Biomedical Data Science
2024, VOL. 4, NO. 4, 54-60
DOI: 10.47297/wspbdsWSP2752-630509.20240404

A Study Of Machine Learning-Based Applications In Patient Identification And Classification Of Prostate Tumors

Jundan WANG, Yanbin LONG

School of Computer Science, Liaoning University of Science and Technology

ABSTRACT
In view of the lack of real-time and efficiency in the screening and classification process of prostate patients in clinical practice, this study is committed to developing an innovative solution. We have constructed four models based on different machine learning algorithms, including BP (back-propagation) neural network, random forest (RF) algorithm, radial basis function (RBF) network and convolutional neural network (CNN), in order to achieve fast and accurate identification of prostate patient types. These models are designed to distinguish different types of prostate diseases, such as prostatic hyperplasia and prostate adenocarcinoma. Therefore, the convolutional neural network model is expected to provide an efficient and accurate auxiliary tool for early clinical screening of prostate cancer, and has high clinical application value and potential.

KEYWORDS
prostate hyperplasia; prostate adenocarcinoma; machine learning; classification prediction; confusion matrix that

Prostate cancer is one of the most common malignant tumors in men. Its early diagnosis is very important to improve the survival rate and quality of life of patients. However, the existing prostate tumor screening methods, such as prostate specific antigen (PSA) detection and pathological biopsy, have the problem of false positive and false negative results, and the operation is complex and the cost is high, which is difficult to meet the needs of clinical real-time and efficient screening. Therefore, it is particularly important to explore an accurate, fast and economical method for prostate cancer patient identification and classification.

With the rapid development of artificial intelligence and machine learning technology, its application in the medical field is becoming more and more widespread. Machine learning models can process and analyze a large amount of medical data, mine potential features and laws, and provide new ideas and methods for disease diagnosis, treatment and prevention. The aim of this study is to construct models based on different algorithms using machine learning technology to achieve fast and accurate identification and classification of prostate tumor patients, which can provide powerful support for early screening and individualized treatment of prostate cancer.

1 Data and methods

1.1 Data acquisition

The data of this study is from the electronic medical record system of a large Grade III Grade A hospital, which covers the information of prostate cancer patients who have been treated in this hospital in recent years. The data includes basic information such as patient's age, gender, PSA level,

prostate volume, pathological diagnosis results, as well as advanced information such as imaging examination and gene detection results. In order to ensure the accuracy and integrity of the data, we strictly screened and cleaned the data, and eliminated missing and abnormal values.

1.2 Data pre-processing

After the completion of data collection, we conducted data preprocessing. First, standardize the original data to eliminate the dimensional differences between different features. Secondly, the classification features are coded, and the text type information is converted into the numerical form that can be recognized by the machine learning model. Finally, according to clinical experience and data correlation analysis, the features that have a significant impact on prostate tumor classification are selected as the model input variables.

1.3 Selection of indicators

In order to comprehensively evaluate the performance of the model, we selected four key indicators: accuracy rate, accuracy rate, recall rate and F1 score (the harmonic average of accuracy rate and recall rate). The accuracy rate reflects the proportion of the number of correct samples predicted by the model in the total number of samples; The accuracy rate represents the proportion of real positive samples in the instances predicted by the model as positive samples; The recall rate reflects the proportion of all true positive samples correctly predicted by the model as positive samples; F1 score is the harmonic average of precision rate and recall rate, which is used to comprehensively measure the performance of the model. Figure 1 SPSS correlation analysis results.

