnosis of patients with essential hypertension[5], the development of intelligent heart disease prediction, erythema squamous disease and breast cancer diagnosis system[6-8].

1.4 State Hypotheses and Their Correspondence to Research Design

In response to the questions raised in this article, the following model assumptions are made:

1. It is assumed that all experimental results of the original data are reasonable and accurate;
2. It is assumed that only the inhibitory effect of the drug's biological activity on cancer cells is considered, without adverse effects on normal cells.

Based on the actual situation, consider starting from the two perspectives of independent variables and sample sets.First clean the sample data, delete the molecular descriptors that have negative correlation coefficients between the independent variable and the dependent variable and the variance expansion coefficient (VIF) is greater than 10; then select samples that meet at least three good properties in ADMET as the new training set.

We have verified through the XGBoost prediction model to prove that the result of the variable selection is true and effective. Then proceed to continuous numerical classification. Finally, by analyzing the significant association between the variables, we can get the value range of the molecular descriptor,

2. Method

This work mainly considers the value range of the molecular descriptor when it has better biological activity and at least three of the five given ADMET properties are better. Starting from the actual situation, consider starting from the two perspectives of independent variables and sample sets. Firstly, the sample data is cleaned a second time, and the molecular descriptors with negative correlation coefficients between the independent variable and the dependent variable and the variance expansion coefficient greater than 10 are eliminated; then the samples that meet at least three good properties in ADMET are selected as the new training set. Combine the remaining independent variables and the new training set to establish a XGBoost molecular activity prediction model, and test it with the actual value. Finally, the significant association analysis between the variables is carried out through the association rule algorithm, and the interval is specified to obtain the value ran ge of the molecular descriptor. All simulations are based on the R project for statistical computing and Statistical Product and Service Solutions (SPSS).

2.1 Pearson correlation coefficient

Pearson correlation coefficient measures the linear correlation. If the value is 0, it can only be said that there is no linear correlation between the independent variable and the dependent variable, not that there is no correlation. The greater the absolute value of the correlation coefficient, the stronger the correlation: the closer the correlation coefficient is to 1 or -1, the stronger the correlation; the closer the correlation coefficient is to 0, the weaker the correlation. Pearson correlation coefficients of each molecular descriptor and pIC50 were calculated, and variables with Pearson correlation coefficients greater than 0.3 were selected. Pearson correlation coefficient calculation formula is as follows:

$$\operatorname{cov}(x_i, y) = \frac{\sum_{j=1}^{n}(x_{ij} - x_{\mu})(y_j - y_{\mu})}{n-1}, \quad i = 1,2\ldots489, j = 1,2\ldots1974 \tag{1}$$

$$\rho_{X_i,Y} = \frac{\operatorname{cov}(X_i, Y)}{\sigma_X \sigma_Y}, \quad i = 1,2\ldots489 \tag{2}$$

By calculating Pearson correlation coefficient, 83 independent variables with correlation coefficient greater than 0.3 were selected from 489 independent variables.

2.2 Screening of variables and samples

After variable screening of Pearson's correlation coefficient, sample screening is now carried out. In order for the sample to satisfy the given five ADMET properties, at least three of the properties are good. The sum I of ADMET values is calculated for the regularized data. If it is greater than or equal to 3, the sample is retained; otherwise, it is discarded. The specific conditions are as follows:

$$VIF = \frac{1}{1 - R_i^2} \tag{3}$$

After this round of sample screening, 632 eligible samples were obtained from 1974 samples.

2.3 Variance Inflation Factor

According to the sorted variables, the coefficient of variance expansion is further calculated. After removing the variables with strong correlation between variables, the main variables of this problem are obtained. The coefficient of variance inflation is mostly used to test the independence of linear relations. It can be expressed as the ratio of the variance of the estimator of regression coefficient to the variance when the assumed independent variables are not linearly correlated. The variance inflation coefficient can measure the severity of multicollinearity in the multiple linear regression model, and its specific formula is as follows:

$$\begin{cases} I = Caco\text{-}2 + CYP3A4 + hERG + HOB + MN \\ I \geq 3, \text{the sample is retained}; \end{cases} \tag{4}$$

The closer the value of VIF is to 1, the lighter the multicollinearity is, and the heavier it is vice versa. VIF = 10 is usually taken as the criterion. When VIF < 10, there is no multicollinearity; When 10 ≦ VIF < 100, there is strong multicollinearity between variables; When VIF ≧ 100, severe multicollinearity is considered. Through 632 samples, VIF calculation is carried out for 83 existing independent variables and pIC50 dependent variables. If they are less than or equal to 10, the independent variable is retained.Finally, 632 sample data with ATSc4, C1SP3, minHBint10, maxssCH2 and MDEC_22 as independent variables and pIC50 as dependent variables were obtained.

