Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 12, No. 2, 2024

A boosting-based prediction model for disease progression

Yixuan Gao[1], Liushuang Hou[1], Yonghao Gao[1], Meng An[1], Wenbao Xiao[2]

[1] School of Biological Science, Qufu Normal University, 273165, China
[2] School of Computer Science, Qufu Normal University ,276626, China

Abstract: With the intensification of the aging population in China, the number of patients with Parkinson's disease is showing a high growth trend. At present, the diagnosis of Parkinson's disease mainly relies on clinical assessment scales, but there are shortcomings such as poor accuracy, large scoring span and poor timeliness. Therefore, based on boosting, this paper uses a cross-validation algorithm to select the best model among 14 machine learning models and constructs the Multimode Parkinson model. We validate the performance of our model in predicting the extent of Parkinson's disease. Experimental results show that this method is superior to traditional algorithms and can be effectively applied to the treatment of Parkinson's disease.

Keywords: Bioinformatics; Machine learning; Boosting.

1. Introduction

Parkinson's disease, also known as tremor paralysis, is a chronic neurodegenerative disease characterized by neuronal degeneration and dopamine neuron loss. Commonly seen in the middle-aged and elderly population, the average age of onset is 60 years old, with males slightly higher than females [1]. It usually leads to motor disorders, muscle stiffness, tremors, and other non motor symptoms, such as cognitive impairment and autonomic dysfunction.

According to experts from the World Parkinson's Organization, the current global number of Parkinson's disease patients is as high as 5.8 million, with China accounting for about half, and the annual number of new cases exceeds 100000. It is predicted that by 2030, there will be 8.7 million Parkinson's disease patients worldwide, with China accounting for about 5 million, making it the largest country in Parkinson's disease. Moreover, with the changes in social environment pressure, the disease is gradually moving towards youthfulness. In recent years, the number of Parkinson's disease patients under the age of 40 has been continuously increasing. However, due to the unclear exact cause of the disease, it is evident that Parkinson's disease may become another major disease in China. Therefore, research and diagnosis of Parkinson's disease have become increasingly urgent [3,4].

In recent years, machine learning has shown excellent performance in processing large amounts of data, attracting widespread attention from various industries [5], [6], [7]. These data-driven processing technologies, with their novel and efficient processing methods, have great appeal for data processing tasks. Multiple machine learning based methods have been proposed for the diagnosis of Parkinson's disease. The mainstream prediction and diagnostic models can usually be divided into three categories: diagnosis based on language data, diagnosis based on clinical data, and diagnosis based on medical imaging. Almeida et al. [8] employed multiple feature extraction and machine learning strategies to detect Parkinson's disease. They found that vocalization is the most effective activity for detecting Parkinson's disease. This study used Multi Layer Perceptron (MLP), Optimal Path Forest, and Support Vector Machine (SVM) as classifiers. Kim et al. [9] used three cognitive screening methods to predict cognitive progression in Parkinson's disease. This literature compared the diagnostic accuracy of three common screening methods: the Montreal Cognitive Assessment (MoCA), the Dementia Rating Scale-2 (DRS-2), and the Mini Mental State Examination (MMSE). Use logistic regression and Cox proportional hazards regression models to test the predictive factors for cognitive decline. This study for the first time indicates that MoCA is a predictive factor for the transition to Parkinson's dementia (PDD). MoCA has been shown to be sensitive to detecting Parkinson's disease-related cognitive impairment (PD-MCI) and progression of Parkinson's disease, but its low specificity reduces its practicality as a diagnostic test. The heterogeneity of Parkinson's disease progression and the lack of objective biomarkers make the clinical trial design and results of Parkinson's disease complex [10]. Therefore, better models or strategies for Parkinson's disease progression are needed to select specific clinical trials.

