International Journal of Applied Mathematics in Control Engineering 6 (2023) 66-73

Contents lists available at YXpublications

International Journal of Applied Mathematics in Control Engineering

Journal homepage: http://www.ijamce.com

A Predictive Model for The Safety of Compounds That Anti-Breast Cancer on The Human Heart Based on Machine Learning

Cunqing Rong[a], Ke Jin[a] , Jincai Chang[a,b*]

[a] *College of Science, North China University of Science and Technology, Hebei Tangshan 063200, China*

[b] *Hebei Key Laboratory of Science and Application, Hebei Tangshan 063210, China*

ARTICLE INFO

Article history:
Received 10 February 2023
Accepted 20 March 2023
Available online 25 March 2023

Keywords:
anti-breast cancer candidate drugs
variance analysis
XGBoost
genetic algorithm
SVM

ABSTRACT

With the increasing number of breast cancer patients, the optimization of candidate anti-breast cancer drugs has become particularly important. Only those with better antagonizing ERα activity and less physical harm to patients can be identified as candidate drugs for the treatment of breast cancer. In this paper, variance analysis and importance analysis are first performed on the compound's data features, and the dimensionality is reduced by PCA. The SVM classification algorithm was used to predict the cardiac safety of compounds. The correct rate of cross-validation is 83 %. This model has guiding significance for the development of anti-breast cancer drugs in the future.

1. Background

Breast cancer is one of the major malignant tumors that seriously endanger women's health. The incidence rate of women in the world is 24.2%, and 52.9% of them occur in developing countries. Treatment options vary depending on the type, size, stage, and distribution of the tumor. Treatments can include chemotherapy, radiation, surgery, and targeted therapy. It is important to get regular screenings to increase the chances of early detection.

Estrogen receptors (ER) are a member of the steroid receptor superfamily, and there are two main types: ERα and ERβ. It is a nuclear transcription factor that can regulate gene expression. Estrogen binds to estrogen receptors on the cell surface to form estrogen receptor complexes to exert corresponding biological functions. For breast cancer patients with ER α expression, anti-hormone therapy is used to regulate estrogen receptor activity. Therefore, ERα is an important indicator for the treatment of breast cancer. A drug candidate for the treatment of breast cancer refers to a compound capable of antagonizing the activity of ERα. A compound that can be a candidate drug must have biological activity on ERα, and the biological activity of the compound on ERα is related to some structures and properties of the compound.

ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity) properties of compounds refer to how a compound interacts in the human body after it has been administered. The effects of ADMET on the human body are vast and varied based on the specific compound. It is important to understand how a compound will interact within the body in order to more accurately assess the safety of said compound The ADMET properties of the compound are related to the pharmacokinetics and safety of the drug. Therefore, while considering the activity of the compound, the impact of the compound on the human body should also be considered.

hERG stands for Human Ether-aGo-Go Related Gene. It is a member of the potassium channel family responsible for regulating heart rhythm. Mutations of this gene lead to Long QT Syndrome, a cardiac arrhythmia, increasing the risk of sudden cardiac death. Knowing the influence of hERG on the human body is essential for the diagnosis, prognosis, and treatment of cardiac disorders. Blocking the hERG potassium ion channel may lead to long QT syndrome, arrhythmia, and even sudden death, so it is necessary to predict the toxicity of hERG -the encoded potassium ion channel.

Machine learning (ML) has the potential to revolutionize medical diagnosis and treatment. ML algorithms can be used to improve predictive accuracy and identify patterns in data, which can help healthcare providers make more informed decisions. By incorporating ML into medical practice, physicians can improve accuracy, reduce false negatives and detect potential diseases faster

* Corresponding author.
E-mail addresses: jincai@ncst.edu.cn (J. Chang)
Doi:

and more accurately than traditional methods.

ML can be applied in different areas of clinical medicine, such as diagnostics, drug research, precision medicine, and surgical procedures. In diagnostics, ML algorithms can be used to analyze medical imaging data and determine the presence of certain diseases or conditions. In drug discovery, ML algorithms can predict interactions between drugs and their targets, helping researchers identify potential treatments for positive outcomes. In precision medicine, ML can identify individuals with specific genetic markers that could make them particularly receptive to novel treatments. In surgical procedures, ML systems can analyze data points to find optimal operations for each particular patient.

ML has broad applications in the medical field. Through the use of these powerful algorithms, practitioners and researchers can gain a better understanding of medical datasets and optimize medical outcomes for patients.

This paper develops a predictive model for the safety of compounds that antagonize ERα activity on the human heart using the machine learning method.

2. Materials and methods

2.1 Sources of information

The data set used in this article comes from the "Molecular_Descriptor.xlsx" and "ERα_activity.xlsx" provided by the 2021 Huawei Cup Mathematical Modeling Competition, with a total of 729 molecular descriptors (features).

