MicroRNA Prediction Based on Machine Learning Algorithm

Shaoyan Zheng [1, a], Miaona Zeng [2, b]

[1]College of Life Science and Technology, Jinan University, Guangzhou City, Guangdong Province 510632, China;

[2]Department of Chemistry, Zhaoqing Medical College, Zhaoqing City, Guangdong Province 526020, China.

[a]imbshao1015@163.com, [b]zengmiaona668@163.com

Abstract

Based on the microRNA data of tumor samples and normal samples, the sample classification model was established using support vector machine（SVM）and extreme gradient boosting（XGBoost） algorithm. During the establishment of the support vector machine model, genetic algorithms and cross validation were used to optimize the parameters c and g. The accuracy of the final SVM model was 53.8462%. At the same time, using the Xgboost method with better performance for rule mining, the established classification model is superior to the SVM model, and the accuracy rate reaches 100%. Studies have shown that XGBoost has a stronger fitting ability in building models.

Keywords

MicroRNA, support vector machine, extreme gradient boosting, classification model.

1. Introduction

MicroRNA is a type of endogenous small molecule RNA with a length of 19-24 bases. The microRNA plays a regulatory role in gene expression at the post-transcriptional level and plays an important role in biological and disease processes[1, 2]. Nearly a third of human genes are regulated by microRNAs, including cell development, tissue differentiation, and cell cycle[3-5]. The dysregulation of microRNAs is closely related to some diseases, especially when more than 50% of human microRNAs are associated with cancer gene segments. Studies have shown that it has a close relationship with cancer. Research on microRNAs helps people understand the relationship between genetically controlled networks, and is more conducive to the study of gene function and the evolutionary exploration of biology. Research on microRNAs is of great significance. MicroRNAs play an important regulatory role in cells. They exist extensively in eukaryotic cells and play important biological functions in all aspects of gene expression regulation. Exogenous siRNA and RNA interference technology has become the first choice for future gene drug design due to their relative simplicity, efficiency, and specificity. In order to further study the regulation of microRNAs on living organisms,the first and foremost function is the effective prediction of microRNAs, and the prediction of microRNAs has become an important topic in the field of bioinformatics research.The machine learning-based miRNA target gene prediction method not only has low research cost, requires less prior knowledge, but also can fully mine the valuable laws between data through learning modeling. Therefore, it has attracted the attention of researchers.

In this work, we analyzed large datasets of small RNA sequences in cervical cancer/normal samples. Based on this sequence data, we described statistical methods for cancer classification, and we proposed a new method for identifying diagnostic miRNAs using sequencing-based miRNA analysis data. This method should be versatile when analyzing differential sequence representations between biological sample sets.

2. Methods

First, 58 samples of microRNA data (including normal and tumor) were randomly divided into a training set and a validation set in a ratio of 3:1. That is, 38 training sets were used to establish the classifier, and the remaining 19 were classified as verification sets. The performance of the model was evaluate by total accuracy.

2.1 Support vector machine (SVM)

SVM was first proposed by Vapnik and originated from statistical learning theory. It is a machine learning algorithm for solving two kinds of problems. In pattern classification, support vector machines follow the structural risk minimization criteria to construct a decision hyperplane, making two types of samples The interval between the largest classification. Support vector machines have the following advantages: 1) versatility: ability to construct functions in a wide variety of function sets; 2) robustness: no need for fine-tuning; 3) effectiveness: always one of the best methods in solving practical problems; 4) simple: The implementation of the method only requires the use of simple optimization techniques; 5) theoretically perfect: the framework of the vapnik-chervonenkis promotion theory.

During the application of SVM, the selection of kernel functions and the choice of parameters are very important to the performance of the model. In this work, the radial basis kernel functions, as well as the genetic algorithm and cross validation (CV) are used to select the c and g parameters. There is a certain relationship between the classification accuracy of the SVM model and the penalty factor c and the kernel function g. In order to obtain the SVM model with the best classification performance, the best c and g values need to be obtained. Obviously this is an optimization problem. If you search the optimal value in an exhaustive way, the amount of calculation will be too large to be realized.

