Feature Diagnosis of Alzheimer’s Disease Based on Stacking Fusion Algorithm and DBSCAN Clustering

Wenchen Wu*, Bin Ma, Jingyi Zhu, Jiajun Huang

Guangdong Ocean University, Zhanjiang, Guangdong, China.

How to cite this paper: Wenchen Wu, Bin Ma, Jingyi Zhu, Jiajun Huang. (2023) Feature Diagnosis of Alzheimer’s Disease Based on Stacking Fusion Algorithm and DBSCAN Clustering. Advances in Computer and Communication, 4(3), 143-148. DOI: 10.26855/acc.2023.06.007

Received: May 28, 2023  
Accepted: June 25, 2023  
Published: July 21, 2023

*Corresponding author: Wenchen Wu, Guangdong Ocean University, Zhanjiang, Guangdong, China.

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disease. It is characterized by a progressive decline in the ability to perform activities of daily living, accompanied by various neuropsychiatric symptoms and behavioral disorders. The disease usually develops progressively in older people, with progressive loss of independent living and death from complications 10 to 20 years after onset. First of all, we eliminate the indicators that the data vacancy accounts for 80% or more of the data volume, and eliminate similar duplicate indicators. Collaborative filtering algorithm and LOF abnormal data detection algorithm are used to process the data and get the original data. The importance of establishing a random forest model to calculate characteristic indexes, and combining Spearman correlation coefficient method to solve the correlation between the characteristics of original data and the diagnosis of Alzheimer's disease. Then divide the data into five categories. XGboost and SVM models with 50% cross-validation in stack fusion are used to classify characteristic indexes, study cognitive behavior and structural characteristics of brain, and design intelligent diagnosis of Alzheimer's disease. And we divided CN, MCI and AD into three categories, and screened the index data in nine categories. Unsupervised classification is carried out by DBSCAN clustering, and three subclasses are further refined into three subclasses by elbow method. Then, according to the classification results, the characteristic indexes and patients were analyzed by single factor, and some indexes were eliminated. The change trend of characteristic indicators with time was studied, and the evolution patterns of different types of diseases with time were obtained by combining descriptive statistics of indicators.

Keywords

Random forest algorithm, XGboost, stacking Fusion algorithm, DBSCAN, Elbow method

1. Introduction

1.1 Background

Alzheimer's disease (AD) is a progressive neurodegenerative disease onset hidden. It is clinically characterized by a spectrum of dementias, including memory impairment, aphasia, language impairment, agnosia, visuospatial skill impairment, executive dysfunction, and personality and behavioral changes, the causes of which are unknown [1]. It is characterized by a progressive decline in the ability to live daily, accompanied by various neuropsychiatric symptoms and behavioral disorders. The disease usually progresses in the elderly, with progression of independent...
living skills and death from complications 10 to 20 years after onset [2].

Preclinical stage of Alzheimer's disease, is also known as mild cognitive impairment (MCI), which is a transition state between normal and severe. Due to limited awareness of the disease among patients and their families, 67% of patients were diagnosed with moderate to severe disease, missing the optimal intervention stage [3]. Therefore, early and accurate diagnosis of Alzheimer's disease and mild cognitive impairment is of great significance.

2. Problem analysis

2.1 Data analysis

Firstly, the indicators in the article are studied. Combined with the relevant knowledge of medicine, the indicators with data vacancies accounting for 80% or more of the data volume are firstly eliminated, and similar duplicate indicators are also eliminated [4]. The collaborative filtering algorithm is used to fill the data, and then the LOF abnormal data detection algorithm is used to detect the abnormal data after filling the data, and finally the original data is obtained.

2.2 Analysis of question one

Based on the original data obtained after processing, a random forest model was established to solve the importance of feature indicators, and the correlation between the features of the original data and the diagnosis of Alzheimer's disease was solved by combining the Spearman correlation coefficient method.

2.3 Analysis of question two

Based on the data processing and problem-solving process, the data is five categories. By straking fusion model, combined with fifty percent cross-check XGboost model and SVM model, classifying characteristic indexes, and study the structure characteristics and characteristics of cognitive behavior and brain to design intelligent diagnosis of Alzheimer's disease [5].