2.2 Method

In this paper, the original data is firstly standardized to eliminate the influence of different data dimensions; then the analysis of variance is carried out and the features with statistical significance are screened out with a false discovery rate of less than 0.01; finally, the feature importance is sorted and compared with the selected features PCA dimensionality reduction. Finally, 8 principal components are selected.

Based on these 8 principal components, the SVM classification is carried out, and the appropriate parameters are adjusted using the genetic algorithm; finally, the appropriate SVM kernel function is selected using the ROC curve. The specific process is shown in Figure 1:

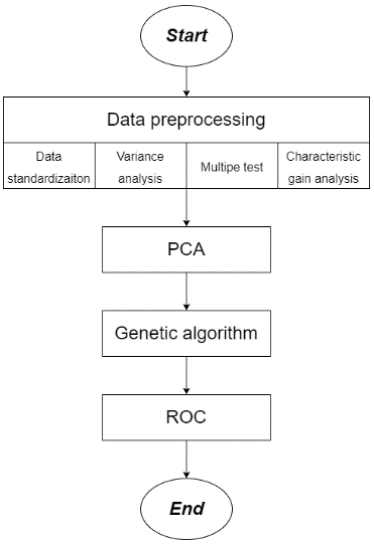

Fig. 1. h ERG prediction mode

3. Data preprocessing

3.1 Data standardization

Data standardization is the process of converting data into a common format that allows data to be accurately compared and contrasted across different sources. It is an important part of data cleaning and data integration processes, as it helps ensure that data from different sources is consistent and can be processed together.

According to the data characteristics provided by the file "Molecular_Descriptor.xlsx", it is found that there is a dimensional gap between the features. To improve the efficiency and accuracy of the machine learning algorithm, the data is standardized and mapped to [- 1, 1], and makes Each feature follows a normal distribution with mean 0 and variance 1. Use the StandardScaler class of python's preprocessing library to standardize the data. The specific formula is as follows:

$$x_n^* = \frac{x_n - \mu_n}{s_n} \tag{1}$$

Where x_n^* is the normalized data x_n of the nth feature, is the original data μ_n of the nth feature, is the mean of the original data of s_n the nth feature, and is the standard deviation of the original data of the nth feature.

3.2 Analysis of variance

Analysis of Variance (ANOVA) is a statistical technique used to compare the means of two or more groups. It is a common technique in many fields, including psychology, economics, and biology. ANOVA can be used to test if the means of different groups are equal or not. It tests the null hypothesis that all group means are equal, by comparing the variance between the groups with the variance within the groups. In ANOVA, each observation belongs to exactly one group, and there is no overlap between the groups. Each observation is correlated with its own group mean, but uncorrelated with the other group means.

Taking the analysis of compound heart safety evaluation (human Ether-a-go-go Related Gene, hERG) as an example, analysis of variance is performed on the 729 features provided, and analysis of variance is used for the significance test of the mean difference of multiple samples. Among the 729 features, the features that have a significant impact on hERG can be found.

When the value of hERG is 1, $x_n^*$the values of hERG form a set S_+, and when the value of hERG is 0, $x_n$the values of hERG form a set S_-.

The first step is to establish a test hypothesis. H_0: $\mu_{S^+} = \mu_{S^-}$. H_1: $\mu_{S^+} \neq \mu_{S^-}$. The value of the given significance level $\alpha$is usually taken as α=0. 05.

The second step is to construct the test statistic F, see formula (2).

$$F = \frac{S_A/(r-1)}{S_E/(n-r)} \sim F(r-1, n-r) \tag{2}$$

S_A, S_E is the variance between groups and intraclass variance, r is the number of independent variables, and n is the number of samples.

The third step is to make a conclusion. If $F > F_{0.05}(r-1, n-r)$it is rejected H_0, it is considered that $\mu_{S^+} \neq \mu_{S^-}$, it x_n^*has a large contribution to the classification prediction. Otherwise, x_n^* the contribution to classification prediction is small.

SelectKBest class of python's sklearn.feature_selection library to perform ANOVA on the data and calculate x_n^* the p -value

(H_0minimum significance level for negation). And the values of each feature are p arranged from small to large, and the values of p365 features with a p-value than 0.05 are drawn as shown in Figure 2:

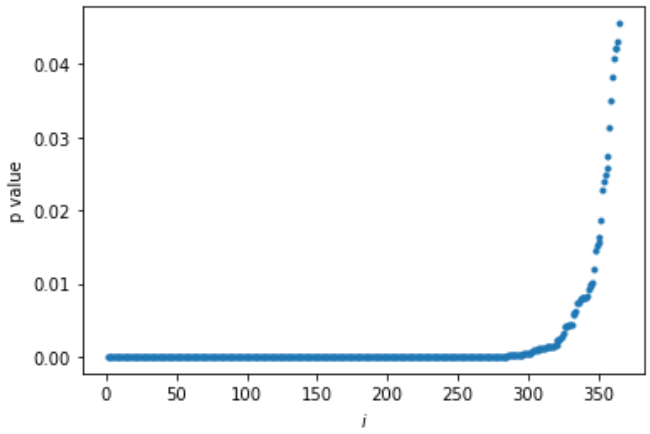

Fig. 2. h ERG ***p***value

3.3 False discovery rate

The FDR (False Discovery Rate) error control method is a method proposed by Benjamin and Hochberg in 1995. The basic principle is to determine p-value range by controlling the FDR value.

