Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 3, No. 1, 2023

Intelligent Diagnosis of Alzheimer's disease

Ruifan Huang[1, a], Huifei Xu[2, b]

[1] School of Sun Yat-sen University, Guangzhou 510000, China
[2] School of Hangzhou Normal University, Hangzhou 310000, China.
[a]1059577592@qq.com, [b]1225604713@qq.com

Abstract: In this paper, we combine the knowledge of decision tree model and K-means clustering algorithm with knowledge of statistics to solve the problems of Diagnosis of Alzheimer's disease using brain structure. In the step of selecting features, this paper considers both data-based and theoretical perspectives for the analysis of variance in Alzheimer's disease. Based on the data perspective, the value of variance represents the magnitude of data volatility, i.e., data with data volatility has more abilities to cause data changes in feature classification. Therefore, in this paper, a total of 28 indicators of Site-faq were selected as necessary indicators and histograms were made to observe the volatility. After the above data processing, we Use the attached structural brain features and cognitive behavioral features to design an intelligent diagnosis of Alzheimer's disease. For multi-classification models, this paper uses the K-means algorithm to cluster and continue to cluster MCI into SMC, EMCI, LMCI to observe the difference between cluster centers. After clustering 53 indicators, this paper analyzes the three types of cluster centers, mainly to analyze the different points in the number of clusters, and analyzes the difference in the mean value of the characteristics of the cluster centers, and at the same time, this paper observes the differences between the characteristics by visualizing the data. In the end, we analyze them in relation to the time points to uncover patterns in the evolution of different categories of diseases over time. At the same time, due to the large amount of data, this paper only selects a small sample of five features for analysis.

Keywords: Decision tree; Random forest; K-means clustering; Pearson coefficient.

1. Introduction

1.1. Problem Backgroud

Alzheimer's disease (AD) is a progressive neurodegenerative disease with an insidious onset.

It is characterized clinically by a full spectrum of dementia, including memory impairment, aphasia, dysfluency, agnosia, impairment of visuospatial skills, executive dysfunction, and personality and behavioral changes, the cause of which is still unknown. It is characterized by a progressive decline in the ability to perform activities of daily living, with various neuropsychiatric symptoms and behavioral disturbances. The disease is usually progressive in the elderly, with progressive loss of independent living skills and death from complications 10 to 20 years after the onset of the disease. The preclinical stage of Alzheimer's disease, also known as mild cognitive impairment (MCI), is a transitional state between normal and severe. Due to the limited cognition of the disease by patients and their families, 67% of patients were diagnosed as moderate to severe and had missed the best intervention stage. Therefore, early and accurate diagnosis of Alzheimer's disease and mild cognitive impairment is of great significance [1]

1.2. Clarifications and Restatements

This paper uses the brain structural characteristics and cognitive behavioral characteristics of the different categories of people provided in the data set to construct an Alzheimer's disease identification model and design an intelligent diagnostic method to accurately diagnose Alzheimer's disease. In this paper, the research is decomposed into four problems shown below.

1). Preprocess the characteristic indicators of the attached data to investigate the correlation between data characteristics and the diagnosis of Alzheimer's disease.

2). Use the attached structural brain features and cognitive behavioral features to design an intelligent diagnosis of Alzheimer's disease.

3). First, cluster CN, MCI and AD into three major classes. Then, for the three subclasses contained in MCI (SMC, EMCI, and LMCI), the clustering was continued to be refined into three subclasses.

4). The same sample in the annex contains features collected at different time points, we will analyze them in relation to the time points to uncover patterns in the evolution of different categories of diseases over time.

2. General Assumptions and Variable Description

2.1. Assumptions

1). Assume that the effect of environmental factors on the composition of ancient glass artifacts will not change the original chemical composition content ratio by 20%.

2). Assume that in the sampling test, there is no human error in the test results, while the error between the test results and the true content should be less than 2%.

3). It is assumed that chemical components not collected in the sampling tests and not studied will not have a significant correlative effect on the content of other chemical components.

4). Assume that there are no interactions between different chemical components or between different representations of the artifacts.

5). For taking into account the impact of ambient temperature and humidity on the test, assuming that the sampling process are in the same environment and the sampling time interval is short.

