Optimal modeling of anti breast cancer drug candidates

Jiaying Li, Guowei Zhang *, Hongxia Yang

School of Mining Liaoning Technical University Fuxin, Liaoning 123000

* Corresponding Author Email: 494280722@qq.com

Abstract. As breast cancer is one of the most common cancers with high mortality rate in the world, this paper studies the optimal screening of anti-breast cancer candidate drugs. First, the data of 729 compound molecular descriptors are preprocessed, rough cleaned and filtered to 253, and then the cluster feature tree and correlation analysis are used to further reduce the dimension of data redundancy information, and 29 representative molecular descriptors are screened. In order to determine the importance and significance variables affecting the activity of compounds, the preliminary results were obtained by Pearson correlation analysis, and the results were further verified by Spearman correlation analysis. The stability of molecular descriptors was determined by coefficient of variation analysis, and the most representative 20 variables were selected from 29 molecular descriptors by principal component analysis. Projection pursuit model (PP) can reduce the difficulty of storage and calculation of high-dimensional data. 253 variable information reflecting biological activity are extracted with this model, and then the projection direction is optimized by using accelerated genetic algorithm (RAGA). Finally, 20 projection function values in the optimal projection direction are calculated to complete the prediction of compound activity index.

Keywords: Clustering feature tree; Correlation analysis; Projection pursuit model; Accelerated genetic algorithm.

1. Introduction

Breast cancer is one of the most common and fatal cancers in the world. The development of breast cancer is closely related to estrogen receptor. Some studies have shown that ERα It plays a very important role in the development of breast. Therefore, ERα is considered as an important target for the treatment of breast cancer and can antagonize ERα active compounds may be candidate drugs for the treatment of breast cancer. In drug research and development, if a compound wants to become a candidate drug, it needs not only good biological activity, but also good pharmacokinetic properties and safety in human body.

In order to save time and cost, the method of establishing compound activity prediction model is usually used to screen potential active compounds. The quantitative structure activity relationship (QSAR) model of compounds is often constructed with the molecular structure descriptor as the independent variable and the bioactivity value of compounds as the dependent variable, and then the model is used to predict new compound molecules with better bioactivity or guide the structural optimization of existing active compounds. Of course, while paying attention to the bioactivity of compounds, The ADMET nature also needs to be considered.

2. Analysis of the importance and significant influence of molecular descriptor variables

According to the files "molecular descriptor." (MD) and ERα the data provided were preprocessed for 729 molecular descriptors of 1974 compounds. First, clean and delete the unavailable MD; Then, the remaining MD were preliminarily screened and some high-quality MD were screened out; Finally, the high-quality MD is screened twice to eliminate the MD with redundant information, and an effective MD sample is obtained. Further analyze its ERα of biological activity according to the final effective MD sample the importance ranking and the top 20 most significant MD are given.

2.1. Data preprocessing of molecular descriptor variables

In this paper, the variable pretreatment of molecular descriptor (MD) includes cleaning and variance filtering. The data in this paper include one-dimensional linear structure expressions of 1974 compounds in MD file and 729 MD. Generally speaking, MD cleaning is manual cleaning based on chemical intuition or characteristics observed in experiments [1]. However, manual cleaning relies heavily on professional knowledge and experience, which brings a high threshold for manual cleaning, and it is difficult to ensure the effectiveness and reliability of the results after manual cleaning and filtration, resulting in the distortion of the final effective MD samples. In this paper, LabVIEW software and manual are combined to perform cleaning. After cleaning, combined with Friedman rank analysis of variance [1], the number of 729 MD is reduced to 253.

2.1.1. Variance filtering and elimination based on Friedman

When there are differences between blocks in the processed sample data MD, the influence of blocks on the results must be considered. Generally, two factor analysis of variance (ANOVA) method can be used, but the analysis of variance requires that the test error is normally distributed. When the data does not meet the normal premise of analysis of variance, Friedman suggests using rank analysis of variance. Friedman test has no requirement for normal analysis of test error, and only depends on the rank observed in each block. Sometimes it is called two-way analysis of variance by rank. The basic principle of Friedman test [1] is as follows:

Assuming that there are k processes and b blocks, the data observation values, such as $X = \left(x_{ij}\right)_{bk}$, are the same as most of the test problems of analysis of variance. Here, the hypothesis test problems of location parameters are:

$$H_0 : \theta = \cdots = \theta_k \leftrightarrow H_i : \exists i, j \in 1, \ldots, k; \theta_i \neq \theta_j \tag{1}$$

Due to the influence of block groups, the ranks in different block groups are not comparable. However, if data is collected according to different blocks, the comparison between different processes in the same block is meaningful. Therefore, the rank of each process should be allocated in each block first, so as to obtain the completely random block rank data table R_{ij}.