False discovery rate is a statistical method used to identify statistically significant results while controlling the rate of Type I errors (incorrectly rejecting null hypotheses). It allows researchers to detect real effects while avoiding too many incorrect rejections of null hypotheses. FDR helps limit this by controlling the rate at which incorrect rejections occur. Generally, it controls the number of false positives a researcher finds compared to the number of true positives. It is a powerful tool that can be used to ensure data accuracy, especially when multiple hypothesis tests are being made. The FDR reduces the chances of selecting variables that are just noise or random correlations, which can lead to overfitting. While the error detection rate (EDR) is often used to detect faults in signs, the FDR is generally better at filtering false positives that could occur if too many variables are included in the model.

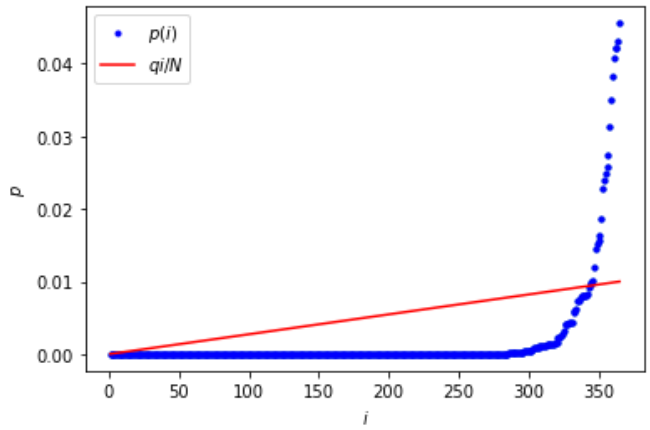

Fig. 3. h ERG $\boldsymbol{qi/N}$ *curve*

Find a p value threshold θ to keep the control error rate below 0.01 in multiple testing. Assume that N experiments are tested with index $i \in 1,2,3 \ldots N$. The threshold of the experiment $\theta = p(i)$. Then in N this experiment, the expected number of false positives F P is expected to be:

$$E(FP) = Np(i) \tag{3}$$

Assuming that the acceptable error rate in multiple testing is q = 0.01, it can be deduced $p(i)$ that the following needs to be met:

$$\frac{Np(i)}{i} < q \Rightarrow p(i) < \frac{qi}{N} \tag{4}$$

Plotted qi/N(red) on Figure3.

qi/N intersection $p(i)$ value of θ and is equal to 0.009268411599394519. A total of 343 features have p-values less than or equal to θ.

It can be concluded that among the 343 features, no more than 4 features do not contribute to the prediction.

3.4 Feature Importance

The extreme gradient boosting algorithm (XGBoost) is optimized on the gradient boosting decision tree (GBDT), which can be used for regression and classification tasks, and is also a boosting tree model. XGBoost is a powerful machine learning algorithm that uses gradient boosting trees to find decision tree ensembles that can be used to classify supervised learning data effectively. The feature selection implemented in XGBoost is an important part of this process, as it helps reduce the noise in the data and also helps to identify the most relevant features which will have the most impact on predictions. By pruning unnecessary features, the model can better focus on the more accurate ones and produce more reliable predictions.

XGBoost has several advantages when it comes to selecting the importance of features:

1. It is an optimized and regularized model which helps to minimize overfitting and the effect of outliers.
2. It uses decision trees for feature selection and uses an objective function to determine the "importance" of each feature, rather than just relying on heuristics.
3. It produces detailed features the importance which can be visualized with graphs or tables.
4. Its speed and scalability make it easy to utilize in large datasets.
5. It provides several methods to control regularization, including shrinkage and pruning, so users can customize the model to choose only the important features.

Additionally, by choosing the right features, XGBoost helps to reduce overfitting and underfitting, leading to more accurate and reliable models. Obtain the gain percentage of a feature in all features through the “feature_importances” attribute of the XGBClassifier object of the xgboost library in python, and the gain is the relative contribution of the feature to GBDT. The formula for calculating the gain is as follows:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{(H_L+H_R)^2+\lambda}\right] - \gamma \tag{5}$$

where $\frac{G_L^2}{H_L+\lambda}$is the gain of the left subtree, is the gain $\frac{G_R^2}{H_R+\lambda}$of the right subtree, and $\frac{(G_L+G_R)^2}{(H_L+H_R)^2+\lambda}$is the gain of not splitting the node.