2.2. Variable Description

Table 1. Symbol specification table

Symbols	Description
$y_i(k)$	The specific value of the k sample of corresponding to the i indicator of the parent series
$x(k)$	Subseries i.e. interpolation rate at the k th sample
β_k	Coefficient of influence of the k variable on the classification
x_i	The i variable
$\hat{y}_i$	denotes the probability that the i th variable is classified as a barium lead variable
$\tilde{x}_i$	Normalized value of a chemical component in the i-th data set
C_i	Cluster i
p	The sample points in the C_i
m_i	Sample Means
$m(i)$	Distance of the point to all other points in the cluster it belongs to
$n(i)$	The average distance from a point to all points in a cluster that does not contain it
$S(i)$	Contour coefficient of the point
d_i	The rank difference, i.e. the position of the number after sorting two sets of data from smallest to largest

3. Problem 1

The process of modeling and solving Problem 1 is shown in Figure 1:

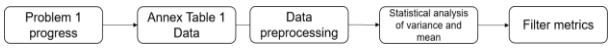

Figure 1. Problem 1 solution process

3.1. Task 1: Analysis of relationship

3.1.1. Data pre-processing

The flow chart for data pre-processing is as follows:

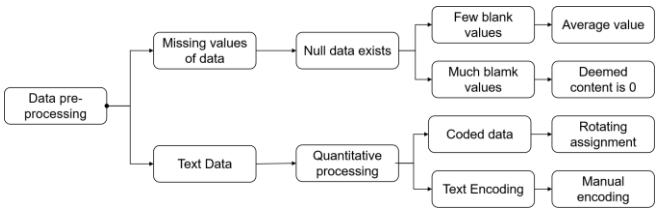

Figure 2. Basic data pre-processing process

For the codified data, this paper quantifies them according to the coding serial number, and for the more important codes, i.e., the text coding part, this paper chooses to encode them manually, for example, for the severity of AD, named as follows.

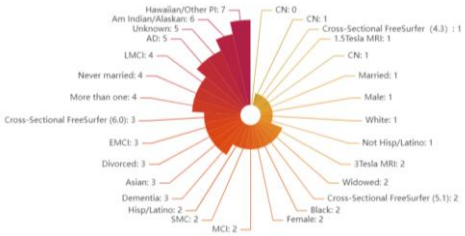

Figure 3. Quantitative processing

3.1.2. Model buiding and solving

After the selection of indicators, in order to further determine the reasonableness of the selection of indicators. In this paper, ANOVA will be used to determine whether there is a significant difference in the change of each statistic over time. First, ANOVA is introduced in this paper as follows.

The statistical inference method used in the one-way ANOVA is to calculate the F statistic and perform the F test. The total sum of squares of variance, denoted as SST, is decomposed into two parts: one part is the variance caused by the control variables, denoted as SSA (sum of squares of variance between groups); the other part is the SSE (sum of squares of variance within groups) caused by the random variables. The relationship between the two variables is shown below [2].

$$\mathrm{SST} = \mathrm{SSA} + \mathrm{SS}$$

In general, assume that there are k levels and establish the following assumptions.

$\mathrm{H_0}: \mathrm{u_1} = \mathrm{u_2}{=}\ldots \mathrm{u_k}$

$\mathrm{cH_1}: \mathrm{u_1}, \mathrm{u_2}, \ldots, \mathrm{u_k}$, not all equal to each other

The error within each overall system, mainly the random error caused by random sampling, is called the intra-group error, and the sum of squared intra-group errors is denoted as SSA.

$$\mathrm{SSA} = \sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{n_i}(\overline{\mathrm{x}_\mathrm{i}} - \overline{\mathrm{x}})^2$$

The errors between the totals, mainly random errors and systematic errors caused by random sampling, are called inter-group errors, and the sum of squared inter-group errors is SSE.

$$\mathrm{SSE} = \sum_{\mathrm{i}=1}^{\mathrm{k}} \sum_{\mathrm{k}}^{\mathrm{n_i}} (\mathrm{x_{ij}} - \overline{\mathrm{x}_\mathrm{i}})^2$$

The total sum of squared errors of the data is denoted as SST.

$$\mathrm{SST} = \sum_{\mathrm{i}=1}^{\mathrm{k}} \sum_{\mathrm{k}}^{\mathrm{n_i}} (\mathrm{x_{ij}} - \overline{\mathrm{x}})^2$$

Where denotes the j^{th} observation at the i^{th} level, denotes the mean value at the j^{th} level, and denotes the mean value of the n observations.

$$\mathrm{SST} = \sum_{\mathrm{i}=1}^{\mathrm{k}} \sum_{\mathrm{k}}^{\mathrm{n_i}} (\mathrm{x_{ij}} - \overline{\mathrm{x}})^2 = \mathrm{SSA} + \mathrm{SSE}$$

The above calculation shows that the total error of the data comes from two main sources: intra-group error on the one hand and inter-group error on the other. It can be shown that when the means of the k overall groups are equal, $\mathrm{SSA}/_{\mathrm{k}-1}/\mathrm{SSE}/_{\mathrm{n}-\mathrm{k}}$ obeys the F distribution with degrees of freedom k-1 in the numerator and n - k in the denominator.

$$\mathrm{F} = \frac{\mathrm{SSA}/(\mathrm{n}-1)}{\mathrm{SSE}/(\mathrm{n}-\mathrm{k})} \leq \mathrm{F}_\alpha(\mathrm{k}-1, \mathrm{n}-\mathrm{k})$$

Each F value will correspond to a p value, and the larger the F value and the smaller the p value, the less likely it is to accept the original assumption, which is considered to be a large difference in characteristics between groups. Therefore, the smaller the p-value, the more this feature should be preserved.

