MCGNet+: An Improved Motor Imagery Classication Based on Cosine Similarity

Yan Li
Zhejiang University
Ning Zhong
Maebashi Institute of Technology
David Taniar
Monash University
Haolan Zhang (haolan.zhang@nit.zju.edu.cn)
Zhejiang University Ningbo Institute of Technology https://orcid.org/0000-0002-5033-1866

Research

Keywords: Graph Convolutional Networks, Electroencephalography(EEG), Brain-computer Interfaces(BCI)

Posted Date: November 18th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1014504/v1

Version of Record: A version of this preprint was published at Brain Informatics on February 1st, 2022. See the published version at https://doi.org/10.1186/s40708-021-00151-3.

RESEARCH

MCGNet$^+$: An improved motor imagery classification based on Cosine similarity

Yan Li[1], Ning Zhong[2], David Taniar[3] and Haolan Zhang[4*]

It has been a challenge for solving the motor imagery classification problem in the brain informatics area. Accuracy and efficiency are the major obstacles for motor imagery analysis in the past decades since the computational capability and algorithmic availability cannot satisfy complex brain signal analysis. In recent years, the rapid development of Machine Learning (ML) methods has empowered people to tackle the motor imagery classification problem with more efficient methods. Among various ML methods, the Graph neural networks(GNNs) method has shown its efficiency and accuracy in dealing with inter-related complex networks. The use of GNN provides new possibilities for feature extraction from brain structure connection. In this paper, we proposed a new model called MCGNet$^+$, which improves the performance of our previous model MutualGraphNet. In this latest model, the mutual information of the input columns forms the initial adjacency matrix for the cosine similarity calculation between columns to generate a new adjacency matrix in each iteration. The dynamic adjacency matrix combined with the spatial temporal graph convolution network(ST-GCN) has better performance than the unchanged matrix model. The experimental results indicate that MCGNet$^+$ is robust enough to learn the interpretable features and outperforms the current state-of-the-art methods.

Keywords: Graph Convolutional Networks; Electroencephalography(EEG); Brain-computer Interfaces(BCI)

1 Introduction

Brain-computer-interface(BCI) technology has drawn much attention globally due to its significant meaning and extensive applications [1]. It enables their users to interact with the machine through the brain signals [2], such as the task of converting the psychological imagination of motion into a command[3], which can be utilized to help people with disabilities as a rehabilitation device[4] and could be considered the only way for people with motor disabilities to communicate[5]. The motor imagery classification based on the features extracted from the EEG imagination data of moving the body parts without actual movement, but the feature extraction process often relies heavily on prior knowledge to exclude certain features[6]. Consequently, more robust feature extraction techniques will continue to drive the development of BCI technologies.

A typical brain-computer interface system consists of four main processes[7]: brain-electric raw data acquisition, data pre-processing, feature extraction and feature classification. The previous studies show that the feature extraction and classification are two important phases, which determine whether the system is effective or not. The feature extraction process is designed to describe EEG signals by relevant values[8], and features should contain the information embedded in the original EEG signals while filtering out the noise and other irrelevant information. The classification phase is critical because an efficient classifier can take advantage of as many extracted features as possible and greatly improve the accuracy of the classification. The motor imagery classification is an EEG-based task that focuses primarily on the feature extraction and classification, which have been studied extensively in previous work. Some research highlights two most common types of features that include frequency band power features and time point features[9], both of which benefit from extracting zone after spatial filtering[10]. Principal component analysis(PCA) and independent component analysis(ICA) are two classic unsupervised spatial filter methods[11], supervised spatial filters include the common spatial patterns(CSP) and filter bank common spatial patterns(FBCSP)[12]. In terms of the classifiers for motor imagery task, many state-of-the-art methods have been proven effective, such as linear discrimination analysis(LDA) and support vector machine(SVM)[13].

*Correspondence: haolan.zhang@nit.zju.edu.cn
[4]Ningbo Institute of Technology, Zhejiang University, Ningbo, China
Full list of author information is available at the end of the article

Nowadays, the deep learning methods have been efficiently applied to various areas. Much recent work has explored the application of deep learning to EEG-based analytical tasks[14]. The deep learning methods improve the analytical efficiency and accuracy and provide end-to-end learning for EEG-based tasks, such as sleep stage detection, anomaly detection, motor imagery classification and so on[15]. In spite of the typical deep learning methods, such as convolution networks, can learn from the raw data without manual feature extraction, they still have some major limitations. For instance, typical deep learning methods require large datasets to train the models, which can be a disadvantage for EEG based tasks because the collection of EEG data usually costs a lot. In addition, EEG datasets represent the unique characteristics of an individual, and the data collected from different areas of the brain. Therefore, the spatial connection between the EEG data can't be ignored. However, existing methods including recent deep learning methods are unable to effectively learn the connections between different brain regions[16].

Graphs are the most appropriate data structure for brain connections, and graph neural networks(GNNs) has been shown to be effective in classifying graph structures[17], the core idea of GNNs is to update each node's embedding iteratively through aggregating the representations of its neighbors and itself. The EEG channels could be represented as nodes in the graph and the connections between the channels correspond to the edges of the graph, but the graph convolutional networks need adjacency matrix to be given in advance which is the representation of the graph connection[18], so determining a suitable brain map structure is still a challenge due to the limitations of cognition of brain structure. And there are some methods that could be used to generate the adjacency matrix, we could utilize the position to calculate the distance between the electrodes as the degree of correlation or utilize the features collected from the electrodes to calculate the correlations. Moreover, the collection of EEG data is usually in chronological order, so in addition to spatial characteristics the temporal characteristics also need to be taken into account.