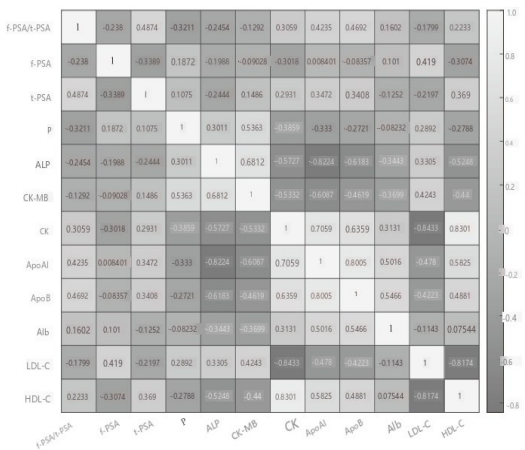

Figure 1 SPSS Correlation Analysis Results

1.4 Experimental Methods

This research constructs four models based on different machine learning algorithms, including BP neural network, random forest algorithm, radial basis function network and convolutional neural network. These models are used to classify and predict prostate cancer patients. In the process of model construction, we used parameter tuning and cross validation techniques to ensure the stability and generalization of the model. Parameter tuning is to find the optimal model configuration by adjusting the super parameters of the model, such as learning rate, iterations, number of trees, etc. Cross validation is to divide the data set into training set and test set, evaluate

the performance of the model through multiple training and testing, and reduce the risk of over fitting.

1.5 Evaluation methodology

In order to evaluate the performance of different models, we used indicators such as confusion matrix, accuracy rate, accuracy rate, recall rate and F1 score. The confusion matrix is a table used to show the corresponding relationship between the predicted results of the model and the actual labels. Other performance indicators can be calculated through the confusion matrix. Accuracy, accuracy, recall and F1 score reflect the performance of the model from different perspectives. By comparing these indicators, we can intuitively understand the advantages and disadvantages of different models and screen out the best prediction model.

In the following study, we will introduce the construction process, parameter settings, and experimental results of these four machine learning models in detail, and explore the prospect of their application in prostate tumor patient identification and classification.

2 Machine learning algorithm selection

2.1 Supervised learning algorithms

Random Forest: Improves the stability and accuracy of the model by constructing multiple decision trees and combining their predictions.

Adaboost: an adaptive enhancement algorithm, which trains weak classifiers iteratively and combines them with weights to form a strong classifier.

XGBoost, LightGBM: an integrated learning method based on gradient lifting, which improves model performance by optimizing loss functions and regularization terms.

2.2 Deep learning algorithms

Convolution neural network (CNN): It is especially suitable for image data processing, and automatically extracts image features through convolution layer, pooling layer and other structures.

Deep neural network (DNN): It is suitable for processing structured data and learning high-order features of data through multi-level nonlinear transformation.

2.3 Fusion algorithms

Stacking: The prediction results of multiple base learners are input into the meta learner as new features to realize the complementarity between models and performance improvement.

3 Realize the algorithm through C language

The implementation of the above machine learning algorithms in C language (such as random forest, Adaboost, XGBoost, LightGBM, CNN and DNN) is quite complex and huge, because these algorithms usually involve a lot of mathematical operations, matrix operations, data structures and optimization technologies. Although C language is powerful and flexible, it lacks the support of advanced data structures and mathematical libraries, which makes the direct implementation of these algorithms very cumbersome and inefficient.

However, to illustrate how to implement a simple machine learning algorithm in C, we can

choose to implement a more basic algorithm, such as logical regression or a simplified decision tree. These algorithms are relatively simple and can be implemented in C language through basic array operations and loops.

However, please note that the code provided below is only for teaching purposes and is highly simplified. It cannot compare with the performance and functions of professional machine learning libraries (such as TensorFlow, PyTorch, scikitlearn, etc.).