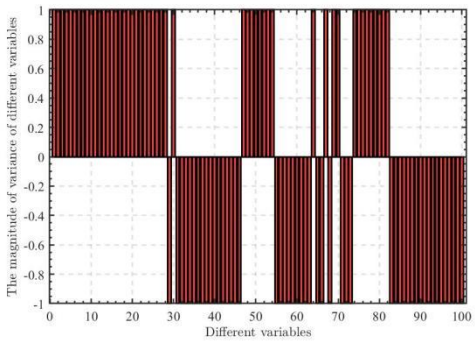

Figure 4. Analysis of variance histogram

Pearson correlation analysis is a method used to study the closeness of association between two or more variables. In order to explore which factors are associated with the cause of Alzheimer's disease, so the correlation analysis algorithm is used, and the Pearson correlation value can be a good representation of the correlation between two variables. Some of the indicators in the experiment have an interaction relationship, and the Pearson correlation value is chosen to calculate the degree of correlation between different human physical characteristics affecting Alzheimer's disease, and the Pearson correlation value expression is shown below [3].

$$f(x,y) = \frac{\sum_{a=1}^{t}(X_a - \overline{X})(Y_a - \overline{Y})}{\sqrt{\sum_{a=1}^{t}(X_a - \overline{X})^2}\sqrt{\sum_{a=1}^{t}(Y_a - \overline{Y})^2}}$$

In the equation, $f(x,y)$represents the similarity between the two variables X and Y, a represents the a^{th} sample of a certain indicator, t is the total number of samples, and $\overline{X}$ and $\overline{Y}$ represent the values of two variables, X and Y, respectively. Pearson correlation values $f(x,y)$ always fall between -1 and 1. When 0.8<a<1, the two variables are extremely strongly correlated; When 0.6<a<0.8, the two variables are strongly correlated; When 0.4<a<0.6, the two variables are moderately strongly correlated; When 0.2<a<0.4, the two variables are weakly correlated; When 0<a<0.2, the two variables are very weakly or not correlated.

Figure 5. Correlation value line graph

Therefore, in this paper, a total of 28 indicators of Site-faq were selected as essential indicators, and histograms were made to observe the volatility, and Pearson correlation analysis was done for the remaining variables with high volatility to screen the indicators with strong correlation to the generation and deterioration of Alzheimer's disease. Based on this, a total of 53 indicators were selected as basic indicators in this paper.

4. Problem 2

Problem 2 model building and solution process is shown in Figure 6;

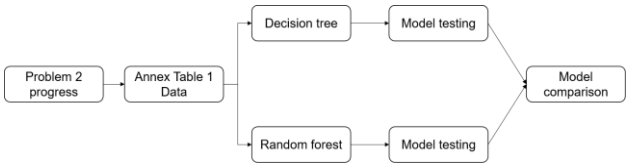

Figure 6. Problem 2 solution process

4.1. Model buiding and solving

Decision trees are supervised learning, a typical tree model, which embodies a functional mapping between features and labels. The non-leaf nodes represent the division of the sample on a feature, and the sample set is divided into several sub-trees according to the different values of the decision tree is divided into several sub-trees according to the different values of the feature, until the leaf node or the condition is satisfied. or a condition is met, the sample set is not split again. corresponds to a classification. The core problem of constructing a decision tree is how to select a suitable feature to split the current sample set.

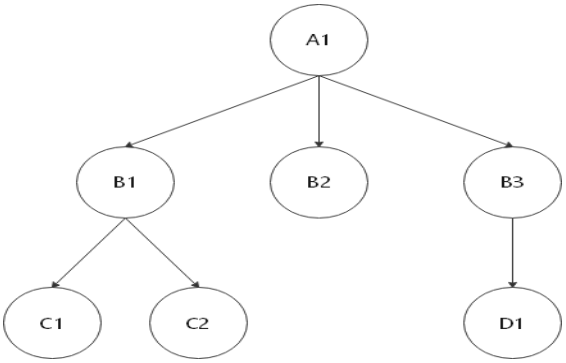

Figure 7. Decision tree structure diagram

There are two types of decision trees, depending on the target variable.