In this paper, we proposed a novel model called MCGNet$^+$ based on the our proposed MutualGraphNet, combined the spatial-temporal filter and graph convolutional networks to learn the temporal and spatial characteristics, which achieved robust performance on the motor imagery classification tasks. The contributions of this paper are as follows:

- The model could realize end-to-end learning. Furthermore, the model is specially designed to adapt to the characteristics of EEG data, so it could be able to utilize the features to the great extend.
- For the first time, we use mutual information to generate the initial adjacency matrix and use cosine similarity to update the adjacency matrix dynamically, and achieve better performance.
- Experimental results demonstrate that the newly proposed model have better performance than state-of-the-art methods.

2 Related Work

A motor imagery classification task is of great significance for people with disabilities. Numerous work has been done to improve classification performance. In earlier studies, traditional machine learning methods were commonly used for motor imagery classification task, such as support vector machine(SVM), K-Nearest-Neighbor(KNN) and artificial neural network(ANN) are frequently used[19], but these traditional methods have limited performance on EEG-based classification tasks. Currently, the deep learning methods are utilized in EEG-based classification tasks, Deep Belief Network(DBN)[20] was proposed to manually extract features from the channels then fed them into the network. Convolutional Neural Networks(CNN) could automatically learn features from EEG data and perform better than DBN due to their regular structure and the degree of ambiguity of the translational structure[21]. Two CNN models were specially designed for motor imagery classification called Shallow ConvNets and Deep ConvNets[14], both of them have better performance than the state-of-the-art methods. Then another CNN model called EEGNet[15] was proposed, which utilizes the Depthwise and Separable convolutions to replace the traditional convolutions for the motor imagery task that have better performance than the ConvNets.

The CNN models can effectively extract the local patterns of data, but it can only be applied for the standard grid data[22], graph convolutional networks have been proven to have better performance on the graph structure data. Much has been done to improve the performance of the graph convolutional networks. So far, GCNs have been applied in many fields, the spatial-temporal graph convolution network(ST-GCN)[23] is proposed to learn the dynamic graphs for the human action recognition tasks, the spatiotemporal multi-graph convolution network(ST-MGCN)[24] is proposed for ride-hailing demand forecast which encodes the non-Euclidean correlations among regions into multiple graphs, GraphSleepNet[16] based on spatial-temporal convolution network(ST-GCN) is proposed for automatic sleep stage classification. When using GCNs, the connection relationship between each electrode need to be given as a prior knowledge, in other words, the adjacency need to be calculated as input.

There are different methods that can be used to generate the adjacency matrix, the distance between two electrodes can be used directly to represent the degree of correlation between electrodes and there are many different vector distance calculation methods such as the euclidean distance[25] which only need the physical position of the electrodes, the chebyshev distance[26] is defined as the maximum difference between two vectors in any coordinate dimesion, hamming distance, manhattan distance and so on. Furthermore, we can use the correlations of vectors to determine the degree of relevance of the different channels, such as cosine similarity[27] that calculates the similarity relationship between the characteristics of different electrode channels, pearson correlation that evaluates the linear relationship between two continuous variables, spearman correlation that evaluates the monotonic relationship between two continuous variables, kendall correlation, Point-Biserial correlation and so on. Also, we could use some machine learning methods, such as the information gain[28] that evaluates the gain of each variable in the context of the target variable and mutual information is the name given to information gain when applied to variable selection that calculates the statistical dependence between two variables.

Motivated by the studies mentioned above, considering the graph structure and the dynamic spatial-temporal characteristic of the EEG data, also the graph structure of different motor imagery may be different, the traditional GCNs models may be not optimal for EEG-based motor imagery classification task. We propose the novel model to best suit to the characteristics of EEG data which uses the mutual information to generate the initial adjacency matrix and use the cosine similarity to update the adjacency matrix after each iteration.

EEG Raw Data

Time	E1	E2	.....	E32	
0	30.62997	8.051328	.....	37.19137	d
0.001	31.00115	8.134437	.....	37.46626	
0.002	31.18674	8.13885	.....	37.53345	
0.003	31.20832	8.074863	.....	37.40517	
0.004	31.09179	7.952038	.....	37.11807	
0.005	30.85872	7.782878	.....	....	
....	.....	.....	.....	.....	
177.001	3.317724	2.016712	.....	176.999	

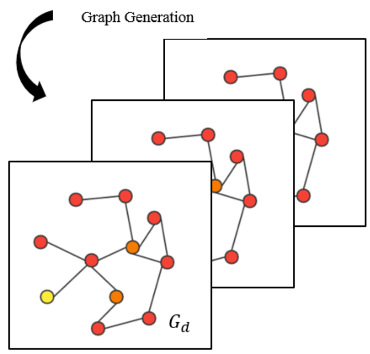

Figure 1 The structure generation of EEG data, where the data at the range of time d forms a graph.

3 Preliminaries

In this study, the EEG data could be defined as an undirected graph $G = (V, E, A)$, where V is a finite set of $|V| = N$ nodes and N represents the number of the EEG data channel; E is a set of edges, indicating the connectivity between different channels; A represents the adjacency matrix of graph G. Figure 1 shows how the graph is generated from the EEG raw data. The recorded EEG signals are divided into several labeled segments called trials, the $d-th$ trial can be denoted as $X^d = (x_t^1, x_t^2, ..., x_t^N)^F \in \mathbb{R}^{N\times F}$, where N denotes the number of the EEG electrodes and F denotes the values of all nodes within the time steps t. The dataset can be described as $D = (X^1, y^1), (X^2, y^2), ..., (X^L, y^L)$, L denotes the number of the trials and y represents the label corresponding to the trial, there are four motor imagery categories including left hand, right hand, feet and tongue, so the label can be denoted by 0-3 respectively. The goal of the task is to learn the mapping relationship between the EEG data and the motor imagery categories represented as labels and the problem can be defined as: given a input trial $X^i \in \mathbb{R}^{N\times F}$, $0 < i < L$ identify the corresponding label y^i.