2.4 Generalized Data

By means-variance classification method, the original numerical data is generalized and a new set is constructed. SPSS software

was used for visual discretization of data. Based on the average sum of scanned cases, the values of variables were divided into 2i+1 group by adding or subtracting I standard deviation values (i = 1,2,3) from the mean value of variables. N partition points will generate N+1 intervals. In this paper, taking into account the actual situation and the maximum value of existing data, 1, 2 and 3 standard deviations plus or minus are selected as the segmentation points on the basis of the mean value, and the actual sample maximum value is combined to further refine each interval. Although the interval values in some intervals were divided into negative values, the sample data were not taken into this interval, so such intervals were not considered in subsequent association analysis.

2.5 Apriori Algorithm

Association rule algorithm is to find the relationship between item sets from known data and get strong association rules. Association rules often use the following indicators (support, confidence and lift) to indicate the significance and correctness of the rules, the calculation formula is as follows:

$$Support(X \Rightarrow Y) = P(X \cap Y) \tag{5}$$

$$Confidence(X \Rightarrow Y) = P(Y \mid X) = \frac{P(X \cap Y)}{P(X)} \tag{6}$$

$$Lift(X \Rightarrow Y) = \frac{P(Y \mid X)}{P(Y)} = \frac{P(X \cap Y)}{P(X)P(Y)} \tag{7}$$

Apriori algorithm can be divided into the following steps:

a. Scan the database quickly, search the project set from bottom to top, compare it with the minimum threshold value of support, if it passes the threshold, it can be regarded as the high-frequency project set, denoted as L_1, and set $K = 1$.
b. Set $K = K+1$, and generate a new candidate K item set. Delete any candidate set of K-1 sub-item set in candidate K item set that does not belong to L_1, and record the filtered candidate item set as C_k.
c. Calculate whether the corresponding support degree of set C_k is not lower than the minimum support degree set in advance. If there are unqualified project sets, they should be deleted, so as to obtain L_k of high-frequency project sets.
d. Determine whether all candidate project sets have been searched.If so, go to the next step; Otherwise, go back to step (b) until the search is complete.
e. Find out significant association rules and make further decisions.

2.6 XGBoost Algorithm

XGBoost (Extreme Gradient Boosting) is a massively parallel TREE tool and the most used open source VTree toolkit.[9] Boosting is also a machine learning algorithm for reducing bias in supervised learning. Most Boosting algorithms consist of iteratively using weak learning classifiers and adding their results to a final strong learning classifier.[10,11] In addition, they are usually given different weights according to their classification accuracy.

After weak learners are added, the data is usually re-weighted to reinforce the classification of previously misclassified data points. The central idea of XGboost algorithm is to perform Taylor's second-order expansion of the objective function at t = 0, and introduce regular terms to control the complexity of the established model.

The objective function can be defined as:

$$\begin{cases} OBj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{t} \Omega(f_i) \\ \Omega(f_i) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \; L_2 \end{cases} \tag{8}$$

The newly generated number needs to fit the residual of the last prediction, so when t trees are generated, the objective function is rewritten as:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + cons\tan t \tag{9}$$

Taylor expansion of the objective function can be obtained:

$$\begin{cases} L^{(t)} \cong \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)} + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \\ g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \\ h_i = \partial^2_{y_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \end{cases} \tag{10}$$

Since the prediction score of the first T-1 tree and the residual difference of Y do not affect the optimization of the objective function, the objective function can be simplified as:

$$\tilde{L}(t) \cong \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{11}$$

Combined with the above formula, the final objective function can be obtained:[12]

$$\tilde{L}^{(t)} \cong \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T \tag{12}$$

3.Results

3.1 Screening of Variables and Samples

By calculating the Pearson correlation coefficient, we selected 83 independent variables with a correlation coefficient greater than 0.3 from 489 independent variables. Then we began to screen the samples. In order to satisfy that the sample satisfies the conditions of at least three good properties among the five given ADMET properties, 632 qualified samples were obtained from 1974 samples. Through these samples, VIF is calculated for the existing 83 independent variables and pIC50 dependent variables. The number of variables after calculation is 5.