If R_{ij} represents the rank of the j-th process in the i-th block group, the rank is summed according to the process as follows:

$$R_j = \sum_{i=1}^{b} R_{ij} \ (j = 1, \ldots, k); \bar{R} = R_j / b \tag{2}$$

When the null hypothesis holds, each average $\bar{R}_j$ has the following properties:

Under the assumption of zero:

$$E\left(\bar{R}_j\right) = \frac{k+1}{2}, \operatorname{var}\left(\bar{R}_j\right) = \frac{k^2-1}{12b}, \operatorname{cov}\left(\bar{R}_i, \bar{R}_j\right) = -\frac{k+1}{12b} \tag{3}$$

According to the above formula, the average sum of each treatment room is:

$$SSt = n\sum\left(\bar{R}_j - \bar{R}_{\cdots}\right)^2 = \sum \bar{R}_j^2 / b - \bar{R}_{\cdots}^2 / bk = \sum R_j^2 / b - bk(k+1)^2 / 4 \tag{4}$$

The correction formula Q is:

$$Q = \frac{k-1}{k} \frac{SST}{\operatorname{var}\left(R_{ij}\right)} = \frac{12}{bk(k+1)} \sum R_j^2 - 3b(k+1) \tag{5}$$

The Q value approximates the χ^2 distribution of the degree of freedom $v = k+1$.

When the data have the same rank, the correction of Q value is as follows:

$$Q_c = \frac{Q}{1 - \frac{\sum g\left(\tau^3 - \tau_i\right)}{bk\left(k^2-1\right)}} \tag{6}$$

Where τ_i is the length of the i-th knot; g is the number of knots.

Conclusion: if $Q < \chi^2_{0.05}; k-1, g$ are the number of knots.

In this paper, the molecular descriptor variable (MD) of compounds is regarded as different block groups. Without considering the block factors, this is a one-way ANOVA.

Table 1: Rank average of local molecular descriptors.

MD	Rank average	MD	Rank average
ALogp	82.94	CrippenMR	241.43
ALogp2	105.48	ECCEN	246.51
AMR	242.10	nHBd	105.22
apol	237.75	nHBa	148.84
.....	....	....	....
nBondsM	203.94	XLogP	125.87
bpol	219.85	Zagreb	243.39
C1SP2	64.11	V255	218.10

Next, call Friedman function for nonparametric Friedman test to analyze whether there are differences between different MD, as shown in table 2. Here, need to eliminate some MD with zero variance and abnormal average rank. The remaining 253 MD return the test p value according to Friedman function: P<0.001, which means that the original hypothesis is rejected at the significance level of 0.05. It can be seen that the remaining different MD have statistical differences. The specific MD have significant differences need to be further analyzed.

2.1.2. Preliminary screening of molecular descriptor variables

The correlation of 253 MD after data preprocessing is complex. In order to further screen and distinguish MD with high correlation from those with low correlation, we need to build a clustering feature tree birch-Balanced Iterative Reduction and Clustering using Hierarchies-to preliminarily screen and classify MD, and then take the results of clustering algorithm as input to further optimize and select high-quality MD according to their species relationship.

Step 1: normalization of variables

Due to the different values of each molecular descriptor (MD), some correlations between the properties of molecules and one-dimensional linear structure expressions are revealed. It is observed that some MD values are less than 10, while some MD values are greater than 100. Due to the large differences in different MD value ranges, which is not conducive to subsequent progress, we use the min-max normalization method to map all MD data between 0 and 1 [1]. The normalization formula of min-max is as follows:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{7}$$

Where X' is the normalized MD, X is the original MD, $X_{\max}$ is the maximum value of the original MD, and $X_{\min}$ is the minimum value of the original MD. Here, the result after unified normalization is six decimal places. The correlation of the normalized 253 MD is complex. In order to further screen and distinguish the MD with high correlation from that with low correlation, and then further optimize and select the high-quality MD according to the species relationship according to the results of the clustering algorithm. The details are as follows:

Step 2: construct the redundant information of dimension reduction variables of clustering feature tree

