The Research of BDPCA Identifying Emotion by EEG

Chao Yan¹³, Jintao Nian¹ and Chao Lv¹²

¹ Changchun University of Science and Technology, College of Electronic Information Engineering, Changchun Jilin, China
² Corresponding Author
³ Author to whom any correspondence should be addressed

Email: 1483776443@qq.com

Abstract. Electroencephalogram (EEG) data contain wealthy information about the brain’s and body’s pathology and physiological state. It’s not easily to identify the truth that EEG contained. The unsupervised learning method don’t need to take label by human. Without subjective feeling, it greatly improve training accuracy. In this current paper, adopted improved Density Peak Clustering Algorithm (DPCA) to train EEG data. To solved the problem that difficult to determined cluster center number, Bayesian Information Criterion (BIC) was introduced. The algorithm was verified feasibility that in EEG processing by experiment which divided fatigue state level in lab. And used SJTU Emotion EEG Data set (SEED) identifying different emotions. Compared with other cluster algorithms, BDPCA accuracy totally raised about 5%. And BDPCA behavior was steadier in different emotion types.

1. Introduction

Electroencephalogram (EEG) is a weak electrical signal which generated by the activity that have Spontaneous and rhythmical ion exchange between neurons in brain. It is representing of comprehensive physiological electrical activity in brain[1]. The EEG contains a wealthy information about the brain’s and even body’s pathology and physiological state[2]. Due to the bioelectricity signal have the characteristics that well real-time difference and difficult to disguise[3] and Brain Computer Interface (BIC) [4], which can translate EEG signal into user intentions and output is implementation. These Provide the conditions that EEG Widely used. With the rapid development of artificial intelligence technology, the methods based on artificial intelligence have attracted more attention in data processing. Jianfeng Hu [5] applied K Nearest Neighbor(K-NN) and Support Vector Machine (SVM) into EEG data set analysis. Saeedi Maryam [6] applied SVM and Multilayer Perceptron (MLP) into the EEG data set processing and analysis. T Waili [7] applied Artificial Neural Network (ANN) into EEG data set to identify different people personal identity verification. Bouallegue Ghaith [8] used Artificial Neural Networks (ANN) analyzed EEG data set, achieved Alzheimer's disease (AD) identifying. Shen lei[9] used convolutional neural networks (CNN) identify epilepsy using EEG data set. However, the knowledgeable of state which could represented by EEG signal is not sufficient. It’s a mainly considered problem that how to fully excavate useful information from EEG signal. The classification algorithm usually divides the state which was known [10]. Comparing with the supervised learning method, the unsupervised learning method like clustering algorithm don’t need take the label by human and it can automatically divide different EEG signal into different categories. DPCA is based on the local density of data and the distance between data points as gist divide.
different cluster, but DPCA can’t determine the number k of cluster center. Therefore, Bayesian Information Criterion (BIC) was introduced to improve DPCA. In the current paper, the feasibility has been verified by experiment of fatigue state dividing. The emotion identifying experiment by SEED was used to analyze BDPCA and other cluster algorithm performance.

2. Emotion identifying mechanism

The EEG consist of abundant rhythmic wave[10]. When the peoples physical state changed, the rhythmic wave accordingly changed[11]. The relationship between the EEG that generated by the brain system and the physical state proved that it was possible that divide different physical state by EEG. The steps of analyzing different moods based on the EEG are as follows: collecting EEG, determining the clustering analysis algorithm, EEG preprocessing, extracting EEG feature, dividing different moods. The methods of EEG feature extraction usually process in frequency domain or time domain. The time domain analyzing methods are simple and intuitive, it extracts parameters such as amplitude, mean value, variance, regularity and synchronization in general. The simple time domain feature extracting methods are not suitable for complex EEG analyzing. The frequency domain analyzing methods mapping the signal into the frequency domain for analyzing. However, EEG have randomness, and most of useful information is transient. According to the "uncertainty principle", signal processing can’t have good time resolution and frequency resolution at the same time, so it is necessary to combine time domain and frequency domain to process. The wavelet packet transform not only considerate the frequency resolution, but also retain the time resolution. The wavelet packet transform can divide the low frequency signal more careful compared with the wavelet transform. Therefore, the wavelet packet transform method was determined to extract the eigenvalue from the EEG data in this current paper. Then, the improved BDPCA was used to train the extracted feature vectors, and the division of different moods was carried out. A complete fatigue state division mechanism flow chart is shown as Figure 1.

![Figure 1. Flowchart of different moods classification mechanism.](image)

3. Extracting feature from EEG

3.1. EEG data preprocessing

The EEG raw data collected by the experiment which contain a lot of interference such as Electromyography (EMG), Electrocardiogram (ECG) and power frequency interference noise[12]. Therefore, it is necessary to preprocessing EEG raw data to eliminate the above noise before extracting feature from EEG data[13]. Firstly, the EEG raw data is re-sampled with 128Hz, which not only reduces the size of the memory that EEG data occupied, but also makes the extraction of energy features parameters by wavelet packet transform easier. Then, the notch filter is used to remove the 50Hz power notch and the 0.45-35Hz band-pass filter is used to remove low frequency noise and high frequency interference. Finally, the independent component analysis (ICA) that in EEGLAB toolbox of MATLAB is adopted to remove obvious artifacts noise components. After the above steps, the EEG data is used for subsequent feature extraction.