(1) Categorical variable decision tree: The target variable is classified according to the input characteristics. For example, suppose you are asked to predict the relative price of a computer into one of the following three categories: low, medium, or high. Features may include CPU, memory size, hard disk size, graphics card, sound card, etc. The decision tree will learn from these features, and after passing each data point to each node, it will eventually reach the 3 classification goals: low, medium, and high.

(2) Continuous variable decision tree: In this case, the features input to the decision tree (e.g. quality of a house) will be used to predict the continuous output (e.g. price of a house).

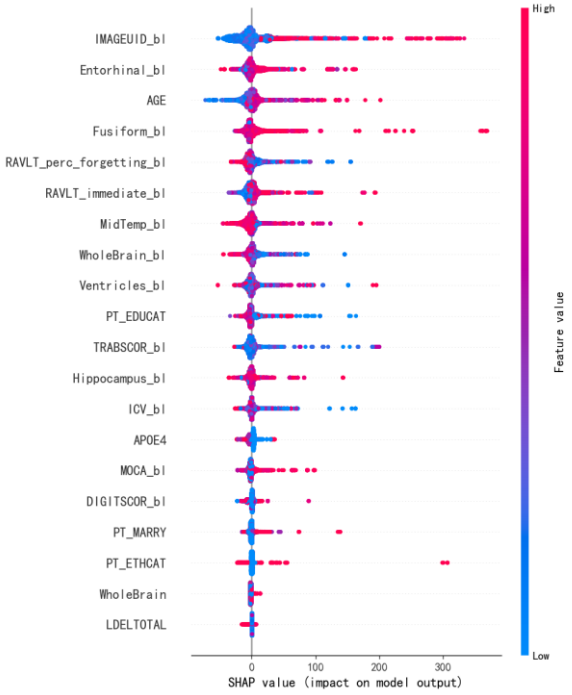

Figure 8. SHAP value (impact on model output)

1). Build a gradient boosted tree (GBDT) classification model from the training set data.

2). Calculate the feature importance from the established Gradient Boosted Tree (GBDT).

3). Apply the established Gradient Boosted Tree (GBDT) classification model to the training and testing data to obtain the classification evaluation results of the model.

4). Since the gradient boosting tree (GBDT) is random, the result of each operation is not the same, if this training model is saved, the subsequent data can be directly uploaded to this training model to calculate the classification.

5). Note: Gradient Boosting Tree (GBDT) cannot get the definite equation like the traditional model, and the model is usually evaluated by testing the data classification effect.

During the development of machine learning models, it is expected that the trained model will perform well on new, unseen data. To simulate new, unseen data, data partitioning is performed on the available data, thus splitting it into 2 parts (sometimes called training-test partitioning). In particular, the first part is a larger subset of the data used as the training set (e.g., 80% of the original data) and the second part is usually a smaller subset used as the test set (the remaining 20% of the data). It is important to note that this data splitting is performed only once. Next, a prediction model is built using the training set, and this trained model is then applied to the test set (i.e., as new, unseen data) for prediction. The best model is selected based on its performance on the test set, and hyperparameter optimization is also performed in order to obtain the best model [4].

As the below figure is a mixed matrix, firstly, all samples are divided into positive samples P (labeled as positive) and negative samples N (labeled as negative), the positive samples that are still positive after model prediction are noted as True Positive (TP), the positive samples that are predicted to be negative are called False Positive (FP), similarly, the negative samples that are predicted to be negative are called True negative (TN), and the negative samples that are predicted to be positive are called False negative (FN). samples that are predicted to be negative are called True negative (TN) and negative samples that are predicted to be positive are called False negative (FN), so after model prediction there are P = TP+FN and N = FP+TN.