2.4 Analysis of question three

CN, MCI and AD are divided into three categories. The index DX is selected in the BI column of the article, and the index SMC\EMCI\LMCI is selected in the H column. The three categories are classified into five unsupervised categories through DBSCAN, and the visualization is combined with the elbow method, and the three sub-categories are further refined into three subcategories [6].

4. Model

4.1 Question One

4.1.1 Data processing based on collaborative filtering algorithm

In order to facilitate the follow-up research, the data processing method based on collaborative filtering algorithm is used to sort out the data in the article. Collaborative filtering algorithm is a widely used recommendation algorithm, which is mainly divided into user-based and item-based collaborative filtering [7]. Based on user collaborative filtering is the main idea by drawing from similar to the user's preferences and ideas which are recommended, and the main idea of the collaborative filtering is the calculation based on the project early electronic products business of commodity type and inventory, at the end of inventory [8], merchandise sales and similarity of goods in order, according to the similarity of e-commerce sales good commodity type.

Cosine similarity calculation formula:

\[
\text{sim}(u,v) = \frac{\sum_{i \in I_u \cap I_v} r_{ui} \times r_{vi}}{\sqrt{\sum_{i \in I_u} r_{ui}^2} \times \sqrt{\sum_{i \in I_v} r_{vi}^2}} \tag{1}
\]

Pearson correlation coefficient calculation formula:
Among them: $r_{ui}$ and $r_{vi}$ respectively represents users $u$ and $v$ for the project $i$ the score; $r_u$ and $r_v$ respectively represents users $u$ and $v$ or the project the average score.

### 4.1.2 LOF abnormal data detection algorithm

The LOF algorithm determines whether a point $p$ is an outlier by comparing the density of each point $p$ with its neighborhood points. If the density of point $P$ is lower, the more likely it is to be identified as an outlier. For density, it is calculated by the distance between points, the farther the points are, the lower the density, and the closer the distance, the higher the density [9]. Because LOF calculates the density by the KTH neighborhood of a point instead of the global calculation, it is named "local" outlier factor. For the two datasets $C_1$ and $C_2$, LOF can correctly deal with it, and will not mistakenly judge normal points as outliers due to different data density dispersion.

### 4.1.3 Importance Evaluation with Random Forests

The construction steps of the classification decision tree are as follows: (1) design the attribute selection strategy for classification, (2) design the construction method of the decision tree, (3) design the pruning strategy, and (4) test the performance of the constructed classification decision tree.

Design classification attribute selection factor ASF is as follows:

$$ASF(A_i) = \frac{(2 \text{Averagegain}(A_i) - 1) \cdot \text{Reduce}(A_i)}{TC(A_i)_{normal}}$$

Among them, $\text{Averagegain}(A_i)$ said the average information gain, $TC(A_i)_{normal}$ show that the cost of standardized tests, $\text{Reduce}(A_i)$ said of misclassification cost.

### 4.2 Question Two

#### 4.2.1 Support vector machine (SVM)

In the application of support vector machine regression, in order to establish the learning model, it is often necessary to define a class of ill-defined target variables, but it is not easy to define these variables accurately. In fact, these types of target variables are just comparison values. They only need a relative value that can be compared, and they do not need to be very accurately defined.

But the target variable can be affected by a number of factors, which are $n$ dimensional vector $X(x_1, x_2, \cdots, x_n)$, will $X \in \mathbb{R}^n$. Map to $Y \in \mathbb{R}$, The target variable $Y$ can be defined as: $X \mapsto Y, (R^* \mapsto R)$.

