Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/15/20230834

Research on the diagnosis of corona virus disease 2019(COVID-19) based on machine-learning

Fang Du

College of Physics and Information Engineering, Fuzhou University, Fuzhou, China.

031901302@fzu.edu.cn

Abstract. In recent years, the Corona Virus Disease 2019(COVID-19) has brought a huge impact on people's daily life and the normal operation of society, and as machine learning research deepens, this technology could help detect viruses such as the novel coronavirus. According to this, how to accurately, quickly and effectively analyze and classify laboratory results with multiple indicators through the method of machine learning is the object of this paper. In order to explore the classification performance of various machine learning classification algorithms on laboratory results, linear and non-linear methods were used respectively to analyze and classify the laboratory results. In linear analysis, distance discriminant and fisher discriminant were used to explore the classification effect of linear classification on laboratory results. The non-linear analysis mainly used Adaptive Boosting(AdaBoost) and Random Forest algorithm which are widely used to test the classification effect under the influence of multiple indexes. In this paper, python and other tools were used to classify samples of different combinations by using the idea of cross validation. By comparing the running time and detection accuracy, it was found that Adaboost algorithm is applicable in most cases and was a relatively fast and accurate classification method. In addition, Random Forest algorithm had similar accuracy, but it might have better performance on a large data set.

Keywords: machine-learning, fisher discriminant, index discriminant, random forest, COVID-19.

1. Introduction

Machine learning, as a multidisciplinary subject, uses computers as the tools to simulate human learning behavior in real time, and uses knowledge structures to divide the content to effectively improve learning efficiency [1]. With the deepening of people's learning and development in the field of artificial intelligence, machine learning has been applied more deeply in the fields of knowledge-based systems, natural language reasoning, machine vision and pattern recognition. In recent years, as a new strain of coronavirus that has never been found in humans before, Corona Virus Disease 2019(COVID-19) has become a member of coronaviruses. Because of the rapid spread of the virus, diversity of transmission routes and the cause of a wide range of serious illnesses, COVID-19 has caused a major impact on people’s daily lives and work, while also bring the severe challenge to the world-wild health system. For this virus, medical laboratory tests play an important role in assisting doctors to diagnose and clustering and discriminant algorithms in machine learning has provided scientific means of judgement.

DOI: 10.54254/2755-2721/15/20230834

In the current research, machine learning has been widely used in drug composition analysis and pathological analysis, such as Hui Xie et al. used Random Forest model, logistic regression model, Adaptive Boosting(AdaBoost), Gaussian Naive Bayes(GaussianNB) and other machine learning methods in the classification model to analyze the proportion of immune cell infiltration in pancreatic cancer [2]. Yixiao Zhai used the Random Forest classification method to classify antioxidant proteins [3]. Bin Tian et al. used a variety of machine learning classification methods such as decision tree, Random Forest, K-Nearest Neighbor(Knn) and support vector machine to analyze the feature recognition of lung images in COVID-19 [4]. Machine learning has a wide range of application and prospects in medicine composition analysis and pathological analysis.

The main task of this paper is to use the discriminant method in machine learning to find the relationship among different factors and judge whether the patient is infected with the COVID-19 according to the data from different cases. By comparing Fisher discriminant, Random Forest and other methods, we can select the best-matched and the highest-speed one and it will help people to judge the unknown virus with a better method.

2. Method

2.1. Problem description

Machine learning is widely used in component analysis and category classification. Among them, the component analysis and result judgement of medical laboratory results are closely related to the current world. In this highly intelligent era, people need to have some better method to classify some unknown things, such as virus and products, to help us find the factors we need and the results of judgment in the huge amounts of data. This paper mainly introduces the principles of clustering, linear discrimination algorithms and their applications in COVID-19 detection. The main analysis systems used are python and matlab.

2.2. Data collection

Since there are few data sets about the medical laboratory test results and detail of COVID-19, the data in this paper comes from a data set about the COVID-19 assay results provided in a Mathematical Contest in Modeling competition. Its main content is the laboratory results of 60 different patients. Each patient's laboratory results contain the patient's medical record number and seven test indicators. The first 30 patients are confirmed cases and the last 30 are healthy. Our goal is to choose the method with the highest accuracy and the fastest computing speed as the model to judge whether a certain person is infected by the COVID-19. The software and platforms used in this project are Pycharm (version 2022.3.2) and MATLAB.