Finally, we get 632 sample data with independent variables ATSc4, C1SP3, minHBint10, maxssCH2, MDEC_22, and dependent variables pIC50.

3.2 Data generalization results

The following is the grouping situation after data generalization.

Table 1. Property description

Attributes	Attributes' corresponding range
ATSc4	[-0.90,-0.42)→{1}; [-0.41,-0.29)→{2}; [-0.28,-0.15)→{3}; [-0.14,-0.01)→{4}; [0.00,0.12)→{5}; [0.13,0.26)→{6}; [0.27,0.39)→{7}; [0.40,0.83)→{8};
C1SP3	[-1.78,0.02)→{3}; [0.03,1.82)→{4}; [1.83,3.63)→{5}; [3.64,5.44)→{6}; [5.45,7.25)→{7}; [7.26,14.00)→{8};
minHBint10	[-3.02,-0.88)→{3}; [-0.87,1.28)→{4}; [1.29,3.44)→{5}; [3.45,5.60)→{6}; [5.61,7.76)→{7}; [7.77,10.58)→{8};
maxssCH2	[-0.37,0.13)→{3}; [0.14,0.64)→{4};[0.65,1.15)→{5}; [1.16,1.66)→{6};
MDEC_22	[0.00,4.61)→{3}; [4.62,11.09)→{4}; [11.10,17.57)→{5}; [17.58,24.04)→{6}; [24.0,30.52)→{7}; [30.53,34.21)→{8};
pIC50	[2.46,2.64)→{1}; [2.65,3.80)→{2}; [3.81,4.96)→{3}; [4.97,6.12)→{4}; [6.13,7.28)→{5}; [7.29,8.45)→{6}; [8.46,9.61)→{7}; [9.62,9.86)→{8};

3.3 The Apriori Algorithm Result

Set the minimum support and confidence thresholds to 0.2 and 0.4, use R software to implement the algorithm, and get 105 rules that meet the minimum support and confidence, and sort them in descending order of support, and output the first 10 rules for viewing. See the table 2 for details

Table 2. Mining results of partial association rules

Number	Condition	Result	Support	Confidence	Lift
1	{ATSc4=[5,8]}	{minHBint10=[4,8]}	0.653	1.000	1.003
2	{minHBint10=[4,8]}	{ATSc4=[5,8]}	0.653	0.655	1.003
3	{maxssCH2=[5,6]}	{minHBint10=[4,8]}	0.560	0.994	0.997
4	{minHBint10=[4,8]}	{maxssCH2=[5,6]}	0.560	0.561	0.997
5	{pIC50=[5,8]}	{minHBint10=[4,8]}	0.498	0.993	0.996
6	{minHBint10=[4,8]}	{pIC50=[5,8]}	0.498	0.500	0.996
7	{C1SP3=[5,8]}	{minHBint10=[4,8]}	0.484	0.993	0.996
8	{minHBint10=[4,8]}	{C1SP3=[5,8]}	0.484	0.485	0.996
9	{maxssCH2=[5,6]}	{ATSc4=[5,8]}	0.458	0.814	1.246
10	{ATSc4=[5,8]}	{maxssCH2=[5,6]}	0.458	0.702	1.246

It can be seen from the above table 2. there is only one rule with the confidence level of the top 15 rule whose result item is pIC50, so it needs to be filtered. Set the rule result item (RHS) to pIC50, and also consider that the Lift should be greater than 1. It means that the rule is valid. The detailed process of screening for significant association rules is shown in Figure 1. Through screening, the final result item is pIC50 and 12 effective rules are obtained. Figure 1 is the output association rule group matrix diagram. Through the overview of this figure, you can analyze the specific items included in the generated association rules, and select the associated rules for specific analysis.

Figure 1. Group matrix diagram of association rules

The vertical column on the right shows the result items of the rule (right-hand-side, RHS), and the horizontal column above lists the rule condition items (left-hand-side, LHS). The intersection of the matrix represents the support of the group rule according to the size of the circular area, and the shade of the prototype color represents the lift range. Given that the filter condition pIC50 value is larger, the better, so the RHS items on the right side of the result column in Figure 3 are all {pIC50=[5,8]}, and the rule condition item above shows that LHS exists {ATSc4=[5,8]}, {ATSc4=[5,8], minHBint10=[4,8]}, {maxssCH2=[5,6]}, etc.