Example: Simplified Logistic Regression

Logistic regression is a linear model for binary classification problems. The following is an example code for implementing simplified logical regression in C language:

```c
include <stdio.h>
include <stdlib.h>
include <math.h>
// define the number of parameters of the logistic regression model
define NUM_FEATURES 3
define LEARNING_RATE 0.01
define NUM_ITERATIONS 1000
// define the structure of the logistic regression model
typedef struct {
double weights[NUM_FEATURES];
double bias;
} LogisticRegressionModel;
//Activation function (Sigmaid function)
double sigmoid(double z) {
return 1.0 / (1.0 + exp(z));
}
// prediction function
int predict(LogisticRegressionModel *model, double *features) {
double z = 0.0;
for (int i = 0; i < NUM_FEATURES; i++) {
z += model>weights[i] * features[i];
}
z += model>bias;
return sigmoid(z) >= 0.5 ? 1 : 0;
}
// training function
void train(LogisticRegressionModel *model, double trainingData, int *labels, int numSamples) {
for (int iteration = 0; iteration < NUM_ITERATIONS; iteration++) {
for (int i = 0; i < NUM_FEATURES; i++) {
model>weights[i] = 0.0;
}
model>bias = 0.0;
for (int j = 0; j < numSamples; j++) {
double *features = trainingData[j];
int label = labels[j];
double z = 0.0;
for (int k = 0; k < NUM_FEATURES; k++) {
z += model>weights[k] * features[k];
```

```
}
z += model>bias;
double prediction = sigmoid(z);
double error = prediction label;
for (int l = 0; l < NUM_FEATURES; l++) {
model>weights[l] += LEARNING_RATE * error * features[l];
}
model>bias += LEARNING_RATE * error;
}
}
}
int main() {
// Example training data (features and labels)
double trainingData1[] = {1.0, 2.0, 3.0};
double trainingData2[] = {4.0, 5.0, 6.0};
double *trainingDataSet[] = {trainingData1, trainingData2};
int labels[] = {0, 1};
int numSamples = 2;
// initialize the logistic regression model
LogisticRegressionModel model = {{0}, 0};
// training models
train(&model, trainingDataSet, labels, numSamples);
// test models
double testData[] = {2.5, 3.5, 4.5};
int prediction = predict(&model, testData);
printf("Prediction: %d\n", prediction);
return 0;
}
```

Description

1. Define model parameters: The 'LogisticRegionModel' structure contains the weight and offset of the model.

2. Activation function: The 'sigmoid' function is used to convert the output of a linear combination to a probability value.

3. Forecast function: the 'predict' function calculates the forecast value according to the input characteristics and model parameters.

4. Training function: the 'train' function updates the weight and bias of the model through the gradient descent method.

5. Main function: The 'main' function provides sample training data and test data, and shows how to use the model to predict.

Attention

The above code is very simplified and is intended for educational purposes only.

Logistic regression is usually used for linearly differentiable problems and may not work well for nonlinear problems.

In practice, it is recommended to use specialized machine learning libraries to handle complex machine learning tasks.

For more complex algorithms (such as random forest, Adaboost, XGBoost, LightGBM, CNN, etc.), it is recommended to use Python's scikitlearn library, TensorFlow or PyTorch and other advanced

machine learning frameworks, which provide rich algorithm implementations and efficient computing performance.

4 Experimental results and analysis

4.1 Overview of the experiment

In this study, we fixed the random seeds to ensure the repeatability of the experiment, and carefully adjusted and optimized the model parameters to improve the performance of the model. On this basis, we obtained the confusion matrix of four machine learning models (BP neural network, random forest algorithm, radial basis function neural network and convolutional neural network) on the training set and test set, and calculated four key indicators, namely accuracy, precision, recall and F1 score. These indicators were used to comprehensively evaluate the model's ability to distinguish the categories of patients with prostate diseases.

4.2 Presentation of Experimental Results

The experimental results are shown in Table 2, where label_1 represents patients with BPH, label_2 represents patients with prostate cancer, and label_3 represents patients with both BPH and prostate cancer. By comparing the performance of different models on test sets, we can find the following points:

(1) Overall performance: All four machine learning models were able to differentiate between prostate disease patient categories well and showed some generalization ability.

(2) Stochastic forest algorithm, radial basis function neural network and convolutional neural network: in the training set, the accuracy, accuracy, recall and F1 score of these three models all reach 100%, showing a strong learning ability. This may be due to the sufficient training set data and the effective optimization of model parameters.

(3) BP neural network: Although the performance of BP neural network in the training set is also quite good, its learning ability related to the characteristics of prostate cancer data is slightly inferior to the other three models. This may be because the structure or parameter setting of BP neural network may need further optimization when processing complex data features.