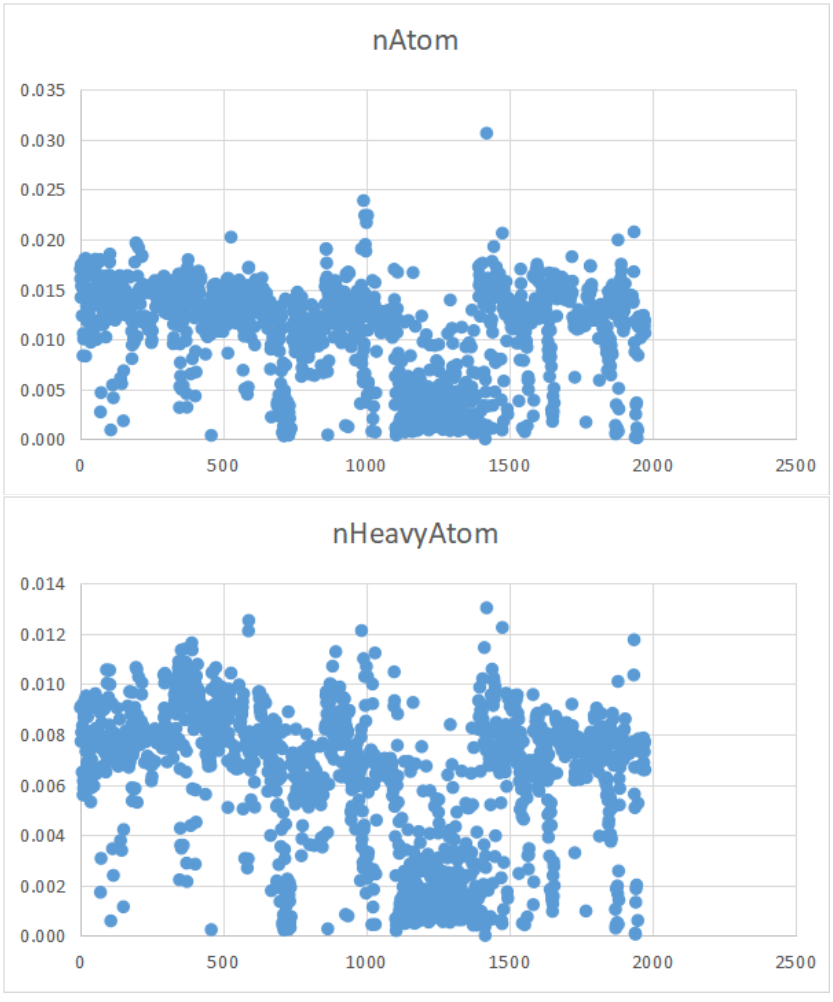
Constructing clustering feature tree is a key part of balancing iterative reduction and clustering use hierarchy (Birch) [1]. The input data is compressed and represented by clustering feature tree. When new data is generated, the clustering feature tree is reconstructed and the compression is enhanced. The leaves in the clustering feature tree are used for clustering.

Further screen the MD excluding the constant vector, distinguish the MD with high correlation from those with low correlation, and then further optimize and select the high-quality MD according to the detailed species relationship of the summary file of "molecular descriptor interpretation" according to the results of the clustering algorithm, as follows:

Table 2: Clustering characteristic tree variable analysis results.

Clustering feature tree	Clustering feature tree analysis results
Maximum tree depth	3
Minimum number of cases in parent node	100
Included arguments	SP-6, WPATH, fragC, ECCEN...
Number of nodes	31
Number of terminal nodes	20
depth	3
Risk estimation	0.047
Risk standard error	0.003
Growth method	CHAID
dependent variable	IC

After data normalization, two potentially intractable problems need to be solved in MD. First, the high relevance of MD leads to the information redundancy contained, which further hinders the screening of high-quality MD. Second, MD close to the constant vector. In order to solve this problem, it is necessary to further screen the MD excluding the constant vector, distinguish the MD with high correlation from those with low correlation, and then further optimize and select the high-quality MD according to the detailed species relationship of the summary of "molecular descriptor interpretation" according to the results of the clustering algorithm.