Previous research has mainly focused on using methods based on language data and clinical data to predict and diagnose Parkinson's disease. However, compared to diagnosing Parkinson's disease based on abnormal information of Parkinson's related proteins and peptides, these methods have the following four significant issues that cannot be ignored. Firstly, language expression can be influenced by individual differences, emotional states, and environmental factors, so prediction results may be subjectively influenced, making it difficult to fully reflect the patient's physiological condition and disease progression. Secondly, the clinical manifestations of Parkinson's disease are highly heterogeneous, and the progression of the disease is complex. Therefore, methods based on clinical data often struggle to capture subtle changes in disease progression. Thirdly, medical imaging technology is often expensive and requires professional knowledge and equipment, which limits its application in large-scale prediction and diagnosis. At the same time, certain medical imaging techniques require contrast agent injections, which may be inconvenient and risky for patients. Therefore, we urgently need objective and sensitive methods for detecting and predicting Parkinson's disease.

This article investigates the application of machine

learning in predicting the progression of Parkinson's disease and proposes a new prediction method. This method is based on abnormal information of Parkinson's related proteins and peptides, and uses cross validation to predict in 14 commonly used machine learning models. When predicting the severity of Parkinson's disease at four levels, we choose the model that performs best in each level as the prediction model. Subsequently, we used these models for boosting and constructed a new model called the MultiModelParkinson model. Compared with previous studies, we have noticed that there have been no studies using artificial intelligence methods to predict the progression of Parkinson's disease based on abnormal information of Parkinson's related proteins and peptides. Our study fills this research gap, making the diagnosis and prediction of the disease more accurate and reliable, enabling non-invasive diagnosis and monitoring, reducing inconvenience and risk to patients. At the same time, our strategy fully utilizes the advantages of multiple models. By comprehensively utilizing the advantages of multiple models, the MultiModelParkinson model will be able to more accurately predict the development of Parkinson's disease. To evaluate the effectiveness of our proposed model in predicting the severity of Parkinson's disease, we compared it in detail with traditional models (including SVM and MLP) and conducted ablation experiments to evaluate their respective predictive abilities. The experimental results show that the proposed method outperforms these traditional algorithms in predicting the severity of Parkinson's disease.

The rest of this article is organized as follows. The second section introduces our proposed model. The third part introduces the experimental section, including experimental details, comparative experiments, and ablation experiments. The fourth section includes conclusions and reference materials.

2. Related work

2.1. Overview of our model

Based on the abnormal information of Parkinson's related proteins and peptides, we propose a corresponding MultiModelParkinson model, as shown in Figure 1. Including preprocessing module, machine learning model selection module, and boosting module. By combining multiple superior models, the overall predictive performance can be improved. The Boosting algorithm iteratively trains multiple weak learners and combines them into a strong learner, reducing the risk of overfitting for a single model and fully utilizing their respective advantages to handle complex relationships and adversarial noise in the data. This can further improve the accuracy and generalization ability of the model and enhance its robustness. Below, we will introduce their detailed details separately.