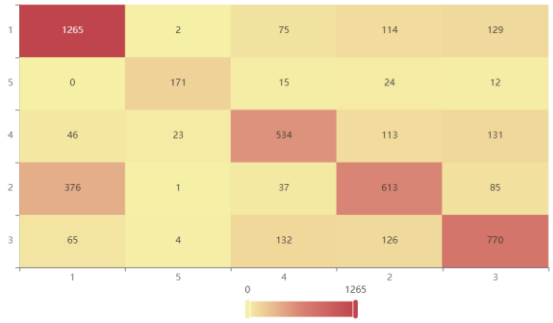

Figure 9. Confusion Matrix Heat Map

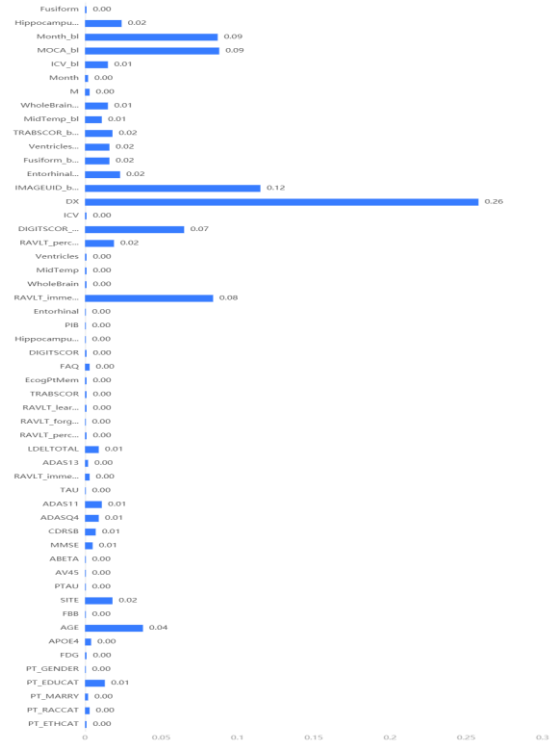

Figure 10. Feature Importance

The below table shows the prediction evaluation metrics of the cross-validation set, training set and test set to measure the prediction effectiveness of GBDT through quantitative metrics. Among them, the evaluation metrics of the cross-validation set can continuously adjust the hyperparameters to obtain a reliable and stable model.

- Accuracy rate: the proportion of correctly predicted samples to the total samples, the larger the accuracy rate, the better.
- Recall rate: The proportion of predicted positive samples out of the actual positive samples, and the larger the recall rate, the better.
- Precision rate: The proportion of the predicted positive samples to the actual positive samples, and the larger the precision rate, the better.

F1: The sum of precision and recall. Precision and recall affect each other, and although a high level of both is the desired ideal, in practice, it is often the case that a high precision rate results in a low recall rate, or a low recall rate results in a high precision rate. If a balance of both is needed, then the F1 metric can be used.

Table 2. Model evaluation results

	Accuracy Recall Accuracy	Accuracy Recall Accuracy	Accuracy Recall Accuracy	F1
Training set	1	1	1	1
Test set	0.689	0.689	0.687	0.687

The random forest algorithm is a highly flexible machine learning algorithm proposed by Leo & Culte et al. in 2001, which is based on statistics and combines Bagging integrated learning theory and random subspace methods. The basic idea is to set many weak classifiers into one strong classifier. The random forest consists of several classification trees, each corresponding to a different training subset. Each training subset consists of multiple random samples of the training set with put-back by the Bootstrap sampling method.

The k trees are constructed as described above, and these decision trees are grouped together to form a random forest. The random forest contains a second level of randomness, where m values are randomly selected from all the feature input values M to construct each decision tree; from these m values, each segmentation node is selected for optimization according to the principle of minimizing the Gini coefficient, thus reducing correlation and improving predictive power. The random forest simulation process does not perform the pruning process as in the decision tree. K predictors are obtained, and each predictor makes one prediction, and finally the predictions are aggregated and the average is calculated. The overall vector space is divided into smaller spaces by building individual decision trees, and the extracted decision trees are used as the training set.

The set of samples that are not extracted is called out-of-bag (OOB) data and is used as the validation set. The best branch is found by randomly selecting less than p alternative branches within each decision tree of the original data set. By adjusting the size of the leaf branches, the tree is stopped when it reaches 4 splits. The growth of the decision tree is controlled by adjusting the size of the leaf branches and stopping the formation of a single decision tree when it reaches 4 splits. A combinatorial model of the random forest algorithm is generated according to certain rules.

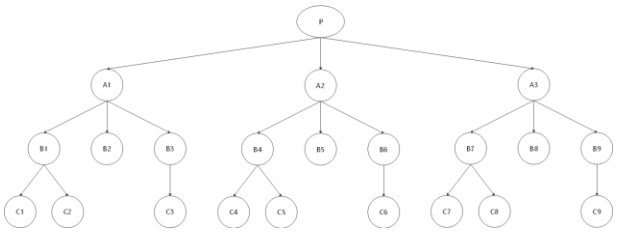

Figure 11. Random forest model structure diagram

When studying machine learning, we expect new, never-before-seen material that has been trained to yield good results. To simulate the new, unseen material, the already existing material is divided into two blocks (sometimes called the training-testing area). Specifically, said first component is a subset of the larger data (e.g., 80% of the initial data), while said second component is typically a much smaller subset (the remaining 20% of the data). It is important to note that such a splitting of the data can be performed only once. Next, this training set is used to construct a forecast model and the model is used for the test set (i.e., new information that has not been seen). By examining its performance, an optimal model was selected and optimized in terms of hyperparameters, and a better mathematical model was obtained.