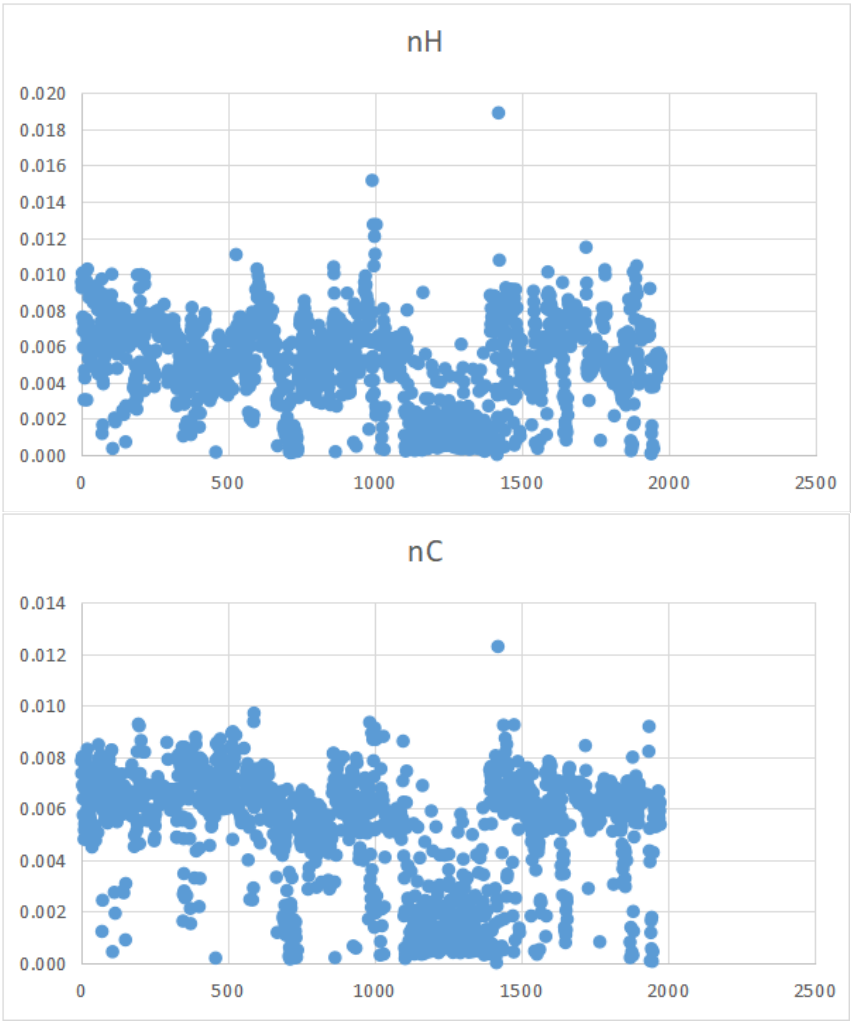


Figure 1: Frequency distribution of AtomCountDescriptor.

According to the detailed species relationship of the summary file, it is not difficult to find that the nAtom, nHavyAtom, nH and nC of the molecular descriptor (MD) in Figure 1 belong to the AtomCountDescriptor in the Descriptor Java Class, but the frequency distribution diagram shown is quite different. Among them, the frequency distribution of nAtom and nH is more concentrated in the constant vector [1], nAtom and nH are almost 17% and 19% respectively, and the ratio of nHavyAtom and nC close to constant vector (CV) is far less than 15%. The selection of all sub samples will distort the final results and have errors.

According to relevant studies, the development of breast cancer is closely related to Estrogen Receptor alpha, (ERα) It is expressed in no more than 10% of normal breast epithelial cells, but about 50% - 80% of breast tumor cells; And for ERα The experimental results of gene deletion mice showed that ERα indeed, it plays a very important role in the process of breast development, so we prefer to choose the offspring with nearly constant vector less than 15%, that is, choose the more excellent offspring nHavyAtom and nC. Next, select high-quality MD representatives based on the CV frequency distribution map with MD.

Step3: elimination of variables close to constant vector (CV)

After data normalization, two potentially intractable problems need to be solved in the selection of MD.

First, the high relevance of MD leads to the redundancy of information contained, which further hinders the screening of high-quality MD.

Second, MD close to the constant vector cannot play an effective role in screening high-quality MD modeling, so it needs to be further eliminated directly, as shown in Figure 2 and figure 3. Here, in order to adjust the normalized digits to three digits after the decimal point

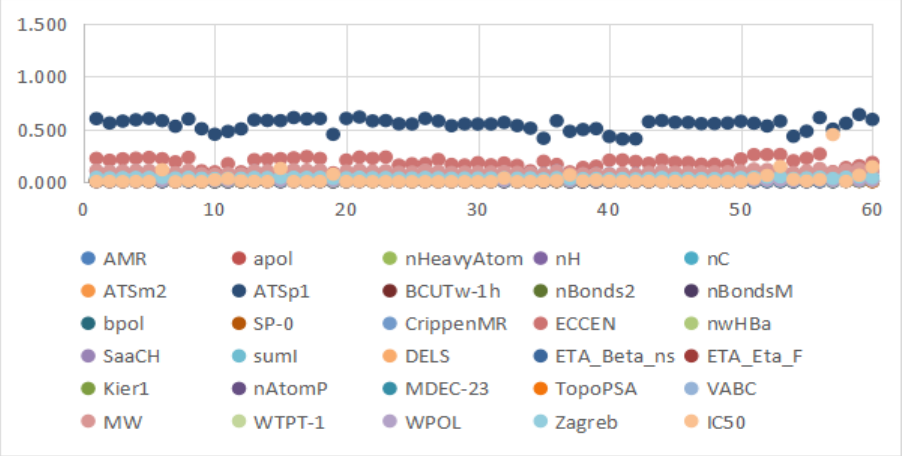


Figure 2. Partial schematic diagram of MD close to constant vector.

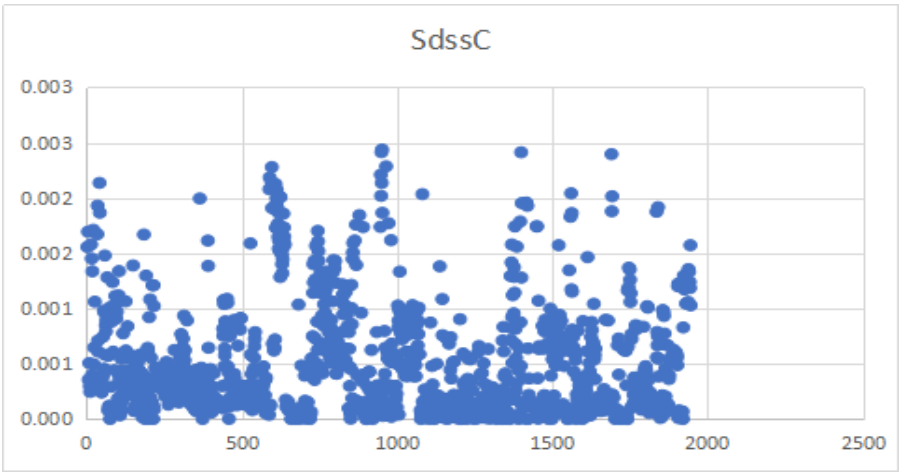


Figure 3. Individual diagram of MD close to a constant vector.