3.2. Extracting feature from EEG

The purpose of extracting feature is to extract the information from the preprocessing EEG data for subsequent analysis. The EEG data mainly contains five wave components with different frequency band ranges, such as delta band (0.5-4Hz), theta band (4-7.5Hz), alpha band (8-13Hz), beta band (14-30Hz) and gamma band (30-45Hz)[14]. Wavelet packet transform was adopted to decompose EEG data in this current paper, and energy parameters in different frequency bands were extracted as the
feature values. The preprocessing EEG data were decomposed into 4-layer wavelet packets, and 16 nodes were obtained. The location and number of nodes were shown as Figure 2.

Figure 2. Four-layer decomposed wavelet packet trees.

The number of nodes which was used frequency band range are shown as Table 1.

Table 1. Wavelet packet nodes correspond to the band range.

| Wavelet packet decomposition node label | Frequency band range (Hz) |
|----------------------------------------|---------------------------|
| S(4,0)                                 | 0.5-4                     |
| S(4,1)                                 | 4-8                       |
| S(4,2)                                 | 8-12                      |
| S(4,3)                                 | 12-16                     |
| S(4,4)                                 | 16-20                     |
| S(4,5)                                 | 20-24                     |
| S(4,6)                                 | 24-28                     |
| S(4,7)                                 | 28-32                     |

The energy parameters of rhythm waves can be calculated by the coefficients of the decomposition sub-bands which related to this rhythm waves, the calculated formula shown as (1).

\[
E_i = \sum_i |P_i(t)|^2
\]  

(1)

Where \(P_i\) represents the decomposition coefficient of the i node in layer j in the wavelet packet decomposition. The energy of delta band, theta band, alpha band and beta band can be calculated by the above formula. According to Wei-Long Zheng[15], the beta band (14-30Hz) and gamma band(30-45Hz) accuracy of mood identify was higher. In this current paper, Select beta band and gamma band energy parameters as feature vector in experiment of different moods dividing.

4. The improvement of BDPCA

4.1. BDPCA flow

(1) Select the truncation distance \(d_c\).

(2) The local density is obtained according to Equation (2) and (3).

\[
\rho_i = \sum_{x_i \in U} X(\text{dist}(x_i, x_j) - d_c)
\]  

(2)

\[
X(x) = \begin{cases} 
0 & x \leq 0 \\
1 & x > 0 
\end{cases}
\]  

(3)

(3) Calculate the cluster center distance according to Equation (4).

\[
\delta_i = \begin{cases} 
\min(dist(x_i, x_j)) & \rho_i = \max(\rho) \\
\max(dist(x_i, x_j)) & \rho_i < \max(\rho) 
\end{cases}
\]  

(4)
(4) Plot decision graph according to the local density and the distance with the cluster center. Determine the type of the data points according to the graph. The cluster center have big local density in general. And cluster center point has further distance with cluster center points than normal points. Noise points far away cluster and lonely. Thus normal points are near the horizontal axis, and noise points are near the vertical axis. The center of clustering is far away from the horizontal axis and the vertical axis.

(5) Clustering division. The data points which belong to the non-clustering center points are divided into the same class of points which the local density is greater than it and has the minimum distance. After all the points were judged and analyzed, the algorithm will be finished.

DPCA propose the selecting principle of the cluster center point according to the decision graph. In generally, the distribution of the data set is unknown in the actual data analysis, the cluster center number is unknown in general and the points distribution in decision graph is not very clearly, so it is easy to appear that the normal points which were suspected to be the cluster center points is determined the cluster center point. It can be seen that it is difficult to accurately determine the point type of suspected points by the decision graph generated by DPCA. Therefore, BIC is used to improve the algorithm to assist in determining the cluster center points. The core idea of BIC is to maximize the probability of the currently selecting model under the current data set. In the current study of EEG, \( X = \{ x_i \}_{i=1}^R \) is the EEG data set, \( R \) is the number of samples in the data set. \( M_i \) is the selected clustering model, \( P \) is the number of model parameters. In the current study, assumed that the data sets of each cluster obey the Gaussian distribution, so the value of \( P \) is the sum of \( MK \) centroid and \( K \) variance estimates.According to BIC, the posterior probability of the model is selected under the current data set is calculated by Equation (5).

\[
P(M_i | x_1, x_2, ..., x_n) = \frac{P(x_1, x_2, ..., x_n | M_i) P(M_i)}{P(x_1, x_2, ..., x_n)}
\]

(5)

