ISSN 2706-6819 Vol.4, Issue 1: 45-49, DOI: 10.25236/IJFM.2022.040108

Factor Research Based on Multi-index Gray Correlation Analysis Correlation Analysis

Bo Tian*, Lin Hu, Ruo Jia

College of Communications Engineering, Army Engineering University, Nanjing, 210007, China
**Corresponding author*

Abstract: *Breast cancer is one of the most common and fatal cancers in the world. Therefore, this paper uses data mining and prediction technology, which is of great significance. In this paper, firstly, the relevant data were collected, and the Pearson correlation coefficients of 729 molecular descriptors in 1974 compounds were calculated respectively, and the correlation coefficient distribution map was obtained. Through observation, all the molecular descriptors were 0 elements. Then the grey correlation analysis method was used to analyze the correlation degree, and the grey correlation value between the information of 729 molecular descriptors and the bioactivity value of ER α was obtained. Then, making use of the advantage of canonical correlation analysis in feature extraction, according to the feature selection of linear combination coefficient, the molecular descriptors with the most significant effect on biological activity were selected.*

Keywords: *Breast cancer, Canonical correlation analysis, Gray correlation analysis*

1. Introduction

Breast cancer is a linear phenomenon in which breast epithelial cells increase out of control under the action of a variety of carcinogenic factors, and the incidence of breast cancer ranks first among female malignant tumors. Drug companies will consider the time and cost in the process of drug research and development, and usually use the method of establishing a compound activity prediction model to select potentially active compounds. The candidate drugs for the treatment of breast cancer must not only have good biological activity, but also meet the kinetic properties and safety requirements of the human body.

2. Grey correlation analysis GRA

First of all, according to the distribution of Pearson correlation coefficient, we get:

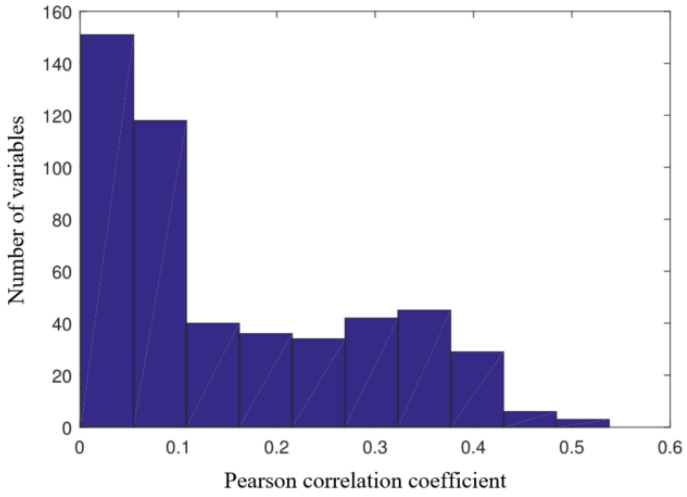

Figure 1: Pearson correlation coefficient distribution diagram of 729 variables and pIC50 column

GRA reflects the correlation degree of the curve and is a quantitative analysis of the dynamic development process. For the comparison series with greater correlation degree with the reference series, the development direction and rate are close to those of the reference series.

ISSN 2706-6819 Vol.4, Issue 1: 45-49, DOI: 10.25236/IJFM.2022.040108

Determine the analytical sequence and take the bioactivity of the compound to ERa as the reference sequence.

$$Y = Y(k) |\ k = 1, 2 \ldots, 1974 \tag{1}$$

The 729 molecular descriptor information of the compound is used as a comparative sequence.

$$X_i = X_i(k) |\ k = 1, 2 \ldots, 1974, i = 1, 2 \ldots, 1974 \tag{2}$$

It is found that the characteristics of some molecular descriptor coefficients of the data set are obvious, and the data set is filtered before data processing. Therefore, this paper uses the method of data standardization to calculate the correlation coefficient.

$$\xi_i(k) = \frac{\min_i \min_k |\ y(k) - x_i(k)| + \rho \max_i \max_k |\ y(k) - x_i(k)|}{|\ y(k) - x_i(k)| + \rho \max_i \max_k |\ y(k) - x_i(k)|} \tag{3}$$

The calculation of the correlation degree coefficient is to compare the correlation degree value of the series and the reference series in each data, and the average value has been used as the index of the correlation degree between the series. Then sorted by size according to the ranking correlation degree, the j molecule describes the effect of the information on the biological activity value.

