Transactions on Computational and Scientific Methods | Vol4, No. 8, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

K-Means and N-TSK Algorithm for EEG-Based Alzheimer's Disease Feature Identification

Ember Novak[1],Lorcan Fintan[2]
Department of Neuroscience, University of South Florida, USA[1],Department of Neuroscience, University of South Florida, USA[2]
ember.novak96@usf.edu[1],Lorcan.F779@gmail.com[2]

Abstract: Alzheimer's Disease (AD) is a prevalent neurodegenerative condition characterized by progressive cognitive and memory decline, significantly impairing daily living activities. Despite advancements in medicine, effective treatments remain elusive. This study explores the application of the N-TSK algorithm to identify features in EEG brain function networks of Alzheimer's patients, combining complex network analysis with a Takagi-Sugeno-Kang (TSK) fuzzy system model to improve identification accuracy and interpretability. While this approach enhances model performance, subjectivity in determining the index weight vector poses challenges, particularly when the feature set is large. Additionally, K-means clustering was employed due to its simplicity and efficiency in clustering EEG data. However, the algorithm's limitation lies in the need for prior knowledge to preset the appropriate K value, as random selection can yield unsatisfactory clustering results. Further research is needed to refine these methods for more accurate and scientific clustering in Alzheimer's disease diagnosis.

Keywords: Alzheimer's Disease; N-TSK Algorithm; K-means Algorithm.

1. Introduction

Alzheimer' s Disease (AD) is a disease with a high prevalence and leads to cognitive impairment[1]. It often exists in the elderly and the cause of the disease is not clear, so it is also named dementia. Although there is a relatively developed medical level in the 21 st century, we still have not found a better treatment for AD. The specific symptoms of Alzheimer' s disease include cognitive decline, memory decline, language function degradation, etc. It is characterized by a gradual decline in activities of daily living, accompanied by various neuropsychiatric symptoms and behavioral disorders, gradually losing independent life skills and dying of complications within 10 to 20 years after the onset[2].

2. Intelligent Diagnosis Method based on N-TSK Algorithm

2.1. Algorithm based on Brian Network

The brain functional network of patients with Alzheimer's disease has pathological changes, the correlation between brain regions is weakened, and the small world degree of the brain network is reduced, indicating that the functional network is an effective method to identify AD. We combine the features of brain functional network with fuzzy classification algorithm, and propose N- TSK algorithm based on brain network for the recognition of Alzheimer' s EEG functional network . According to the characteristics of structural changes in the brain network of patients with Alzheimer' s disease, combined with complex network method and fuzzy learning theory, a new supervised fuzzy classification algorithm based on the brain network is proposed to

accurately identify the features of AD brain functional network. First, the PSI phase synchronization index is used to measure the functional connection between EEG leads, and the EEG feature extraction method is constructed by setting the threshold value. This question applies the PSI algorithm to EEG to capture the brain network features of AD, and uses the fuzzy algorithm to improve the brain network recognition performance[3]. PSI based on signal entropy distribution is defined as:

$$\rho = (Q_{\max} - Q)/Q_{\max} \tag{1}$$

Where $Q = -\sum_{i=1}^{k} P_i In(P_i)$, $P_i\left(\hat{\varphi}\right)$ is the statistical of the phase difference $\left(\hat{\varphi}\right)$ in the ith box. The instantaneous phase difference related to the two signals obtained by using Hilbert transform is:

$$\hat{\varphi}(n) = \hat{\varphi}_1(n) - \hat{\varphi}_2(n) \tag{2}$$

Where $Q_{\max} = InK$, K is the total number of boxes divided. The range of PSI values is [0,1], PSI = 1 indicates that the rhythms of two signals are completely synchronized, PSI= 0 indicates that there is no synchronous rhythm trend between the two systems. For EEG records containing M leads, the M× M adjacency matrix can be obtained by calculating the PSI value of each pair of EEG leads.