After the above setp1 and setp2 steps, the number of satisfied molecular descriptor variable (MD) individuals is reduced from 253 to 49, as shown in Table 4. After the setp3 step, the satisfied MD are reduced from 51 to 29, as shown in Table 5 and figure 3.

Table 3. List of high quality variables after setp1 and setp2.

AMR	apol	nAtom	nHeavyAtom	nH	nC	ATSm1
ATSm2	ATSm3	ATSm4	ATSm5	ATSp1	ATSp2	BCUTw-1h
nBonds	nBonds2	nBondsS	nBondsS2	nBondsM	bpol	SP-0
CrippenMR	ECCEN	nwHBa	nHaaCH	nHother	SHBa	SwHBa
SaaCH	maxHBa	sumI	gmax	DELS	DELS2	ETA_Beta
ETA_Beta_	TA_Beta_	ETA_Eta	ETA_Eta_R	ETA_Eta_F	Kier1	nAtomP
MDEC-23	TopoPSA	VABC	MW	WTPT-1	WTPT-3	WPOL

Table 4. List of high quality variables after setp3.

AMR	apol	nHeavyAtom	nH	nC
ATSm2	ATSp1	BCUTw-1h	nBonds2	nBondsM
bpol	SP-0	CrippenMR	ECCEN	nwHBa
SaaCH	sumI	DELS	ETA_Beta_ns	ETA_Eta_F
Kier1	nAtomP	MDEC-23	TopoPSA	VABC
MW	WTPT-1	WPOL	Zagreb	

2.1.3. Secondary screening of variables based on correlation analysis

Firstly, the correlation between molecular descriptor variable (MD) and Estrogen Receptor alpha, (ERα) was analyzed The importance of influence variables on ERα was obtained; Reuse 29 MD and ERα indicators pIC_{50} was verified because of ERα Indicator IC_{50}_nM with negative logarithmic relationship (the IC_{50} value can be converted to pIC_{50}). In order to prove the effectiveness of significance, Spearman correlation was further analyzed, and finally the top 20 molecular descriptors with the most significant impact on biological activity were given.

Step 1: correlation analysis

In the process of data analysis, it is often necessary to analyze the causal relationship between two or more variables. Pearson correlation analysis is usually used, which does not need to distinguish between independent variables and dependent variables. There is an equal relationship between two or more variables. 29 molecular descriptor variables (MD) and Estrogen Receptor alpha, (ERα) can be understood through correlation analysis

Step 2: in depth analysis Spearman correlation analysis

In statistics, Spearman Rank Correlation Coefficient is a nonparametric index to measure the dependence of two variables. It uses monotone equation to evaluate the correlation of two statistical variables. If there are no duplicate values in the data and the two variables are completely monotonically correlated, the Spearman correlation coefficient is + 1 or − 1.

For samples with sample size n, n original data are converted into hierarchical data, and the correlation coefficient ρ is

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{8}$$

In practical application, the connection between variables is irrelevant, so ρ can be calculated through simple steps the difference between the grades of the two observed variables, then ρ is:

$$\rho = 1 - \frac{6\sum_i d_i^2}{n\left(n^2 - 1\right)} \tag{9}$$

Generally, the Spearman correlation coefficient is useful in the case of three or more conditions. One way to determine whether the ρ value of the observed data is significantly non-zero (r always has $1 \geq r \geq -1$) is to calculate the probability that it is greater than R as the original hypothesis and test it with the hierarchical arrangement test. The advantage of this method is that it considers the number of data in the sample and the risk of using the sample to calculate the rank correlation coefficient.

Step3: coefficient of variation analysis

The coefficient of variation reflects the stability of the content of this component in candidate drugs [2]. The smaller the coefficient of variation, the more stable and representative its content in drugs Figure. 4 further shows that although the contents of ATSp1 and nBonds2 vary greatly in different compounds. However, the final stability reflects the result of the comprehensive action between the components, which ensures the relative stability of the molecular descriptor through the balance and coordination between the components.

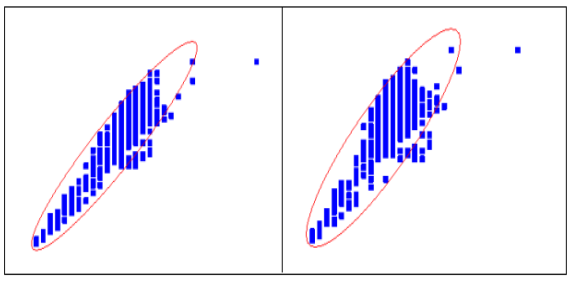


Figure 4. Variation of ATSp1 and nBonds2 variables in different compounds.

