A new discriminant model for Parkinson's disease based on logistic regression

Pengcheng Guo, Wei Li and Zhongyu Su
School of Science, Dalian Ocean University, 116023 Dalian, China

Abstract. Although Parkinson’s Disease has significantly affected patients’ normal family and life, it’s difficult to find the correlations of many kinds of indicators and an effective way of diagnosing timely. In this paper, an efficient Discriminant Analysis model based on Logistic Regression is provided to give the reduction of Parkinson’s patients’ indicators and diagnose whether or not you have the disease. Firstly, Principal Component Analysis and Clustering Analysis are implemented to reduce the data dimension of Parkinson’s indicators and give the reductive clusters to complete the feature extraction. The 22 physical indicators data of 195 patients is efficiently 3 main components and 4 clusters. Secondly, according to the extracted features and discriminant analysis based on Logistic Regression, the new discriminant function is established. At last, Python programming is designed to input the characteristic values and output whether or not have the disease to finish the efficient computer implementation of this model. Its result that the accuracy of 195 Parkinson cases of discrimination is 100% has verified the efficiency of this model.

1. Introduction
Parkinson’s Disease (PD) is a common neurological disease in elderly patients. It mainly causes symptoms such as quivering, muscular rigidity and body position instability, and leads to impaired motor function, which brings greater influence to daily life and brings greater economic and mental burden to the family [1]. The incidence of PD is improving as the increasingly aging population [2]. At present, there are few and inaccurate methods to diagnose PD, so it’s necessary to find a efficient way to solve this problem timely. In the existing data from the Parkinson’s dataset of UCI Machine Learning Repository [3], there are more than 22 kinds of physical indicators of PD and the correlations between many variables can’t be ignored, and there is a way that can only complete reducing dimensions [4] and can’t complete factor extraction. It increases the difficulties of data analysis. In addition, not only the misjudgment [5] of exiting discriminant analysis’ model is not avoidable and controlled, it is easy to lead to the prediction results inconsistent with the facts. In this paper, a new and efficient discriminant model based on Logistic Regression and Bayes Discriminant Analysis is provided to give the reduction of physical indicators data of PD and discriminate the patients to overcome the above limitations. Based on the extracted factors of PCA and CA, the discriminant model is established, which the accuracy of model is guaranteed by the trend prediction advantages of the regression model. So, the model of the discrimination accuracy of the 195 Parkinson cases is 100%, which verified the efficiency and accuracy of the model and methods above.

This paper consists of four sections. Section 1 is the introduction, including the research purpose and significance; section 2 is the theoretical introduction of the discriminant model including principal
component analysis (PCA), cluster analysis, Logistic Regression and Bayes Discriminant Analysis; in Section 3, based on the theoretical guidance in section 2, the PD is efficiently analyzed using this discriminant model; Section 4 is the summary.

2. Method
An efficient discriminant model using Bayes Discriminant Analysis based on Logistic Regression is provided to give the reduction of physical indicators data of PD and discriminate whether a patient is PD or not. Firstly PCA is implemented to reduce the data dimension of physical indicators of PD. Based on the principal component data, clustering analysis is used to give the reductive clusters to complete the feature extraction. Secondly, according to the extracted features, Bayes Discriminant Analysis model and Logistic Regression are established respectively. Thirdly, a new discriminant model using the value of Logistic Regression and the standardized data of Bayes function is established. At last, Python programming is used to complete the efficient computer implementation of this model to discriminate whether a patient is PD or not.

2.1. principal component analysis
PCA is a multivariate statistical method, which reduces the dimension of multivariate variables through linear transformation and selects several comprehensive indexes [6-8]. The original variables transform into the principal components by the linear transformation and combination. The algorithm steps are as follows:

1. Standardize the original data, and obtain the standardized matrix \( X \).
2. Use \( X \) to calculate the correlation coefficient matrix \( R \).
3. Find the eigenvalue and eigenvector of \( R \) matrix by jacobian algorithm.
4. When the characteristic value is greater than 1 and the cumulative contribution rate is up to 80-95\%, the principal component is determined and the principal component expression is obtained.
5. Establish the initial factor loading matrix and explain the principal components.