Based on PSI adjacency matrix, the brain functional network of AD and normal control was constructed by setting threshold. A geometric model containing 16 nodes and several edges is established to represent the brain functional network. Each node represents a lead in EEG measurement and its corresponding brain region, while the edge represents a strong phase synchronization relationship between the two nodes. By setting a proportional threshold, we can retain a certain proportion of the highest strength functional connections to ensure that there are no isolated nodes in the network, and the main functional connections can be retained. The rest of the connections are weak, so we can ignore their impact on the network topology and properties. The characteristics of global efficiency, local efficiency, clustering coefficient, node intermediate and edge intermediate are extracted respectively to characterize the characteristics of the brain network. They are input into the TSK fuzzy system and used as the input variables of the antecedent and the consequent of the fuzzy system. The linear parameters β_g of the consequent variables are determined by the regression algorithm. The optimization process of the consequent parameters based on ridge regression is as follows:

For the specified regression task, first construct the training data set $D_s = \{x_i, y_i | x_i \in R^d, y_i \in R^c\}$, where x_i is the d dimension input vector of the ith sample, and y_i is the C dimension label vector of the ith sample (when the sample i belongs to the jth class, y_{ij} = 1 , otherwise y_{ij} = 0), N_s is the total number of training set samples, and C is the number of categories of data classification. This question
only distinguishes and normal EEG signals, thus C= 2. The optimization objective function of the construction model is as follows:

$$\min_{\beta_z} j(\beta_g) = \frac{1}{2}\sum_{j=1}^{c}\sum_{i=1}^{N_s} ||\beta_{gj}{}^T x_{gi} - y_{ij}||^2 + \frac{\lambda}{2}\sum_{j=1}^{c}\beta_{gj}{}^T\beta_{gj} \tag{3}$$

Where, λ is the regularization parameter. The optimal solution is:

$$\beta_{gj} = \left(\lambda_1 I_{(d+1)*k\times(d+1)*k} + \sum_{i=1}^{N_s} x_{gi}x_{gi}{}^T\right)^{-1}\left(\sum_{i=}^{Ns} x_{gi}y_{ij}\right) \tag{4}$$

The number of fuzzy rules and ridge parameters of the main parameters of the model are determined by cross validation. N-TSK is applied to brain feature recognition. The training data set is divided into five sub samples of equal size. Four sub samples are selected for the training data set in each training process, and the remaining sub samples are used as the validation data set. The training set and validation set are rotated five times in turn. The PSI value between each pair of EEG leads of AD and normal control groups was calculated.

For each subject, a 16 × 16 adjacency matrix is obtained by calculating each subject. Due to the non-normal distribution of PSI values in each group, Wilcoxon rank sum test was used to analyze the difference between AD and the control group. The p value of Wilcoxon rank sum test represents whether there is significant difference between groups, and represents the significance level in statistical analysis. The average value of PSI adjacency matrix of each subject was calculated and statistically analyzed. The average PSI of AD subjects (0.3881 ± 0.0618) less than that of normal control subjects (0.5653 ± 0.1018), Wilcoxon rank sum test $p < 0.05$, indicating that the functional connectivity of AD group is weaker than that of normal control group, and phase synchronization can be used to construct brain functional network. By setting the threshold, the brain functional network of AD and normal control was constructed. As shown in Figure 5.4, the number of non-isolated nodes in the brain network of all subjects under different proportional thresholds. At the proportional threshold $T > 0.3$, there is no isolated node in the brain network of all subjects. Therefore, the threshold value of brain network proportion based on PSI synchronization adjacency matrix is set to 0.3, and the first 30% of the connections in the PSI adjacency matrix are retained. This method reasonably ignores the weak connections in the network, and ensures that the two groups of networks are moderately connected and the structure is clearly distinguished.