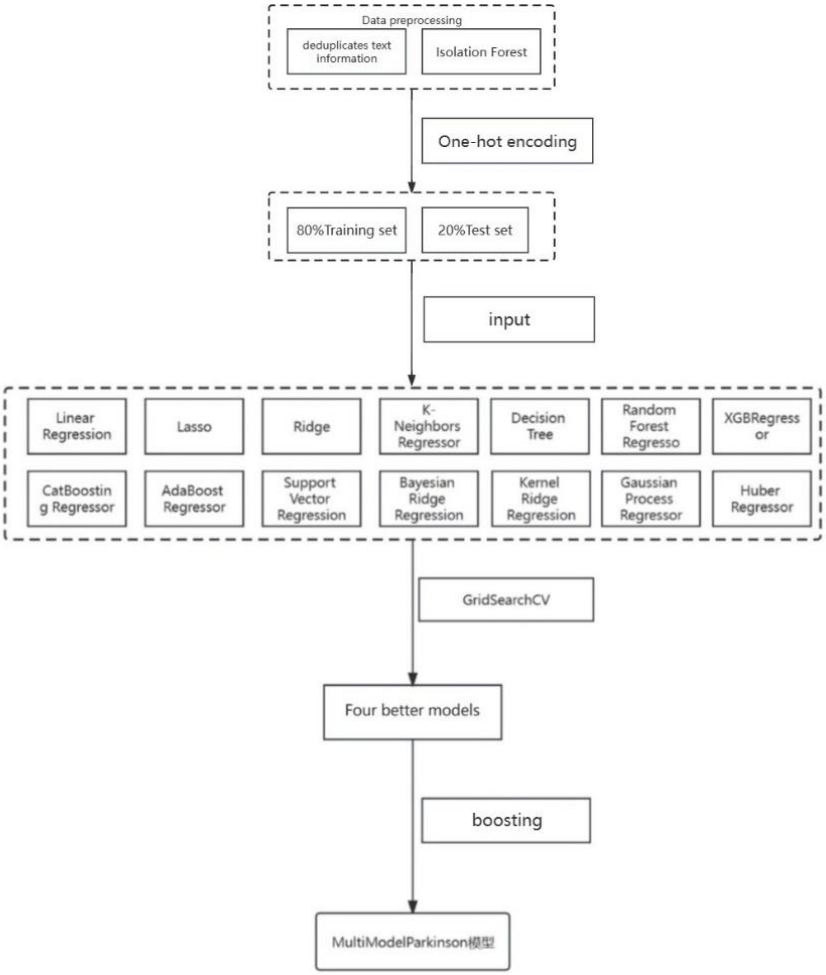

Figure 1. MultiModelParkinson model technology flowchart

2.2. B preprocessing module and machine learning model selection module

In the preprocessing module (Figure 2), first, the dataset is deduplicated. Repetitive text information will have a certain impact on the results, so use Python to deduplicate the text information. Secondly, the Isolation Forest algorithm is used to remove outliers from the dataset and select key biomarkers.

The main idea of IsolationForest is to consider normal samples as common data points, while abnormal samples have a lower density. It evaluates the degree of anomaly by measuring the degree of separation of samples in the tree, thereby identifying potential outliers. During the training process, IsolationForest will construct multiple random trees, each of which will randomly partition the data. By calculating the average number of segmentation times required for a sample in the tree, the anomaly score of the sample can be calculated. Finally, the discrete features in the dataset are transformed into numerical representations that can be processed by machine learning algorithms using unique hot encoding.

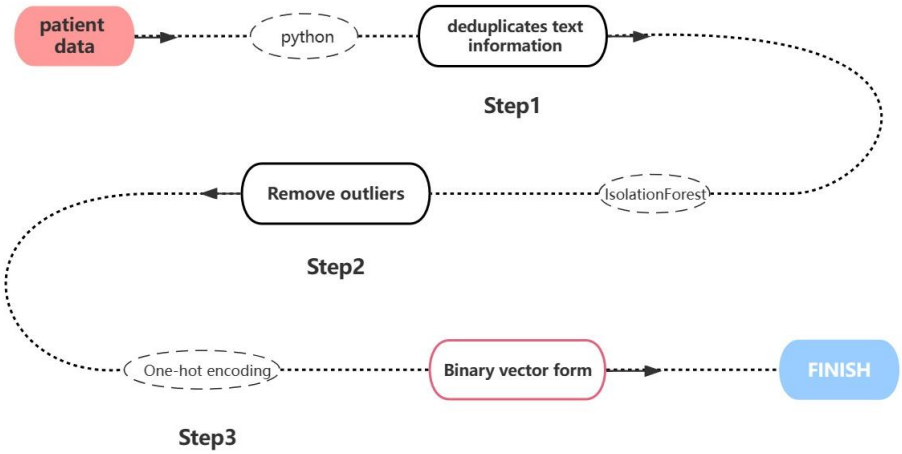

Figure 2. Data preprocessing flowchart

In the model selection module, we compared fourteen models (Linear Regression, Lasso, Ridge, K-Neighbors Regression Decision Tree, Random Forest Regression, XGBRegression, CatBoosting Regression, AdaBoost Regression, Support Vector Regression, Bayesian Ridge Regression, Kernel Ridge Regression, Gaussian Process Regression, and Huber Regression) using grid search cross validation. Based on four levels of severity of Parkinson's disease, we conducted four experiments and selected the model that performed best in each level as the prediction model.