Step4: principal component analysis

Principal component analysis (PCA) is to recombine multiple indicators into a new set of unrelated comprehensive indicators, and select as few comprehensive indicators as possible according to the actual needs. In order to achieve the analysis method of reflecting the original index information as much as possible. After the principal component analysis of the indicators of IC_{50}, it is found that the 29 molecular descriptors screened are highly concentrated, as shown in Figure. 5, which once again shows that they are sorted according to the importance of variables on biological activity, and the error of the molecular descriptors that have the most significant impact on biological activity is within the controllable range.

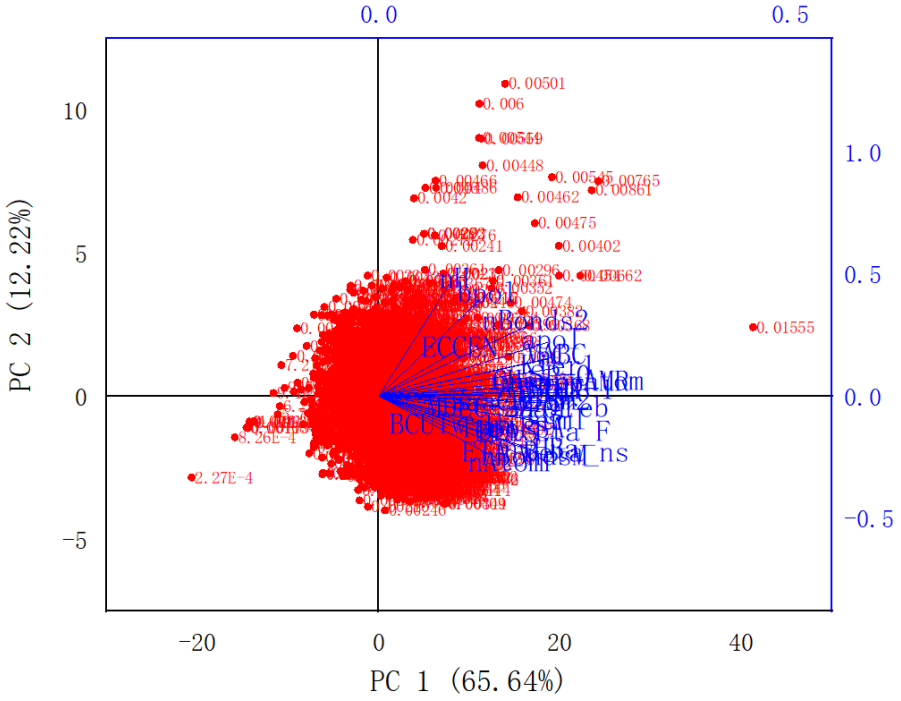


Figure 5. Pincipal component analysis of 29 variables of IC_{50}.

3. Prediction model of compound bioactivity based on pp-raga

Firstly, the projection pursuit *(PP)* model is used to extract 20 indexes that can reflect the main characteristic variables, and then the accelerated genetic algorithm [1] (RAGA) is used to optimize the projection direction, calculating the projection function value in the optimal projection direction, and complete the prediction of compound activity indexes in test.

3.1. basic principle and method of projection pursuit (PP)

Projection pursuit model (*PP* model for short) [5] [6] is a method for dimensionality reduction analysis of high-dimensional data. It combines the local cohesion degree with the overall dispersion degree to form a new index for clustering and various decision-making analysis. Projection pursuit model can overcome the difficulty of data analysis caused by the "Curse of dimensionality" of high-dimensional data. It can project high-dimensional data into low-dimensional space, and then further analyze the projected low-dimensional data to find the optimal projection direction.

Due to the basic principle and method of *PP* method, it's modelin as follow:

Step 1: normalization of decision index setg is divided into the following steps:

Let the decision index sample set be $X=\left(x_{ij}\right)_{n\times p}\ i=1,2,..n;j=1,2,..,p$, where Z_{ij} is the j-th index value of the i-th sample, and n and p are the number of samples and indicators respectively. In order to eliminate the dimension of each index and unify the variation range of each value of each index, the following formula is used for normalization.

For the bigger and better indicators:

$$y(i,j)=\frac{x(i,j)-x_{min}(j)}{x_{max}(j)-x_{min}(j)} \tag{10}$$

For the smaller and better indicators:

$$y(i,j)=\frac{x_{min}(j)-x(i.j)}{x_{max}(j)-x_{min}(j)} \tag{11}$$

Where, $x_{\max}(j)$ and $x_{\min}(j)$ are the maximum and minimum values of the j-th index value respectively, and $y(i,j)$ is the normalized sequence of index eigenvalues.