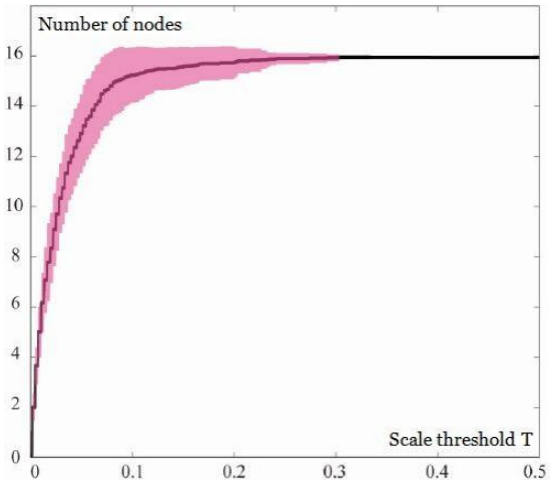


Fig 1. The number of nodes connected to the brain network changes with the threshold

2.2. Morphological Feature Extraction

Next, morphological feature extraction is used to simulate cognitive and behavioral features to provide auxiliary judgment basis for patients' disease diagnosis [4]. in this paper, gray matter images are used to analyze and extract morphological features. Let the gray value of one voxel in the image be x_{ij} , i,j representing the jth image of the ith group of samples. With the above processing process, the samples conform to the normal distribution, and the calculation process of the sum of two groups of samples x_{1j} and x_{2j} can be expressed by formula (5):

$$x_{ij} = \mu_{ij} + \varepsilon_{ij} \; i = 1, 2 \; j = 1, 2, \cdots\cdots, n \tag{5}$$

On this basis, two variables are further added to unify the samples in order to fit the GLM, as shown in the following formula (6):

$$x_{ij}=x_{1j}\mu_1+x_{2,}\mu_2+\varepsilon_{ij},x_{1ij}=\begin{cases}1,i=1\\0,i=2\end{cases},x_{2ij}=\begin{cases}0,i=1\\1,i=2\end{cases} \tag{6}$$

The following equation (7) shows the matrix form of the above equation:

$$x=y\beta+\varepsilon,\quad y=\begin{bmatrix}1\cdots1 & 0\cdots0\\0\cdots0 & 1\cdots1\end{bmatrix}^T,\quad \beta=[\mu_1,\mu_2]^T \tag{7}$$

Based on this, it is assumed that the illuminance is added to the formula, as shown in Formula (8):

$$H_0:c^T\beta=0,c=(1,-1)^T \tag{8}$$

The following equation (9) can give an estimate of the variance of sum:

$$\beta'=(y^Ty)^{-1}y^Tx=\left(\bar{x}_1,\bar{x}_2\right)^T \tag{9}$$

$$\sigma^2=\frac{\varepsilon^T\varepsilon}{n_1+n_2-2} \tag{10}$$

Based on this, the expression of statistic t is shown in equation (11):

$$t=\frac{c^T\beta'}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \tag{11}$$

In order to find out the part or area with large difference in the image through analysis, it is necessary to judge whether the original hypothesis is not satisfied after calculating the value and threshold.

Since the above assumption is the calculation and analysis of independent voxels, if you want to analyze the difference significance of the whole region, you need to run FDR correction to achieve it. According to the Gaussian random field theory, if the number of voxels is greater than the preset threshold, it can be treated as cluster, and FDR will be corrected in clusters. That is, after the region with large difference is found through FDR correction, the region information can be further calculated as the morphological feature, because the information in this region reflects the difference between the two subjects.

The features used in this question are mainly N-TSK algorithm based on brain network and morphological feature extraction to simulate brain structure features and cognitive behavior features for intelligent diagnosis of Alzheimer's disease. Using these two features together is more effective than using them alone.

3. Gabor Transform and K-means Clustering Model

3.1. Feature Fusion based on Gabor Transform

The Alzheimer's disease classification model based on Gabor transform feature fusion divides CN, MCI and AD into three categories. Gabor wavelet is widely used in the field of brain medical images, and the detection of Alzheimer's disease mainly depends on the diagnosis of subjects' brains. Gabor wavelet is sensitive to the edge information of the image when it is used for feature extraction calculation. It can construct different filters to extract the local features of the image by changing the parameter settings [5].