Set up a set of data $\{(x_i, y_i)\}_{i=1}^n$, $x_i$ is the input vector, $y_i$ is the corresponding output. The support vector regression function is as follows:

$$f(x) = w \cdot \theta(x) + b$$

Where, $w$ is the weight vector, $b$ is the long-term partial quantity, $\theta(x)$ is the nonlinear mapping. $\varepsilon$ the insensitive loss function is as follows:

$$L_\varepsilon(f(x), y) = \begin{cases} |f(x) - y| - \varepsilon & \text{if } |f(x) - y| \geq \varepsilon \\ 0 & \text{if } |f(x) - y| < \varepsilon \end{cases}$$

The minimization function:

$$Q(C) = \frac{1}{2} ||w||^2 + C \sum_{i=1}^n L_\varepsilon(y)$$

Where, $C$ is used to evaluate the tradeoff between model smoothness and empirical risk.
By introducing two positive relaxation variables, $\xi_i$, $\xi_i^*$, the above optimization problem is transformed into the following constraint form:

$$\begin{align*}
\text{Minimize:} & \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \\
\text{Subject} & \quad \begin{cases} 
y_i - w \cdot \theta(x) - b \leq \varepsilon + \xi_i \\
w \cdot \theta(x) + b - y_i \leq \varepsilon + \xi_i^* \\
\xi_i, \xi_i^* \geq 0
\end{cases}
\end{align*}$$ (7)

Then, convert the above equation into Lagrangian form:

$$\begin{align*}
\text{Maximize:} & \quad \sum_{i=1}^{n} y_i (a_i - a_i^*) - \varepsilon \sum_{i=1}^{n} (a_i + a_i^*) \\
k(x_i, x_j) = \theta(x_i) \theta(x_j) \\
\text{Subject} & \quad \sum_{i=1}^{n} (a_i - a_i^*) = 0, 0 \leq a_i, a_i^* \leq C
\end{align*}$$ (9, 10, 11)

Among them, $k(x_i, x_j)$ is a kernel function, $a_i$, $a_i^*$ is the Lagrange multiplier. Therefore, the general form of support vector regression can be written as follows.

$$f(x, a_i, a_i^*) = \sum_{i=1}^{n} (a_i - a_i^*) k(x_i, x) + b$$

Adaptive learning support vector regression can improve the generalization ability of traditional support vector regression. The objective function of its performance evaluation is defined as follows:

$$Y(y_i, y_j) = \sqrt{\sum_{i=1}^{n} [y_i - y_i^*]^2}$$ (12)

### 4.2.2 Extreme Gradient Boosting (XGboost)

XGBoost the whole idea is directly add up loss function and the regularization synthesis are an integral part of the loss function, the second derivative of the loss function, get the final obj, calculated by obj a score, the score as small as possible, finally through the obj calculated scores determine the structure of the tree and the strong points of learning [10]. So XGBoost does this not by fitting the residuals, but by computing the tree structure directly derived from the obj function.

$$\text{Obj}(\theta) = L(\theta) + \Omega(\theta)$$ (13)

XGBoost is an improvement of the gradient boosting algorithm. When solving the extreme value of the loss function, Newton's method is used, the loss function is expanded to the second order by Taylor, and the regularization term is added to the loss function. The objective function during training consists of two parts, the first part is the loss of the gradient boosting algorithm, and the second part is the regularization term.

### 4.3 Question Three

#### 4.3.1 DBSCAN clustering algorithm

DBSCAN algorithm forms class clusters for spatial samples based on density. In the n-dimensional space, the threshold radius Eps and the threshold size Minpts were taken as the clustering limiting conditions, and the sample points in the space were iteratively calculated to discover the arbitrary shape cluster, and the noise points in the sample data set were filtered to obtain the density clustering results [11]. In the n-dimension sample set $D = \{X_1, X_2, \ldots, X_n\}$, the corresponding steps of the algorithm are listed:

**Step 1:** Eps neighborhood for any sample point p in the n-dimensional space, with p as the center, Eps is the set of all sample points within the radius. Which is:
\[ N_{\text{eps}}(p) = \{ q \in D | \text{Dist}(p, q) \leq \text{Eps} \} \]  

(14)

Where \( \text{Dist}(p, q) \) represents the distance from \( p \) to \( q \). In general, \( \text{Dist} \) is calculated by Minkowski distance formula, and the formula is as follows:

\[ \text{Dist}(p, q) = \left( \sum_{k=1}^{n} |p_k - q_k|^\frac{1}{\text{p}} \right)^{\frac{1}{p}} \]  

(15)

**Step 2:** Core point in the \( n \)-dimensional space for any sample point \( p \), the neighborhood value of \( p \) as the center point is not less than \( \text{Minpts} \) sample points. That is, it is expressed as:

\[ \epsilon N_p \leq \text{Minpts} \]

calls \( p \) the core point.