4 Methodology

The overall framework of the model proposed in this paper is presented in Figure 2, it includes three main parts: feature extraction and adjacency matrix generation part, spatial-temporal attention part and spatial-temporal graph convolution part. Spatial-temporal attention part puts more attention on the more valuable spatial-temporal information, then spatial-temporal graph convolution part extracts both spatial and temporal features. And the complete algorithm can be seen as follows:

4.1 Adjacency matrix generation

4.1.1 Relevance calculate methods

The relevance of different electrodes can be obtained through calculating the correlations or the information gain of the features of the electrodes. The correlations of different channels can be represented by the distances of the channels. The euclidean distance of the

Algorithm 1 The process of motor imagery classification

Input: The input data $X \in \mathbb{R}^{N*M}$, label $Y \in \{1, 2, 3, 4\}$.
Output: The corresponding classification $\hat{y}$.
Calculate the mutual information of the columns of X and get the adjacency matrix $A \in \mathbb{R}^{N*N}$.
repeat
  Put the A and X in to the spatial-temporal attention block and get the get the attention matrix S.
  Put A, X, S into a GCN layer and get the embedding $\hat{X} \in \mathbb{R}^{N*L}$.
  Calculate the cosine similarity of the column of the embedding, get a new matrix $\hat{A} \in \mathbb{R}^{N*N}$.
  Update the adjacency matrix $A = \hat{A}$ and the input $X = \hat{X}$.
until The repeat times are equal to the number of ST-layers.
Then the output $\hat{y} = softmax(linear(X))$.

electrodes can be represented as:

$$\rho = ((x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2)^{1/2} \quad (1)$$

the euclidean distance can be understood as the straight-line distance between two points, but the electrodes are distributed on the surface of the cerebral cortex, so it is not suitable to directly express the relationships between the electrodes. The chebyshev distance is defined as the maximum difference between two vectors in any coordinate dimension, it's the maximum distance along an axis, and the chebyshev distance of the electrodes can be denoted as:

$$\rho = max(|x_2 - x_1|, |y_2 - y_1|, |z_2 - z_1|) \quad (2)$$

the calculation of the distances of the electrodes only utilizes the positions of the electrodes, we can also use the features of the electrodes to obtain the correlations. The cosine similarity of two vector can be defined as:

$$cos(x, y) = \frac{x.y}{\|x\|\|y\|} \quad (3)$$

However, the cosine similarity does not consider the magnitude of the vectors, but only consider the directions. The jacquard index, also known as the intersection ratio and jacquard similarity coefficient, can be used to compare the similarity and diversity of sample sets:

$$J(x_1, x_2) = \frac{|x_1 \bigcap x_2|}{|x_1| + |x_2| - |x_1 \bigcap x_2|} \quad (4)$$

one of the main disadvantage of the jacquard index is that it is greatly affected by the size of the data. Large data sets have a great impact on the index, because it can significantly increase the union while maintaining similar intersection. Moreover, we could use information gain between the feature vectors to obtain the

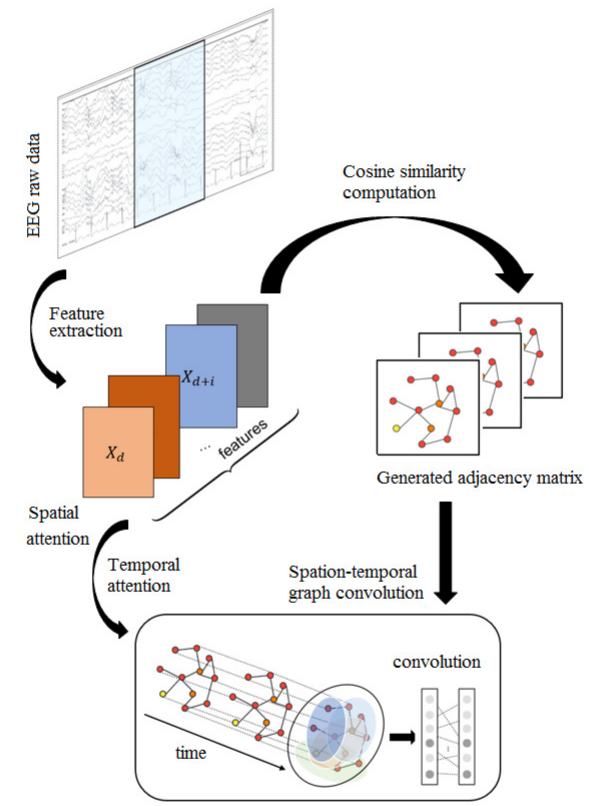

Figure 2 The overall structure of the proposed model consists of three parts : the feature extraction and the mutual information computation part, the spatial-temporal attention mechanism part and spatial-temporal graph convolution part.

degree of relevance, information gain is calculated by comparing the entropy of the dataset before and after a transformation. The mutual information calculates the statistical dependence between two variables and is the name given to information gain when applied to variables selection.

4.1.2 Adjacency matrix update

In order to make full use of and adjust the input prior knowledge in time according to the embedding learned by GCNs, we use the mutual information to generate the initial adjacency matrix and use the cosine similarity to update the adjacency matrix during the training process.

Mutual Information(MI)[29] is used to indicate whether there is a relationship between two variables and the strength of the relationship. The mutual information of two variables X and Y can be defined as:

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) log \frac{p(x, y)}{p(x)p(y)} \quad (5)$$

Mutual information is related to entropy, which is the expected or mean value of the information of all variables. The entropy of X is defined as:

$$\begin{aligned} H(X) &= \sum_{x \in X} P(x) log \frac{1}{P(x)} \\ &= -\sum_{x \in X} P(x) log P(x) = -ElogP(X) \end{aligned} \tag{6}$$

Then MI of X and Y can be computed by the equations:

$$\begin{aligned} I(X,Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) = H(Y) - H(Y|X) \\ H(X,Y) &= \sum_{x \in X} \sum_{y \in Y} p(x,y) log \frac{1}{p(x,y)} = -ElogP(X,Y) \\ H(Y|X) &= \sum_{x \in X} \sum_{y \in Y} p(x) p(y|x) log \frac{1}{p(y|x)} \\ &= -ElogP(Y|X) \end{aligned} \tag{7}$$

where $H(X,Y)$ is the joint entropy of X and Y, and $H(Y|X)$ is the conditional entropy that X is given in advanced. Thus, $I(X,Y)$ is the reduction in the uncertainty of the variable X by the knowledge of another variable Y, equivalently, it represents the amount of information that Y contains about X.