The function defined by Gabor is mainly the introduction of time window, which is an improvement of Fourier transform, so Gabor transform can be regarded as windowed Fourier transform. Equation (12) describes the calculation method of Gabor transform:

$$G_f(a;b,w)=\int_{-\infty}^{+\infty} f(t)\,g_a{}^*(t-b)\,e^{iwtdt} \tag{12}$$

In the above formula, $g_a(t-b)$ is the sliding window function, which is calculated by the formula:

$$g_a(t-b)=\frac{1}{2\sqrt{\pi a}}\exp\left(-\frac{(t-b)^2}{4a}\right) \tag{13}$$

The definition of two-dimensional Gabor wavelet is shown in (13), while equation (14) is its Fourier transform, as follows:

$$g(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}e^{\left(-\left(\frac{x^2}{2\sigma^2{}_x}+\frac{y^2}{2\sigma_y{}^2}\right)\right)} \tag{14}$$

$$H(x,y)=e^{\left(-\left(\frac{(\mu-\sigma)^2}{2\delta^2{}_x}+\frac{v^2}{2\sigma^2{}_r}\right)\right)} \tag{15}$$

In the formula, w is the modulation frequency, σ_x and σ_y is the standard deviation, u is the direction, v is the frequency. A variety of filters can be obtained by transforming parameters.

Use the wavelet of formula (14) as the original wavelet to transform. The main transformation direction and scale can obtain multiple Gabor wavelets. The definition is shown in formula (16):

$$\begin{aligned} &g_{mn}(x,y)=a^{-m}g'(x',y'),a>1\\ &x'=a^{-m}(-x\cos\theta+y\sin\theta),\theta=\frac{n\pi}{k}\\ &y'=a^{-m}(-x\sin\theta+y\cos\theta),\theta=\frac{n\pi}{k} \end{aligned} \tag{16}$$

The two-dimensional Gabor filter is a complex function, but generally speaking, it will only retain the amplitude and eventually give up the Gabor linear characteristics. After completing the above calculation, in order to extract the final texture features, it is necessary to further calculate the variance and mean value corresponding to this amplitude, and finally take them as features. CN, MCI and AD are divided into three categories and further refined based on these three categories.

3.2. Clustering Analysis based on K-means Algorithm

Then, we use the clustering analysis based on K-means algorithm, continue to cluster and refine the three subcategories (SMC, EMCI and LMCI) contained in MCI, divide them into three cluster centers according to Euclidean distance, and finally continue to refine the cluster into three subcategories. The main idea of K-means algorithm is to set the K value according to the demand and data type, and randomly select the initial cluster center points after initialization, also known as the center of mass. Finally, select the distance formula, run the algorithm, calculate the distance between other points and the center of mass according to the distance formula, cluster according to the preset conditions, and divide into different classes (also known as clusters). After all points are allocated, different clusters recalculate the center of mass position of the cluster according to the points in a cluster, and then iterate the above steps, Until the preset number of iterations is reached or the centroid in each cluster basically does not change or the change value is ignored, the clustering can be ended indefinitely [6]. K-means clustering algorithm is mainly divided into four steps:

(1) The first step is to initialize data, and randomly select k initial centroids $M=\{m_1,m_2,\cdots\cdots m_k\}$;

(2) The second step is to calculate the distance from each point to the centroid, $\forall i,j=\{1,2,\cdots\cdots k\},i\neq j,\ \forall s\in S$. If s is cluster centroid c_i $\sum_{s\in c_i}\frac{S}{\|C_i\|}$,the distance

between s and c_i is less than the distance from the point to c_j , s will be divided into clusters C_i ;

(3) The third step is to calculate the average value of all points in each cluster, $\forall i \in \{1,2, \ldots, k\}$, and set this value as the new center of mass in the cluster, and recalculate the center of the cluster C_i;

(4) According to the preset conditions, execute (2) and (3) iteratively until the preset number of iterations is reached or the centroid in each cluster basically does not change or the change value can be ignored to end the clustering without timing.

3.3. Cluster Analysis of MCI

For the three subcategories (SMC, EMCI and LMCI) included in MCI, call the K-means package in Python, set the parameters, and calculate the number of clusters and the cluster center. The parameters are set as follows: the clustering effect is evaluated according to the data and the characteristics of the problem, as shown in Figure 2. When k is 7 or 8, the clustering effect is good, but the contour coefficient is locally optimal at 7. As shown in Figure 3, when the elbow coefficient is 6-8, the change rate of the inflection point is large, and when the elbow coefficient is 7, it is the maximum inflection point. Therefore, the best clustering result is when k is 7.