2.3. C boosting module

Boosting is an ensemble learning method that improves model performance by iteratively training multiple weak learners and combining them into one strong learner. Boosting can significantly improve the predictive performance of the model by combining multiple weak learners. It can reduce bias, adapt to various types of data and problems, dynamically adjust the complexity of the model based on the complexity and noise level of the data, and reduce variance by focusing on samples with previous model prediction errors, thereby improving the accuracy and generalization ability of the model.

The following is the boosting process:

Step 1: Initialize weights for each sample in the training data to make them equal.

Step 2: For each iteration step, train a weak learner (usually a decision tree) and train based on the current sample weight. During the training process, weak learners will focus on samples that were previously predicted incorrectly by the model to improve its performance.

Step 3: Combine the prediction results of multiple weak learners. Using weighted voting, the weight of each model is determined by its performance during the training process.

Step 4: Output the final trained strong learner as the final model for predicting new, unseen data.

In each iteration step, the weak learner will focus on the samples that were previously predicted incorrectly by the model, thereby enhancing the model's ability to handle these samples. By iteratively training multiple models and combining them, the Boosting method can improve the predictive performance and generalization ability of the model.

3. Experimental result

In this section, we first introduced the evaluation indicators used in the experiment. Subsequently, we outlined the hyperparameter configurations used during the training process. Finally, we will apply the trained MultiModelParkinson model to the prediction task. All experiments were conducted on a PC with an Intel Core i5-12400F CPU, 16GB main memory, and NVIDIA GeForce RTX 3060 GPU.

3.1. Hyperparameter setting

When building the MultiModelParkinson model, we used some hyperparameter configurations. Specifically, we set the number of subsamples in isolated forests to a certain proportion of the dataset size, where we chose 0.5 as the sample ratio. In order to improve the stability of the model, we set up 100 isolated trees. In addition, we used 5-fold cross validation to evaluate the performance of the model. When selecting the number of weak learners, we set it between 3 and 8. Although having more weak learners may lead to better performance, it can also increase computational overhead. In the model selection module, we used the default parameters of weak learners, while in the boosting module, we chose the optimal parameter configuration. Finally, we adopted a weighted voting approach to combine the prediction results of weak learners. This configuration is expected to improve the performance and generalization ability of the MultiModelParkinson model.

In the model selection module, the measurement metrics include Mean Squared Error (MSE) and Mean Absolute Error (MAE).

Mean square error is a commonly used indicator to measure the prediction accuracy of regression models. It measures the square of the average difference between the predicted values of the model and the actual observed values. The formula is as follows.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (1)$$

Where $(Y_i - \hat{Y}_i)^2$ represents the square of the difference between the predicted value and the actual value, and n is the total number of observed values.

The average absolute error measures the average absolute difference between the predicted values of the model and the actual observed values. The formula is as follows.

$$\eta = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \quad (2)$$

Where is $|x_i - \bar{x}|$ the absolute value of the difference between the predicted value and the actual value

To evaluate the effectiveness of the MultiModelParkinson model, we chose mAP as the evaluation metric. MAP is the average value of AP, representing the average accuracy of the object detection model. Represented by formula (4)

$$mAP = \frac{\Sigma_{i=0}^{n} - \int_0^1 P(R)dR}{n} \quad (4)$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (5)$$

Recall (R) refers to the proportion of actual positive class samples among all correctly predicted sample sizes.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (6)$$

3.2. Experiment on Model B Selection

We used collaboration from Kaggle and Accelerating Medicines Partnership ® Parkinson's Disease, abbreviated as AMP ® The Parkinson's disease dataset of PD. The dataset includes approximately 12000 visits from Parkinson's patients, with the core including protein abundance values obtained from mass spectrometry readings of cerebrospinal fluid (CSF) samples collected from hundreds of patients, with each patient providing multiple samples annually. Each data includes patient ID codes, UniProt ID codes for relevant proteins, amino acid sequences contained in peptides, frequency of amino acids in samples, standardized protein expression (NPX), patient scores for the N part of the Unified Parkinson's Disease Rating Scale, and whether patients took drugs such as levodopa during UPDRS evaluation.