Because the genetic algorithm (GA) [6] has implicit parallelism and powerful global search ability, it can search the global best point in a very short time. Therefore, this work uses genetic algorithm to optimize the parameters of SVM classification model. First, the SVM classification model penalty factor c and kernel function g values are binary-encoded, and the initial population is randomly generated; secondly, each chromosome in the population is decoded to obtain c and g values, and a part of the training sample set data is used to train the SVM classifier. The model uses the trained SVM classifier to calculate the recognition rate (RR) of the test sample set data. According to the cross-validation principle, the recognition rate RR reflects the generalization ability and classification ability of the SVM model to a certain extent, so the genes can be constructed accordingly. The fitness of the string Fitness = RR; then determine whether the stop criterion of the genetic algorithm is satisfied, if it is satisfied, stop the calculation and output the optimal parameter; otherwise, perform operations such as selection, crossover, and mutation to generate a new generation of population, and start a new generation.

2.2 Extreme Gradient Boosting

Extreme gradient boosting (XGBoost)[6], which is an extension of Gradient Boosting Machine. The Boosting classifier belongs to the ensemble learning model. Its basic idea is to combine hundreds of tree models with lower classification accuracy into one model with higher accuracy. The model continues to iterate. Each iteration generates a new tree. How to generate a reasonable tree at each step is the core of the boosting classifier. The gradient boosting machine algorithm uses the idea of gradient descent to generate each tree, based on all the trees generated in the previous step, and proceeds in a direction that minimizes the given objective function. With reasonable parameter settings, a certain number of trees need to be generated to achieve the desired accuracy. When the data sets are large and complex, the gradient boosting machine algorithm has a huge amount of calculations. Xgboost is an implementation of the gradient boosting machine that automatically uses the CPU's multithreading for parallelism and improves the algorithm to improve accuracy. Xgboost's base learner has both a tree (gbtree) and a linear classifier (gblinear), resulting in linear regression or

logistic regression with L1+L2 penalty. Its loss function is a second-order Taylor expansion with high accuracy and is not easily over-fitted. Features such as scalability and scalability can deal with high-dimensional sparse features in a distributed manner. Therefore, under the same conditions, Xgboost algorithm is more than 10 times faster than similar algorithms. The model parameters of Xgboost model are shown in Table 1.

Table 1 The parameter of XGBoost model

Parameter	Value	Parameter	Value
nthread	None	eta	0.1
seed	None	Max_depth	3
objective	Binary:logistic	subsample	1
nrounds	100	Colsample_bytree	1

3. Result and discussion

3.1 Data description

Data set is gene expression profiling data from cervical cancer tumor and matched normal samples (29 each)，which from the research team of Witten, D.[8]. The data are the raw read counts (not normalized) from sequencing of microRNA.

3.2 The SVM model

In order to verify the validity of the method used in this paper, the SVM classifier model was selected using the training set validation set using a combination of 5-CV and GA. CV is a statistical analysis method used to verify the performance of the classifier. The basic idea is to group the original data in a certain sense, one part as a training set and the other part as a verification set. The method is to first train the classifier with the training set, and then use the verification set to test the trained model, and use the obtained classification accuracy as the performance index of the evaluation classifier. First of all, the genetic algorithm selects, crosses, and mutates to determine the optimal solution, then decodes and outputs the optimal solution (bestc=0.01602, bestg=23.3942). From Figure 1, it can be seen that optimal fitness is 53.8462%. However, the accuracy of the model is not enough and a better model is needed to establish the classification model.