Considering the features of EEG data $X = \{x^1, x^2, ..., x^N\} \in \mathbb{R}^{N \times F}$, we could compute the mutual information $m_i j$ of x^i, x^j and use it as the weight of the connection of x^i, x^j, then we could generate a $N \times N$ weight matrix which could be used as the input adjacency matrix of the graph convolution networks. In our proposed work[30], we kept the initial adjacency matrix unchanged during the training process. However, the embedding changes after each iteration, so we update the adjacency matrix after each iteration synchronously to improve the performance of the model. Here we compute the cosine similarity of two columns of the embedding as the weight of the adjacency matrix. The cosine distance of two vector x, y is defined as:

$$cos(x,y) = \frac{x.y}{\|x\|\|y\|} \tag{8}$$

the updated weight can be defined as:

$$a_{i,j}^{l+1} = \frac{e_i^l . e_j^l}{\|e_i^l\|\|e_j^l\|} \tag{9}$$

where the $a_{i,j}^{l+1}$ denotes the element of the $i-th$ row and $j-th$ column of the adjacency matrix at the $l+1th$ iteration, and e_i^l, e_j^l represents the $i-th, j-th$ column of the embedding at $l-th$ iteration. The process of generating and updating the adjacency matrix can be seen in Figure 3.

4.2 Spatial-temporal attention

The spatial-temporal attention mechanism could capture the dynamic spatial and temporal correlations of the motor imagery network. In the spatial dimension, the activities of one brain region has influence on other brain regions and generally different brain activities convey different information, so the dynamic spatial-temporal capture mechanism is required. We use a spatial attention mechanism[31], which could be represented as:

$$\begin{aligned} S &= V_p * \sigma((\chi^{(r-1)} W_1) W_2 (W_3 \chi^{(r-1)})^T + b_p) \\ S_{i,j}^{'} &= \frac{exp(S_{i,j})}{\sum_{j=1}^{N} exp(S_{i,j})} \end{aligned} \tag{10}$$

where S denotes the spatial attention matrix, which is computed by current layer. $V_p, b_p \in \mathbb{R}^{N \times N}$, $\chi^(r-1) = (X_1, X_2, ..., X_{T_{r-1}} \in \mathbb{R}^{N \times C_{r-1} \times T_{r-1}}$ C_{r-1} is the number of channels of the input data in the $r^t h$ layer. $W_1 \in \mathbb{R}^{T_{r-1}}, W_2 \in \mathbb{R}^{C_{r-1} \times T_{r-1}}, W_3 \in \mathbb{R}^{C_{r-1}}$, $S_{i,j}$ in S represents the correlation strength between node i and j, then a softmax function is used to normalize the attention weights. Combine the adjacency matrix and the spatial attention matrix, the model could adjust the impacting weights between nodes dynamically.

In the temporal dimension, there are correlations during each motor imagery trial, since that the brain waves are transmitted in the cerebral cortex and the active areas of the brain will change over time, so the collected EEG data also changes over time. Therefore, a temporal attention is utilized to capture dynamic temporal information. The temporal attention mechanism is defined as:

$$\begin{aligned} E &= V_e * \sigma(((\chi^{(l-1)})^T M_1) M_2 (M_3 \chi^{(l-1)}) + b_q) \\ E_{m,n}^{'} &= \frac{exp(E_{i,j})}{\sum_{j=1}^{T_{r-1}} exp(E_{i,j})} \end{aligned} \tag{11}$$

where $V_e, b_q \in \mathbb{R}^{T_{l-1} \times T_{l-1}}$, $M_1 \in \mathbb{R}^N, M_2 \in \mathbb{R}^{C_{l-1} \times N}$, $M_3 \in \mathbb{R}^{C_{l-1}}$, $E_{m,n}$ denotes the strength of the correlation between motor imagery network m, n, and E is normalized by the softmax function, so the temporal attention matrix can be directly applied to the input.

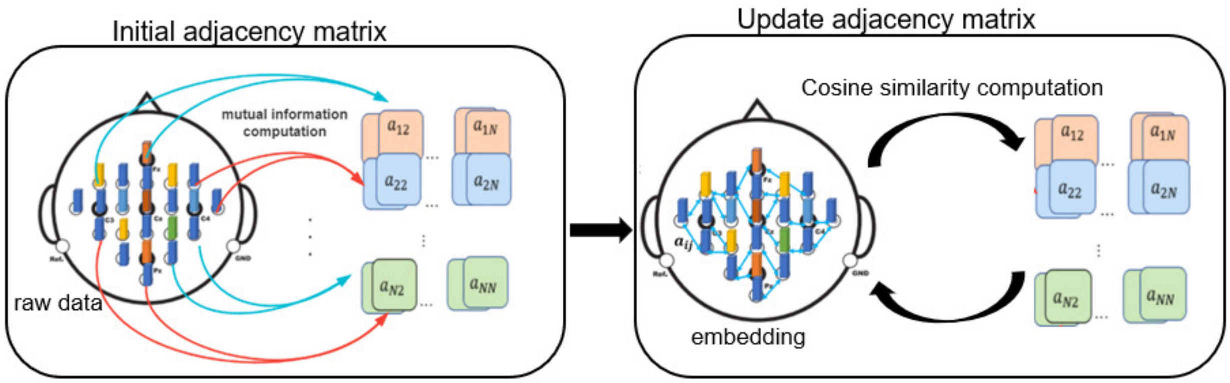

Figure 3 The process of generating and updating the adjacency matrix.

