Problem Chosen	2022 ShuWei Cup Summary Sheet	Team Control Number
C		2022090512506

An intelligent diagnosis model for Alzheimer's disease based on BP neural network

summary

When the memory disappears, the soul disappears with it. Data show that every 3 seconds, there will be one more Alzheimer's disease patient in the world. In this paper, we build a deep learning model through BP neural network to achieve an intelligent judgment model for Alzheimer's disease.

For problem 1,the data in the appendix were processed for missing values, and data with more than 50% missing values of features were not considered. The remaining data were cleaned to remove the abnormal data, and the missing values were filled in with sample bar interpolation and the variables were dimensionless, and finally the correlation between the data features was construc-ted with principal component analysis, and it was concluded that there was a correlation between the study data features and Alzheimer's disease at 99% confidence level.

For problem 2,based on the first question preprocessing, BP neural network simulation was establis-hed with structural brain features and cognitive-behavioral features as input, and finally validated using a support vector machine model, obtaining a correctness of 86.7% under this model.

For problem 3,a large number of samples were generated using the smote algorithm, followed by random forest classification and logistic regression classification to classify CN, MCI, and AD, and again, the large number of samples generated by smote were clustered by fuzzy cluster analysis to reduce the MCI subclasses, and finally, the clusters were evaluated by cross-validation method, and it was found that EMCI, SMC, Finally, the clustering was evaluated by cross-validation, and the EMCI, SMC, and LMCI ratios were 45.3%, 68.85%, and 38.32%, respectively, and it was found that the combination of biomarker and neurocognitive scale data with brain image data could improve the classification accuracy to 78.3%.

For problem 4, by establishing a multiple regression model for fitting, judging whether the feature information has a good linear relationship and establishing the trend of feature information over time, the accuracy of fitting was obtained as 98.13%.

For problem 5, an intelligent and smart diagnostic system for Alzheimer's disease based on the second question and a review of the relevant literature to come up with targeted recommendations for early intervention and diagnostic criteria for the five disease categories of CN, SMC, EMCI, LMCI and AD.

Finally, the model used in this paper is evaluated and analyzed to make it more perfect, scientifically reliable, and in line with the actual needs.

Keywords: Cluster analysis, grey correlation, support vector machines, time series

Content

1 Introdurtion

1.1 Background of the problem

Alzheimer's disease is a neurodegenerative disease that develops slowly and worsens over time. The number of people with Alzheimer's disease is increasing (see Appendix A for an increase in numbers).

The most common early symptom of Alzheimer's disease is the loss of short-term memory, and it is clinically characterized by a range of dementias, including memory impairment, aphasia, language impairment, disorientation, mood instability, loss of motivation, inability to care for themselves and many behavioral problems. Advanced Alzheimer's patients can lose their memories and even forget their family and friends. They need to be cared for completely, which can be both physical and emotional. The entire course of Alzheimer's disease can take 8 to 10 years, with gradual physical and mental decline, until the loss of mental ability and eventually death. So it's often said that people with Alzheimer's disease die twice, a mental death and a physical death. But if early symptoms are recognised and diagnosed and treated as early as possible, there is a chance of delaying progression and improving quality of life for patients. Therefore, early and accurate diagnosis of Alzheimer's disease and mild cognitive impairment is of great significance.

1.2 Our Approach

For problem 1: Missing value processing should be carried out first. Since the missing value of some features is above 50%, we will not consider these features with high missing value. Then we cleaned the remaining data, cleared the abnormal data, filled in the missing values with spline interpolation and carried out dimensionless processing of variables. Finally, principal component analysis was used to build the correlation of data features.

For problem 2: The brain structural features and cognitive behavior features are taken as input, while CN, SMC, EMCI, LMCI and AD are taken as output for complex function mapping problem, BP neural network simulation is carried out with sigmod function as activation function, and finally the support vector machine model is used for verification.

For problem 3: First, use smote to produce plenty of sample data, second, then categorise CN,MCI,AD using random-forest classification and logistic regression classification. Then, a large number of samples generated by smote were clustered by fuzzy cluster analysis to reduce the dimensionality of MCI subclasses, and finally, the clusters were evaluated by cross-validation method.

For problem 4: By establishing a multiple regression model for fitting, we judge

whether the feature information has a good linear relationship, establish the trend of feature information over time, and analyze the feature evolution law under different time points

For problem 4: Early intervention and diagnostic criteria for the five categories of CN, SMC, EMCI, LMCI and AD by using the intelligent smart diagnostic system for Alzheimer's disease established in the second question and reviewing the relevant literature, and making some suggestions.

Based on the above analysis we need to solve the following problems :

- Preprocess the characteristic indicators of the attached data to investigate the correlation between data characteristics and the diagnosis of Alzheimer's disease.
- Use the attached structural brain features and cognitive behavioral features to design an intelligent diagnosis of Alzheimer's disease.
- 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.
- The same sample in the annex contains features collected at different time points, please analyze them in relation to the time points to uncover patterns in the evolution of different categories of diseases over time.
- Please consult the relevant literature to describe the early intervention and diagnostic criteria for the five categories of CN, SMC, EMCI, LMCI, and AD.

2 Problem assumption

- It is assumed that some of the outliers provided in the annexes of this paper are due to errors in drawing samples or recording data.
- It is assumed that each sample is independent and unrelated to each other, and that their respective indicators do not interfere with each other.
- Assuming the generated feature matrix, the rows and columns are not invertible and are independent of each other.
- It is assumed that the data provided in the annex of this question are true and reliable.

3 Symbol Description

Important notations used in this paper are listed in Table 1,

Table 1：Notations

Symbol	Descriptions
x	Sample Mean
S^2	Sample variance
N	Number of samples in the overall population
t	t-test statistic
b	Skewness
a_1	Kurtosis
a_2	Threshold
a_3	Correlation degree
a_4	Connection weights
σ	Standard deviation of samples
v_b	Residual error
R	Correlation coefficient matrix
R^2	Sample variance
j	Contribution margin
r_s	Spearman's correlation coefficient

4 The model

4.1 Model establishment and solution of problem 1

Firstly, the data given in the attachment is preprocessed. Too much data in the attachment will lead to increased errors in the process of data recording. Therefore, this paper adopts the to clean these data.Let the measured data be meas-ured with equal

precision and independently obtain $x_1, x_2 \ldots, x_n$, calculate its arithme-tic mean x and the remaining error $v_i = x_i - x(i = 1, 2, \cdots, n)$,and in accordance with Bessel's formula,the standard deviation is calculated as δ[1],if a measured value x_b theresidual error of $v_b (1 \le b \le n)$, then the following equation is satisfied.