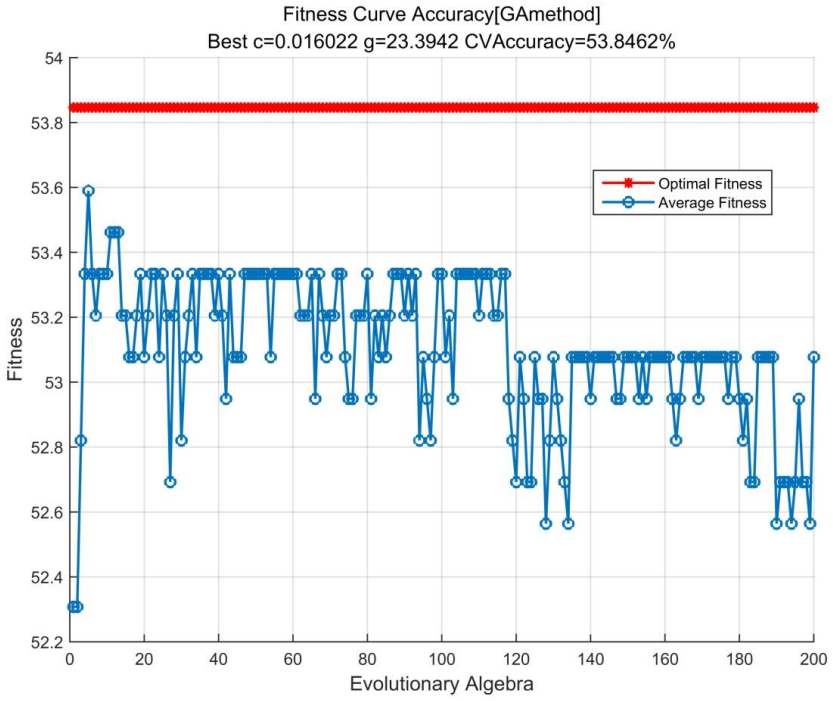

Fig.1 Using GA to find the fitness (accuracy) curve of the best parameters

3.3 XGBoost model

Based on the data analysis in this paper, the accuracy of the model is as high as 100%. The XGBoost model has a good fitting effect and can effectively classify the normal samples and tumor samples of experimental data. The algorithm effectively overcomes the "dimensional hazard" problem of experimental data and enables the classification algorithm to achieve rapid iteration.

The classification of tumors and normal samples by the two algorithms shows that the model established by XGBoost is superior to SVM. The accuracy of the model validation set established by SVM is 47.37%. The model established by XGBoost achieves 100% accurate prediction in the data sample. The data analyzed in this paper is typical of a small sample size and high dimensionality. The XGBoost algorithm approaches the target by constructing a tree. Each tree is constructed based on the error of the previous tree, and the established model is more robust.

4. Conclusion

In this paper, the SVM and XGBoost algorithm are used to classify the microRNA data of tumor/normal samples. The experimental results show that, XGBoost has stronger fitting ability in establishing models and can improve the classification accuracy compared with SVM. It provides a new idea for diagnosing diseases based on microRNA data.

References

[1] D Baltimore, M P Boldin, R M O'Connell, et al. MicroRNAs: NewRegulators of Immune Cell Development and Function[J]. Nature Immunology. 2008, 9(8): 839-845.
[2] L Song, R S Tuan. MicroRNAs and Cell Differentiation in Mammalian Development[J]. Birth Defects Research Part C: Embryo Today: Reviews. 2006, 78(2): 140-149.
[3] MicroRNAs: Genomics, Biogenesis, Mechanism, and Function[J].
[4] S M Hammond. MicroRNAs as Oncogenes[J]. Current opinion in genetics & development. 2006, 16(1): 4-9.
[5] J E A Brennecke. Bantam Encodes a Developmentally Regulated MicroRNA That Controls Cell Proliferation and Regulates the Proapoptotic Gene Hid in Drosophila[J]. Cell. 2003, 113(1): 25-36.
[6] L Davis. Genetic Algorithms in Search, Optimization, and Machine Learning[M]. Reading: Addison-Wesley, 1989.
[7] T Chen, C Guestrin. Xgboost: A Scalable Tree Boosting System[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining[Z]. 2016.
[8] D Witten, R Tibshirani, S G Gu, et al. Ultra-high Throughput Sequencing-based Small RNA Discovery and Discrete Statistical Biomarker Analysis in a Collection of Cervical Tumours and Matched Controls[J]. 2010, 8(1): 58.