A mixing matrix is shown below, where all samples are divided into P (marked as positive) and N (marked as negative), and samples that remain positive (TP) after prediction by the model are called false positive (FP) if they are predicted to be negative (FP), and if they are predicted to be negative (FP). The sample with negative prediction is called "TN", while the prediction for "FN" is "FN", so that after the model prediction, there will be P= TP+ FN, N= FP TP+ FN, N= FP+ TN.

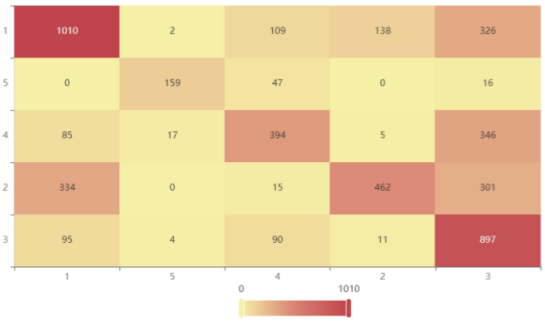

Figure 12. Feature Importance

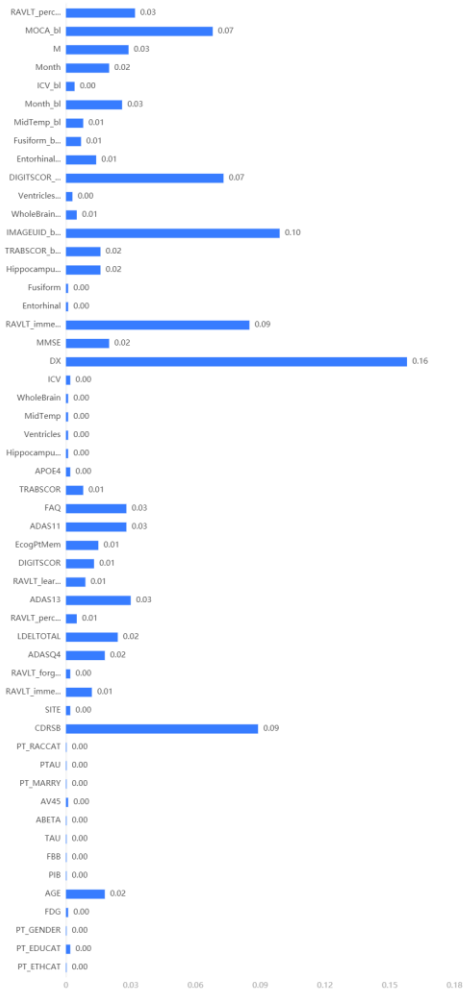

Figure 13. Feature Importance

The below table shows the classification evaluation metrics for the training and test sets, and the quantitative metrics are used to measure the classification effectiveness of the

decision tree on the training and test data.

● Accuracy: the proportion of correctly predicted samples to the total samples, the larger the accuracy, the better.

● Recall rate: The proportion of predicted positive samples out of the actual positive samples, the larger the recall rate, the better.

● Precision rate: The proportion of the predicted positive samples to the actual positive samples, and the larger the precision rate, the better.

F1: The sum of precision and recall. Precision and recall affect each other, and although a high level of both is the desired ideal, in practice, it is often the case that a high precision rate results in a low recall rate, or a low recall rate results in a high precision rate. If both are needed, then the F1 metric can be used.

● oob_score: For classification problems, oob_score is the accuracy rate of out-of-bag data. If there is put-back sampling selected during the tree building process, about 1/3 of the records are not extracted. The ones that are not extracted naturally form a control data set that can be used for model validation. Therefore, the random forest does not need to set aside additional data for cross-validation; its own algorithm is similar to cross-validation, and the out-of-bag error is an unbiased estimate of the prediction error.

Table 3. Model evaluation results

	Accuracy Recall Accuracy	Accuracy Recall Accuracy	Accuracy Recall Accuracy	F1
Training set	0.842	0.842	0.852	0.844
Test set	0.601	0.601	0.639	0.597

Based on this, this paper uses decision tree model and random forest model to classify the data, and constructs model accuracy indexes: confusion matrix, acc, etc. so as to judge the advantages and disadvantages of the classification model. By comparing the judgment indexes of decision tree model and random forest model, it is found that decision tree model has better classification effect.

5. Problem 3

The problem 3 model building and solution process is shown in Figure 13.