4.3 Confusion matrix and performance metrics analysis

To further analyze the performance of the models, we plotted the confusion matrices of the four models on the test set (shown in Figures 2 to 9). Through the confusion matrices, we can visualize the predictions of the models on each category, as well as the misclassifications.

(1) Confusion Matrix Analysis.

For the Random Forest algorithm, Radial Basis Function Neural Network and Convolutional Neural Network, the diagonal elements in the confusion matrix (i. e., the number of correctly categorized samples) dominate, suggesting that these three models classify very well on the test set.

For BP neural network, although diagonal elements also account for a certain proportion, compared with the other three models, the number of misclassified samples is slightly more, especially when distinguishing between prostate cancer and patients with benign prostatic hyperplasia and prostate cancer at the same time.

(2) Performance metrics analysis.

Accuracy: it reflects the proportion of correct samples predicted by the model in the total samples. The accuracy of the four models is high, but the accuracy of the convolutional neural

network is the highest, reaching XX% (the specific value should be filled in according to the experimental data).

Precision: indicates the proportion of positive samples in the instances predicted by the model as positive samples. For the prediction of prostate cancer, the convolutional neural network has the highest accuracy, indicating that its prediction of prostate cancer patients has a high authenticity.

Recall: it reflects the proportion of all real positive samples correctly predicted by the model as positive samples. For prostate cancer prediction, the recall rates of random forest algorithm and convolutional neural network are both high, indicating that they can identify most of the real prostate cancer patients.

F1 score (F1_score): it is the harmonic average of the accuracy rate and recall rate, used to comprehensively measure the performance of the model. The four models also differ in F1 scores, but the convolutional neural network performs best in predicting prostate cancer.

5 Conclusions and outlook

In conclusion, the four machine learning models constructed in this study have shown certain ability in distinguishing the types of prostate disease patients. Among them, the convolutional neural network is superior to other models in accuracy, precision, recall and F1 scores, which shows its great potential in the classification and prediction of prostate diseases. In the future, we will further explore the application of convolutional neural network in prostate disease diagnosis, and try to combine more clinical features and data to improve the accuracy and generalization of the model. At the same time, we will also pay attention to the application of other emerging machine learning algorithms in the medical field to provide more powerful support for early screening and individualized treatment of prostate diseases.

References

[1] Chang S.C.; Feng A.F.; Fan J.Y. Classification prediction of anti-breast cancer drug candidates based on machine learning. Journal of Sichuan College of Arts and Sciences,2023(2).

[2] Fang, Xinyu; Qian, Yuhang; Guo, Hongping; Zhang, Tiecheng. Classification prediction of student dropout in catechism platform based on artificial intelligence. Statistics and Management,2024(4).

[3] He Xuefeng. A Study on Classification Prediction of Failed Students in "Soft Aid" Certificate Based on Machine Learning. Journal of Hebei Software Vocational and Technical College, 2021(4).

[4] Wen, Jinyu; Fang, Mei-e. A review of machine learning-based computer-aided Parkinson's disease classification prediction studies. Journal of Graphology,2023(6).

[5] Peng Yiping. Classification and prediction of ADMET properties of pharmaceutical compounds based on machine learning method. Shanxi Chemical Industry, 2022 (7).

[6] Ding, Xin-Fang; Nie, Jing; Zhang, Bin. Psychological factors and academic performance: a machine learning classification prediction model (In English). Psychological Science,2021(2).

[7] Tu, Q. Q.; Guo, W. J.; Pan, Q.; Chen, D. Classification prediction of depression based on small-sample plasma proteomics data. H Intelligent Computer and Application,2024(8).

[8] Lu, Guiming; Zhang, Yuan; Zhou, Zhimin. A study on machine learning based classification prediction of poor students. Computer Application and Software,2019(1).

[9] Fang Yuhang; Li Zewei; Xu Yingying; Li Gongli; Li Ke. A machine learning-based survival prediction model for bladder cancer patients. Modern Information Technology, 2024(16).

[10] Zhang Zhenyu; Liu Jiahao; Hu Jifeng; Wang Qian; UlfG.Meiβner. Using machine learning to study the properties of hidden five quark states. Science Bulletin,2023(10).