Proceedings of the 2nd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/32/20240846

Comparison and analysis of the effect of various machine learning algorithms on cancer prediction

Zonghao Wang[1], Dan Wang[1], Guomin Si[2, 3]

[1]College of Traditional Chinese Medicine, Shandong University of Traditional Chinese Medicine, Jinan, 250355, China.
[2]Department of Traditional Chinese Medicine, Provincial Hospital Affiliated to Shandong First Medical University, Jinan, 250355, China

[3]sgm977@126.com

Abstract. Cancer is a serious disease that can have a big impact on the physical and mental health of patients. The incidence and mortality rate of cancer are increasing globally. Therefore, predicting the occurrence and treatment effect of cancer has become a hot research topic in the medical field. Machine learning algorithms can use vast amounts of data and algorithmic models to predict aspects of cancer occurrence, progression, and treatment effectiveness. This algorithm can learn and find patterns from large amounts of medical data, thereby improving the diagnosis and treatment of cancer. In this paper, naive Bayes model, logistic regression model, random forest model, KNN model and Ridge regression model were used to predict the data of cancer patients. The accuracy of naive Bayes model and logistic regression model is the highest, reaching 95%. The accuracy of random forest model and KNN model is 94%. The worst performing model was Ridge regression, with an accuracy of 93%. Naive Bayes algorithm and logistic regression algorithm perform well in solving binary classification problems because they are both probability-based classification algorithms. In this problem, the number of features is large, so the ridge regression algorithm can effectively process the data, but due to the small amount of data, it is easy to appear underfitting, so the performance is poor. In this paper, a variety of machine learning algorithms are used to predict cancer patients, and the accuracy, accuracy, recall rate and f1 parameters of each model are calculated and compared, which provides a basis for subsequent research.

Keywords: Machine learning algorithms, Cancer prediction, KNN.

1. Introduction

Cancer is a serious disease that can have a significant impact on the physical and mental health of patients [1]. The incidence and mortality rate of cancer are increasing globally. Therefore, predicting the occurrence and treatment effect of cancer has become a research hotspot in the medical field [2,3].

Machine learning algorithms can use a large amount of data and algorithm models to predict the occurrence, development and treatment effect of cancer [4]. This algorithm can learn and find patterns from large amounts of medical data, thereby improving the diagnosis and treatment of cancer. Machine learning algorithms have made some achievements in predicting cancer, but there are still many challenges and problems to be solved [5,6].

DOI: 10.54254/2753-8818/32/20240846

Research on machine learning algorithms in predicting cancer has achieved certain results [7]. For example, researchers can use support vector machine (SVM) algorithms to predict the occurrence and progression of breast cancer. This algorithm can learn and find rules from a large number of clinical data and imaging data, thus improving the diagnosis and treatment of breast cancer [8]. Researchers can use Random Forest algorithms to predict the onset and progression of liver cancer. This algorithm can learn and find rules from a large number of clinical data and imaging data, thus improving the diagnosis and treatment of liver cancer [9]. Researchers can use convolutional neural network algorithms to predict the onset and progression of lung cancer. This algorithm can learn and find rules from a large number of clinical and imaging data, thereby improving the diagnosis and treatment of lung cancer [10]. This algorithm can use a large amount of data and algorithmic models to predict aspects of cancer occurrence, development, and treatment effects. In the medical field, machine learning algorithms have been widely used to improve the diagnosis and treatment of cancer.

2. Data sets and machine learning algorithms

2.1. Decision tree regression model

Health is an essential aspect of everyone's life. Cancer is found in men or women when cells in the body begin to grow out of control. These cells often form tumors that can be felt or seen on X-rays. Cancer can be classified as benign or malignant. This article selects the kaggle competitions of public data sets (https://www.kaggle.com/datasets/vijayaadithyanvg/breast-cancer-prediction). Patient data contains multiple input features: Sample code number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal The output characteristics of Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses were Class. The class is divided into non-cancerous and cancerous.

2.2. Naive Bayes algorithm

Naive Bayes algorithm is a classification algorithm based on Bayes theorem. Its basic idea is to use training samples of known categories to estimate the conditional probability between each feature and each category, and then calculate the probability that the samples to be classified belong to each category according to Bayes' theorem, and finally classify the samples to be classified into the category with the greatest probability.