2.3. Data analysis

In order to observe the distribution and characteristics of data in the data set more directly and prepare for subsequent data classification and analysis, using a line chart to display the data set is necessary. It’s not difficult to find from the statistical graph drawn by the seven indicators that the data of the diagnosed and health people fluctuated greatly in chart 1 and chart 7. In chart 2, chart 3, chart 5 and chart 6, there was no significant difference in the detection results between the two groups except for several peaks (Figure 1). In chart 4, there was a significant difference between diagnosed people and healthy people. These chart and analysis may help with the initial setting of some hyperparameters.

DOI: 10.54254/2755-2721/15/20230834

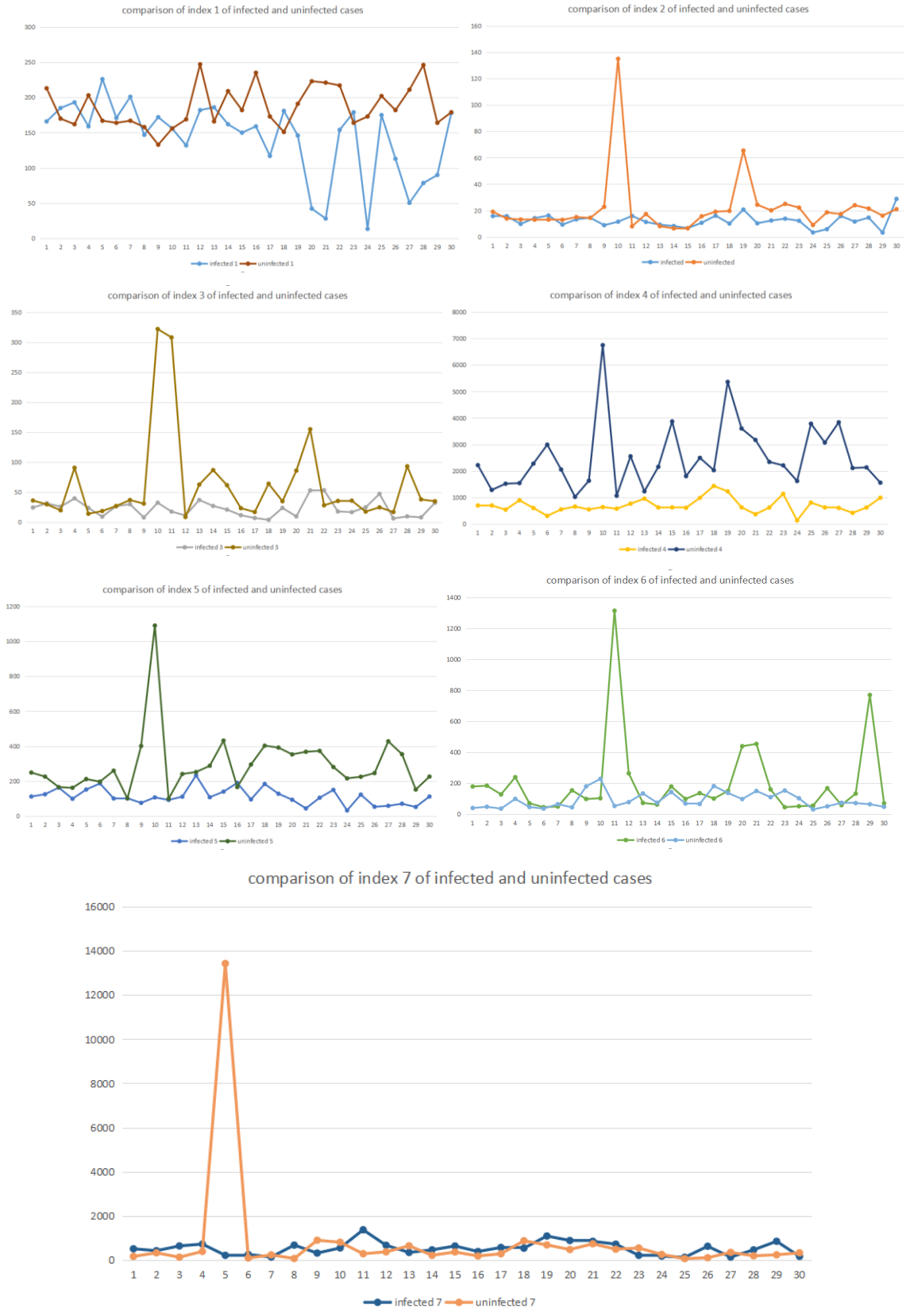

Figure 1. Seven indicators affecting COVID-19 in the data set.