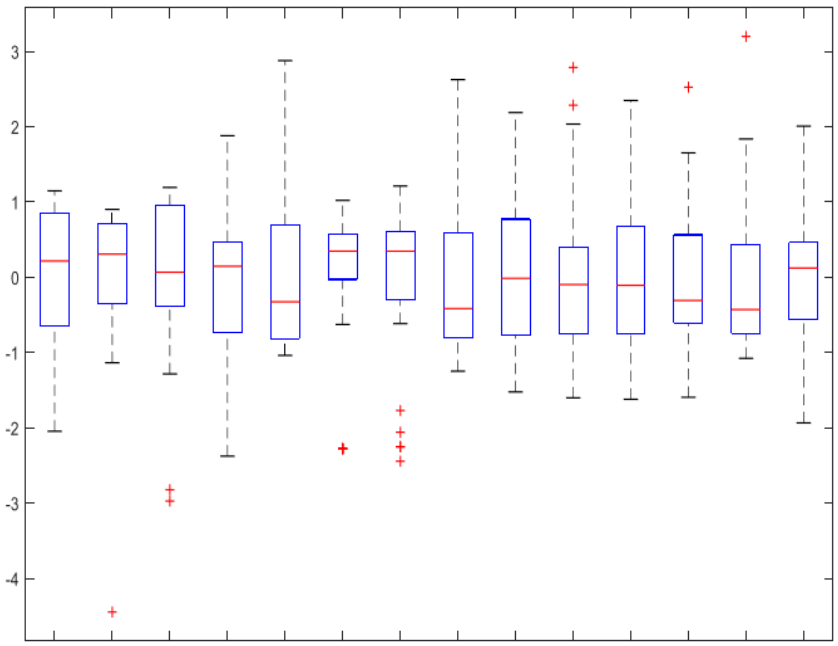
$$|\nu_b| = |\chi_b - x| > 3\delta \tag{1}$$

We consider x_b to be a bad value containing a large error value and should be excluded.

According to hypothesis 1, the Raida's rule， It can be assumed that the range of the given values between $x \pm 3\delta$, So the data in this interval we can consider as the normal range of values, then the data outside this interval are abnormal values should be modified or rejected.

By some data distribution in the attachment does not conform to the normal distribution, the common methods used to detect outliers are the Law of 3δ and the Z distribution method cannot be used. By observing that some data distributions in the characteristic information of Alzheimer's disease do not conform to a normal distribution, box line plots, which are not required for the data distribution, were selected for the detection of their data.

The principle of the box line diagram is to calculate the quartile plus or minus 1.5 times the value of the



quartile distance, i.e., to calculate the value of $Q_1 \pm 1.5IQR$. We specify that the value that falls within the $(Q_1 - 1.5IQR, Q_1 + 1.5IQR)$ interval is the value that conforms to common sense, and the value that falls outside the interval is an outlier. In this paper, the upper quartile will be used to represent data greater than $Q_1 + 1.5IQR$ and the lower quartile will be used to represent data less than $Q_1 - 1.5IQR$, and then box line plots will be drawn based on these data.

Figure 1：Box plots

4.1.1 Spline interpolation and its application

After the pre-processing of the data, some abnormal data are not eliminated, which will lead to the extreme case of missing data, so a new method of spline interpolation is used to complete the missing data. The sample interpolation method is as follows：

We first take n+1 interpolated nodes on [a,b]

$$a = x_0 < x_1 < \Lambda < x_n = b \tag{2}$$

The function f (x) is known to have the function value $y_k = f(x_k)$ at these $n+1$ points, then the mth sample-valued interpolation function S(x) of the function $y_k = f(x_k)$ on the interval [a b] satisfies the following condition: S(x) is continuous on (a,b) up to m - 1 order derivatives. $S(x_k) = y_k (k = 0, 1, 2, 3\Lambda, n)$ In the interval $x_k \in [x_k, x_{k+1}] (k = 0, 1, 2, 3\Lambda, n-1)$, S(x) is an mth order polynomial.

We Annex I missing data using three times spline interpolation, through the historical data, to establish the interpolation fit curve, the specific results obtained. As shown in Table2.

Table 2：Cubic spline interpolation results

Missing value	1	2	3	4	5
Fitted value	0.75	1.38	1.85	1.21	3.26

There may also be some size differences between the attributes of the various ratability metrics due to their size differences within the unit length of time they are measured or possibly between magnitudes. Therefore, in order to eliminate the influence of these unknown influences that may lead to the resulting training data and to ensure the best practical availability of the training statistics set data, we must also

consider the pre-processing of the normalized mapping between the training set data and the test set data according to the relevant laws of numerical model optimization and computational optimization, using the following normalized mapping:

$$f:\ x \to y = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3}$$

The dimensionless data is reduced to the range [0,1], which is conducive to the establishment of statistical evaluation models in the same system. Where $x_{\min} = \min(x)$ $x_{\max} = \max(x)$ $x, y \in R$. The effect of normalization is that the original data is structured into the range [0,1].

4.1.2 Canonical correlation analysis

Assume that AD is U_i namely diagnosed Alzheimer's disease, and each influencing index factor is V_i. Select several representative comprehensive variables U_i and V_i from the two groups of variables, so that each comprehensive variable is a linear combination of original variables, namely：

$$\begin{aligned} U_i &= a_1^{(i)} X_1^{(1)} + a_2^{(i)} X_2^{(1)} + \cdots + a_p^{(i)} X_p^{(1)} \triangleq \boldsymbol{a}^{(i)} \boldsymbol{X}^{(1)} \\ V_i &= b_1^{(i)} X_1^{(2)} + b_2^{(i)} X_2^{(2)} + \cdots + b_q^{(i)} X_q^{(2)} \triangleq \boldsymbol{b}^{(i)} \boldsymbol{X}^{(2)} \end{aligned} \tag{4}$$

Where the assumption is that：

$$\begin{aligned} X^{(1)} &= (X_1^{(1)}, X_2^{(1)}, \cdots, X_P^{(1)}) \\ X^{(2)} &= (X_1^{(2)}, X_2^{(2)}, \cdots, X_q^{(2)}) \end{aligned} \tag{5}$$

Of course, the number of groups of synthetic variables is uncertain. In order to have accurate data, it is necessary to ensure that the first and second sets of data are not correlated. that is：

$$cov(U_1, U_2) = cov(V_1, V_2) = 0 \tag{6}$$

- **Step 1.** Firstly, the influencing factor matrix is established

$$X=\begin{bmatrix} x_{11} & x_{12} & \Lambda & x_{1p} \\ x_{21} & x_{22} & \Lambda & x_{3p} \\ x_{31} & x_{32} & \Lambda & x_{3p} \\ M & M & M & M \\ x_{p1} & x_{p2} & \Lambda & x_{pp} \end{bmatrix} \tag{7}$$

- **Step 2.** Standardize the raw data

$$\overset{*}{x}_{ij}=\frac{x_{ij}-\bar{x}_j}{\sqrt{Var(x_j)}}(i=1,2,...,n;j=1,2,...p) \tag{8}$$

among $\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}, Var(x_j)=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)2,(j=1,2,...p).$

- **Step 3.** Calculate the sample correlation coefficient matrix.

$$R=\begin{bmatrix} r_{11} & r_{12} & \Lambda & r_{1p} \\ r_{21} & r_{22} & \Lambda & r_{2p} \\ M & M & M & M \\ r_{p1} & r_{p2} & \Lambda & r_{pp} \end{bmatrix} \tag{9}$$

For convenience, it is assumed that the original data is still represented after standardization, then the correlation coefficient of the data after standardization is:

$$r_{ij}=\frac{1}{n-1}\sum_{i=1}^{n}x_{ij}x_{ij}(1,2,...p) \tag{10}$$

- **Step 4.** The Jacobian method is used to obtain the eigenvalue # and the corresponding eigenvector of the correlation coefficient matrix.
- **Step 5.** Select the important principal component and write the principal component expression. Importance is reflected by contribution rates:

$$\frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i} \tag{11}$$

The larger the contribution rate is, the more important the main component is and the stronger the information of the original variable it contains.