4.3 Spatial-temporal graph convolution

The spatial-temporal convolution consists of a graph convolution in the spatial dimension and a normal convolution in the temporal dimension, which could extract both the spatial features and the temporal features.

The spatial features are extracted by aggregating information from neighbor nodes, we use graph convolution to extract the spatial features. The graph convolution is based on laplacian matrix and Fourier transform, the graph laplacian can be defined as:

$$L = I - D^{-1/2} A D^{-1/2} \tag{12}$$

where $A \in \mathbb{R}^{N \times N}$ is the adjacency matrix associated with the graph, $D \in \mathbb{R}^{N \times N}$ is the diagonal degree matrix, $I \in \mathbb{R}^{N \times N}$ is the identity matrix. L is a real symmetric positive semidefinite matrix, it can be decomposed as $L = U \Lambda U^T$ and $\Lambda \in \mathbb{R}^{N \times N}$ is the diagonal matrix of eigenvalues that represent the frequencies of their associated eigenvectors. Let $x \in \mathbb{R}^n$ be a signal defined on the vertices of a graph G, the graph fourier transform of the signal is defined as $\hat{\mathrm{x}} = U^T x$. The graph convolution uses the linear operators that diagonalize in the flourier domain to replace the classical convolution operator, the graph convolution can be defined as:

$$g_\theta(L)x = g_\theta(U \Lambda U^T)x = U g_\theta(\Lambda) U^T x \tag{13}$$

where θ is a vector of fourier coefficients, g_θ is the filter that could reduce the computational complexity, g_θ can be approximated by a truncated expansion in the terms of Chebyshev polynomials[32]:

$$g_\theta(\Lambda) = \sum_{p=0}^{k-1} \theta_p T_p(\tilde{\Lambda}) \tag{14}$$

where k is the order of the Chebyshev polynomials, $\theta_p \in \mathbb{R}^k$ is the vector of Chebyshev coefficients, $T_p(\tilde{\Lambda}) \in \mathbb{R}^{N \times N}$ is the Chebyshev polynomial of order k and $\tilde{\Lambda} = 2\Lambda/\lambda_{max} - I$ ranges in $[-1, 1]$. Then the $j - th$ output feature can be calculated as:

$$y_i = \sum_{i=1}^{F_{in}} g\theta_{i,j}(L) x_i \tag{15}$$

where x_i denotes the $i - th$ row of input matrix, F_{in} equals to the input dimension, the outputs are collected into a feature matrix $Y = [y_1, y_2, ..., y_{F_{out}}] \in \mathbb{R}^{N \times F_{out}}$. In this work, we generalize the above definition to the nodes with multiple channels, the $l - th$ layer's input is $X^{(l-1)} = (x_1, x_2, ..., x_{(T_{l-1})}) \in \mathbb{R}^{N \times C_{l-1} \times T_{l-1}}$, $C_{(l-1)}$ denotes the channel's number and T_{l-1} denotes the $l - th$ layer's temporal dimension.

After the graph convolution having captured the neighboring information for each node in the spatial dimension, a standard convolution layer is used in the temporal dimension, we use a standard two-dimension convolution layer to extract the temporal information, the $r - th$ convolution layer could be defined as:

$$\chi_h^{(r)} = ReLU(\Phi * (ReLU(g_\theta * G\hat{\chi}_h^{(r-1)}))) \tag{16}$$

where Φ is the parameter of the temporal dimension convolution kernel, and $*$ represents the convolution operation, ReLU is the activation function.

5 Experiment

In order to evaluate the effectiveness of our model, we carried out the comparative experiments on a public dataset BCI Competition IV dataset 2a(SMR) for motor imagery task.

5.1 Dataset description

The BCI Competition IV dataset 2a consists of EEG data from nine subjects, there are two sessions recorded, one for training and the other one for testing. Each session includes 288 trials, which are recorded with 22 EEG electrodes and 3 electrooculogram channels, we only utilize the 22 EEG channels in this experiment and the distribution of the EEG electrodes can be seen in Figure 4. There are four types of labels in this dataset, corresponding to movements of the left hand, right hand, feet and tongue. The original dataset is sampled at 250Hz and bandpass-filtered between 0.5Hz and 100Hz, and we low-pass filter the dataset to 4-40Hz. Also in our experiment, we set the length of each trial to 4.5s which starts from 500ms before the start cue of each trial until to the end cue, then we extract 11 differential entropy features(DE) for each channel and double fold the features to make it have the same shape as the adjacency matrix, and combine the two as the input of the graph convolutional network, then we standard scale the data to make it suitable for the machine learning model. To show the effectiveness of our proposed model learning from the raw data and ensure the model could be used for wider range of tasks, we don't do much more preprocessing of the raw EEG data.

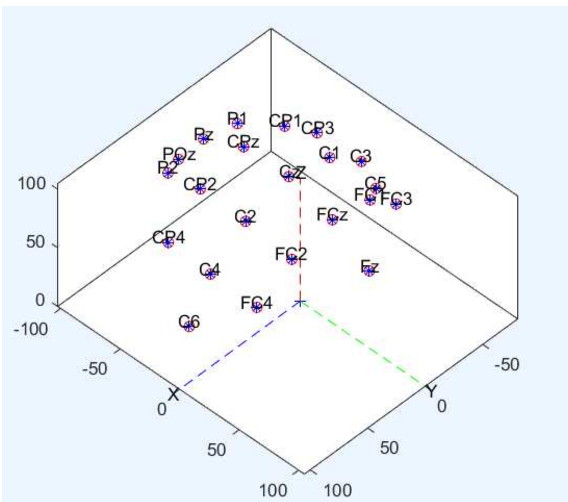

Figure 4 The distribution of the electrodes in 3D space.

5.2 Experiment settings

We compare our model with some state-of-the-art methods as well as the proposed MutualGraphNet, the baseline methods are listed as follows:

1 Filter Bank Common Spatial Patterns(FBCSP)[33]: It extracted the band power features of EEG ,then use the features to train the classifier to predict the labels.