Table 3. Mining results of association rules with setting result items and promotion

Number	Condition	Result	Support	Confidence	Lift
1	{ATSc4=[5,8]}	{pIC50=[5,8]}	0.367	0.561	1.119
2	{ATSc4=[5,8],mi HBint10=[4,8]}	{pIC50=[5,8]}	0.367	0.561	1.119
3	{maxssCH2=[5,6]}	{pIC50=[5,8]}	0.310	0.550	1.097
4	{minHBint10=[4,8],maxssCH2=[5,6]}	{pIC50=[5,8]}	0.306	0.548	1.092
5	{ATSc4=[5,8],maxssCH2=[5,6]}	{pIC50=[5,8]}	0.265	0.579	1.154
6	{ATSc4=[5,8],minHBint10=[4,8] , maxssCH2=[5,6]}	{pIC50=[5,8]}	0.265	0.579	1.154
7	{C1SP3=[5,8]}	{pIC50=[5,8]}	0.253	0.519	1.035
8	{C1SP3=[5,8],minHBint10=[4,8]}	{pIC50=[5,8]}	0.250	0.516	1.029
9	{MDEC_22=[5,8]}	{pIC50=[5,8]}	0.231	0.563	1.123
10	{minHBint10=[4,8],MDEC_22=[5,8]}	{pIC50=[5,8]}	0.227	0.560	1.117
11	{C1SP3=[5,8],maxssCH2=[5,6]}	{pIC50=[5,8]}	0.205	0.562	1.121
12	{C1SP3=[5,8],minHBint10=[4,8], maxssCH2=[5,6]}	{pIC50=[5,8]}	0.202	0.558	1.114

3.5 Interpretation of Significant Association Rules

From Table 3, it can be seen that when the value ranges of the 5 independent variables are used as conditions, there is an association relationship with the result item {pIC50=[5,8]}, and the gain (Lift) is greater than 1, and the above 12 rules are all valid. In order to ensure that the compound has better biological activity for inhibiting (ERα), it is necessary to determine the value range of the independent variable, and the larger the pIC50 value, the higher the biological activity, so the value ranges in all conditions have positive reference significance. It can be considered that the value of each variable is in this interval {ATSc4=[5,8]}; {C1SP3=[5,8]}; {minHBint10=[4,8]}; {maxssCH2=[5,6]}; {MDEC_22=[5,8]}, and the pIC50 value is at a higher level has a significant correlation.

3.6 Variable Suggestion Interval

From Section 3.5, the optimal interval for each variable to be effective for the pIC value has been obtained. Combined with the attribute description in Table 2, the final variable recommended interval can be obtained, as shown in the following table.

Table 4. The optimal value range of variables

Variable name	Optimal value range
ATSc4	[0.00,0.083]
C1SP3	[1.83,14.00]
minHBint10	[-0.87,10.58]
maxssCH2	[0.65,1.66]
MDEC_22	[6.13,9.86]

3.7 Model Checking

After obtaining the suggested selection interval of the variables and establishing the model based on the selected variables, 632 sets of selective samples are used for prediction, and the comparison between the results and the real values is shown in Figure 2.It can be clearly seen in the figure that the predictive model constructed by the selective screening variables is very close to the true value, and the negative influence part of the value is removed.

4. Discussion

Screening of potential active compounds has become an important step in drug design and development. In this process, it is necessary to maintain high biological activity and meet certain kinetic properties as much as possible. In this paper, based on the idea of machine learning, a reasonable feature dimension reduction and importance ranking of molecular descriptors are carried out by judging the weight of single variable and the degree of correlation between variables. Pearson correlation coefficient and VIF between variables were calculated to obtain significant variables, and at least three effective samples with excellent properties were obtained in ADMET through logical screening of positive samples. Then, the numerical samples were classified through data generalization. Finally, the significant association rules were obtained through Apriori algorithm, and then the recommended value categories of each variable were obtained by interpreting the association rules. Finally, the molecular descriptors and their recommended value ranges were obtained by materializing the suggested interval of variables.