Specifically, the naive Bayes algorithm assumes that each feature is independent from other features, that is, the naive Bayes algorithm believes that each feature has an independent influence on the classification result. In practical applications, naive Bayes algorithm is often used in text classification, spam filtering, sentiment analysis and other fields.

$$JB = \frac{n}{6}\left[S^2 + \frac{(K-3)^2}{4}\right] \tag{1}$$

The steps of naive Bayes algorithm are as follows: First, the training data is collected and the category of each sample is marked; For each feature, the conditional probability under each category is calculated; Calculate the prior probability for each class, i.e. P(class); For the samples to be classified, calculate the probability that they belong to each class, that is, P(class | samples to be classified); Finally, the samples to be classified are classified into the category with the greatest probability. The advantages of naive Bayes algorithm are simple, fast and easy to implement, and it is suitable for multiple classification problems. However, the assumption that the features are independent of each other is often not valid in practical application, which may affect the accuracy of classification results.

2.3. Ridge regression algorithm

Ridge regression is a linear regression algorithm, which aims to avoid the problem of overfitting the model by adding a L2 regularization term to the parameters of the model. In ridge regression, we limit the complexity of the model by introducing a regularization term to better fit the data set. The

DOI: 10.54254/2753-8818/32/20240846

regularization term of ridge regression is controlled by a free parameter alpha, and the larger the parameter, the greater the regularization term's influence and the less complex the model.

The advantage of ridge regression is that it can process data efficiently in the presence of multicollinearity. Multicollinearity refers to the situation where there is a high correlation between multiple independent variables in the data set, which leads to the instability of the parameter estimation of the model, and the regularization term of ridge regression can effectively reduce this instability.

Ridge regression algorithm can be realized through the following steps: Firstly, the data is preprocessed, including data cleaning, feature selection, feature scaling, etc.; The data set is divided into training set and test set. The model parameters were obtained by ridge regression fitting to the training set. Use test sets to evaluate the performance of the model, typically measured using mean square error (MSE) or R2; Finally, if the performance of the model is not good enough, you can try to adjust the alpha value and re-fit and evaluate.

2.4. Random forest algorithm

Random forest is an ensemble learning algorithm based on decision trees, which integrates the results of multiple decision trees to improve the accuracy and robustness of the model. The main feature of a random forest is the introduction of both randomness and diversity, which allows it to effectively avoid overfitting problems and perform well on complex data sets.

The algorithm of random forest can be implemented through the following steps: Firstly, a part of samples and features are randomly selected from the original data set to build a decision tree; Repeat the above steps to build multiple decision trees; For new samples, each decision tree is used to make predictions and the results are integrated, for example by voting or averaging to get the final prediction result.

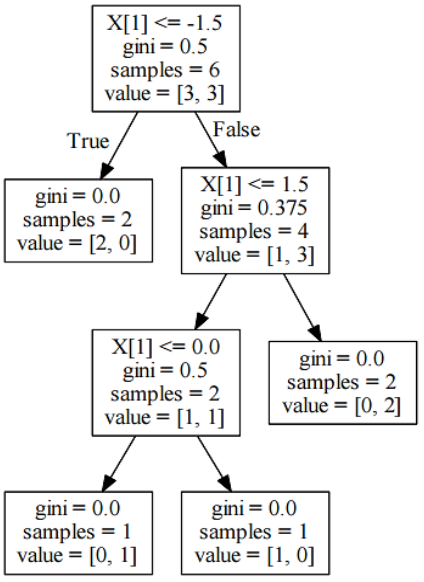

Figure 1. Decision tree algorithm schematic.(Photo credit : Original)

The randomness of random forest is reflected in two aspects: First, it randomly selects some features in the construction process of each decision tree, so as to avoid the excessive influence of some features on the model; Secondly, it randomly selects a part of the samples in the training set of each decision tree, which can avoid the overfitting of the model to specific samples.

The advantage of random forest is that it has good robustness and accuracy, and can handle high-dimensional, large-scale data sets. In addition, it can also be used for problems such as feature selection, anomaly detection and clustering. However, the disadvantage of a random forest is that it is not as model-explanatory as a single decision tree, and in some cases, overfitting problems may arise. Random forest

DOI: 10.54254/2753-8818/32/20240846

is a powerful ensemble learning algorithm, which can effectively improve the accuracy and robustness of the model. It has been widely used in practical applications, especially in classification and regression problems.

2.5. Logistic regression algorithm

Logistic regression is a widely used classification algorithm, which can be used for binary classification and multi-classification problems. The goal of logistic regression is to predict the probability of a binary output variable given the input characteristics. Logistic regression is a linear classification algorithm that maps a linear combination of input features to a range of [0,1] via a sigmoid function, representing the probability that the output variable is 1.