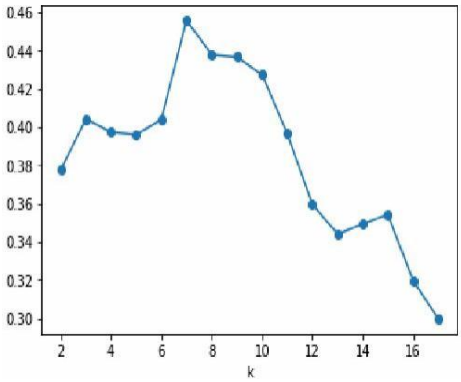


Fig 2. Profile coefficient diagram of K-means algorithm

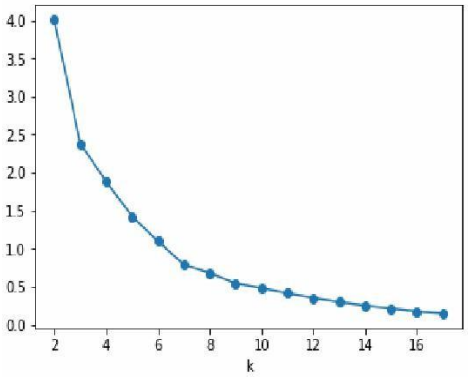


Fig 3. Elbow coefficient of K-means algorithm

The data is brought into K-Means cluster analysis model, and the distribution map of MCI classification is obtained through python. We use different colors to represent it, and the same color represents that it will be divided into the same category. The following figure 4 is the analysis chart after clustering analysis:

Among them, red represents SMC, green represents EMCI, and blue represents LMCI. Finally, K-means algorithm is used to find the cluster center, and Euclidean clustering is calculated to re cluster and refine MCI.

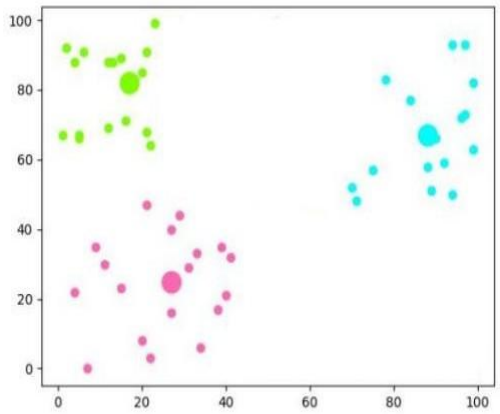


Fig 4. K-Means Cluster Analysis

4. Conclusion

N-TSK algorithm is used to identify EEG brain function network features of Alzheimer's patients. The method of combining complex network method with TSK fuzzy system model is adopted to improve the identification accuracy and interpretability of the model. However, the determination of index weight vector is subjective. When the index factor set is large, the relative membership weight coefficient is often small, and the result will be super fuzzy.

The biggest advantage of K-means algorithm lies in its simplicity and rapidity. The time complexity of the algorithm is; However, the K value of this algorithm must be preset in advance. How to preset an appropriate value requires rich prior knowledge and experimental verification. If it is not set randomly according to the actual situation, scientific clustering results will not be obtained.

References

[1] Yan Qi Research on differential privacy decision tree method based on Pearson correlation coefficient [D]. Guangxi Normal University, 2021.

[2] Wen Bingmei Sparse principal component regression of binary data [D]. Southwest Jiaotong University, 2021.

[3] Lei Xinyu Brain network construction and feature recognition of Alzheimer's disease [D]. Tianjin University, 2019.

[4] Luo Chuanji Research on detection method of Alzheimer's disease based on feature relationship [D]. University of Electronic Science and Technology of China, 2019.

[5] Liu Jifeng Research on Hyperspectral Image Compression and Perceptual Reconstruction Method Based on K-means Clustering [D]. Jiangxi University of Technology, 2019.

[6] Tang Shuyi Prediction of China's total health expenditure based on PCA-BP neural network model [D]. Beijing Jiaotong University, 2021.