DOI: 10.54254/2755-2721/15/20230834

2.4. Technical approach

2.4.1. Distance discriminant.

Distance discriminant is a classification method by calculating the distance between the center of the origin point and the surrounding categories. The basic idea of this algorithm is to calculate the distance between the point to be measured and various categories, and select the category with the shortest distance from the point as the point's classification.

Compared with other algorithms, Distance discriminant is simpler and easier to observe. This method is suitable for discriminating random variables with continuous distribution and is applicable to almost all probability distributions of variables. In Distance discriminant, Euclidean Distance discriminant and Mahalanobis Distance discriminant are usually used.

Euclidean Distance discriminant: If x, y are two points in n-dimensional space, the Euclidean distance between x and y is:

$$\mathrm{d}(\mathrm{x},\mathrm{y}) = \| \mathrm{x}-\mathrm{y} \|_2 = \sqrt{(\mathrm{x}-\mathrm{y})^{\mathrm{T}}(\mathrm{x}-\mathrm{y})} \tag{1}$$

Mahalanobis Distance discriminant: If x and y are two samples taken from a population X with a mean of μ and a covariance matrix of Σ, the Mahalanobis distance between x and y is:

$$\mathrm{d}(\mathrm{x},\mathrm{y}) = \sqrt{(\mathrm{x}-\mathrm{y})^{\mathrm{T}}\Sigma^{-1}(\mathrm{x}-\mathrm{y})} \tag{2}$$

In this equation, x, y, μ are all vectors. According to the Distance discriminant, if the distance (Euclidean distance or Mahalanobis distance) between the point x and the categories G1 and G2 is d1 and d2 respectively, we can discriminate it with:

$$\begin{cases} x \in G1, d1 < d2 \\ x \in G2, d1 > d2 \end{cases} \tag{3}$$

The algorithm process of Distance discriminant is given in Table 1.

Table 1. Distance discriminant.

Algorithm 1: Distance discriminant
1. Select the sample which is used for test from the sample set 2. Calculate the distance between the sample and each center of categories 3. Choose the category which has the shortest distance from the sample is selected as the classification of the sample

2.4.2. Fisher discriminant.

Fisher discriminant is a classification method of machine learning which select an appropriate projection direction and project the sample in this direction to make the projected sample points be separated as far as possible. This method can effectively reduce the dimension of data and maintain the necessary characteristics of sample set.

Suppose the set of training samples is $\mathrm{X} = \{\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_\mathrm{N}\}$ and each sample is a d-dimensional vector. The sample of class W_1 is $\mathrm{X}_1 = \{\mathrm{x}_1^1, \mathrm{x}_2^1, \cdots, \mathrm{x}_{\mathrm{N}_1}^1\}$, the sample of class W_2 is $\mathrm{X}_2 = \{\mathrm{x}_1^2, \mathrm{x}_2^2, \cdots, \mathrm{x}_{\mathrm{N}_2}^2\}$. the next step in this algorithm is to find a projection direction:

$$y_i = w^T x_i \, , i = 1,2, \cdots, N \tag{4}$$

By reducing the dimension of the sample data to one dimension and using the Fisher criterion (Rayleigh quotient), the projection direction is discriminated to be the best projection direction:

$$\max \ \mathrm{J_F(w)} = \frac{\widetilde{\mathrm{S_b}}}{\widetilde{\mathrm{S_w}}} = \frac{(\widetilde{\mathrm{m_1}}-\widetilde{\mathrm{m_2}})^2}{\widetilde{\mathrm{S_1}}^2+\widetilde{\mathrm{S_2}}^2} \tag{5}$$

In this equation, $\widetilde{\mathrm{S_b}}$ is the between-class scatter after projecting and $\widetilde{\mathrm{S_w}}$ is the within-class scatter. $\widetilde{\mathrm{m_1}}$ and $\widetilde{\mathrm{m_2}}$ are the mean vectors after projecting. $\widetilde{S_1}$ and $\widetilde{S_2}$ are the within-class scatter matrix [5].