- **Step 6.** Establish the correlation coefficient matrix of variables

Correlation analysis refers to the quantitative analysis of different indicators, so as to judge the relationship between them. The matrix expression of correlation coefficient is as follows

$$R=\begin{bmatrix} r_{11} & r_{12} & \Lambda & r_{1p} \\ r_{21} & r_{22} & \Lambda & r_{2p} \\ M & M & M & M \\ r_{p1} & r_{p2} & \Lambda & r_{pp} \end{bmatrix} \tag{12}$$

In this paper, the linear correlation coefficient is selected,,that is

$$r_{ij}=\frac{\sum_{k=1}^{n}\left(x_{ki}-\bar{x}\right)\left(x_{kj}-\bar{x}_{j}\right)}{\sqrt{\sum_{k=1}^{n}\left(x_{ki}-\bar{x}_{i}\right)^{2}\sum_{k=1}^{n}\left(x_{kj}-\bar{x}_{j}\right)^{2}}} \tag{13}$$

Rinciple of correlation analysis: The value of correlation coefficient r is between -1and 1. When the value of r is -1 or 1, the two variables are linearly dependent. The closer |r| is to 1, the greater the degree of linear correlation between variables; On the contrary, the degree of linear correlation is less. The positive and negative of r indicates the positive and negative correlation between variables, that is, if one variable increases with the increase of another variable, it is positive correlation, and if one variable decreases with the increase of another variable, it is negative correlation. In particular, if r=0, the variables are linearly independent. MATLAB solution can be obtained:

Table3：Canonical correlation

Correlation	Eigen value	Wilk's statistics	F	Molecular degrees of freedom	Denominator degree of freedom	Conspicuousness
0.844	2.483	0.194	17.970	9.000	168.078	0.000
0.562	0.463	0.675	7.590	4.000	140.000	0.000
0.110	0.012	0.988	0.870	1.000	71.000	0.354

The H0 for Wilks test means that the correlation in both the current and subsequent rows is zero

Assuming that the significance is p, and observing the P-value of the last column, we find that the first pair and the second pair can further explore the specific relationship between the variables. Check the "typical correlation table". In the table, the typical correlation of the first pair and the second pair of typical variables is not 0, and the significance P=0.000<0.05, while the significance of the remaining group is greater than 0, indicating that a pair of typical variables are extracted from the two groups of variables, indicating that there is a correlation between structural variables and product performance at the 99% confidence level. Moreover, the correlation between the first and the second pair of typical variables is extremely significant, which is in an extreme state, so the second pair is selected for analysis. The third pair had a P-value of 0.354. The correlation of the third pair of typical variables is not significant. Check the standardized canonical correlation coefficient of "Set 1" and "Set 2". The standardized canonical coefficient of the first pair of canonical variables in Set 1 is (-1.092, 0.021, -0.156), and the corresponding relationship can be obtained:

$$CV_{1-1} = -1.092 * y_1 + 0.021 * y_2 - 0.156 y_3 \tag{14}$$

Similarly, by looking at the pair of typical variable numbers in set 1 and set 2, we can get:

$$\begin{aligned} CV_{1-2} &= 1.642 * y_1 - 1.216 * y_2 + 0.437 * y_3 \\ CV_{2-1} &= 0.859 * Z_1 + 0.640 * Z_2 + 0.437 * Z_3 \\ CV_{2-2} &= 0.587 * Z_1 - 1.452 * Z_2 - 0.847 * Z_3 \end{aligned} \tag{15}$$

After the second pair is selected, the following table is obtained after the data of the first pair and the third pair is deleted:

Table4：Eigenvector matrix

Indicatoreig Envector	1	2	3	4	5	6
SITE	-O.451	-O.OO7	0.462	-0.447	0.072	-O.615
AGE	-0.392	0.210	0.510	0.717	-0.092	0.136
PTEDUCAT	-0.310	0.557	-0.593	0.113	-0.332	-O.346
FDG	-0.485	-0.061	-0.329	0.058	0.790	0.160
ADAS13	-0.510	-0.052	-0.005	-0.428	-0.410	0.628
DIGITSCOR	0.245	0.800	0.258	-0.294	0.290	0.253

Finally, the contribution rate formula is used： Contribution rate $=\frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i}$,

The contribution rate of each component is calculated as follows：

Table5： Contribution rate of each component schematic table

Principal component	1	2	3	4	5	6
Rate of contribution	0.571	0.179	0.145	0.063	0.024	0.018

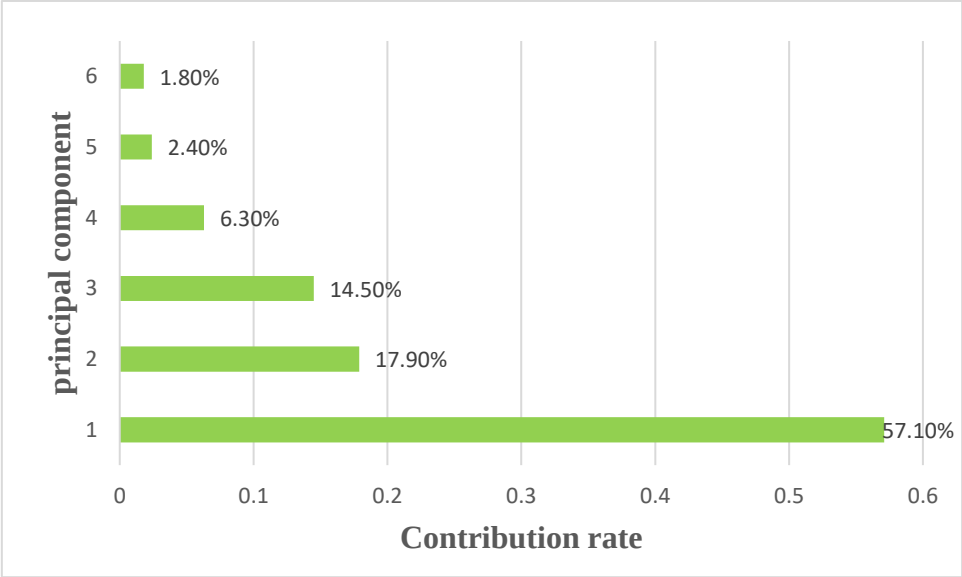


Figure2: Contribution rate of each component schematic table

4.1.3 Result analysis

The eigenvector matrix and contribution rate of each component were obtained by principal component analysis. Further analysis of Table 2 and Table 3 above shows that：The contribution rate of the first three principal components is as high as 89.5%, so they are the main components affecting AD Alzheimer's.

The first principal component is

$$y_1 = -0.451x_1 - 0.392x_2 - 0.310x_3 - 0.485x_4 - 0.501x_5 + 0.245x_6.$$

The second principal component is

$$y_2 = -0.007x_1 + 0.210x_2 + 0.557x_3 - 0.061x_4 - 0.052x_5 + 0.800x_6.$$

The third principal component is

$$y_3 = 0.462x_1 + 0.510x_2 - 0.593x_3 - 0.329x_4 - 0.005x_5 + 0.258x_6.$$

Among them, FDG and SITE were the main influencing factors in the first principal component.