Using the prediction model of xgboost, adjust the threshold of the gain percentage, and filter out different feature numbers. The results are shown in Figure 4:

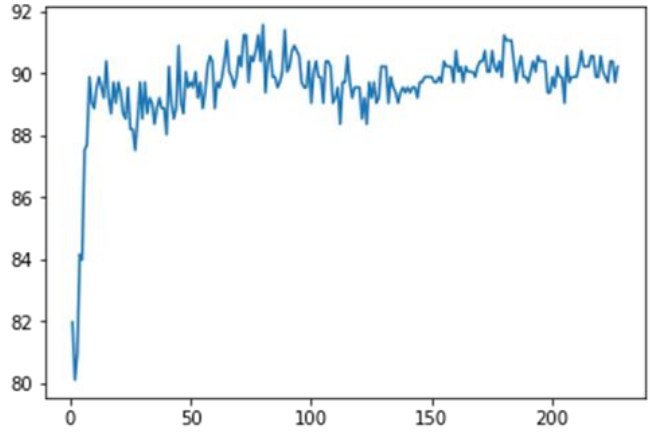

Fig. 4. The relationship between the number of features and the prediction accuracy

It is found that when the threshold of the gain percentage is 0.008583, 17 important features are screened out, and the descriptions of these 17 features are shown in table 1:

Tab. 1. Description of important features

feature	illustrate	gain percentage
nTG12Ring	Number of rings with more than 12 members (including a count of fused rings)	0.01149517
nFRing	Number of fused rings	0.01863866
MDEO-11	Molecular distance edges between all primary oxygens	0.01853098
Kier2	second kappa shape index	0.02466917
nHBAcc3	Number of hydrogen bond acceptors	0.00942097
max HsOH	Maximum atomic type HE states: -OH	0.01325267
wxya	Maximum electron state of weak hydrogen bond acceptor	0.00868131
wxya	Maximum electron state of a (strong) hydrogen bond donor	0.01417888
minaaS	The smallest atomic electronic state: aSa	0.00955727
minsssN	Minimum atomic type electronic state: >N-	0.02279783
SsssN	The sum of atomic electronic states: >N-	0.02558094
wxya	The sum of atomic E states: -NH-	0.0123142
SH	The sum of atomic HE states: -NH-	0.0087998
ECCEN	Topology Descriptor Combining Distance and Adjacency Information	0.22428963
VP-0	Valence Path, Level 0	0.09467269
C2SP2	Double bonded carbon bonded to two other carbons	0.02220235
bpol	The sum of the absolute values of the differences in atomic polarizability of all bonded atoms (including implicit hydrogens) in the molecule	0.02602787

Sort the gains of features from large to small, as shown in Figure 5:

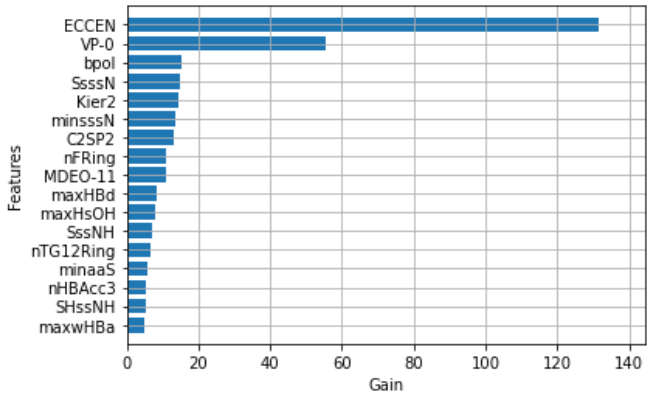

Fig. 5. Feature Gain Score

4. Principal component analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique that is used to reduce a large set of variables into a smaller set that still contains most of the information in the original set. It is an unsupervised learning algorithm, meaning that it doesn't make use of any label or outcome data to do its work. The principal components of a dataset are a set of variables that account for most of the variability in the data. PCA finds the directions, called principal components, that have the largest variance in the dataset. It then projects the data in those directions. Therefore, PCA can be thought of as projecting a dataset onto a smaller, lower dimensional subspace so that the variance of the data along each principal component is maximized. The advantages of PCA include reducing the complexity of a dataset as well as being able to visualize the data in a lower-dimensional space. Additionally, it can be used to detect outliers and noise more easily.

The principal component analysis is a multivariate statistical method widely used in the medical field to identify patterns, which can classify factors that affect a specific phenomenon, or by cutting principal components (PC) with smaller variance. In order to reduce the dimensionality, the PC that can be used to develop new models can be screened out, and the weight of PC can be used to calculate the contribution of each factor in the data. In this study, principal component analysis was performed on the 17 features screened by xgboost , and the result of the variance contribution rate is shown in Figure 6:

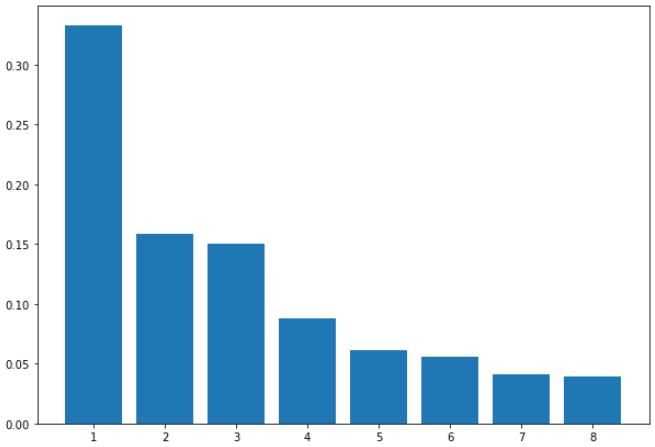

Fig. 6. Variance contribution rate of each principal component

Taking the first 8 principal components (PC1~PC8), the cumulative variance contribution rate is 92.54180787375579%.