### 5. Solution of each problem

#### 5.1 Question one Result

Study the index data in the article, and use origin to make box plot visualization of the attached data. The collaborative filtering algorithm and LOF abnormal data detection algorithm were used to process the data, eliminate some indicators with low correlation, and obtain the final original data for data visualization.

The data after data processing is used as the original data, and the indicators in the original data are studied to screen out the feature engineering. Random forest model was established to calculate the importance of each feature index, and Spearman correlation coefficient method was used to calculate the correlation between data features and Alzheimer's disease diagnosis. Using 0.05 as the dividing line, the importance of different key indicators was obtained.

The importance of the index to the target variable is solved by machine learning through random forest, and the correlation coefficient of the index is obtained by combining the Spearman correlation coefficient method. It can be concluded that the correlation between data characteristics and the diagnosis of Alzheimer's disease is ranked as follows.

\[ \text{LDEL/TOTAL}_{\text{BL}} > \text{mPACCdigit} > \text{FAQ} > \text{ADAS}_1 > \text{AGE} > \text{MMSE} > \text{ADAS}_2 > \text{PTEDUCAT} > \text{RAVLT}_{\text{perc_forgetting}} \]

#### 5.2 Question two Result

Based on the original data obtained in problem 1, the accompanying outcome brain features and cognitive behavioral characteristics were screened to obtain the corresponding indicators. Origin was used for data visualization.

The data were processed into five categories, and two machine learning models, XGboost and SVM model, were established, and the cross test of 50% was added. In the data set of problem 2, the training set and the test set are divided, and the stacking fusion model is used to combine the results obtained after XGboost and SVM machine learning, so as to study the structural characteristics and cognitive behavior characteristics of cognitive behavior and brain.

By stacking and fusing XGboost and SVM model, it can be found that the accuracy can reach 74.8889% after machine learning classification of the data. At the same time, combined with the importance of the characteristic index obtained in Annex II, through the study of structural brain characteristics and cognitive behavior characteristics, the intelligent diagnosis of Alzheimer's disease is finally designed.

Intelligent diagnosis of Alzheimer's disease: through FAQ, AGE, PTEDUCAT, DX_bl and other four characteristic indicators, based on the data changes of these four indicators in patients, as the method of intelligent diagnosis of Alzheimer's disease.

#### 5.3 Question three Result

Based on the results of problems 1 and 2, data is first processed to obtain the data needed for problem 3 and then visualized, as shown in the figure below.

After obtaining the data required for problem 3, the unsupervised classification of DBSCAN is firstly used to classify the data into five categories, and CN, MCI and AD are divided into three categories. The following results are obtained [13].

Select the index DX in the BI column of the article and the index SMC, EMCI, LMCI in the H column. The three index data of SMC, EMCI, LMCI were processed into nine categories, and the three categories were unsupervised classified by DBSCAN. Combined with the elbow method, the three sub-categories were refined and further refined.
into three sub-categories. The result of question three is as follows:

6. Sensitivity Analysis

When building a stacking model based on machine learning, it has high requirements on the data processed, so the results of data processing must have high accuracy to make the established model more accurate. Therefore, this paper carries out sensitivity analysis on the results of question 2 and the evaluation results of changes in the results of data processing within a certain range.

7. Weakness: The promotion of the model

This problem involves a large amount of work in data processing. In the process of data processing, the collaborative filtering algorithm is used to process missing values, and the LOF algorithm is used to detect and process outliers, so that the data is more accurate [10]. In question 2, different machine learning algorithms are used to make the results more accurate, and stacking algorithm is used to fuse different machine learning algorithms. The model and method in question 2 can be used to solve most non-time series based data, and the fusion has high accuracy in predicting chemical element content.

References