Step 2: construct projection index function $Q(a)$

PP method is to synthesize p-dimensional data $\{y(i,j)|j=1,2,\ldots p\}$ into one-dimensional projection value $Z(i)$ with $a=\{a(1),a(2),\ldots,a(p)\}$ as projection direction:

$$Z(i)=\sum_{j=1}^{p}a(j)y(i,j)\ (i=1,2,\ldots n) \tag{12}$$

Then the classification and decision are made according to the one-dimensional scatter diagram of $\{Z(i)|i=1,2,\ldots n\}$. Where a is the unit length vector? When synthesizing the projection index value, it is required that the dispersion feature of the projection value $Z(i)$ is that the local projection points are as dense as possible, preferably condensed into several point clusters, and as a whole, the projection point clusters are scattered as far as possible. Therefore, the projection scaling function can be expressed as:

$$Q(a)=S_z*D_z \tag{13}$$

In the above formula, S_z is the standard deviation of the projection value $Z(i)$, and D_z is the local density of the projection value $Z(i)$.

$$D_z=\sum_{i=1}^{n}\sum_{j=1}^{n}(R-r(i,j))*u(R-r(i,j))S_z=\sqrt{\frac{\sum_{i=1}^{n}(Z(i)-E(z))^2}{n-1}} \tag{14}$$

In the above formula: $E(z)$ is the average value of sequence $\{Z(i)|i=1,2,\ldots n\}$; R is the window radius of local density. Its selection should not only make the average number of projection points contained in the window not too small, avoid too large sliding average deviation, but also not make it increase too high with the increase of samples. R can be determined according to the test, and generally can be taken as $0.1*S_z;r(i,j)$; Represents the distance between samples $r(i,j)=|Z(i)-Z(j)|$; the function $u(t)$ is a unit step function. When $t\geq 0$, its value is 1; when $t<0$, the function value is 0.

Step 3: optimize projection index function

When the sample set of each index value is given, the projection index function $Q(a)$ [7-9] changes only with the change of projection direction a. Different projection directions reflect different data structure characteristics. The best projection direction is the projection direction that is most likely to expose a certain type of feature structure of high-dimensional data. Therefore, the best projection direction can be determined by solving the problem of maximizing the projection index function that is, maximizing the objective function:

$$\max: Q(a)=S_z*D_z \tag{15}$$

The constraints are:

$$s.t = \sum_{j=1}^{p} a^2(j) = 1 \tag{16}$$

This is a complex nonlinear optimization problem with $\{a(j) | j = 1, 2, \ldots, p\}$ as the optimization variable, which is difficult to deal with by traditional optimization methods. Therefore, the real coded accelerated genetic algorithm (RAGA), which simulates the survival of the fittest and the exchange mechanism of chromosome information within the population, is used to solve its high-dimensional global optimization problem, so as to obtain the best projection direction.

Step 4: classification and prioritization

The best projection values of the three points can be obtained by $Z(i) = \sum_{j=1}^{p} a(j) y(i, j)$ substituting the projection directions of the samples. By sorting the values from large to small, the effect of molecular descriptor (MD) on Degree of impact of biological activity ERα can be judged. The coding method of standard genetic algorithm (SGA) often adopts binary coding, and its individual genotype is a binary coding symbol string. Although the binary code is simple, and the genetic operation procedures such as crossover and mutation are easy to realize and use patterns to make in-depth theoretical analysis of the algorithm, it is not easy to reflect the structural characteristics of the problem, especially for the optimization of some continuous functions. Because of its randomness, it will lead to poor local search ability, and for some continuous function optimization problems with high dimension and high precision, there will be many disadvantages when using binary coding to represent individuals. Moreover, the search and optimization function of selection operator operation and crossover operator operation gradually weakens with the increase of the number of evolutionary iterations. In practical application, it often appears that the search and optimization work is far from the global optimum. Therefore, using the change interval of excellent individuals generated by evolutionary iteration as the new initial change interval of variables, rerun the standard genetic algorithm to form accelerated operation, then the excellent individual interval will gradually shrink and become closer to the optimal individual, until the optimization criterion function value of the optimal individual is less than a certain set value or the algorithm operation reaches the predetermined acceleration times, and the whole operation operation will be ended; At this time, the best individual in the current population is designated as the result of the accelerated genetic algorithm. This method is called real coding based on accelerating genetic algorithm (RAGA).