4.2 Model establishment and solution of question 2

4.2.1 Establishment and solution of model

Problem 2 can be regarded as a complex function mapping problem with brain structural features [X1] and cognitive behavioral features [X2] as inputs,CN[Y1], SMC[Y2], EMCI[Y3] and AD[Y4] as outputs, with five layers of hidden layer and sigmod function as activation function for BP neural network simulation. The process is as follows：

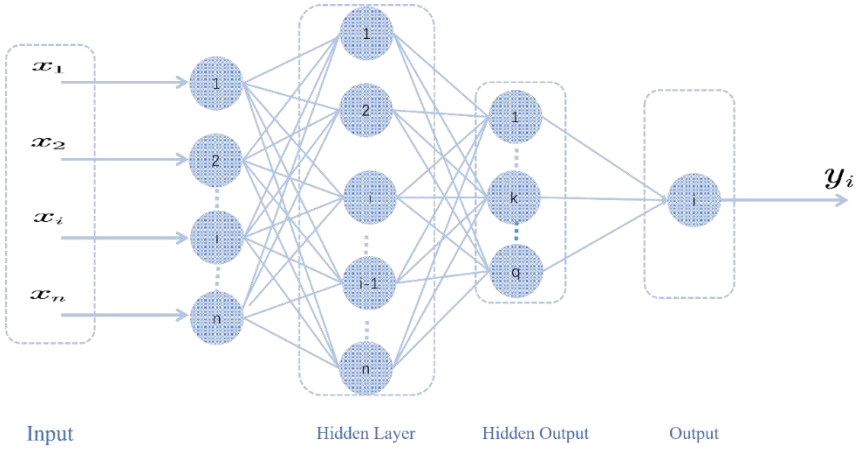


Figure3：Neural network flow chart

Step 1. Training set and test set

After the elimination and normalization of outliers, the training set is the first 75% data after the shuffled order, and the verification set is the last 25% data.

Step 2. Construction of BP neural network

BP neural network includes input layer, five hidden layer and output layer, the number of neurons is 2, the output node is 3, from the output layer to the hidden layer and between the hidden layer, sigmoid function is used as the activation function, from the last hidden layer to the output layer between the perelin function. Get the expression between the network input data and the output data as shown below

$$y_k{}' = \sum_{j=1}^{r} v_j . f\left[\sum_{i=1}^{m} w_{ij} . p_i + \vartheta_j\right] \tag{16}$$

Where (k=1,2... N), W_{ij} is the link weight, θ_i is the threshold, y_k is the expected output value, and y_k is the actual output value of the network. The neural network structure diagram is shown in the figure below. The brain structure feature [X1] and cognitive behavior feature [X2] are the input,CN[Y1], SMC[Y2], EMCI[Y3], AD[Y4] are the output.

Step 3. Parameters of BP neural network

BP neural network mainly includes the following parameters for prediction: Maximum training steps net.tranParam.epochs, training result interval steps net.trainParam.show, learning rate net.trainParam.1r, training target error net.trainParam.goal. The parameters set in this model are as follows:

Table6:

Maximum number of training steps	Training results Interval steps	learning rate	Training target error	Number of training
1000	1	0.0000001	0.000001	1000

It is not difficult to see that the prediction result of neural network is good from the goodness of fit and norm.

Table7：BP neural network results

ADASQ4	ADAS13	RAVLT_perc_forgetting	DIGITSCOR	ADAS11_bl
38	850	2.8044	96.2449	84.6934
33	950	2.5615	96.0928	87.7469
28	1150	2.5911	96.1860	87.5895
23	1250	2.3932	95.8179	85.1191
38	1250	3.4366	96.7624	83.7734
33	1150	2.9657	96.5570	84.3172
28	950	2.2212	95.3950	88.1711
23	850	1.6978	93.9765	86.3594

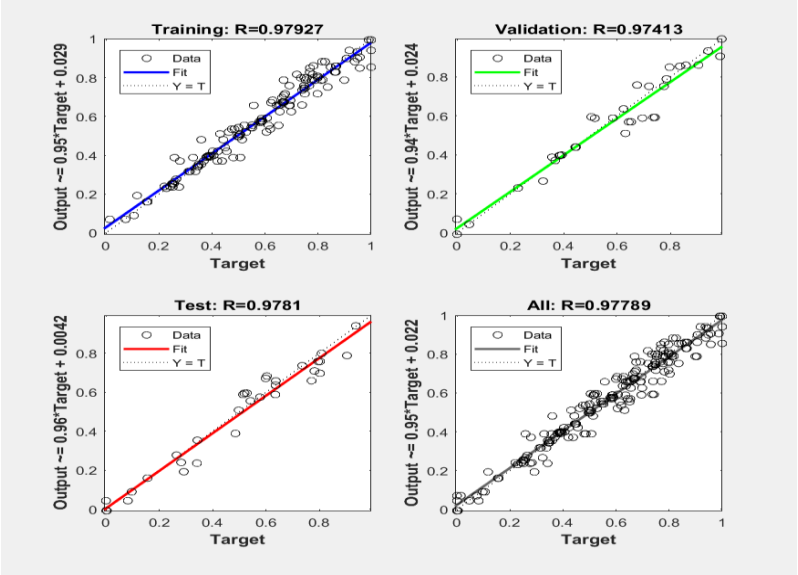


Figure4：Prediction of regression effect of BP neural network model

It can be seen from the regression effect that the correlation coefficient R value is above 0.97, indicating that the BP neural network is more accurate for CN[Y1], SMC[Y2], EMCI[Y3], AD[Y4], and the norm is 0.7, the accuracy is above 90%, showing good performance and accurate results.

4.2.2 Support vector machine model supplemented by verification

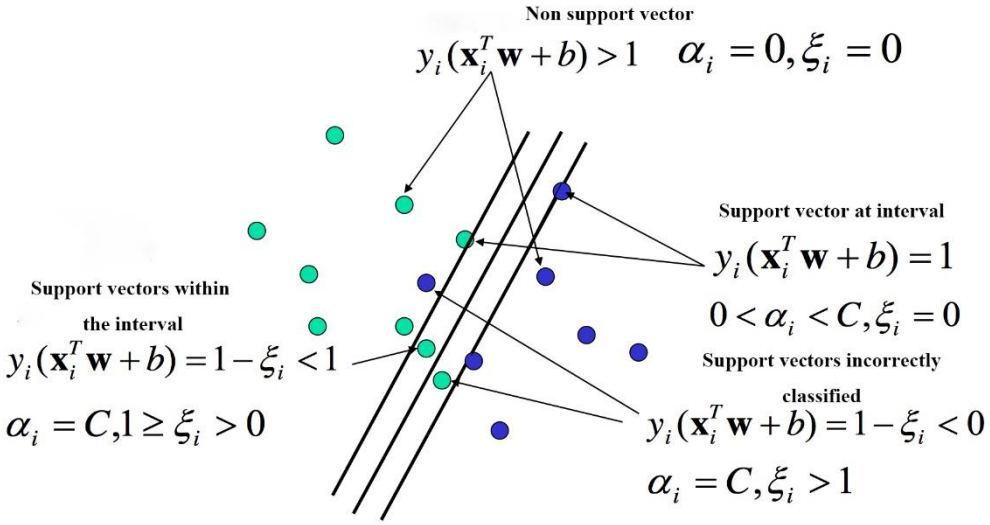


Figure5：Schematic diagram of support vector machine

SVM is a binary classification model, which is a linear classifier defined in the feature space with the largest interval. The basic idea of using SVM is to solve the separation hyperplane that can correctly partition the training data set and has the largest geometric interval. The flow chart is as follows.