2 Shallow ConvNet[14]: An end-to-end learn method, which use convolutional networks to do all the computations.

3 Deep ConvNet[14]: It has more convolution-pooling blocks and is much deeper than Shallow ConvNet.

4 EEGNet[15]: It uses the depthwise and separable convolution and has two convolution-pooling blocks.

In addition to the above baseline methods, we also compared traditional machine learning methods, support vector machine(SVM)[34] and random forest(RF)[35].

In order to prove that the model can effectively extract features and has the ability to eliminate the influence of individual differences, we no longer conduct experiments on each subject separately, we mixed the experimental data of nine subjects, and a total 2592 training trials and 2592 testing trials, and we use four-fold cross-validation to evaluate the performance. Since that the training set is not big enough, so in order to reduce the impact of over-fitting, we adopt a loss flooding strategy[36] during the training process, which is defined as:$\tilde{R}(g) = |\hat{R}(g) - b| + b$ and $\hat{R}(g)$ is the loss of the model, b is a constant called loss flooding level, here we set b as 0.5. The hyper-parameters are shown in Table 1. As for the baseline methods, in order to evaluate the performance of the models more reasonably, we use 250Hz sampling 4.5s EEG data for all experiments. Since that the EEGNet[15] used the 128Hz resampled data to conduct experiment in the original paper, so we double the lengths of temporal kernels and average pooling size of the original model for double sampling rate to better adapt the input which proven to have better performance than the original model. In response to changes in the length of the sampling time, we also adjusted the parameters of each model accordingly, conducted experiments and selected the best model performance. The training parameters of other baseline methods are the same as in the paper [15].

Table 1 The hyper-parameters of the model and their corresponding values

Hyperparameter	Value
Learning rate	9.6e-4
Learning rate decay	0
Dropout rate	0.5
Optimizer	Adam
L1,L2 regularization	0.002, 0.001
Training epochs	500
Batchsize	32
Chebyshev polynomial	2

5.3 Results and discussion

We compare our model with the six baseline methods on SMR, we use the accuracy, f1-score and precision as the evaluation metrics to evaluate the performance of the models. Table 2 shows the performance of the different models on the SMR dataset, the results show that our model have better performance compared to the other baseline methods and the proposed MutulaGraphNet. For the traditional methods, the random tree(RF) has better performance than the support vector machine(SVM), but both of them aren't good enough. The FBCSP cannot extract and utilize complex features in multi-subject tasks, though it has good performance in single-subject tasks. And the results show that the traditional machine learning methods can't learn the complex features well, the deep learning models EEGNet and ShallowConvNet all outperform the traditional methods which demonstrate the effectiveness of deep convolutional neural networks for EEG-bask classification tasks, however the performance of DeepConvNet demonstrates that the deeper convolutional network doesn't work better.

Table 2 The performance comparison of the state-of-the-art approaches on the SMR dataset

Model	Accuracy	F1-score	Precision
SVM	0.3488	0.3485	0.3486
Deep ConvNet	0.3507	0.3191	0.4148
FBCSP	0.3511	0.3366	0.3714
RF	0.4008	0.3996	0.4004
EEGNet	0.4616	0.4838	0.5095
Shallow ConvNet	0.4857	0.4789	0.4978
MutualGraphNet(ours)	0.5190	0.5175	0.5208
MCGNet$^+$(ours)	**0.5227**	**0.5239**	**0.5278**

In order to evaluate the effect of the depth of network, we study the impact of the layers of ST-GCN in Figure 5. The horizonal axis in Figure 5 represents the layers of ST-GCN and the vertical axis represents the corresponding performance of the model. The results show that the MCGNet$^+$ with more ST-GCN layers doesn't work better, the best performance is achieved with 4 layers and with the number of layers increases the performance gets worse. That is because the increase in the number of layers leads to an increase in training parameters, but the training data set is too small to train the model with more parameters.

In this paper, we extract differential entropy(DE) feature as the input of the model, and in EEG-based tasks there are other five different features[37]: power spectral density(PSD), differential asymmetry (DASM), rational asymmetry(RASM), asymmetry(ASM) and differential caudality(DACU) features from EEG. The DASM and RASM can be expressed as:

$$DASM = DE(X_{left}) - DE(X_{right}) \tag{17}$$

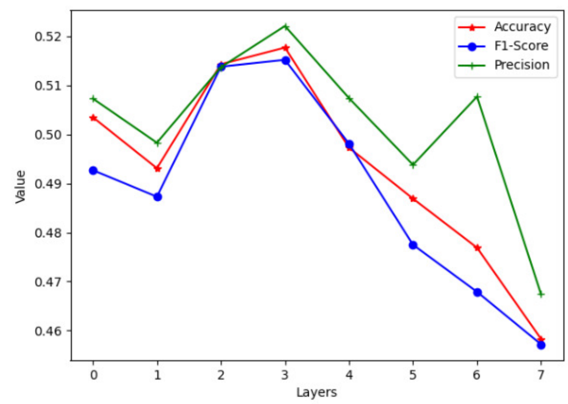

Figure 5 Performance of the proposed model with different ST-GCN layers.

$$RASM = DE(X_{left})/DE(X_{right}) \tag{18}$$

ASM features are the direct concatenation of DASM and RASM features. DCAU features are the difference between DE features of frontal-posterior electrodes, which can be defined as:

$$DCAU = DE(X_{frontal}) - DE(X_{posterior}) \tag{19}$$

We also evaluate the performance of our models on these features. All the experiments are performed with 4-fold cross-validation and the training settings are the same as above. The results are presented in Table 3, the PSD feature still has the worst performance and the DE feature outperforms the other features, DCAU feature also achieves comparable performance, but ASDM and DSAM feature contain less information which leads to limited performance. All the features have better performance with the new model, which indicated the effectiveness of the newly proposed method. Moreover, the results indicate that there exists some kind asymmetry of the brain which has dis-