DOI: 10.54254/2755-2721/15/20230834

The algorithm process of Fisher discriminant is given in Table 2.

Table 2. Fisher discriminant.

Algorithm 2: Fisher discriminant
1. Use the known sample observation matrix to calculate the sample mean vector of each population $\overline{x}^{(i)}$ and total mean vector of each population $\overline{x}$ 2. Calculate the between-class scatter matrix and the within-class scatter matrix respectively 3. Use the within-class scatter matrix and the between-class scatter matrix to search for the projection vector u and minimize the within-class distance and maximum the between-class distance 4. Discriminate the distance between the sample points and the categories on the projection, choose the closest category as the classification

2.4.3. Random Forest algorithm.

Random Forest algorithm is an algorithm that integrates multiple trees through the idea of ensemble learning. This method uses the decision tree as the basic learner and work effectively on classification and regression.

Decision tree is an important classification and regression method in machine learning and data mining. It is a model which use a tree-like structure to represent the predictive analysis. It is constructed in a recursive order from root to leaves and divide the sample into different subsets through selecting the main features. If these subsets can be classified correctly, the leaf nodes are built. Repeat the above steps until each subset is divided into leaves. The decision tree will be constructed at last.

In the process if decision tree construction, information entropy is the most common index to measure the sample set. Suppose that the proportion of class k in the current sample set D is p_k:

$$\mathrm{Ent(D)} = -\sum_{k=1}^{|y|} p_k \log_2 p_k \tag{6}$$

The purity of D increased with decreasing Ent(D) value. after that, Use the information entropy to calculate the information gain of each index:

$$\mathrm{Gain(D, index)} = \mathrm{Ent(D)} - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v) \tag{7}$$

By comparing the information gain, the appropriate feature dividing can be selected [6,7].

In addition to information gain, Gini coefficient is often used to divide decision trees in Random Forest. It's because the Gini coefficient reflects the probability that two randomly selected samples of D are inconsistent. The Gini coefficient is calculated by:

$$\mathrm{Gini(D)} = \sum_{k=1}^{|y|} \sum_{k' \neq k} p_k p_k' = 1 - \sum_{k=1}^{|y|} p_k^2 \tag{8}$$

$$\mathrm{Gini\_index(D, index)} = \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Gini}(D^v) \tag{9}$$

P_k is the proportion of class k in the current sample set D, V is the number of the indexes.

The coefficient used in this paper is the Gini coefficient. The algorithm process of Random Forest algorithm is given in Table 3.

DOI: 10.54254/2755-2721/15/20230834

Table 3. Random forest algorithm.

Algorithm 3: Random Forest algorithm
1. Randomly extract m samples from the original data set with the returned samples to generate m training set 2. Use the training set to train m decision tree models 3. For the decision tree model in random forest, the best features are selected to partition the data set by comparing the information gain or Gini coefficient. 4. Use the generated decision trees to construct a random forest and final classification of the test samples is decided by voting according to the multiple trees Decision tree algorithm: 1. Calculate the information entropy of the root node 2. Calculate the information entropy of each index and select the index with maximum information entropy as the dividing index 3. Reduce over-fitting risk by pruning 4. Repeat the preceding steps for the divided child nodes until no further dividing is possible

2.4.4. AdaBoost algorithm. AdaBoost algorithm is an improvement on Boosting algorithm and the way it is used to train weak learners is to train with all the data in the data set. The training samples will be given a weight again in each iteration and a more effective classifier will be constructed on the basis of the last weak learner. By increasing the weight of misclassification, the model pays more attention to misclassified samples and predicts the final result through voting.