The algorithm of logistic regression can be realized through the following steps: Firstly, the data is preprocessed, including data cleaning, feature selection, feature scaling, etc.; The data set is divided into training set and test set. The model parameters were obtained by logistic regression fitting to the training set. Use test sets to evaluate the performance of the model, typically measured using metrics such as accuracy, accuracy, recall, and F1-score; Finally, if the performance of the model is not good enough, you can try to adjust the hyperparameters of the model, such as regularization parameters and learning rate. The advantage of logistic regression is that it is simple, fast, easy to implement and explain. In addition, it can be used for problems such as feature selection and model interpretation. The disadvantage of logistic regression is that it is a linear classification algorithm and may not perform well for data with non-linear relationships. In addition, the performance of logistic regression suffers in the presence of class imbalances. Logistic regression is a common classification algorithm, which can be used for binary classification and multi-classification problems. It is simple, fast, easy to implement and interpret, and is suitable for processing small and medium-sized data sets.

2.6. KNN algorithm

KNN algorithm is an instance-based machine learning algorithm that can be used for classification and regression problems. The basic idea of KNN algorithm is that for a new sample, find K samples that are most similar to it in the training set, and then predict the label of the sample according to the label of the K samples.

The KNN algorithm can be realized through the following steps: firstly, the data is preprocessed, including data cleaning, feature selection, feature scaling, etc.; The data set is divided into training set and test set. For a new sample, K samples that are most similar to it are found in the training set. Predict the label of the sample according to the label of the K samples, for example, get the final prediction result by voting or average; A test set is used to evaluate the performance of the model, typically measured using metrics such as accuracy, accuracy, recall, and F1-score.

The advantage of KNN algorithm is that it is simple, easy to understand and implement. In addition, the KNN algorithm is suitable for dealing with data with nonlinear relations, and can handle multiple classification problems, and it does not need to make assumptions about the data or perform parameter estimation, so it does not require assumptions about the distribution of data.

The disadvantage of KNN algorithm is that it is sensitive to noise and outliers in the training set, and it needs to calculate the distance between each test sample and all training samples, so the computational complexity is high when dealing with large-scale data sets. KNN algorithm is a simple, easy to understand and implement machine learning algorithm, which is suitable for dealing with nonlinear relationship data and multi-classification problems. It has been widely used in practical applications, such as image recognition, speech recognition, recommendation system and so on.

DOI: 10.54254/2753-8818/32/20240846

3. Comparison of prediction results of multiple machine learning algorithms

Table 1. Model evaluation.

model	Traning Accuracy	Testing Accuracy
Naive Bias	96%	95%
Ridge Classifier	93%	93%
Random Forest	100%	94%
Logistic Regression	97%	95%
K Nearest Neighbour	97%	94%

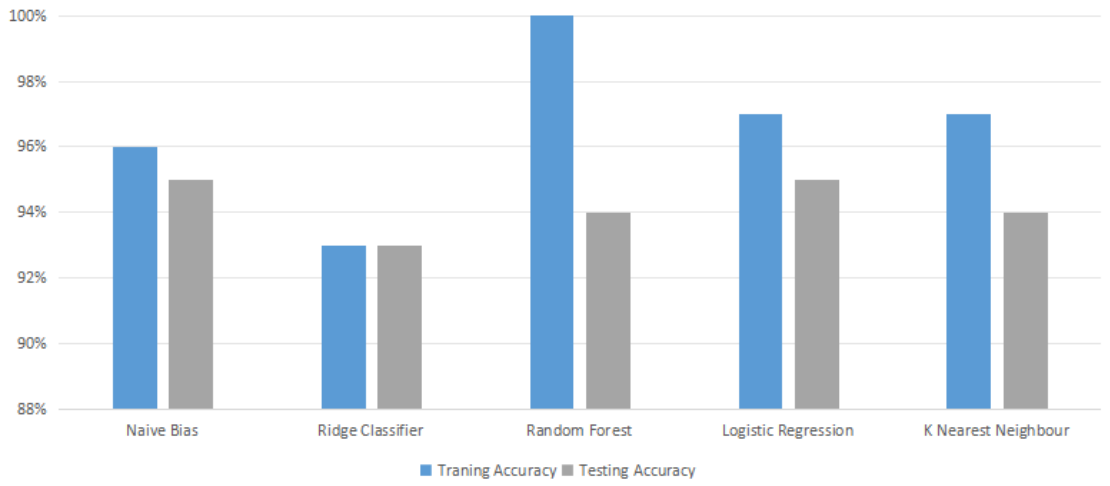

Figure 2. Dataset image.(Photo credit : Original)