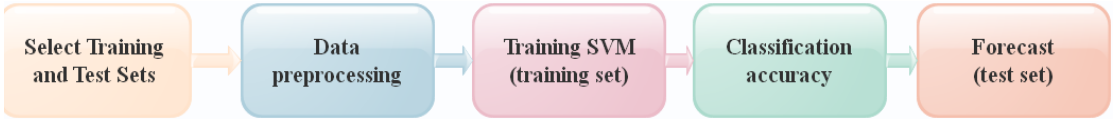


Figure6：flow chart

First of all, the training set and test set should be extracted from the original data, and then the predetermined preprocessing (feature extraction is required when necessary). After that, the training set is used to train the SVM, and finally the obtained model is used to predict the classification label of the test set.

(a) Selection of training and test sets

Now divide each category into two groups and recombine the data, one as a training set (train_wine) and the other as a test set (test.wine).

(b) Training and forecasting

The training set train_wine was used to train the SVM classifier, and the obtained model was used to predict the label of the test set. Finally, the classification accuracy was obtained.

Table8：Results of model

Arithmetic	Number of old people	AD quantity	Accuracy
SVM	1738	1507	86.7%

4.3 Establishment and Solution of Problem 3

4.3.1 Preparation of model

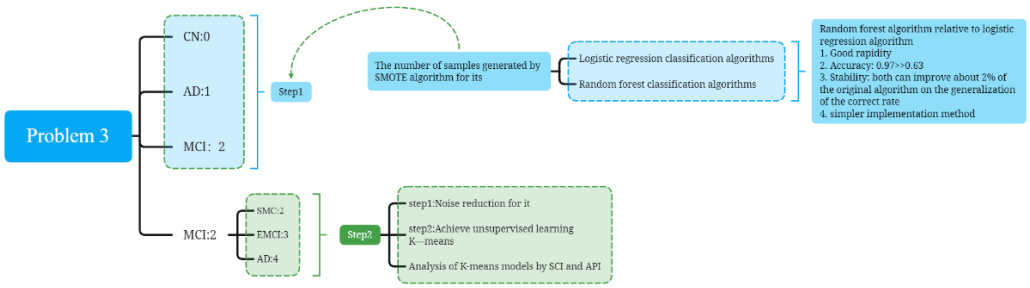


Figure7：Process diagram based on Process three

4.3.2 Establishment of model

- **Step 1.**Realization of SMOTE

Firstly, samples are extracted from the pre-treated samples of the first question, and the Euclidean distance is used as the standard to calculate the distance between them

and all samples in a minority sample set to obtain their K-nearest neighbors. Where, the Euclidean distance is:

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \quad (17)$$

Secondly, a sampling ratio is set according to the sample imbalance ratio to determine the sampling multiplier N. For each minority sample, a number of samples are randomly selected from its K-nearest neighbors, assuming that the selected nearest neighbors are.

Finally, for each randomly selected neighbor, a new sample is constructed through the following formula.

$$x_{new}=x+rand(0,1)\times\left(\tilde{x}-x\right) \quad (18)$$

- **Step 2.**

CN,MCI and AD were regarded as dependent variables, and the brain structural characteristics and cognitive behavior of the second question were regarded as independent variables to establish the logistic regression classification function:

$$g(z)-\frac{1}{1+e^{-z}} \quad (19)$$

$$z=\theta_0+\theta_1x_1+\cdots+\theta_nx_n=\sum_{i=1}^{m}\theta^Tx \quad (20)$$

Where z is the predicted value

In this way, any input can be mapped to the interval [0,1]. In other words, a predicted value can be obtained from the regression analysis, and then mapped to the sigmoid function. In this way, the conversion from value to probability can be realized and the task of classification can be realized. The accuracy of sigmoid function graph and logistic regression classification obtained by python:

Table9： RandomForestClassifier was used to implement the classification

	precision	recall	f1-score	support
0	0.88	0.64	0.74	2152
1	0.18	0.42	0.25	181
2	0.33	0.54	0.41	553
accuracy			0.67	2886
macro avg	0.46	0.53	0.47	2886
weighted avg	0.73	0.60	0.64	2886

- **Step 3.** Random Forest Classifier is used to implement the classification

The samples generated by step1 are randomly extracted, repeated N times, and k eigenvalues are randomly selected to train the decision tree, and a decision forest is formed through continuous training of the decision tree. The flow chart is as follows:

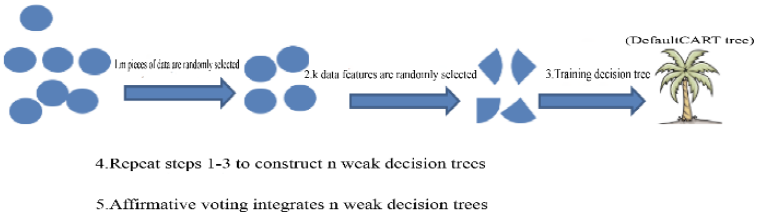


Figure8: RandomForestClassifier was used to implement the classification

Pandas is used in python to create the following table:

Table10：Random forest classification effect table

	precision	recall	f1-score	support
0	0.93	0.98	0.95	1474
1	0.89	0.90	0.89	423
2	0.94	0.86	0.90	989
accuracy			0.97	2886
macro avg	0.92	0.91	0.91	2886
weighted avg	0.93	0.93	0.92	2886

Step 4. Through the comparison of logistic regression classification and random forest classification, the accuracy of forest is more stable than that of random forest. The following improvements have been made:

1.It has good rapidity

2. Accuracy：0.97>>0.63

3. Stability: the accuracy of generalization can be improved by about 2% on the original algorithm

4. The implementation method is simpler

4.3.3 Reduced dimension clustering of MCI subclasses based on fuzzy clustering analysis

The flow chart and steps are as follows:

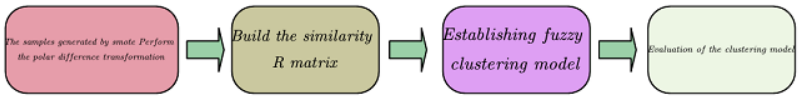


Figure9：Fuzzy clustering algorithm flow

- **Step 1.**Pretreat the samples first: by sampling the samples generated by the third ask smote algorithm, to establish the evaluation of SMC,LMCI,EMCI information matrix

$$A=\begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{pmatrix} \tag{21}$$

- **Step 2.** On the sampling data generated by smote, use range transform:

$$y_{ij}=\frac{x_{ij}-\min(x_i)}{\min(x_i)+\max(x_i)} \tag{22}$$

Where is the known evaluation index and is the standardized value.

- **Step 3.** The similarity matrix is established. The maximum and minimum method is adopted to establish the similarity matrix:

$$R=\begin{bmatrix} r_{11} & r_{12} & \Lambda & r_{1p} \\ r_{21} & r_{22} & \Lambda & r_{2p} \\ M & M & M & M \\ r_{p1} & r_{p2} & \Lambda & r_{pp} \end{bmatrix} \tag{23}$$

$$\text{Among，}\ r_{jk}=\frac{\sum_{i=1}^{n}\min(x_{ij},y_{jk})}{\sum_{i=1}^{n}\max(y_{ij},y_{ik})}$$

- **Step 4.** The transitive closure of similar matrix is solved by the flat method and the results are clustered by transitive closure. For transformation, find its transitive closure. The solution of similarity matrix is consistent with the solution of correlation in problem 1.

The transfer closure $\hat{R}=R^4$ can be obtained by the flat method, and the specific data is shown in Table9.