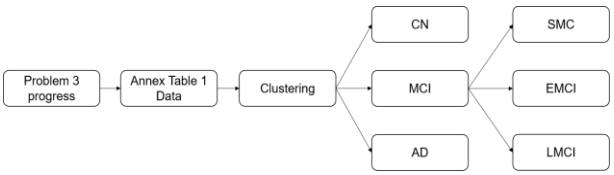

Figure 13. Problem 3 solution process

5.1. Model Establishment and Solving

The K-means clustering algorithm is a clustering algorithm that classifies based on a functional criterion, based on minimizing the clustering criterion function. The clustering criterion function is the sum of squares of the distances from each node in the cluster set to the center of the cluster of that class. For all M pattern classes criterion function is defined.

$$K = \sum_{k=1}^{M} \sum_{i=1}^{N_j} \left\| X_i - Z_j \right\|$$

Where S_j is the j^{th} cluster set with cluster center Z_j; N_j is the number of samples contained in the jth cluster set Sj. The clustering criterion of KM: choose a cluster center Zk Let the criterion function K be extremely small, i.e., make the value of Kk extremely small, we get

$$Z_k = \frac{1}{N} \sum_{i=1}^{N_k} X_i$$

Let there be N nodes, and the calculation steps are as follows:

Any M initial cluster centers $Z_1(1), Z_2(1), \ldots Z_M(1)$, M < N, and the serial numbers in parentheses represent the number of iterations.

The remaining nodes are assigned to M cluster centers according to the shortest distance principle to one of the M cluster centers, i.e.

$$\min\{\|X - Z_i(m)\|, i = 1,2,\ldots,M\} = \|X - Z_k(m)\|$$

Where m is the order number of the iterative operation.

Calculate the new vector values for each cluster center.

$$Z_k(m+1) = \frac{1}{N_k} \sum_{X \in S_k(m)} X$$

If $Z_k(m+1) \neq Z_k(m), j = 1,2,\ldots,M$, then return to reclassify the nodes one by one and repeat the iterative computation, and vice versa, the algorithm converges and the computation is completed.

Table 4. Model evaluation results

	DX_bl	SITE	AGE	PT_GENDER	PT_EDUCAT	PT_ETHCAT	PT_RACCAT	PT_MARRY	APOE4
1	106.52	70.86	2.00	16.26	1.18	1.34	1.64	0.40	0.13
2	80.51	73.16	2.00	16.05	1.04	1.13	1.40	0.53	0.41
3	81.93	73.41	2.00	16.20	1.04	1.13	1.39	0.48	0.04

Therefore, in this paper, we first clustered the overall into three classes, denoted as MCI, CN\AD. Then, this paper used K-means algorithm for clustering, and continued to cluster MCI into SMC\EMCI\LMCI thereby observing the difference between the clustering centers.

In this paper, we first clustered the overall into 3 classes, denoted as MCI, CN, and AD. Then, this paper used K-means algorithm for clustering and calculated the distribution of the values of 53 indicators under the three clustering centers. Under different clustering categories, the important factors affecting the classification are judged by comparing the magnitude of the values of the same indicators.

The above visualization reveals that among the 53 underlying indicators, a large number of indicators have little effect on clustering, which means that there is little difference between the values of the same indicators under different clustering categories, or even almost identical, but there are significant differences in variables such as FDG, AV45, FBB, and ABETA. Although the values differed little, their standardization revealed significant differences between the values of the same indicators under different clustering categories, which also provides a good basis for finding the etiology of Alzheimer's disease.

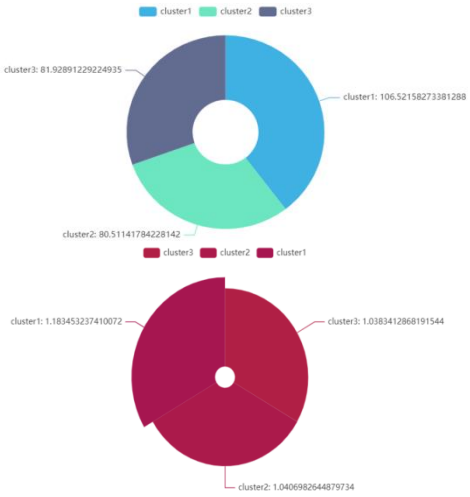

Figure 14. Visualization of clustering effects

5.2. Model Establishment and Solving

The clustering was continued by clustering the MCI into SMC, EMCI, and LMCI to observe the difference between the clustering centers.

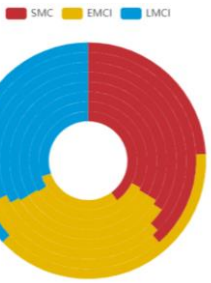

Figure 15. Visualization of clustering effects

From the above visualization, it is clear that many of the 53 basic indices have a small effect on the clusters, i.e.: in the classification, the differences are small and almost identical for the same index, but in FDG, AV45, FBB; different factors such as ABETA differ significantly. Despite the small difference between the two data, after normalization, we can see that there is a significant difference in the same index in each cluster under the same classification, which can find a good basis for the pathogenesis of Alzheimer's disease.