2.2. Clustering analysis
The main principle of clustering analysis is to divide data sources into several categories based on their similarity size, so that the significant objects can be approximately represented the most of objects, greatly improving the efficiency of data analysis [9-11]. In this paper, the system clustering method is implemented to analyze the main components obtained by PCA.

The algorithm steps are as follows [12]:
1. regard each sample as each category and calculate the distance of each group.
2. merge two categories into one with the shortest distance and become a new group.
3. repeat the above operation until ideal categories.

2.3. Logistic regression
Logistic Regression is a special kind of Linear Regression with only two values of dependent variables [13], the functional form of it is: \( \log \frac{q_1}{q_2} = bX \). The algorithm steps are as follows:
1. Determine training data and construct a prediction function between 0 and 1.
2. Construct a loss function to balance the error of Logistic Regression.
3. Find a function to represent the deviation between the predicted value of all training data and the actual category.

2.4. Bayes discriminant analysis
If there are two populations, their prior probabilities are \( q_1 \) and \( q_2 \) respectively, and the density functions of each population are and, when a sample is observed, the posterior probability from the first \( K \) population can be calculated through the Bayesian formula constructed. Then comparing their probabilities and discriminate which class it belongs to [14].
Determine the prior and posterior probabilities of Bayes Discriminant Analysis.

(2) Construct the Bayesian discriminant function.

(3) Determine the type of sample by comparing the distance to the function.

3. Results

3.1. Principal component analysis

The original data is from the dataset of UCI Machine Learning Repository [3], it is 22 physical properties of 195 patients of Parkinson’s Disease as Table 1.

| name          | phon_R01_S01_1 | ... | phon_R01_S07_2 | phon_R01_S07_3 |
|---------------|----------------|-----|----------------|----------------|
| MDVP:Fo       | 119.992        |     | 199.228        | 198.383        |
| MDVP:Fhi      | 157.302        |     | 209.512        | 215.203        |
| ...           | ...            | ... | ...            | ...            |
| PPE           | 0.284654       |     | 0.068501       | 0.09632        |
| status        | 1              |     | 0              | 0              |

There are 147 patients and 48 healthy cases, in status, 1 represents a PD’s patient, 0 represents a healthy person, and then PCA is implemented to reduce the dimension of data.

3.1.1. Factor extraction. Finding the principal components can complete factor extraction. The basis of finding them is that the cumulative contribution rate of variance is over 85%, which shows in Table 2.

|       | Total Variance percentage | Accumulation % |       | Total Variance percentage | Accumulation % |
|-------|---------------------------|----------------|-------|---------------------------|----------------|
| 1     | 188.5                     | 96.711         | 1     | 188.5                     | 96.711         |
| 2     | 5.033                     | 2.581          | 2     | 5.033                     | 2.581          |
| 3     | 1.245                     | 0.639          | 3     | 1.245                     | 0.639          |
| 4     | 0.132                     | 0.068          | 4     | 0.132                     | 0.068          |
| ...   | ...                       | ...            | ...   | ...                       | ...            |
| 194   | -5.52E-15                 | -2.83E-15      | 194   | -5.52E-15                 | -2.83E-15      |
| 195   | -4.75E-14                 | -2.43E-14      | 195   | -4.75E-14                 | -2.43E-14      |

According to Table 2, the first three principal components account for 99.931% of all the majority of all information. From the fourth principal component, it’s basically flat. Therefore, we obtain 3 principal components to express the vast majority of the information that can be expressed by all the original indicators.