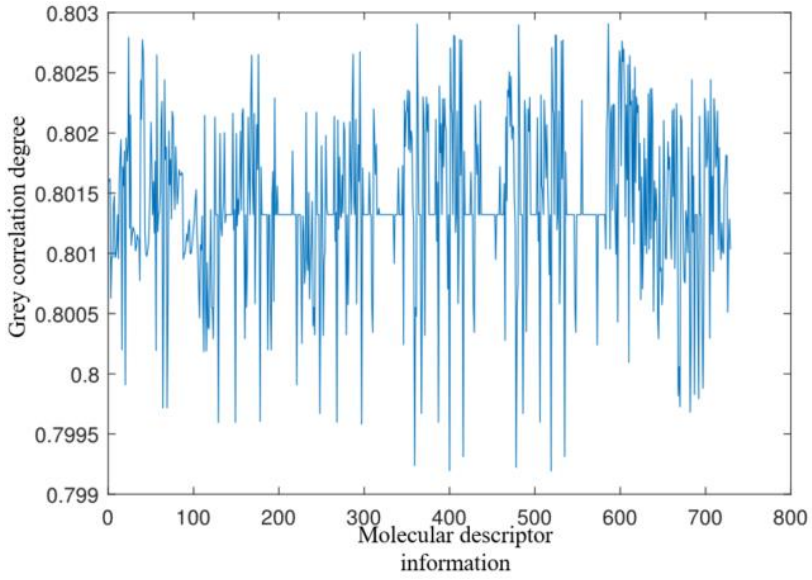

Figure 2: Grey correlation analysis

3. Canonical correlation analysis CCA

In order to study the relationship between the two groups of variables, using a method similar to principal component analysis, several representative variables are selected to form a representative comprehensive index, through the study of the correlation between the two groups of comprehensive index scaffolds. to replace the correlation between the two groups of variables.

The algorithm is extended to the matrix X, that is, it is linearly represented, and the corresponding projection vector is B. for the homology matrix Y, the projected data is

$$X' = a^T X, Y' = b^T Y \tag{4}$$

The projection vector corresponding to a, b

$$\arg_{a,b} \max \rho(X', Y') = \arg_{a,b} \max \frac{\operatorname{cov}(x', Y')}{\sqrt{DX}\sqrt{DY}} \tag{5}$$

The mean value is 0 and the variance is 1.

ISSN 2706-6819 Vol.4, Issue 1: 45-49, DOI: 10.25236/IJFM.2022.040108

$$\begin{aligned} \operatorname{cov}\left(X', Y'\right) &= \operatorname{cov}\left(a^T X, b^T Y\right) = a^T E\left(XY^T\right) b \\ DX' &= D\left(a^T X\right) = a^T DXa = a^T E\left(XX^T\right) a \\ DY' &= D\left(a^T Y\right) = a^T DYa = a^T E\left(YY^T\right) a \end{aligned} \tag{6}$$

Let

$$S_{XY} = \operatorname{cov}\left(X, Y\right) \tag{7}$$

The problem is changed into:

$$\arg_{a,b} \max \rho\left(X', Y'\right) = \arg_{a,b} \max \frac{a^T S_{XY} b}{\sqrt{a^T S_{XXa}} \sqrt{b^T S_{YY} b}} \tag{8}$$

Because the numerator denominator increases by the same multiple and the optimization goal remains the same, the optimization prevention similar to SVM can be used to fix the denominator and optimize the numerator, which can be transformed into:

$$\arg_{a,b} \max \rho\left(X', Y'\right) = \arg_{a,b} \max a^T S_{XY} b \tag{9}$$

According to the Lagrange multiplier method, there are

$$J\left(a, b\right) = a^T S_{XY} b - \lambda_0 \left(a^T S_{XX} a - 1\right) - \lambda_1 \left(b^T S_{YY} b - 1\right) \tag{10}$$

Then we can get:

$$\begin{aligned} S_{XX}^{-1} S_{XY} b &= \lambda a \\ S_{YY}^{-1} S_{YX} a &= \lambda b \end{aligned} \tag{11}$$

$$S_{XX}^{-1} S_{XY} S_{YY}^{-1} S_{YX} a = \lambda^2 a \tag{12}$$

Then the maximum value of the trivial root of the eigenvalue can be obtained first.