Table11 ： Transitive closure matrix

1	0	0.28	0.28	0.28
0	1	0	0	0
0.28	0	1	0.4517	0.4517
0.28	0	0.4517	1	0.7863
0.28	0	0.4517	0.7863	1

4.3.4 Cluster result analysis

The transitive closure $\hat{R}=R^4$ is the fuzzy equivalence relation matrix, and the clu-ster analysis is carried out on the three sub-factors of mci. The fuzzy clustering is carried out on the three sub-factors according to different threshold level λ, and different classi-fication results will be obtained. The clustering diagram is shown as follows:

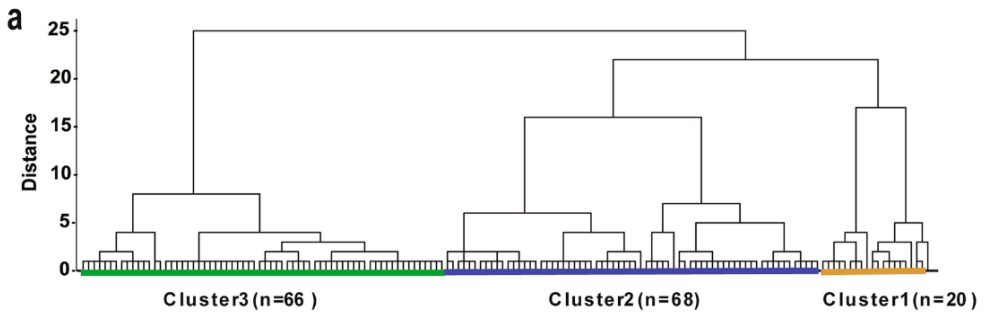


Figure10: Cluster analysis diagram

4.3.5 Evaluation of the model：

The cross-validation method is used to train and evaluate the fuzzy clustering algorithm. The steps are：

1. At random, take 70% of the smote generated sample as the training set and 30% as the test set.

2. The model evaluation indexes mainly include accuracy rate, recall rate, accuracy rate, F1 score and AUC under the receiving operation curve.

3. 100 training sets and verification sets were generated by random division, and the average value of each index above was calculated as the final result.

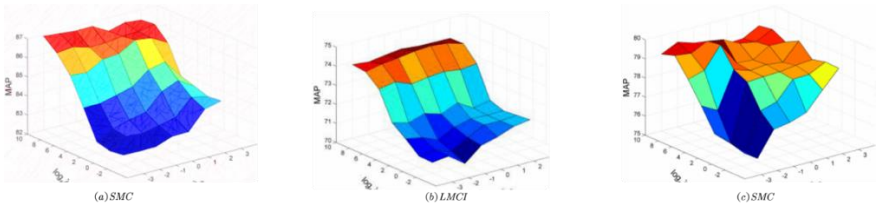


Figure11：model evaluation

Using brain image data to simultaneously classify three disease states, the overall classification accuracy was only 55.66%. The proportion of SMC and LMCI groups correctly classified as EMCI group was 45.33%, 68.85% and 38.32%, respectively. Combining the brain image data with biomarkers and neurocognitive scale data can improve the accuracy of classification.

4.4 Establishment and Solution of Problem 4

4.4.1 Preparation of model

- **Our research ideas：**

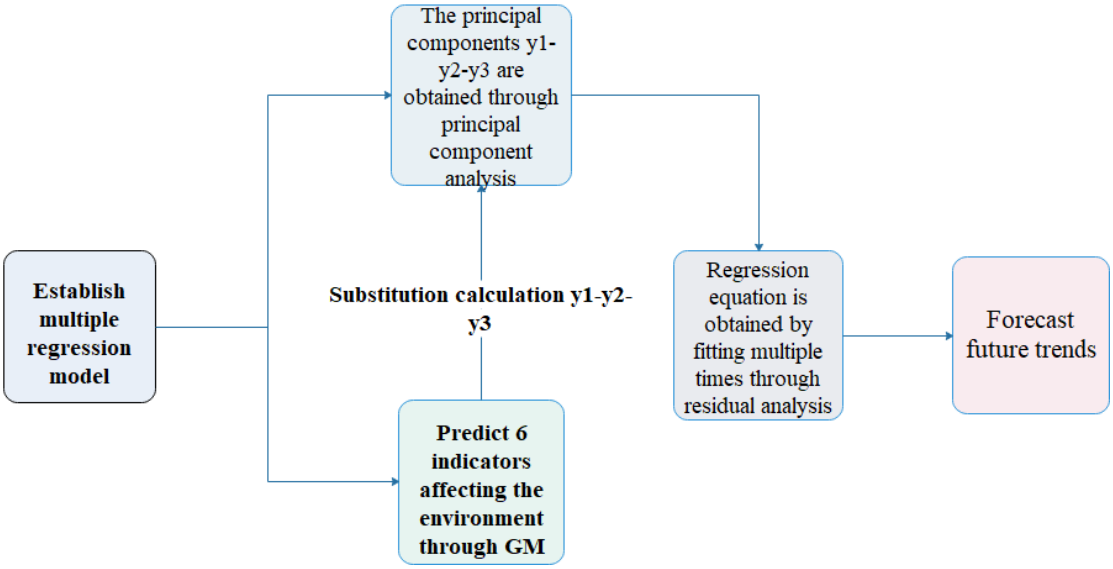


Figure12：Establish multiple linear regression equation model

The principal component y_1、y_2、y_3 affecting Alzheimer's was obtained by principal component analysis in the first question y_1、y_2、y_3。Let A represent the health state of the elderly at the time point，$\beta_i(i=0,1,2,3)$ # is the regression coefficient. Therefore, the linear regression equation is established as follows:

$$A=\beta_0+\beta_1 y_1+\beta_2 y_2+\beta_3 y_3 \tag{24}$$

among：

First principal component

$$y_1=-0.451x_1-0.392x_2-0.310x_3-0.485x_4-0.501x_5+0.245x_6.$$

Second principal component

$$y_2=-0.007x_1+0.210x_2+0.557x_3-0.061x_4-0.052x_5+0.800x_6.$$

Third principal component

$$y_3=0.462x_1+0.510x_2-0.593x_3-0.329x_4-0.005x_5+0.258x_6.$$

4.4.2 Solution of model

Use multiple linear regression relationship

The regress function in MATLAB is used for the first linear regression to determine the coefficient of the linear regression equation. These include regression coefficient, parameter confidence interval, correlation coefficient, value and corresponding probability,the specific coefficients obtained are shown in Appendix B

Further, the residual analysis is carried out using MATLAB. The diagram is as follows:

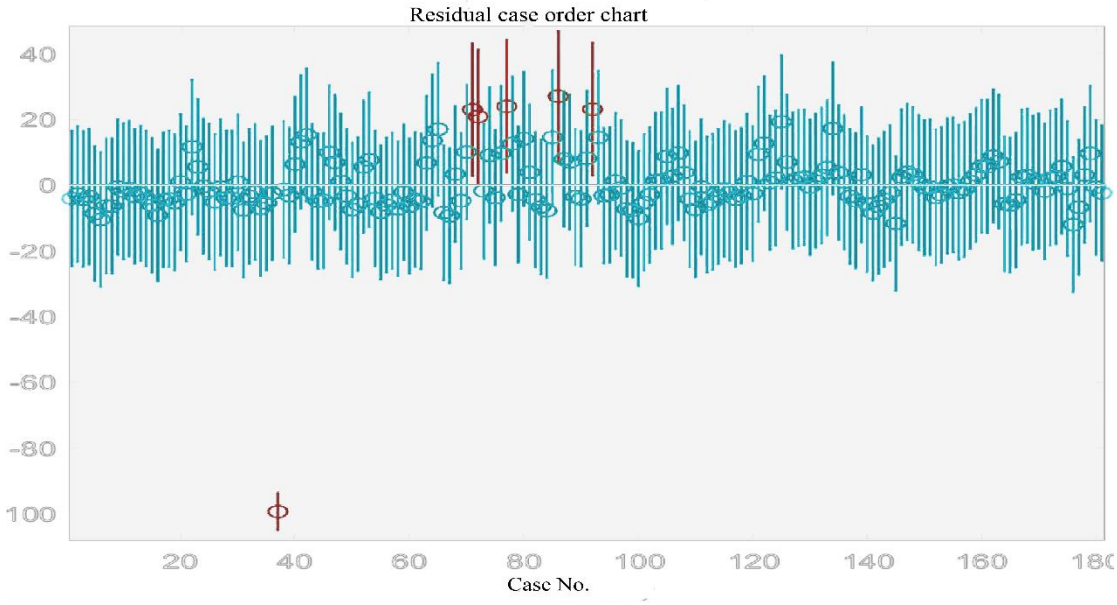


Figure 13:Diagram of the first linear regression residue

The multiple linear regression equation can be written as:

$$A = 19.0706 - 0.5296y_1 + 0.0867y_2 + 0.2128y_3 \quad (25)$$

According to Appendix B, the correlation coefficient R^2 is 0.8222, indicating that the fitting accuracy is 82.22%. The closer the R^2 value is to 1, the higher the fitting degree is, and the first fitting result is not ideal. The reason was the interference of outliers in the original data. In order to obtain the optimal multiple linear regression model, the outliers were removed in this question: 37,71,772, 77, 86, 92, and the second fitting was performed. The residual diagram is as follows:

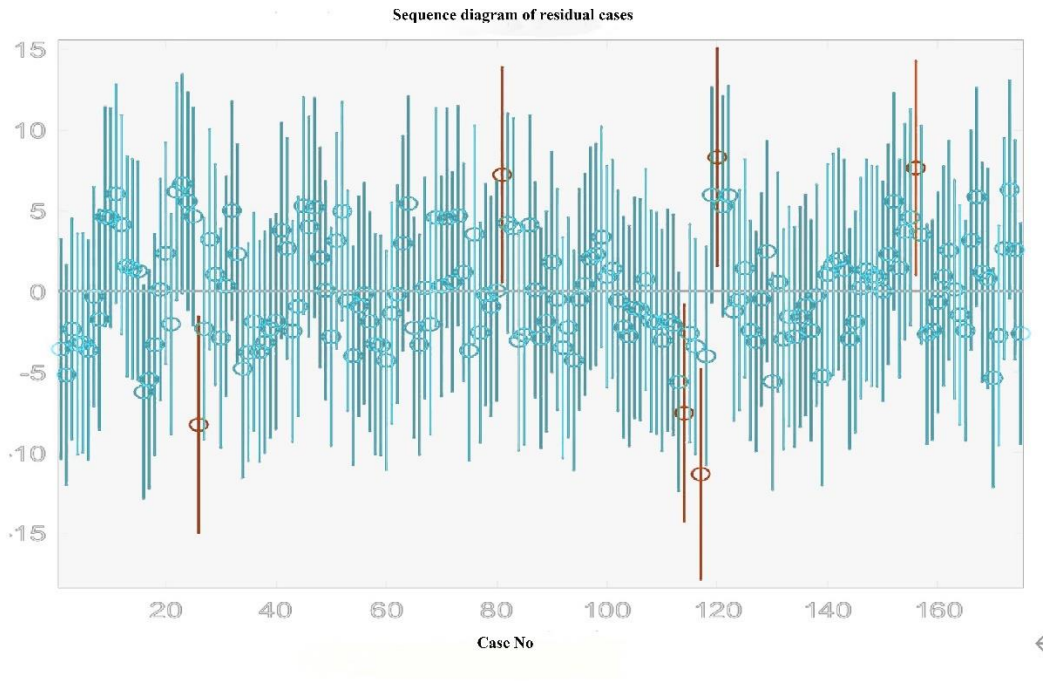


Figure 14 Diagram of the second linear regression residue

This is repeated until the optimal multiple linear regression equation is obtained. The operation result is as follows:

Table12: After fitting the concrete result data table

Number of fit	The equation with outliers removed	Accuracy of fit
1	$A = 19.0706 - 0.5296\,y_1 + 0.0867\,y_2 + 02128\,y_3$	82.22%
2	$A = 13.0620 - 0.3836\,y_1 - 0.0976\,y_2 + 0.6813\,y_3$	97.70%
3	$A = 12.1990 - 0.3960\,y_1 - 0.0888\,y_2 + 0.6762\,y_3$	98.13%

Thus, the multiple linear regression equation between the health status of the elderly and the three principal components y_1, y_2, and y_3 of the time point is：

$$A = 12.1990 - 0.3960y_1 - 0.0888y_2 + 0.6762y_3 \quad (26)$$

In the first question, the relationship between the principal component and the 6 influencing factors obtained by the principal component analysis is shown in Formula

(27). After the substitution, the final relationship is obtained：

$$A = 12.1990 + 0.4916x_1 + 0.4814x_2 - 0.3277x_3 - 0.0250x_4 + 0.1996x_5 + 0.0064x_6 \quad (27)$$

MATLAB was used to draw the comparison between the estimated curve and the true value as shown below:

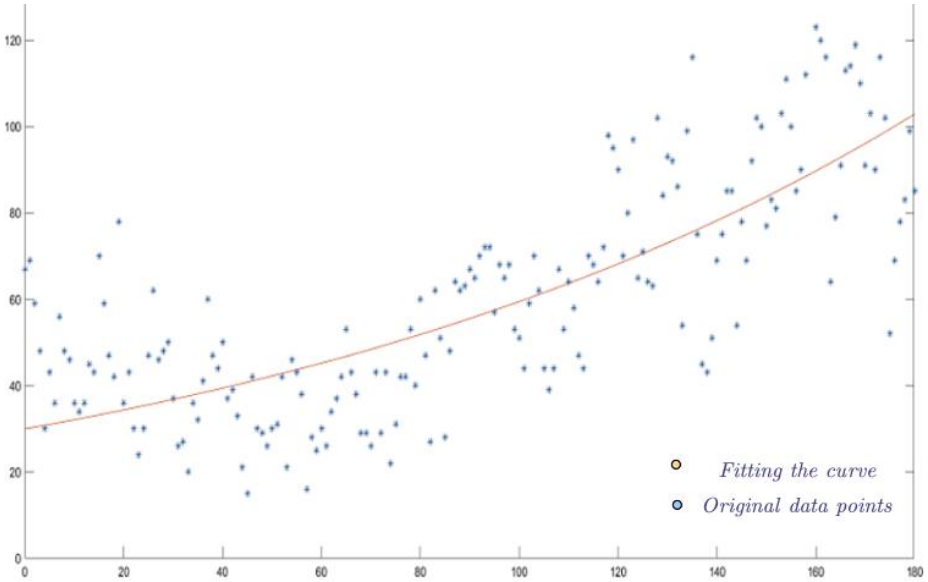


Figure15：Curve fitting diagram

4.5 Solution to problem 5

Problem 2: Establishing an intelligent diagnosis of Alzheimer's disease, making an intelligent diagnosis of an elderly person with normal cognition, measuring the characteristics of 'EXAMDATE', 'DX_bl', 'PTGENDER', 'PTETHCAT', 'PTRACCAT', 'PTMARRY', Combined with the diagnostic criteria for protein deposition of β-amyloid protein by imaging technology [5], it is determined that this cognitively normal old man may develop Alzheimer's disease in the future, so his eventual development of Alzheimer's disease can be roughly summarized as five stages: Preclinical Alzheimer's disease, cognitive impairment (MCI) due to Alzheimer's disease, mild dementia due to Alzheimer's disease, moderate dementia due to Alzheimer's disease, severe dementia due to Alzheimer's disease. Once an elderly person is affected by the severe dementia caused by Alzheimer's disease, he or she will experience severe memory impairment and loss of ability to perform daily activities. So while the old man is still in the CN stage, some advice for him and his family is as follows:

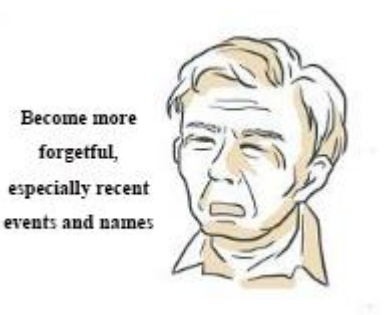
1. Adjust lifestyle and maintain good sleep habits.

2. Keep exercising and follow a healthy diet.

3. Family to give the elderly some companionship, more care for the elderly.

We considered pharmacologic approaches, but our team does not recommend pharmacologic therapy because of the limited effectiveness of current pharmacologic therapies.

However, unfortunately, due to the slow development of Alzheimer's disease, family members fail to find it in time. When the elderly suffer from mild cognitive impairment caused by Alzheimer's disease, they often complain that they always forget things, which can be determined as further deterioration of Alzheimer's disease, which is known as subjective memory complaint (SMC).



Over time, it will develop into early mild cognitive impairment (EMCI), then into late mild cognitive impairment (LMCI), and finally into the preclinical stage of Alzheimer's disease, known as mild cognitive impairment (MIC). If the family can find the initial symptoms of the elderly in time, go to the hospital in time, conduct a comprehensive evaluation and diagnosis, and receive hospital treatment, it is still a correct choice, which will greatly prolong the development of the disease. Of course, the most important is the company of family members and the correct understanding of Alzheimer's disease, as studies have shown that 67% of patients miss the optimal stage of intervention due to limited knowledge of the disease by family members.

5 Advantages and disadvantages of the model

5.1 Strengths

1. **Accuracy**. In question 1, we use typical correlation performance to accurately calculate the correlation of each factor.

2. **Convenienc**. The regression analysis method used in the analysis of the multifactor model makes the analysis of the characteristic information of Alzheimer's disease more convenient and concise.

3. **Robustness**. Our model shows great robustness to most of the parameters.

5.2 Weaknesses

1. **Restrictive**. If the feature dimension of the SVM model used is much larger than the number of samples, the performance of SVM is average. When the sample size of SVM is very large and the kernel function mapping dimension is very high, the calculation amount is too large and it is not suitable for use.

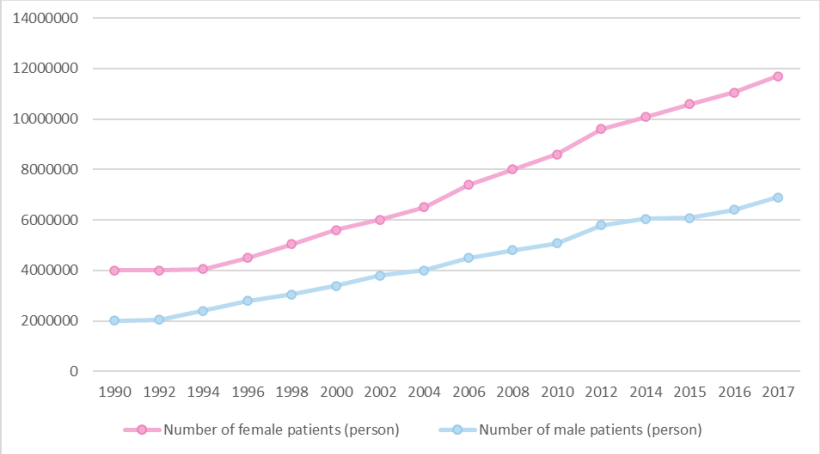
2. **Lack of other potentially relevant factors**. When analyzing the specific scores and descriptions, we only make judgments and predictions based on the data in the annex. Alzheimer's disease is a disease with complex pathology and long onset time, so the intelligent diagnosis system established is difficult to be convincing.

Reference

[1]Qu M, Xiao Jun. Research progress in early diagnosis and intervention of Alzheimer's disease [J]. Chinese Journal of Clinicians: Electronic edition, 2012, 6(2):4.
[2] Li Shiyu, Wang Feng, Cao Bin, et al. Review on the application of artificial intelligence in neuromedicine. Computer Science, 2017, 44(B11):5
[3] Mo Y M. Progress in early diagnosis and treatment of mild cognitive dysfunction [J]. Chin J Clinical New Med, 2011, 4(9):4. (in Chinese)
[4] Jiang Yan, Hu Tao, Yang Ning. Application of artificial intelligence in medicine [J]. Modern Preventive Medicine, 2009, 36(8):4.
[5] Wu Zhangying, Zhong Xiaomei, Chen XR, et al. Analysis of the characteristics of mild cognitive impairment and psychobehavioral symptoms in Alzheimer's disease patients [J]. Chinese Journal of Psychiatry, 2016(3):7.

Appendices

Appendix A Number of patients with Alzheimer's disease



Appendix B Comparison between predicted value and actual value

			拟合度								
拟合统计	平均值	标准误差	最小值	最大值	5	10	25	50	75	90	95
平稳R方	0.514		0.515	0.515	0.515	0.515	0.515	0.515	0.515	0.515	0.515
R方	-0.062		-0.063	-0.063	-0.063	-0.063	-0.063	-0.063	-0.063	-0.063	-0.063
RMSE	0.073		0.073	0.073	0.073	0.073	0.073	0.073	0.073	0.073	0.073
MAPE	6.232		6.232	6.232	6.232	6.232	6.232	6.232	6.232	6.232	6.232
MaxAPE	24.871		24.872	24.872	24.872	24.872	24.872	24.872	24.872	24.872	24.872
MAE	0.054		0.054	0.054	0.054	0.054	0.054	0.054	0.054	0.054	0.054
MaxAE	0.179		0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179
正态化	-5.048		-5.04	-5.04	-5.04	-5.04	-5.04	-5.04	-5.04	-5.04	-5.04

BIC	8	8	8	8	8	8	8	8	8

Appendix C Coefficient parameters of multiple linear regression

第一次线性回归所得数据表

回归系数	回归系数估算值	回归系数置信区间
β_0	19.0706	[11.9390, 26.2021]
β_1	-0.5296	[-0.6603, -0.3989]
β_2	0.0867	[-0.0641, 0.2375]
β_3	0.2128	[0.0292, 0.3963]

R^2=0.8222 F=272.8620 P=0.0000 S^2=111.2442

Appendix D Supporting code

There is no source code in this paper, and the relevant code is implemented on the basis of the original model