In this method, the method of updating new weight ω' is [8,9]:

$$\varepsilon = \omega \times (\widehat{y_1} == y) \tag{10}$$

$$\alpha_j = 0.5 \times log\frac{1-\varepsilon}{\varepsilon} \tag{11}$$

$$\omega' = \omega \times e^{(-\alpha_j \times \widehat{y_1} \times y)} \tag{12}$$

The algorithm process of AdaBoost algorithm is given in Table 4.

Table 4. AdaBoost algorithm.

Algorithm 4: AdaBoost Algorithm
1. Initialize the sample weights that have the same initial value, and apply the constraint that the sum of the sample weights is 1 2. In the m boosting, do the step 3 to 5 for the j boosting 3. Train a weak leaner with a weight: $C(j) = \text{train}(X, y, \omega)$ 4. Predicted the sample and calculate the error rate of the weight 5. Calculate the parameter, update and normalize the weights 6. Complete the final forecast

3. Experimental procedure

3.1. Data selection and partitioning

In this work, the data set used had 7 indicators and had divided into two different labels. This study expects training and testing different algorithms by dividing the data set into training and test data sets. In this experiment, Distance discriminant, Fisher discriminant, Random Forest algorithm and AdaBoost algorithm were used to classify the test data set. The evaluation metrics are the time used to process the data and the accuracy of the classification. There are 30 infected and 30 uninfected cases in the total

DOI: 10.54254/2755-2721/15/20230834

data set. In this experiment, 10 data will be randomly selected form each of the infected and uninfected cases and these 20 data will be used as the test data set. The remaining data will be used as training data set to train the algorithm model.

In experiment, random data set partitioning was repeated 20 times, and the average accuracy of classification and the average running time were used as evaluation metrics for various algorithms.

3.2. Results

The experiment repeated the classification process several times, only the classification results of the first three times in 20 simulations of various classification algorithms are presented here. According to the results, compared with linear algorithm, AdaBoost algorithm and Random Forest algorithm consumed more time but had higher accuracy. In the following results, T means the classification is correct and F means the classification is wrong, as illustrate in Table 5 to table 8.

The results of Distance discriminant are shown in the table 5:

Table 5. Classification results of test data sets by distance discriminant.

Test sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
test 1	T	T	T	T	T	T	T	T	T	T	F	F	T	T	T	T	T	T	T	T
test 2	T	T	T	T	T	T	T	T	T	T	T	T	F	T	T	T	F	T	F	T
test 3	T	T	T	T	T	T	T	T	T	T	T	T	T	F	T	T	F	T	T	F

The results of Fisher discriminant are shown in the table 6:

Table 6. Classification results of test data sets by Fisher discriminant.

Test sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
test 1	T	T	T	T	T	T	T	T	T	T	T	F	F	T	T	T	T	T	T	T
test 2	T	T	T	T	T	T	T	T	T	T	T	T	F	F	T	T	T	T	F	T
test 3	T	T	T	T	T	T	T	T	T	T	T	T	T	F	T	T	T	F	T	F

The result of Random Forest algorithm is shown in the table 7:

Table 7. Classification results of test data sets by random forest algorithm.

Test sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
test 1	T	T	T	T	T	T	T	F	T	T	T	T	T	T	T	T	T	T	T	T
test 2	T	T	T	T	T	T	T	F	T	T	T	T	T	T	T	T	T	T	T	T
test 3	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T

The results of Adaboost algorithm are shown in the table 8:

Table 8. Classification results of test data sets by AdaBoost algorithm.

Test sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
test 1	T	F	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
test 2	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
test 3	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T

3.3. Result analysis

Finally, the accuracy and running time of the four algorithms are summarized in Table 9 and Table 10, Figure 2 and Figure 3. The results are analyzed through horizontal comparison of the accuracy and running time of different algorithms.

DOI: 10.54254/2755-2721/15/20230834

Table 9. Comparison of accuracy of several algorithms.