The results of grid search cross validation are shown in Figure 3.

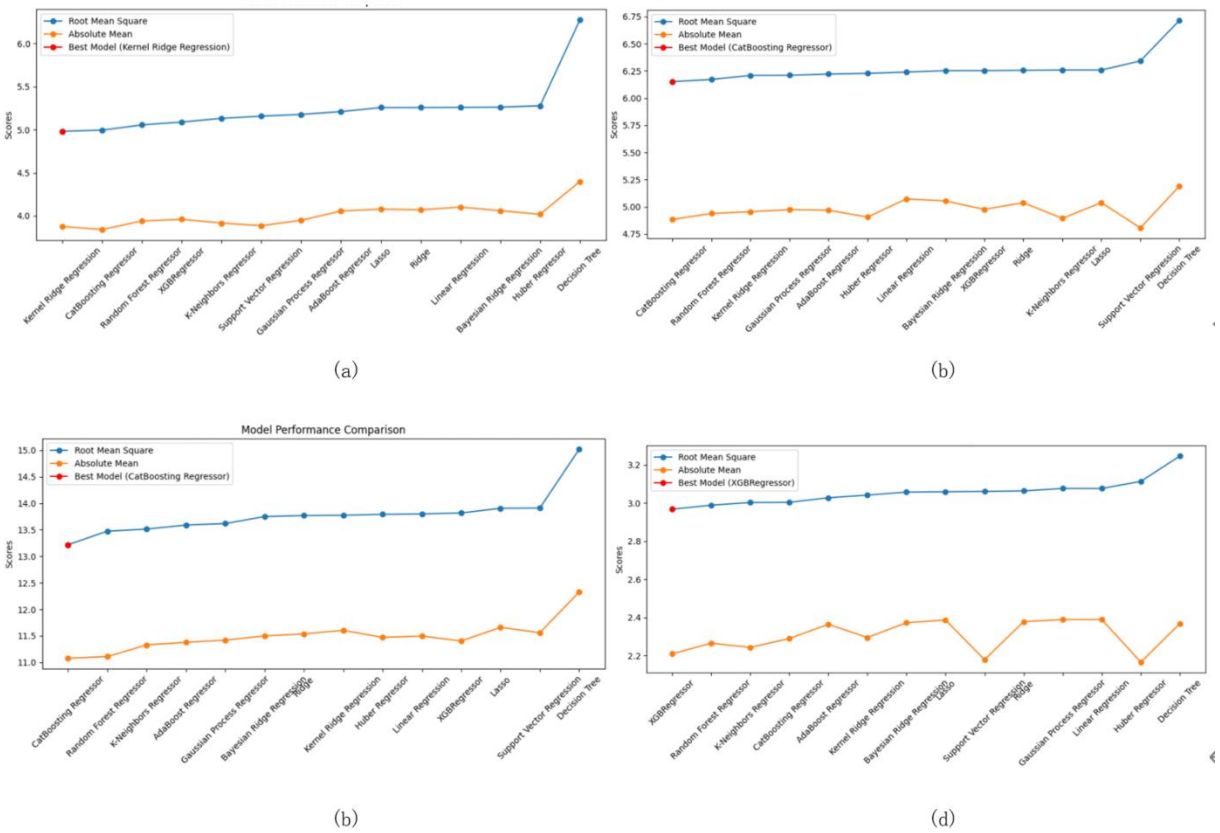

Figure 3. Grid search cross validation results: Figures (a) - (d) correspond to the severity of Parkinson's disease at four levels, respectively

So we selected the optimal three machine learning models for our next research, namely: Kernel Ridge Regression, CatBoosting Regression, and XGBRegression

3.3. Comparison of different methods

We compared our model with other methods and the results are shown in Table I. The experiment shows that the model outperforms Kernel Ridge Regression by 6.5%, CatBoosting Regression by 3.3%, XGBRegression by 6.5%, MLP by 6.5%, and SVM by 6.5% in mAP performance, proving that this method is superior to other classical models.