The weight information in Table 2:

Tab. 2. Principal component weight information

project	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
variance contribution rate	0.332	0.158	0.149	0.087	0.060	0.055	0.040	0.039
standard deviation	5.653	2.701	2.549	1.489	1.036	0.946	0.694	0.667
cumulative probability	0.332	0.491	0.641	0.728	0.789	0.845	0.886	0.925

The eigenvector composition of Table 3:

Tab. 3. Principal component feature weight table

feature	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
bpol	0.37	0.12	-	0.06	-	0.06	-	-

	5	0	0.157	7	0.034	9	0.003	0.068
C2SP2	0.139	-0.374	-0.221	0.053	0.245	-0.07	0.298	-0.374
VP-0	0.391	0.051	-0.129	0.107	0.079	0.046	0.071	-0.157
ECCEN	0.396	0.028	-0.003	0.101	0.069	0.096	0.198	-0.064
SH	0.274	0.162	0.360	-0.144	-0.033	0.154	0.003	0.274
wxya	0.292	0.163	0.343	-0.131	-0.050	0.161	0.033	0.243
SsssN	0.187	-0.088	-0.447	-0.064	-0.139	-0.098	-0.315	0.372
minsssN	0.191	-0.132	-0.461	-0.015	-0.154	-0.056	-0.211	0.321
minaaS	-0.043	0.056	-0.033	0.136	-0.904	0.173	0.167	-0.256
wxya	-0.044	-0.366	0.228	0.392	0.033	0.284	0.038	0.409
wxya	-0.072	0.190	-0.191	-0.014	0.190	0.778	-0.380	-0.239
max HsOH	-0.057	-0.430	0.014	0.462	-0.01	0.227	0.053	0.082
nHBAcc3	0.307	-0.073	0.192	0.143	-0.054	-0.210	-0.079	-0.031
Kier2	0.405	-0.015	-0.005	0.026	0.024	0.095	0.126	-0.132
MDEO-11	0.126	-0.103	0.308	0.267	-0.045	-0.256	0.713	-0.343
nFRing	-0.075	0.427	-0.124	0.476	0.121	-0.128	0.100	0.095
nTG12Ring	-0.055	0.452	-0.100	0.471	0.068	-0.107	0.045	0.075

5. SVM Classification

5.1 Principlejie

Support Vector Machine (SVM) is a supervised machine learning technique used for classification and regression analysis. It is a powerful method for classifying data, especially when there are multiple classes. An SVM model takes a dataset and finds an optimal hyperplane that divides the dataset into two parts. It then draws a line along that divide and classifies each point on one side or the other of the line. The goal is to find the best possible hyperplane that separates two or more classes of data points in a high-dimensional space. The main idea behind SVM is that it works by constructing a hyperplane in a multidimensional space that separates classes of data points. It then assigns new observations to the class that aligns with the hyperplane. To create good classification models, SVM adjusts the parameters of the model to fit the training data. This is done by optimizing an objective function which includes regularization parameters to prevent overfitting and control model complexity.

Advantages of Support Vector Machines (SVM) Classification:

1. SVM can handle high-dimensional data very efficiently.
2. It works really well in complex domains where there is a clear margin of separation.
3. In addition to performing linear classification, it can also successfully carry out non-linear classifications using the kernel trick.
4. SVMs are effective in cases where the number of dimensions is greater than the number of samples.
5. It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
6. Due to its versatility, it can be used to solve both classification and regression problems.
7. The regularization parameter of SVMs allows us to control the trade-off between smooth decision boundaries and classifying the training points correctly.
8. SVMs are very robust against overfitting, as they inherently use the structural risk minimization principle.