The best projection value of molecular descriptor can be calculated according to the best projection direction. The projection value is shown in the following figure:

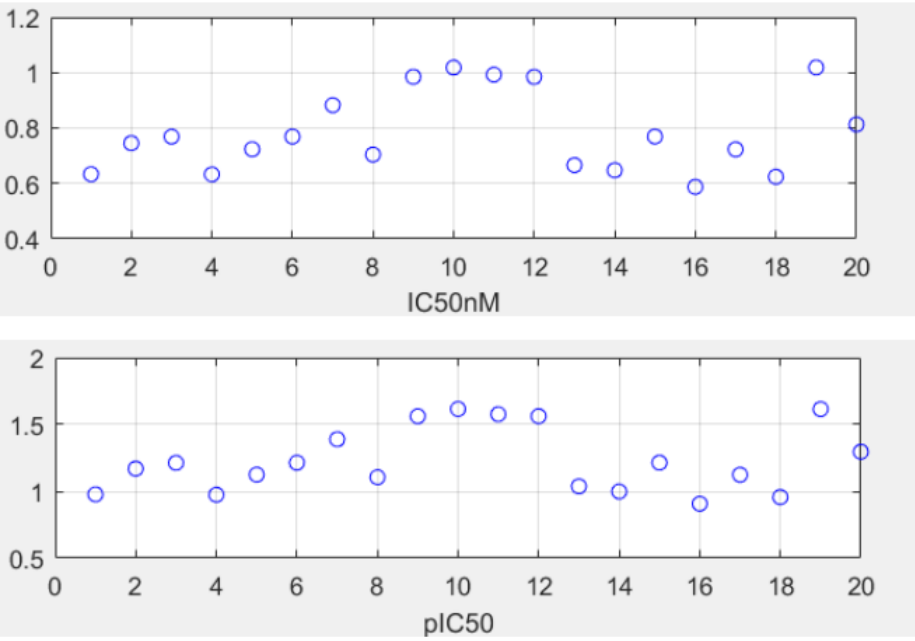


Figure 6. Optimal projection values of IC5 and pIC_{50}

4. Conclusion

In the data preprocessing stage, this paper cleans and deletes invalid variables; then, the remaining variables are further screened to eliminate the invalid variables with redundant information, so as to obtain an effective variable sample. Lay a solid foundation for the establishment and prediction of the following models; When the data does not meet the normal distribution, Friedman rank analysis of variance is used to improve the further dimensionality reduction effect of high-dimensional data; Give full play to the good scalability of clustering feature tree, effectively compress the data through clustering feature tree, and reduce the difficulty of high-dimensional data processing; RAGA is an improvement of standard genetic algorithm, which has nothing to do with the setting of initial value, can converge in a short time, and is not easy to fall into local optimization; In the quantitative prediction of compound biological activity, PP and RAGA model are combined to reduce the processing problem of high-dimensional data, accelerate the convergence speed, fast and efficient, and provide new ideas and methods for the prediction and evaluation of compound activity safety.

References

[1] JINGSHAN L, DEHAN L, TENGTENG W, et al. Represeantative feature selection of molcular descriptors in QSAR modeling [J]. Jounal of Molecular Structure, 2021, 1244.

[2] Gu Rongyan- "A new analysis method of personnel evaluation results -- Friedman rank two-way ANOVA and Kendall consistency test [J]. Shanghai Education and scientific research, 19906, 05): 35-7

[3] Wu Taosheng, Zhang Xinqin, Zhou Tao, et al Genetic diversity analysis of phenotypic characters of cultivated white rice and Germplasm Resources in Guizhou [J] Molecular plant breeding, 1-23

[4] Chen Shuo, Li Feifan, sun Guohui, et al. QSAR modeling and its research progress in antiviral drug design and screening [J] Chemical reagent, 2021, 43 (07): 895-905

[5] Feng Tugen, Liu Hanlong, GAO Yufeng, Yang jiangui. Application of accelerated genetic algorithm in slope seismic stability analysis [J] Journal of water conservancy, 2002 (9): 89-94

[6] Jin Juliang, Yang Xiaohua, Ding Jing, improvement scheme of standard genetic algorithm -- accelerated genetic algorithm [J] system engineering theory and practice, 2001 (04): 9-13

[7] Zhang Yuanyuan, Zhang Yu, Wei Huabo Simulation Research on fault detection of urban traffic intersection, small [J] computer simulation, 2012, 29 (10): 323-6

[8] Li Yun, Li Jiming, Jiang Zhongjun, application of statistical analysis in wine quality evaluation [J] Brewing technology: 1001 - 9262009) 04-0079-04

[9] Feng Jingchun, Chen Limin, Hu zhaoshu Study on comprehensive scoring method and its model for bid evaluation of water conservancy projects [J] Journal of Hehai University (NATURAL SCIENCE EDITION), 2003,31 (4): 461-465

[10] Li Xiao, Li Da, Zhou Xuesong, et al. Construction of compound ADMET property prediction platform [J] Bioinformatics, 2017, 15 (03): 179-85

[11] Gu Yaowen, Zhang Bowen, Zheng Si, et al Construction method of drug ADMET classification prediction model based on graph attention network [J] Data analysis and knowledge discovery 2021, 5 (08);: 76-85.

[12] Zhou Zhihua. Machine learning [J] China Civil and commercial, 2016, 03): 93

[13] Zhang Mingyue. Analysis on the confidence interval of estimating the expected value of any function by Montero simulation [J] Times Finance 2018, 06): 214+22.