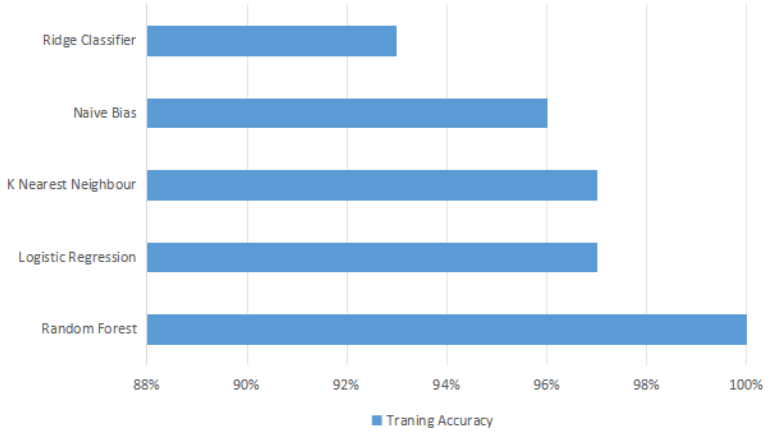

Figure 3. Accuracy.(Photo credit : Original)

As can be seen from the above figure, in the actual prediction effect of the test set, the accuracy of naive Bayes model and logistic regression model is the highest, reaching 95%; The accuracy of random forest model and KNN model is 94%. The worst performing model was Ridge regression, with an accuracy of 93%.

4. Conclusion

In this paper, naive Bayes model, logistic regression model, random forest model, KNN model and Ridge regression model were used to predict the data of cancer patients.

DOI: 10.54254/2753-8818/32/20240846

The accuracy of naive Bayes model and logistic regression model is the highest, reaching 95%. The accuracy of random forest model and KNN model is 94%. The worst performing model was Ridge regression, with an accuracy of 93%.

Naive Bayes algorithm and logistic regression algorithm perform well in solving binary classification problems because they are both probability-based classification algorithms. In this problem, the two algorithms may be affected by the correlation between features, but this effect is minimized due to the large amount of data, so the two algorithms perform better. Both random forest algorithm and KNN algorithm are decision tree-based classification algorithms, which can deal with nonlinear problems and perform well when dealing with a large number of features. In this problem, there may be complex nonlinear relationships between features, so these two algorithms perform better. Ridge regression algorithm is a linear regression algorithm for processing high-dimensional data, which prevents overfitting by regularizing the parameters. In this problem, the number of features is large, so the ridge regression algorithm can effectively process the data, but due to the small amount of data, it is easy to appear underfitting, so the performance is poor.

References

[1] Erkan F M M T C S K E A S S .The Value of Prostate-Specific Antigen Density in Combination with Lesion Diameter for the Accuracy of Prostate Cancer Prediction in Prostate Imaging-Reporting and Data System 3 Prostate Lesions.[J].Urologia internationalis,2023,1-6.

[2] Expression of Concern: Eysenck, H. J. (1988). Personality, stress and cancer: Prediction and prophylaxis. British Journal of Medical Psychology, 61(1), 57-75. https: //doi.org/10.1111/j. 2044-8341.1988.tb02765.x.[J].Psychology and psychotherapy,2023,

[3] Xin W W .Prostate cancer prediction model: A retrospective analysis based on machine learning using the MIMIC-IV database[J].Intelligent Pharmacy,2023,1(4):268-273.

[4] I C M J N B G J N L M S P K A O M C S .Artificial Intelligence-Driven Mammography-Based Future Breast Cancer Risk Prediction: A Systematic Review.[J].Journal of the American College of Radiology : JACR,2023,

[5] New research on AI cancer prediction platform reveals groundbreaking 93% accuracy rate[J].M2 Presswire,2023,

[6] Jin Z Y J L L .Nomogram based on multiparametric analysis of early-stage breast cancer: Prediction of high burden metastatic axillary lymph nodes.[J].Thoracic cancer,2023,

[7] DongWon S C L .CNN-Based Inspection Module for Liquid Carton Recycling by the Reverse Vending Machine[J].Sustainability,2022,14(22):14905-14905.

[8] Yunan X S L W .iPCa-Net: A CNN-based framework for predicting incidental prostate cancer using multiparametric MRI.[J].Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society,2023,110102309-102309.

[9] Janez S M R J B P T O .Breast cancer risk prediction using Tyrer-Cuzick algorithm with an 18-SNPs polygenic risk score in a European population with below-average breast cancer incidence.[J].Breast (Edinburgh, Scotland),2023,72103590-103590.

[10] Cai T M Z W C H .Breast Cancer Prediction Based on Differential Privacy and Logistic Regression Optimization Model[J].Applied Sciences,2023,13(19):