The core of SVM classification is to generate a hyperplane for dividing data, and introduce a kernel function to calculate the interval between data and the classification plane, so as to seek a hyperplane to separate different types of data with the largest interval. Its cost function is:

$$J(\theta) = C\ \sum_{i=1}^{m}[y^{(i)}cost_1(\theta^T x^{(i)}) + (1-y^{(i)})cost_0(\theta^T x^{(i)})] + \frac{1}{2}\sum_{j=1}^{n}\theta_j^2 \quad (6)$$

Purpose of the SVM is to determine suitable θ^T such that $J(\theta)$minimum. The discriminant function of the SVM classification model is:

$$f(x) = sgn(\sum_{i=1}^{n} a_i K(x_i, x) - b) \quad (7)$$

$K(x_i, x)$There are four commonly used kernel functions in the SVM model, as shown in Table 4:

Tab. 4. Common kernel functions of SVM

kernel function name	kernel function expression
Sigmoid kernel function	$K(x, x_i) = \tanh(gx^T x_i + r)$
radial basis function	$K(x, x_i) = \exp(-\frac{\|x - x_i\|^2}{2\sigma^2})$
polynomial kernel function	$K(x, x_i) = (gx^T x_i + r)^d$
linear kernel function	$K(x, x_i) = x^T x_i$

5.2 Genetic Algorithm Tuning

A genetic algorithm (GA) is a search heuristic that mimics the process of natural selection. This heuristic is routinely used to generate useful solutions to optimization and search problems. Some of the most common example areas include robot control, data mining, and machine learning.

Genetic algorithms are based on the idea of natural selection and inheritance. They use methods such as mutation, crossover, and selection in order to generate new solutions from existing ones.

The process starts with a population of randomly generated solutions and then evaluates each of the solutions (or chromosomes).

The best solutions (with the highest fitness values) are selected for mating. The selection process is done according to the weighted probability distribution. Mating is done by combining the best chromosome with another one. The resulting offspring will have characteristics from both its parents. Finally, the mutation is applied to produce new solutions. This process is repeated several times until the optimal solution is found.

In summary, genetic algorithms are search heuristics that use the principles of natural selection to generate solutions to optimization and search problems. They work by creating a population of potential solutions and then selecting and breeding the fittest individuals to generate newer, more fit solutions.

Using genetic algorithms (GA) to select machine learning hyperparameters is the process of creating an artificial intelligence algorithm to optimize a system's performance. This usually involves making trade-off decisions between competing objectives, such as accuracy and runtime. GA works by creating an initial population of possible solutions and then running simulated evolution cycles over them to find the best ones. During each cycle, it simulates evolutionary processes such as mutation and crossover, which are used to create new generations of better solutions. The newly created solutions are then evaluated to check their performance. Those that perform the best are rewarded by being chosen to form the new population while the rest are removed. This process is repeated until a satisfactory solution is found or until it reaches a predetermined termination criteria.

In SVM, the two parameters that need to be adjusted are the kernel function parameter σ and the penalty coefficient C. Therefore, the decision variables of the genetic algorithm are two-dimensional.

Encode the parameters of SVM into the chromosomes of the genetic algorithm, then initialize the "chromosome" population, and divide the training set and test set. Kernel function parameters $\sigma$and penalty coefficients $C$are called chromosome genes. The measure of the pros and cons of each chromosome is the average cross-validation score. In each iteration of the genetic algorithm, a part of chromosomes with poor performance will be deleted, and chromosomes with good performance will be left. In order to imitate the mating process in the biological world, chromosomes with good performance can exchange genes with each other. The selection strategy of the exchanged genes is Russian roulette, and the probability is:

$$P(x_i) = \frac{f(x_i)}{\sum_{j=1}^{N} f(x_j)} \tag{7}$$

Population size is N, and the fitness of the individuals i(mean cross-validation score) is $f(x_i)$.

If only chromosomes exchange genes with each other, then the kernel function parameters σ and penalty coefficients C will be quickly limited to a range. Therefore, the genetic algorithm introduces the concept of mutation. Generally, commonly used mutation methods include basic bit mutation, uniform mutation, edge mutation, non-uniform mutation, Gaussian mutation reversal mutation, etc.

The kernel function of the SVM is set as the radial basis function, and the average score of the iterative 950 cross-validations tends to be stable. The process is shown in Table 5:

Tab. 5. rbf kernel function genetic algorithm

iterations	score maximum	score average	minimum score	Score standard deviation
50	7.43753E-01	6.48318E-01	6.02735E-01	4.71105E-02
100	8.27918E-01	6.96797E-01	6.02735E-01	8.15751E-02
150	8.31453E-01	7.28063E-01	6.02735E-01	8.13659E-02
200	8.31943E-01	7.57357E-01	6.15882E-01	7.47674E-02
250	8.31958E-01	7.90875E-01	6.22448E-01	5.97834E-02
300	8.32976E-01	8.06122E-01	6.22448E-01	4.47744E-02
350	8.32976E-01	8.21886E-01	7.05820E-01	2.43591E-02
400	8.32976E-01	8.26232E-01	7.05820E-01	1.73848E-02
450	8.32999E-01	8.29089E-01	8.23854E-01	2.58913E-03
500	8.32999E-01	8.29780E-01	8.25385E-01	2.34148E-03
550	8.33007E-01	8.30495E-01	8.25385E-01	1.94520E-03
600	8.33007E-01	8.31076E-01	8.25385E-01	1.59124E-03
650	8.33512E-01	8.31391E-01	8.25385E-01	1.40531E-03
700	8.33512E-01	8.31686E-01	8.28392E-01	1.11294E-03
750	8.33512E-01	8.31900E-01	8.28392E-01	9.51355E-04
800	8.33512E-01	8.32175E-01	8.29433E-01	7.75344E-04
850	8.33512E-01	8.32266E-01	8.29969E-01	7.18399E-04
900	8.33512E-01	8.32419E-01	8.30940E-01	6.74876E-04
950	8.33512E-01	8.32553E-01	8.30987E-01	6.12578E-04