3.4. Ablation experiment

To demonstrate the effectiveness of each part of our model, we conducted ablation studies. Specifically, we gradually removed the Kernel Ridge Regression, CatBoosting

Regression, and XGBRegression components, and observed the mAP changes of the model after each step. As shown in Table II. The experiment showed that after subtracting the Kernel Ridge Regression, the mAP range of the model decreased from 88.3 to 84.5. This indicates that Kernel Ridge Regression plays an important role in the model and contributes significantly to the results. After subtracting CatBoosting Regression, the mAP range of the model decreased from 88.3 to 83.9, which further confirms the importance of CatBoosting Regression in the model and its contribution to the results cannot be ignored. After subtracting XGBRegression, the mAP range of the model decreased from 88.3 to 83.1, indicating that XGBRegression plays a crucial role in the model and has a significant impact on the results. Our ablation research results clearly indicate that each component (Kernel Ridge Regression, CatBoosting Regression, and XGBRegression) has a positive impact on overall model performance. This further validates the effectiveness of our model.

Table I. Comparison of Our Model with Other Models

Model	MAP50 (%)
MultiModelParkinson model	eighty-eight point three
Kernel Ridge Regression	eighty-eight point five
CatBoosting Regression	eighty-three
XGBRegression	eighty-seven
MLP	seventy-six point one
SVM	seventy-five point eight

Table II. Results of ablation experiments

Model	MAP50 (%)
MultiModelParkinson model	eighty-eight point three
Remove Kernel Ridge Regression	eighty-four point five
Remove CatBoosting Regression	eighty-three point nine
Remove XGBRegression	eighty-three point one

4. Conclusion

We propose a boosting based model for predicting the progression of Parkinson's disease. Through five fold cross validation, we selected three models with better performance from 14 and combined them for boosting. This model has shown excellent performance in predicting Parkinson's disease. However, we note that the model's ability in biomarker selection still needs to be strengthened and further research is needed. Despite these limitations, our model has great potential for application in predicting the severity of Parkinson's disease.

References

[1] Chen Jiani Overview of Parkinson's disease [J] Biology Teaching, 2010, 35:2-4.

[2] Wang Dan, Chen Xiaofang, Wang Huiqin Research progress on shame in Parkinson's disease patients [J] Nursing and Rehabilitation, 2019, 18:38-41.

[3] Liu Shuying, Chen Biao The current prevalence of Parkinson's disease [J] Chinese Journal of Modern Neurological Diseases, 2016, 16:98-101.

[4] Bai Xue, Hu Fengyun Diagnosis and treatment of juvenile Parkinson's disease [J] Chinese Journal of Practical Medicine, 2013, 40:89-90.

[5] M. I. Jordan and T. M. Mitchell, 'Machine learning: Trends, perspectives, and prospects,' Science, vol. 349, no. 6245, pp. 2015 255-260.

[6] B. Mahesh, Int. J. Sci Res., vol. 9, pp. 381-386, Jan. 2020.

[7] J. Wei, X. Chu, X. Y. Sun, K. Xu, H. X. Deng, J. Chen, Z. Wei, and M. Lei, 'Machine learning in materials science,' InfoMat, vol. 1, no. 3, pp. 338-358, 2019 32 (2), 106-116.

[8] Jefferson S Almeida, Pedro P Rebou ç as Filho, Tiago Carneiro, et al. Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques [J] Pattern Recognition Letters, 2019, 125:55-62.

[9] Hojoong M Kim, Carter Nazor, Cyrus P Zabetian, et al. Prediction of cognitive progress in Parkinson's disease using three cognitive screening measures [J] Clinical Parkinson's disease & related disorders, 2019, 1: 91-97.

[10] Liu Xia Protein shape helps detect Parkinson's disease [N] Science and Technology Daily, 1:30 (004), 2022.