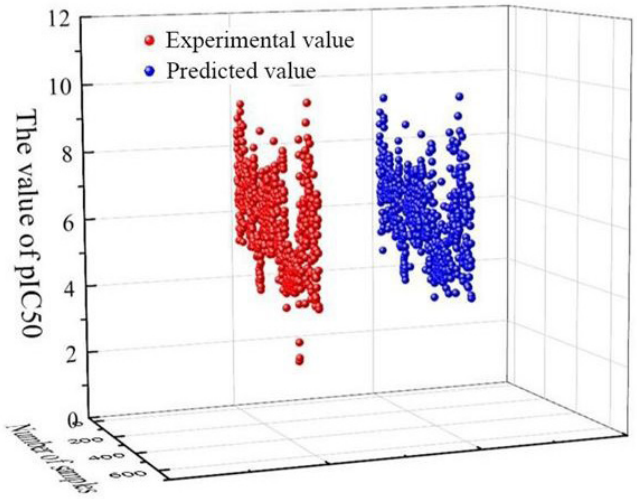

Figure 2. Scatter plot of model predicted value and true value

The quantitative prediction model of XGBoost compound bioactivity was established by using the selected variables. These results provide reliable suggestions for optimizing the biological activity of Erα inhibitors and predicting the properties of ADMET, as well as for accelerating the screening and development of drugs for breast cancer.

Conflict of interest

We have no conflict of interests to disclose and the manuscript has been read and approved by all named authors.

Acknowledgments

This work was supported by the Philosophical and Social Sciences Research Project of Hubei Education Department (19Y049), and the Staring Research Foundation for the Ph.D. of Hubei University of Technology (BSQD2019054), Hubei Province, China.

References

1. Hengwei Chen. The study of deep learning based multi-drug synergy prediction model [D]. Jiangsu University of Science and Technology, 2020, Jiangsu.
2. Yao Luo, Yanlin Song, Jinling Shang, et al. Prediction of PI3K inhibitors based on naive bayesian machine learning [J]. Chinese Journal of new Drugs, 2019, 28(1): 73-80.
3. Lei Chen, Biqing Li, Mingyue Zheng, et al.. Prediction of Effective Drug Combinations by Chemical Interaction [J]. Biomed Res Int, 2013, 2013(2): 1–10.
4. A. AvciE. A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier [J]. Expert Systems with Applications, 2009, 36(7): 10618-10626.
5. A. M. Shin, I. H. Lee, G. H. Lee, et al. Diagnostic analysis of patients with essential hypertension using association rule mining [J]. Healthc Inform Res, 2010, 16(2): 77-81
6. M. Karabatak, M. C. Ince. An expert system for detection of breast cancer based on association rules and neural network [J]. Expert Systems with Applications, 2009, 36(2): 3465-3469.
7. S. Palaniappan, R. Awang. Intelligent heart disease prediction system using data mining techniques [J]. IEEE/ACS International Conference on Computer Systems and Applications, 2008, 8(8): 343-350.
8. J. Xie, C. Wang. Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases [J]. Expert Systems with Applications, 2011, 38(5): 5809-5815.
9. R. Mitchell, E. Frank. Accelerating the XGBoost algorithm using GPU computing [J]. PeerJ Computer Science, 2017, 3(1): 127-157.
10. K. Song, F. Yan, T. Ding, et al. A steel property optimization model based on the XGBoost algorithm and improved PSO [J]. Computational Materials Science, 2020, 174:109472.
11. H. Jiang, Z. He, G. Ye, et al. Network Intrusion Detection Based on PSO-Xgboost Model [J]. IEEE Access, 2020, 8:58392-58401.
12. J. Lin, C. Q, H. Wan, et al. Prediction of Cross-Tension Strength of Self-Piercing Riveted Joints Using Finite Element Simulation and XGBoost Algorithm [J]. Chinese Journal of Mechanical Engineering, 2021, 34(1):36-47.

Appendix A

The main process for perdorming analysis

```
data<-read_excel("datalast.xlsx") dim(data)
data summary(data)
frequentsets=eclat(data,parameter=list(support=0.2,maxlen=10)) inspect(frequentsets[1:10]) rules=apriori(data,parameter=list(support=0.2,confidence=0.4,minlen=2)) summary(rules)
inspect(head(sort(rules,by="support"),30))
write.table(rules, file ="datalast.xlxs", sep=" ", row.names =TRUE, col.names =TRUE, quote =TRUE) png(file = "datalast.png",width =980,height=1180, units = "px", res=120)
a=plot(rules,method="grouped",cex=1.2)
print(a) dev.off();

x=subset(rules,subset=rhs%pin%"pIC50"&lift>=1);x A<-inspect(sort(x,by="support")[1:47])
write.table(A, file ="xrules.xlsx", sep=,, row.names =TRUE, col.names =TRUE, quote =TRUE) png(file = "xrules.png",width=960,height=680,units = "px", res=120)
x1<-plot(x,method="grouped",cex=1.2) print(x1)
dev.off();
```