$$S_{YY}^{-1} S_{YX} S_{XX}^{-1} S_{XY} b = \lambda^2 b \tag{13}$$

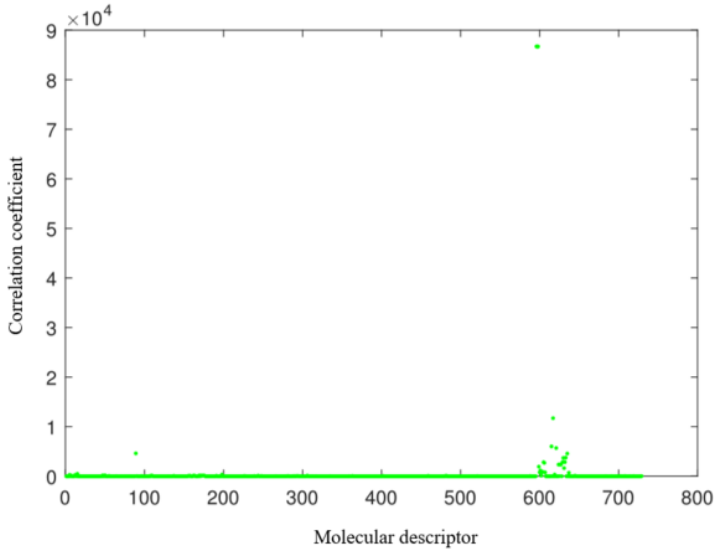

Figure 3: Scatter diagram of correlation coefficients of molecular descriptors

In this paper, the contribution of most of the molecular descriptions to the biology or properties of ERa is reduced, which means that the data are redundant, from which 24 variables with coefficients greater than 500 are preliminarily selected.

ISSN 2706-6819 Vol.4, Issue 1: 45-49, DOI: 10.25236/IJFM.2022.040108

Table 1: 30 high correlation numerical sorting

Serial number	Correlation coefficient	Serial number	Correlation coefficient	Serial number	Correlation coefficient
1	86722.03	11	2852.485	21	887.8519
2	86705.38	12	2848.877	22	788.472
3	86687.06	13	2832.692	23	777.791
4	11718.72	14	2706.153	24	768.0594
5	6021.901	15	2361.605	25	704.9338
6	5696.718	16	2361.559	26	480.5917
7	4615.79	17	2361.31	27	332.8687
8	4570.458	18	1940.872	28	426.0841
9	2740.824	19	1647.101	29	265.9657
10	3687.571	20	1177.872	30	248.3762

Table 2: The top 20 molecular descriptors of significant impact

Molecular descriptor	Correlation coefficient	Molecular descriptor	Correlation coefficient
ETA_EtaP_B	116241.9	nBondsS2	4270.515
ETA_EtaP_L	116222.5	ETA_EtaP_R	4157.058
ETA_dEpsilon_A	116217	nsssN	4105.062
ETA_dAlpha_B	31595.63	nS	3017.068
ndO	23939.34	nC	2881.689
nAromBond	7656.09	ndNH	29.56451
SP-1	7655.752	ETA_AlphaP	27.85037
nsCl	7168.504	ETA_Epsilon_5	12.26559
nssO	4313.893	ETA_Epsilon_1	2.984372
naasN	4310.93	ETA_aEpsilon_1	0.060492

4. Model analysis and evaluation

In this paper, 729 molecular descriptors are screened based on correlation analysis, and two kinds of correlation analysis are used to deal with them successively. There is no significant difference in the correlation between each molecular description and the biological activity of ERa in the results of grey correlation analysis. Screening directly according to the sort size will cause a large error.

From the above results, we can see that the molecular descriptor coefficients are quite different, and the ergodic search carried out by the secondary screening exhaustive method gives priority to the screening quality, and the exhaustive method is adopted.

5. Conclusion

This paper focuses on the process of cancer and anticancer, mainly on the screening of molecular descriptors. In this paper, the Pearson correlation coefficients of more than 1900 compounds are calculated, then the gray correlation algorithm is used to analyze and study, and the feature selection is obtained by using the advantage of canonical correlation analysis in eigenvalues.

References

[1] Wen Qi. The grey relational analysis is used to analyze the reasons after the economy of Guyuan [A]. Geographical Society to build a well-off Society in an all-round way-- Abstract of the Ninth Symposium on Geographic work of Chinese Youth [C], 2003.

[2] Wu Hongyan, Donna, Gui Chuanfeng. Prediction of Stock Price Index based on Grey Relational-BP Network Model [A]. Proceedings of the 9th China Youth Conference of Information and Management Scholars [C], 2007.

[3] Li Jian-Geng, Gao Zhikun. Random forest sets the class weight for small sample data [J]. Computer Engineering and Application, 2009, Magazine 45 (26): 131-134.

[4] Master Dong, Huang philosophy. Analysis of Stochastic Forest Theory [J]. Integrated Technology,

ISSN 2706-6819 Vol.4, Issue 1: 45-49, DOI: 10.25236/IJFM.2022.040108

issue 1, 2013.
[5] Peng Jiawen, Peng Jiahong. Research on a new algorithm for incremental updating multi-level association rules [J]. Microelectronics and computer, 2007 (05): 1-3.
[6] Tang Liang, du Junping. Research on Association Rule Mining in Tourism Emergency Prediction [J]. Journal of Beijing Industrial and Commercial University (Natural Science Edition), 2008 (01): 59-62.