3.1.2. Score of each component. Calculating the score matrix of each component with SPSS yields the following results in Table 3:

| sample         | composition |
|----------------|-------------|
| phon_R01_S37 6 | 1           |
| phon_R01_S37 4 | 0.999       |
| phon_R01_S34 5 | 0.999       |
| ...            | ...         |
| phon_R01_S24 6 | 8.33E-01    |
| phon_R01_S49 4 | 0.829       |
3.2. Clustering analysis of Table 3

Based on the principal component results, with SPSS software, clustering analysis is implemented to further analyze the data. The number of cluster for the data of 3 principal components of 195 PD is 4, as Table 4:

| Case | 7 clusters | Case | 7 clusters |
|------|------------|------|------------|
| 1    | 1          | 13   | 4          |
| 2    | 2          | 14   | 4          |
| 3    | 3          | ...  | ...        |
| 4    | 4          | 21   | 4          |
|      | ...        | 22   | 4          |

As Table 4, based on the three principal components, 22 physical attribute features in the data are successfully classified into 4 categories. As above, the first three clusters are significant to represent all indicators information, which is respectively: MDVP: Fo(Hz), MDVP: Fhi(Hz), MDVP: Flo(Hz), Now the feature extraction operation is completed.

3.3. Logistic regression and bayes discriminant analysis

Based on the three features extracted by PCA and Cluster Analysis can represent the majority of all information, which has reduced the dimension and difficulties of data analysis to conveniently established the discriminant model. The above three characteristics are used as independent variables to predict the dependent variable (Status). a, b, c respectively denote the independent variables. Then Logistic Regression function as Table 5 and initial Fisher linear discriminant function as Table 6 are respectively performed.

| Variables | B       | Standard error | Wald | Freedom | significance | Exp(B) |
|-----------|---------|----------------|------|---------|--------------|--------|
| MDVP:Fo   | -0.013  | 0.006          | 4.621| 1       | 0.032        | 0.987  |
| MDVP:Fhi  | -0.002  | 0.002          | 0.735| 1       | 0.391        | 0.998  |
| MDVP:Flo  | -0.011  | 0.005          | 5.253| 1       | 0.022        | 0.989  |
| Constant  | 4.974   | 0.81           | 37.684| 1       | 0            | 144.543|

As Table 5, Logistic Regression function is $y_0=-0.013*a-0.002*b-0.011*c+4.974$.

| status     | status |
|------------|--------|
| 0          | 1      |
| MDVP:Fo(Hz) | 0.088  | 0.073  |
| MDVP:Fhi(Hz)| 0.013  | 0.011  |
| MDVP:Flo(Hz)| 0.045  | 0.028  |
| Constant   | -13.367| -8.585 |

As Table 6, firstly, initial Fisher discriminant function are respectively $y_1=0.088*a+0.013*b+0.045*c-13.367$, $y_2=0.073*a+0.011*b+0.028*c-8.585$. Secondly, we standardize the two functions to unify their dimensions, and obtain two standardized functions, which is respectively, $y_1=y_1/(y_1+y_2)$, $y_2=y_2/(y_1+y_2)$. Finally, based on Logistic Regression function and standardized functions to discriminate analysis again, a new and efficient discriminant model with two discriminant functions is founded, which is $Y_1=306.255*y_1+6.976*y_2-93.99$ and $Y_2=27.975*y_1+3.421*y_2-4.235$. The new model is used to discriminate the 195 cases of Parkinson’s Disease to verify the efficiency of its. And the misjudgment of the new and traditional model is shown in Table 7, which result is verified the absolutely accuracy of new model:
Table 7. Comparison of the two models.

| Sample | Actual | Predict | Sample | Actual | Predict |
|--------|--------|---------|--------|--------|---------|
| 1      | 1      | 1       | 1      | 1      | 1       |
| 2      | 1      | 2       | 1      | 1      | 1       |
| ...    | ...    | ...     | ...    | ...    | ...     |
| 20     | 1      | 0       | 20     | 1      | 1       |
| 21     | 1      | 1       | 21     | 1      | 1       |
| ...    | ...    | ...     | ...    | ...    | ...     |
| 38     | 1      | 0       | 38     | 1      | 1       |
| ...    | ...    | ...     | ...    | ...    | ...     |
| 195    | 0      | 0       | 195    | 0      | 0       |

As Table 7, when the predicted value is different from the actual value, it is misjudgment. After calculation, previous model’s misjudgment rate is 27.7%, and improved model’s misjudgment rate is almost 0. The results verify the efficiency of this model and overcome the limitations of the traditional prediction methods.