Algorithm	Distance discriminant	Fisher discriminant	Random Forest	AdaBoost
accuracy rate	0.8525	0.8775	0.9175	0.9300

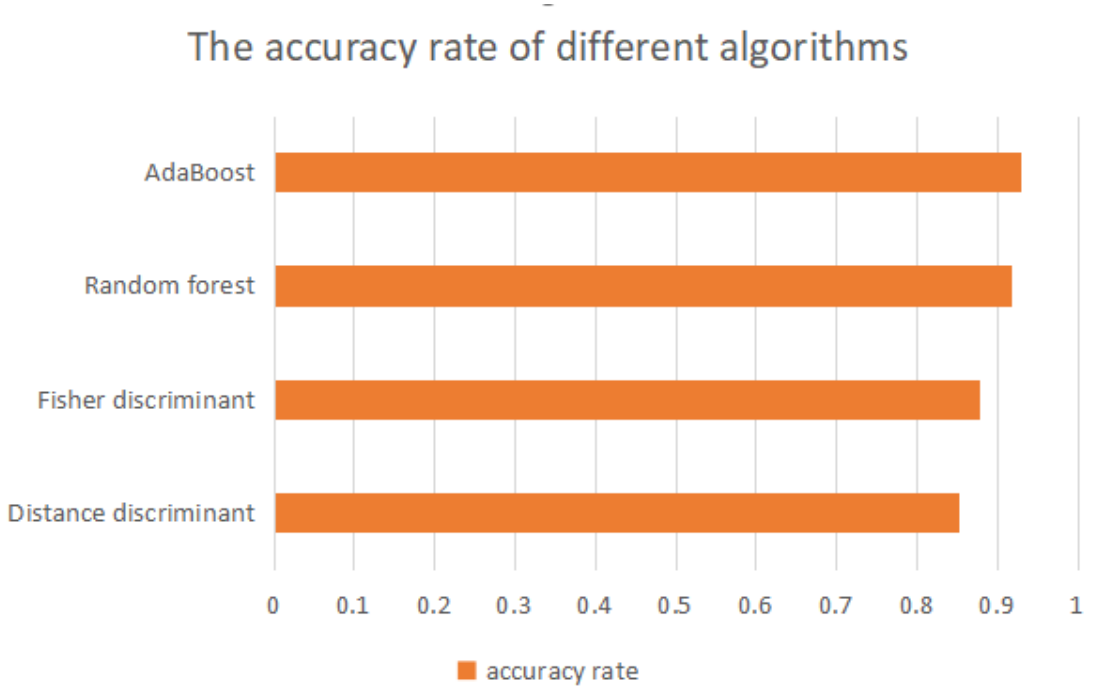

Figure 2. Comparison of accuracy of several algorithms.

By comparing the accuracy of several algorithms, it can be found that among various classification methods, AdaBoost algorithm and Random Forest algorithm are the ones with the highest accuracy. The specific accuracy rate sorting is: AdaBoost > Random Forest > Fisher discriminant > Distance discriminant. According to the above analysis of 7 different indicators, some indicators of infected and uninfected samples overlapped a lot. The overlap in the sample space is also the main reason for the poor classification effect of Distance discriminant and Fisher discriminant on this data set. Compared with AdaBoost algorithm, Random Forest algorithm is more suitable for a large-sample data set. The total amount of data contained in this sample data set is limited, which can not show the advantages of Random Forest algorithm. Therefore AdaBoost learning based on false classification has the best partition accuracy on this data set.

Table 10. Comparison of running time of several methods

Algorithm	Distance discriminant	Fisher discriminant	Random Forest	AdaBoost
running time (s)	0.000302	0.000750	0.012018	0.001594