Table 3 The performance of models for different features

Model	Feature	Accuracy	F1-score	Precision
MCGNet$^+$	PSD	0.2716	0.2695	0.2726
	DSAM	0.4124	0.4049	0.4052
	ASM	0.4039	0.3842	0.3877
	ASDM	0.3973	0.3881	0.3881
	DCAU	0.4375	0.4381	0.4435
	DE	**0.5227**	**0.5239**	**0.5278**
MutualGraphNet	PSD	0.2604	0.2286	0.2595
	DSAM	0.3646	0.3523	0.3541
	ASM	0.3815	0.3820	0.3879
	ASDM	0.3811	0.3777	0.3764
	DCAU	0.4162	0.4144	0.4191
	DE	**0.5190**	**0.5175**	**0.5208**

criminative information and our knowledge of the human brain is still very limited, the deeper understanding of brain is still required to obtain more effective and valuable information from EEG data. The new approach is compared with the several different adjacency matrixes that we designed:

1 KNN: For each channel, select the nearest N channels to establish a connection.
2 The Euclidean distance(ED): According to the actual distance of each electrode on the brain, select adjacent points to establish a connection.
3 Random: Randomly select channels and establish connections between channels.
4 Mut_Euclidean : Use the euclidean diatance to establish connections and calculate the mutual information.
5 Mut_KNN: Use KNN to establish connections and calculate mutual information between connected channels.
6 Mut_ED: Use the euclidian distance to conform connection and calculate mutual information between connected channels

The results of classification with different kind of adjacency matrix are shown in Figure 6. It can be seen that

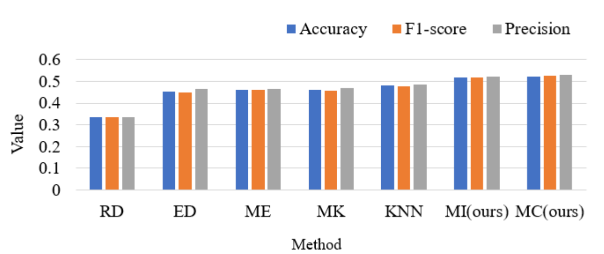

Figure 6 The performance of the proposed model with different kind of adjacency matrix. RD represents the random, ED represents the educlidean distance, ME represents Mut_Euclidean, MK denotes the Mut_KNN, MI denotes Mutual Information and MC denotes Mutual_Cos.

the MI_cos adjacency matrix has better performance than the MI adjacency matrix, Mul_KNN and Mul_ED are better than KNN and ED which means that mutual information could provide valuable information for ST-GCN. Furthermore, the adjacency matrix surly could effect the performance of classification.

6 Conclusion

In this paper, we improve the original model for motor imagery classification task based on our previous work[30]. Instead of using the stable adjacency matrix, we calculate the cosine similarity of the columns of the embedding to generate the dynamic adjacency matrix. The main advantage of the new model is that it could adjust the input matrix during the training process to utilize the features fully. The experiment results demonstrate that the new model outperforms the state-of-the-art methods as well as our previous model. Furthermore, the adjacency matrix has much more impact on the performance of the GCNs, and more suitable adjacency matrix can still be explored.

The current understanding on brain mechanisms is still limited, more influencing factors will be taken into account to further improve the forecasting accuracy. Moreover, motor imagery EEG data presents individual differences, such as FBCSP has different performances when experimenting with EEG data that from different subjects, and it can achieve good results when using the same subject's data for training and testing, but it does not perform well in mixed data of multiple subjects. Individual differences also affected the development of solutions for the classification task of motor imagery. How to eliminate individual differences and extract valuable features is still key for wider application of EEG-based tasks. Some current transfer learning methods may be deployed to eliminate individual differences and further expand the scope of EEG applications.

Acknowledgment

The authors would like to thank Prof.Sanghyuk Lee from XJLU for his support on this project.

Availability of data and materials

The data and code are available upon direct request to the authors.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

The experiment is designed and performed by Yan Li and Haolan Zhang. The paper is written by Yan Li, modified and optimized by David, Ning Zhong and Haolan Zhang. All the authors read and approved the final manuscript.

Authors' information

[1]Zhejiang University, Hangzhou, China, ly21121@zju.edu.cn.[2]Maebashi Institute of Technology, Maebashi, Japan,zhong@maebashi-it.ac.jp.[3]Monash University, Melbourne, Australia, David.Taniar@monash.edu. [4]Ningbo Institute of Technology, Zhejiang University, Ningbo, China, haolan.zhang@nit.zju.edu.cn.

Declarations

Funding

This work is partially supported by Humanity and Social Science Foundation of the Ministry of Education of China (21A13022003), Zhejiang Provincial Natural Science Fund (LY19F030010), Zhejiang Provincial Social Science Fund (20NDJC216YB), Ningbo Natural Science Fund (No. 2019A610083), Zhejiang Provincial Educational Science Scheme 2021 (GH2021642) and National Natural Science Foundation of China Grant (No.72071049).

Author details

[1] Zhejiang University, Hangzhou, China. [2]Maebashi Institute of Technology, Maebashi, Japan. [3]Monash University, Melbourne, Australia. [4]Ningbo Institute of Technology, Zhejiang University, Ningbo, China.