3.4. Python programming

Using Python to discriminate whether or not have the disease when input the value of a, b and c, the part of the original code has put in github and shown as Figure 1:

![Python code](image)

Figure 1. Implementation code.

The above three values can be input to discriminate the class of the new sample. The results verifies that the discriminant data of this model is more effective and accurate.
4. Summary
At present, Parkinson’s disease is diagnosed with Webster score and Parkinson’s disease comprehensive score scale (UPDRS), which is complex and inaccurate [15]. In this paper, a new and efficient discriminant model using Logistic Regression and Bayes Discriminant Analysis based on feature extraction is provided to give the reduction of PD physical indicators and discriminant whether or not a person have the disease. With PCA and cluster analysis, the 22 physical indicators data of 195 PD’s patients is efficiently reduced to 3 main components and 4 clusters. Based on the above extracted features and the new discriminant analysis model, with Python programming, the misjudgment of the model is almost 0, which shows it is successful and accurate. The results verified that this model is more efficient and overcomes the limitations of the traditional discriminant methods. The new model can be not only used in traditional discriminant cases, but also artificial determined cases like stepwise discriminant analysis and so on [16], which is also efficient and accurate.

References
[1] Li J L 2016 Effects and adverse reactions of dopa combined with praxole in the treatment of Parkinson's disease[J]. Contemporary medical 164-166
[2] Xu E, Zhang S and Xu Y Q 2015 Genotyping of hepatitis b virus in uyugur and han populations of xinjiang uyugur autonomous region[J]. Chinese journal of experimental and clinical virology 316-318
[3] Little M A, McSharry P E, Roberts S J, Costello D A E and Moroz I M 2007 Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection[J]. BioMedical Engineering OnLine 6 23
[4] Baliarsingh S K, Vipsita S, Muhammad K and Bakshi S 2019 Analysis of high-dimensional biomedical data using an evolutionary multi-objective emperor penguin optimizer[J]. Swarm and Evolutionary Computation
[5] Zeng C B, Miao Q P, Gao Y H and Li L X 2016 Simulation study on the discriminant effect of Bayes discriminant[J] Journal of mathematical medicine 1272-1275
[6] He X Q 2015 Multivariate Statistical Analysis[M]. Renmin University of China Press
[7] Su W H 2000 Study on the theory and method of multi-index comprehensive evaluation[M] Xiamen University
[8] Guo Y J 2002 Comprehensive evaluation theory and method[M]. Beijing: Science Press
[9] Du Q, Jia L Y and Yan X F 2014 Spss statistical analysis from introduction to mastery[M] People's Post Office
[10] Fang K T and Pan E P 1982 Clustering analysis[M]. Geological Publishing House
[11] Li P Z 2008 The application of cluster analysis and principal component analysis in the evaluation of regional comprehensive consumption level[J]. Nankai University
[12] Zhang R C 2006 Multivariate statistical analysis [M]. Science press 9 244-255
[13] Li Z R 2017 The principle and application of logistic regression method[J]. China's strategic emerging industries 114-115
[14] Ju Q D, Hu Y B and Zhang S Y 2018 Study on the identification method of mine water inrush source based on principal component analysis and bayesian discrimination[J] Coal engineering 90-94
[15] Lu Z H 2008 Analysis and clinical study on early diagnosis and treatment of Parkinson's disease [J]. Chinese journal of practical medicine 13(22) 101-103
[16] Guo P C, Li W and Su Z Y 2019 A prediction model for superconductor critical temperature using stepwise discriminant analysis based on feature extraction[C]. International Conference on Applied Physics and Mathematics (ICAPM 2019), Chulalongkorn University, Bangkok, Thailand