The maximum value and average value of the score are drawn into a curve as shown in Figure 7:

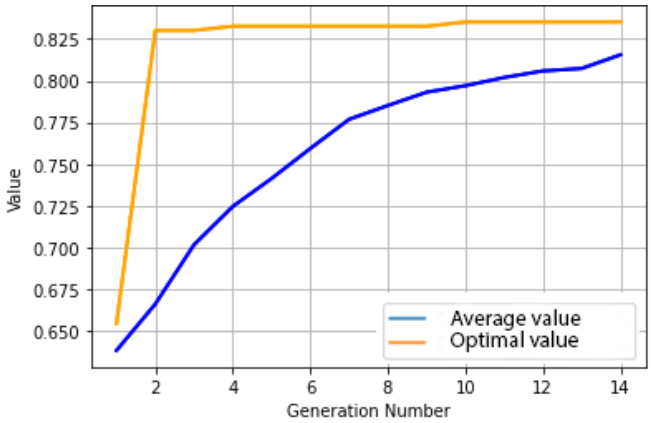

Fig. 7. rbf kernel function iteration curve

It can be obtained that when the penalty coefficient $C$is 8.114378195255995, the kernel function parameter $\sigma$is 0.00390625, the highest correct rate of cross-validation is 83.35120435120434%.

Using the linear kernel function, when the penalty coefficient $C$is 28.34722375869751 and the kernel function parameter is σ = 150.07192885875702, the highest correct rate of cross-validation is 82.7964257964258%. Using the Sigmoid kernel function, the penalty coefficient $C$is 99.00044253468513, and the kernel function parameter σ is 0.00390625. The highest correct rate of cross-validation is 83.50738150738154 %.

5.3 ROC curve

An ROC Curve stands for Receiver Operating Characteristic Curve and is a measure of the area under a graph of true positive rate (TPR) against false positive rate (FPR). The ROC Curve is used to evaluate how well a binary classifier (a model which can classify data into two distinct classes) can distinguish between the two classes.

In more detail, the ROC Curve is a graph that plots the true positive rate (TPR) on the y-axis and the false positive rate (FPR) on the x-axis. The TPR is calculated by dividing the number of true positives (TPs) by the sum of true positives and false negatives (FNs). The FPR is computed by the number of false positives (FPs) divided by the sum of false positives and true negatives (TNs).

The area under the ROC Curve (AUC) can be used to measure

the performance of a binary classifier. A model with an AUC of 1 can perfectly differentiate between the two classes, while an AUC of 0.5 indicates the model has no predictive power. By comparing multiple models, it is possible to decide which performs better based on their respective AUC values.

In conclusion, the ROC Curve is a measure of a binary classifier's ability to distinguish between two classes, based on the area under the curve. The AUC value is used to compare and assess different models, with a value of 1 indicating the perfect classification and 0.5 suggesting no predictive power.

Evaluate the fitting effect of these three kernel functions, using the receiver operating characteristic (ROC) curve of the training set. The results are shown in Figure 8. It can be seen that the area under the curve of the rbf kernel (area under curve, AUC) is a maximum of 0.94, so the fitting effect of the rbf kernel function is the best:

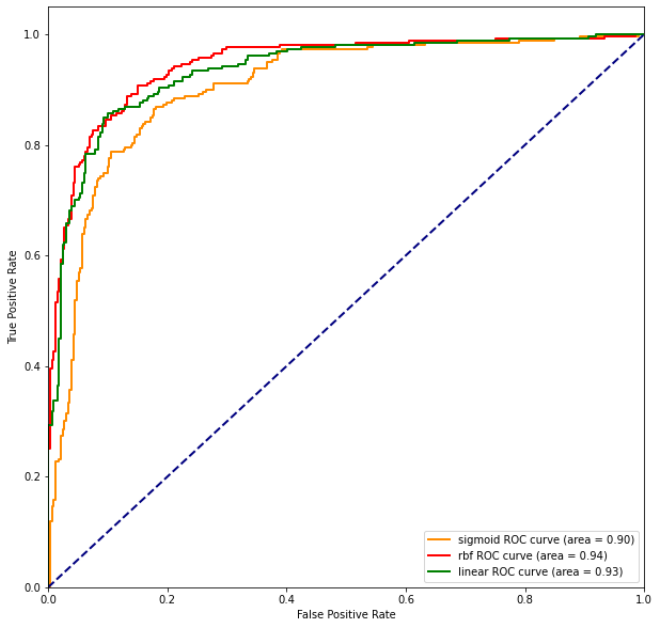

Fig. 8. ROC curves of different kernel functions

6. Discussion

This paper proposes a predictive model for the safety of compounds that antagonize ERα activity on the human heart. The model reduces the dimensionality of the 729 -dimensional features of the original data set to 8 dimensions. And predict by SVM classification algorithm. The correct rate of cross-validation is 83.351204351204 34%, and the AUC is 0.94. The prediction effect is better. This has a guiding significance for the development of anti-breast cancer drugs in the future.