References

1. Song, Y., Wang, D., Yue, K., Zheng, N., Shen, Z.-J.M.: Eeg-based motor imagery classification with deep multi-task learning. 2019 International Joint Conference on Neural Networks (IJCNN), 1–8 (2019)
2. Wolpaw, J.R., Birbaumer, N., Mcfarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Brain-computer interfaces for communication and control. Supplements to Clinical Neurophysiology **113**(6), 767–791 (2002)
3. Blankertz, B., Dornhege, G., Krauledat, M., Müller, K., Curio, G.: The non-invasive berlin brain-computer interface: Fast acquisition of effective performance in untrained subjects. NeuroImage **37**(2), 539–550 (2007)
4. Wolpaw, J.R., Mcfarland, D.J., Neat, G.W., Forneris, C.A.: An eeg-based brain-computer interface for cursor control. Electroencephalography and Clinical Neurophysiology **78**(3), 252–259 (1991)
5. Kübler, A., Kotchoubey, B., Kaiser, J., Wolpaw, J.R., Birbaumer, N.: Brain-computer communication: unlocking the locked in. Psychological Bulletin **127**(3), 358–375 (2001)
6. Mcfarland, D.J., Anderson, C.W., Muller, K.R., Schlogl, A., Krusienski, D.J.: Bci meeting 2005-workshop on bci signal processing: feature extraction and translation. IEEE Trans Neural Syst Rehabil Eng **14**(2), 135–138 (2006)
7. Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, Rakotomamonjy: A review of classification algorithms for eeg-based brain-computer interfaces: a 10 year update. Journal of neural engineering (2018)
8. Bashashati, A., Fatourechi, M., Ward, R.K., Birch, G.E.: A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals. Journal of Neural Engineering **4**(2), 32 (2007)
9. Makeig, S., Kothe, C., Mullen, T., Bigdely-Shamlo, N., Zhang, Z., Kreutz-Delgado, K.: Evolving signal processing for brain–computer interfaces. Proceedings of the IEEE **100**(13), 1567–1584 (2012)
10. Lotte, F.: A tutorial on eeg signal processing techniques for mental state recognition in brain-computer interfaces. In: Miranda E., Castet J. (eds) Guide to Brain-Computer Music Interfacing, 133–161 (2014)
11. Kachenoura, A., Albera, L., Senhadji, L., Comon, P.: Ica: A potential tool for bci systems. Signal Processing Magazine IEEE **25**(1), 57–68 (2008)
12. Keng, A.K., Yang, C.Z., Wang, C., Guan, C., Zhang, H.: Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b. Frontiers in Neuroscience **6**, 39 (2012)
13. Woehrle, H., Krell, M.M., Straube, S., Su, K.K., Kirchner, F.: An adaptive spatial filter for user-independent single trial detection of event-related potentials. IEEE transactions on bio-medical engineering **62**(7), 1696–1705 (2015)
14. Schirrmeister, R.T., Springenberg, J.T., Fiederer, L., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T.: Deep learning with convolutional neural networks for eeg decoding and visualization. Human Brain Mapping **38**(11), 5391–5420 (2017)
15. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: Eegnet: A compact convolutional network for eeg-based brain-computer interfaces. Journal of Neural Engineering **15**(5), 056013–105601317 (2016)
16. Jia, Z., Lin, Y., Wang, J., Zhou, R., Zhao, Y.: Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification. In: Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence IJCAI-PRICAI-20 (2020)
17. Zhou, K., Song, Q., Huang, X., Zha, D., Hu, X.: Multi-channel graph neural networks. In: Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence IJCAI-PRICAI-20 (2020)
18. Bruna, J., Zaremba, W., Szlam, A., Lecun, Y.: Spectral networks and locally connected networks on graphs. Computer Science (2013)
19. Haltaş, K., Erguzen, A., Erdal, E.: Classification methods in eeg based motor imagery bci systems, pp. 1–5 (2019). doi:10.1109/ISMSIT.2019.8932947
20. Lu, N., Li, T., Ren, X., Miao, H.: A deep learning scheme for motor imagery classification based on restricted boltzmann machines. IEEE Transactions on Neural Systems and Rehabilitation Engineering (2016)
21. Aggarwal, S., Chugh, N.: Signal processing techniques for motor imagery brain computer interface: A review. Array **s 1–2** (2017)
22. Guo, S., Lin, Y., Feng, N., Song, C., Wan, H.: Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. Proceedings of the AAAI Conference on Artificial Intelligence **33**, 922–929 (2019). doi:10.1609/aaai.v33i01.3301922
23. Li, C., Cui, Z., Zheng, W., Xu, C., Yang, J.: Spatio-temporal graph convolution for skeleton based action recognition (2018)
24. Geng, X., Li, Y., Wang, L., Zhang, L., Liu, Y.: Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. Proceedings of the AAAI Conference on Artificial Intelligence **33**, 3656–3663 (2019)
25. Elmore, K.L., Richman, M.B.: Euclidean distance as a similarity metric for principal component analysis. Monthly Weather Review **129**(3), 540–549 (2010)
26. Klve, T., Lin, T.T., Tsai, S.C., Tzeng, W.G.: Permutation arrays under the chebyshev distance. IEEE Transactions on Information Theory **56**(6), 2611–2617 (2010)
27. Dongen, S.V., Enright, A.J.: Metric distances derived from cosine similarity and pearson and spearman correlations. Computer ence (2012)
28. Kent, J.T.: Information gain and a general measure of correlation. Biometrika (1), 163–173 (1983)
29. Maes, F., Collignon, A.: Multimodality image registration by maximization of mutual information. IEEE Trans Med Imaging **16**(2), 187–198 (1997)
30. Li, Y., Zhong, N., Taniar, D., Zhang, H.: Mutualgraphnet: A novel model for motor imagery classification. arxiv preprint arxiv:2109.04361 (2021)
31. Feng, X., Jiang, G., Bing, Q., Liu, T., Liu, Y.: Effective deep memory networks for distant supervised relation extraction. In: Twenty-Sixth International Joint Conference on Artificial Intelligence (2017)
32. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering (2016)
33. Keng, A.K., Yang, C.Z., Wang, C., Guan, C., Zhang, H.: Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b. Frontiers in Neuroscience **6**, 39 (2012)
34. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**(3, article 27) (2007)
35. Liaw, A., Wiener, M.: Classification and regression by randomforest. R News **23**(23) (2002)
36. Ishida, T., Yamane, I., Sakai, T., Niu, G., Sugiyama, M.: Do we need zero training loss after achieving zero training error? (2020)
37. Zheng, W.L., Zhu, J.Y., Lu, B.L.: Identifying stable patterns over time for emotion recognition from eeg (2016)