DOI: 10.54254/2755-2721/15/20230834

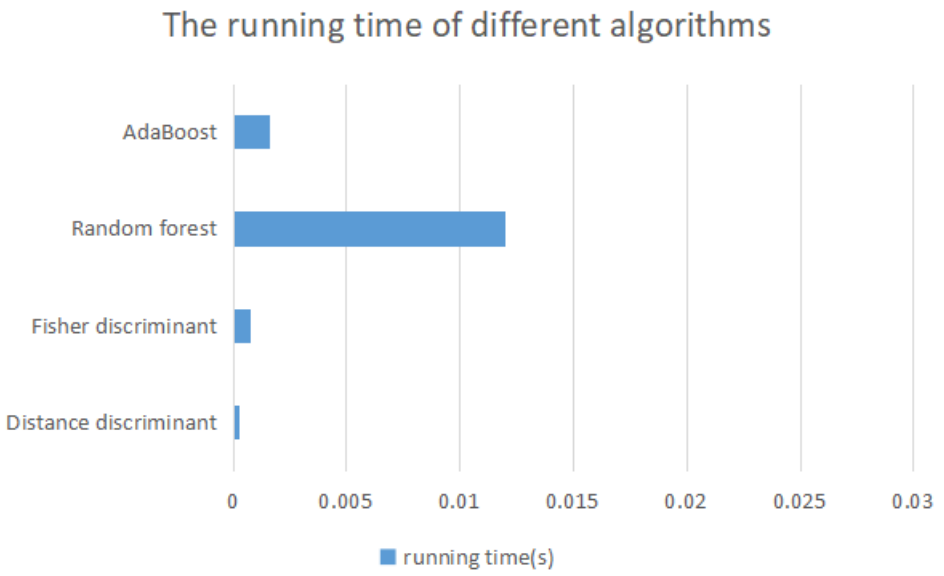

Figure 3. Comparison of running time of several methods.

By comparing the time cost of several algorithms running on the same size sample set, it can be found that the running time is sorted as follows: Random Forest > AdaBoost >Fisher discriminant > Distance discriminant. Random Forest algorithm takes the longest time, because it can accurately run on high_dimensional large data set. However, the space and time occupied by the model training through the training set will increase with the increase of the number of decision trees. In this experiment, the total data set isn't very large, so the time consumed by AdaBoost and other algorithms is relatively small compared with the Random Forest algorithm [10].

In the remaining three algorithms, Distance discriminant is the fastest because of the smallest computation required and it is the same for Fisher discriminant. Generally, AdaBoost costs more time, but in this experiment, it is not obvious due to the small data set.

4. Conclusion

In this paper, linear, non-linear and other machine learning classification algorithm were used in different combinations of COVID-19 test cases, and the idea of cross validation was applied to realize the classification of multi-index results of novel coronavirus test samples. In this study, multiple test indicators of the novel coronavirus test results were analyzed to determine whether a case was infected or not. In the process of testing, this study believes that the AdaBoost model can accurately determine whether a patient is infected with the novel coronavirus through a serious of case indicators in most case, which is a more accurate and rapid classification method. On the other hand, Random Forest algorithm can also achieve similar classification accuracy as AdaBoost, but it cannot work effectively as AdaBoost in data sets with a large amount of data. Therefore, it is more suitable for use in data sets with a large amount of data.

In future studies, algorithms combining various classification algorithms such as AdaBoost and Random Forest will be considered for processing and experiments on data sets with large data and more abundant indicators. In addition, the combination of images, image feature recognition and detection can be applied to the classification and judgment of more medical detection results.

References

[1] Haopeng Li. Intelligent robot exploration based on machine learning method. 2019, Telecom World, **26(4)**:241-242.

DOI: 10.54254/2755-2721/15/20230834

[2] Hui Xie, Zuliang Deng. Pancreatic cancer patients survival Value analysis of machine-learn-based immune cell infiltration classification model in predicting survival. 2021 *Journal of Xiangnan University (Medical Sciences)*, **23(04)**:19-27.
[3] Yixiao Zhai. Classification of antioxidant proteins based on machine learning and sequence information. 2022 *Northeast Forestry University*.
[4] Bin Tian,Hui Yu,Jigang Ren et al. Effectiveness of multiple classification models based on machine learning in the differential diagnosis of novel coronavirus pneumonia and community acquired pneumonia. 2021, *Radiologic Practice*, **36(05)**:590-595.
[5] DeeGLMath. Fisher linear discriminant analysis. https://blog.csdn.net/linjing_zyq/ article/details/ 120515566
[6] Mi Tu. Random Forest algorithm. https://blog.csdn.net/m0_46926492/article/details/122798056
[7] VernonJsn. Decision tree algorithm. https://blog.csdn.net/qingxiao__123456789/article/ details/122530376
[8] Xiu lian zhi lu. Detailed introduction about the AdaBoost algorithm https://blog.csdn. net/sinat_29957455/article/details/79810188
[9] Yoav Freund, Robert E Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. 1997, *Journal of Computer and System Sciences*. **55, 1** 119-139.
[10] Jiashilin. Advantages and disadvantages of Random Forests. https://blog.csdn.net /qq_35290785/article/details/100561148