References

Heldring N, Pike A, Andersson S, et al. Estrogen receptors: how do they signal and what are their targets. Physiol Rev. 2007;87:905–31.

Sun Y S, Zhao Z, Yang Z N, et al. Risk factors and preventions of breast cancer[J]. International journal of biological sciences, 2017, 13(11): 1387.

Osareh A, Shadgar B. Machine learning techniques to diagnose breast cancer[C]//2010 5th international symposium on health informatics and bioinformatics. IEEE, 2010: 114-120.

Amrane M, Oukid S, Gagaoua I, et al. Breast cancer classification using machine learning[C]//2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT). IEEE, 2018: 1-4.

Fatima N, Liu L, Hong S, et al. Prediction of breast cancer, comparative review of machine learning techniques, and their analysis[J]. IEEE Access, 2020, 8: 150360-150376.

S Safe, K Kim. Non-classical genomic estrogen receptor (ER)/specificity protein and ER/activating protein-1 signaling pathways[J]. Journal of Molecular Endocrinology, 2008.

Gal M S, Rubinfeld D L. Data standardization[J]. NYUL Rev., 2019, 94: 737.

Larson M G. Analysis of variance[J]. Circulation, 2008, 117(1): 115-121.

St L, Wold S. Analysis of variance (ANOVA)[J]. Chemometrics and intelligent laboratory systems, 1989, 6(4): 259-272.

Kaufmann J, Schering A G. Analysis of variance ANOVA[J]. Wiley Encyclopedia of Clinical Trials, 2007.

Storey J D. False Discovery Rate[J]. International encyclopedia of statistical science, 2011, 1: 504-508.

Benjamini Y. Discovering the false discovery rate[J]. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2010, 72(4): 405-416.

Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting[J]. R package version 0.4-2, 2015, 1(4): 1-4.

Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

Abdi H, Williams L J. Principal component analysis[J]. Wiley interdisciplinary reviews: computational statistics, 2010, 2(4): 433-459.

Vidal R, Ma Y, Sastry S S, et al. Principal component analysis[J]. Generalized principal component analysis, 2016: 25-62.

Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach[C]//Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. IEEE, 2004, 3: 32-36.

Huang S, Cai N, Pacheco P P, et al. Applications of support vector machine (SVM) learning in cancer genomics[J]. Cancer genomics & proteomics, 2018, 15(1): 41-51.

Shida Caia,Liangliang Sun,Mantong Zhao,et al.Prediction Of Milling Cutter Wear Based on ASO-BP Neural Network.International Journal of Applied Mathematics in Control Engineering[J],2022,5(1):109-113.

Giuliani A. The application of principal component analysis to drug discovery and biomedical data. Drug Discov Today. 2017 Jul;22(7):1069-1076. doi: 10.1016/j.drudis.2017.01.005. Epub 2017 Jan 19. PMID: 28111329.

Xiaoxuan Chena,Renlong Zhang.Target Recognition and Detection Based on Convolutional Neural Network Deep Learning.International Journal of Applied Mathematics in Control Engineering[J],2021,4(1):107-112.

Garcia-Retortill o S, Javierre C, Hristovski R, Ventura JL, Balagué N. Principal component analysis as a novel approach for cardiorespiratory exercise testing evaluation. Physiol Meas. 2019 Sep 3;40(8):084002. doi: 10.1088/1361-6579/ab2ca0. PMID: 31239421.

Yongqiang Zhang,Yanni Song,Hongbin Gao.Design of Image Classification Model by Logistic Regression Weighted Fusion Based on Tensor Flow.International Journal of Applied Mathematics in Control Engineering[J],2020,3(2):128-134.

Cunqing Rong He was born in Ji Ning, Shandong province, China in 1997. Now, he is a graduate student of Graduate School of North China University of science and technology, studying in cyberspace security, mainly researches on machine learning and steganography.

Ke Jin He was born in Zhen Zhou, Henan province, China in 1997. Now, he is a graduate student of Graduate School of North China University of science and technology, studying in network and information security, mainly researches on data security and privacy protection.

Jincai Chang received his B.Sc. degree in 1996 from Ocean University of China, received his M.Sc. degree in 2005 from Yanshan University, received his Ph.D. degree in 2008 from Dalian University of technology, now he is Professor in North China University of Science and technology. His main research interests include theories and methods in mathematical modelling and scientific computation, numerical approximation and computational geometry, etc.