Table 5. Model evaluation results

	DX_bl	SITE	AGE	PT_GENDER	PT_EDUCAT	PT_ETHCAT	PT_RACCAT	PT_MARRY	APOE4
SMC	107.04	70.1	2.00	16.15	1.09	1.37	1.73	0.35	0.12
EMCI	79.75	72.3	2.00	16.08	1.04	1.14	1.38	0.56	0.41
LMCI	80.97	72.5	2.00	16.15	1.05	1.15	1.37	0.54	0.04

6. Problem 4

The problem 4 model building and solution process is shown in Figure 16;

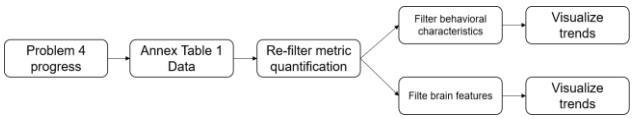

Figure 16. Problem 4 solution process

6.1. Model Establishment and Solving

This paper first tests the model based on temporal indicators, which are the degree of change in the patient's condition from different times of testing, respectively. Therefore, this paper screens the tabular data so that it can be analyzed using statistical knowledge. Also, for further analysis, this paper argues that the evolutionary pattern of different categories of diseases over time can be seen as changes in brain characteristics and cognitive characteristics over the time of testing. Also, due to the large size of the data, only a small sample of the five characteristics is selected for analysis in this paper.

Figure 17. Trend line graph of different brain features

Figure 18. Trend line graph of different brain features

From the above discounted graph of data trends, it can be seen that most of the indicators have a tendency to increase, and the magnitude of the values of a few indicators maintains within a certain fluctuation range, which proves that along with the passage of time, there is a characteristic change process in the degree of disease symptoms, and it can be seen from the graph that the values of RAVLT_immediate, RAVLT_learning, RAVLT_ forgetting and other indicators obviously have a decreasing trend, so when studying the changes of brain characteristics and cognitive characteristics with the detection time, this paper chooses to divide the data into behavioral characteristics and brain characteristics, and analyze different indicators under different types, where the data trend data of cognitive characteristics are shown below.

Table 6. Trend data for different brain characteristics

IMAGEUID	Ventricles	Hippocampus	WholeBrain	Entorhinal	Fusiform	MidTemp	ICV
118697	38712	4495	972134	2239	19818	20622	1524760
118672	38374	0	964788	0	0	0	1513350
35725	41869	0	971882	0	0	0	1514160
59854	44002	0	931047	0	0	0	1527030
83343	44139	0	945435	0	0	0	1523690
129868	49536	0	952495	0	0	0	1504090
205579	55599	0	912309	0	0	0	1524340

7. Summary

In this paper, we first determine the importance index of different indicators with Alzheimer's disease by constructing a correlation analysis through pre-processing of the data. We then use structural brain features and cognitive behavioural features to design an intelligent diagnostic model of Alzheimer's disease in which we choose to use a decision tree as a model for classification.

Secondly, we used a time series model to analyse the relevant data by arguing that the evolutionary patterns of different categories of disease over time can be seen as changes in brain features and cognitive characteristics over the time of detection.

Finally, we used the offering to review the etiology and onset symptoms of Alzheimer's symptoms, summarise and analyse the distribution of different indicators at different onset stages, and provide different early intervention and diagnostic criteria for the five categories of CN, SMC, EMCI, LMCI and AD.

References

[1] Du Jia, Fu Yan, Fan Jia, Zhou Miaorong & Zhang Yeting: Aerobic exercise intervention with Aβ_(1-42) induced changes in hippocampal synaptic structure and synaptic proteins in rats with Alzheimer's disease. Chinese Journal of Tissue Engineering Research.

[2] Lv Romance & Li Aiping. (2022). Role of perineuronal matrix network in regulating cognitive function and Alzheimer's disease. Progress in Physiological Science (02),100-104.

[3] Qu Wenying, Xie Jianguo, Liang Anxin & Shang Xuefeng. (2021). Effects of astragalus polysaccharides on nerve cell activity, cognitive function and Caspase-9 expression levels in rats with Alzheimer's disease. Stroke and Neurological Diseases (05), 543-549.

[4] MENG Kaitao, ZHANG Jianguo, LIU Chong, MA Xiuhong & SONG Yuqiang. (2021). Correlation analysis between serum miR-128 and miR-223 expression levels and inflammatory response and cognitive function in patients with Alzheimer's disease of different severity. Stroke and Neurological Diseases (06), 667-671.