Under the initial conditions, each models have the same probability that were selected, finding the model with the highest posterior probability is equivalent to finding the model with the highest edge probability.

\[
P(x_1, x_2, ..., x_n | M_i) = \int_{\theta_i} L(\theta | x_1, x_2, ..., x_n) g_i(\theta) d(\theta)
\]

(6)

Where \( \theta \) is the parameter vector of the model, \( L \) is the maximum likelihood function, and \( g_i(\theta) \) is the probability density function of the parameter vector. The calculation of BIC parameters and the maximum likelihood function is simplified into Equation (7)[16][17].

\[
BIC = P \ln(n) - \ln(L)
\]

(7)

\( P \) is the number of model parameters. The above BIC parameters represent the loss between the model and the real data. The absolute value of BIC parameters is smaller, the selected model is better.

4.2. The improved BDPCA

The BDPCA which was described in 4.1 has the following disadvantages: the \( \rho_i \) and \( \delta_i \) of different data points are discrete values, it may lead to that \( \rho_i \) or \( \delta_i \) of different data points is equal. Therefore, there are too many suspected cluster center points due to the points with the highest local density. If there was several points both have the max Local density, the distance of these points is difficult to calculate because it is difficult to determined which point distance calculated by max distance. If all points which have the max local density distance calculating by using the max distance equation it would generate the mistake that could avoid. And it would generate obstacles in cluster center points determined. Although the number of optimal cluster classes can be determined by calculating BIC
parameters, it will greatly increase the compute time, and will make difficult to partition the subsequent data points.

Therefore, to improve the situation, use similarity to redefine the values of $\rho_i$ and $\delta_i$ in the current paper. Calculate the optimized local density and distance according by the Equations (8) and (9).

$$\rho_i = \sum_{j=1}^{n} e^{-\frac{\|x_i-x_j\|^2}{2d_i^2}}, \quad \delta_i = \begin{cases} \max(\omega_j) & \rho_i = \max(\rho) \\ \min(\omega_j) & \rho_i < \max(\rho) \end{cases}$$

(8)

(9)

The original $\rho_i$ and $\delta_i$ are replaced by calculating the similarity between the points, it greatly reduces the situations of different data points have the same $\rho_i$ and $\delta_i$. However, the distance between two data points is simply calculated by using the similarity. It is easy to misjudge when the edge data points of one cluster is closer to the edge data points of another cluster.

In order to avoid the above situation, common neighbor parameter (CNN) was introduced to optimize the calculation of similarity[18]. The optimized calculated formula is shown as (10):

$$\omega_j = e^{-\frac{\|x_i-x_j\|^2}{2d_i^2(CNN(s, s+1))}}$$

(10)

Where CNN represents the number of data points co-existing in the $d_i$ neighborhood of data points $x_i$ and $x_j$. Different from the traditional similarity calculation method, the type of data points after optimization is determined by the distance and the local density between two points. The data points are divided into the same cluster with the nearest data points in the data set whose local density is greater than itself. In this way, the misjudgment that caused by the close distance between the edge points of two different clusters could be avoided.

5. BDPCA Applying in EEG

BDPCA has been used in some engineering data processing research, such as aero-engine fault data marking, and the simulation results shown that the algorithm has got a wonderful accuracy[19]. The EEG data have random, non-stationary characteristics like other engineering signals. In this current paper, BDPCA was applied to analyze EEG data.

5.1. Feasibility analyze of BDPCA in EEG

To verified the feasibility, using BDPCA to divide fatigue level by using EEG which collected by the emotive. The roughly process of experiment as follows. The experiment was carried out in the afternoon (13:30-16:00), which was easier to generate drowsiness. Before formally recorded the EEG data, the experimenters have took a 10min test by operating the driving simulation and avoiding the barriers to familiarize the operate system. After the formal experiment began, the experimenters was required to drive at 90 Km/h in a straight line. During the period of driving, the experimenters controlled the steering wheel to avoid obstacles which randomly appearing on the screen in front of him. The experiment was lasted for 150min. During the whole experiment period, experimenters could stop the experiment if the experimenters feels unwell, and the experiment would end after stopping. If didn’t appearing special situations, the drivers would continue driving for 150 minutes without rest. During the driving period, the portable EMOTIV device collected the driver's EEG at the frequency of 256Hz. During the experiment, there were also two staff, one of whom recorded the number of blink frequency of the driver every 2 minutes on the side, and the other recorded the time of the obstacles appearance and the results of obstacle avoidance. The recording results would be used the criteria to classify the fatigue state level.
Firstly, the EEG raw data was divided into 20S data segments, and 10S of the data segment was intercepted to extract feature. Then, used the wavelet packet transform to extract the energy parameters with the sample entropy of the signal as the input vector. After which trained the input vector by the BDPCA model. The experiment result proposed a method to divide different fatigue level by energy ratio parameters of alpha band and theta band with delta band. When feature of energy parameters located in 0-0.9, it represents awake state. When feature of energy parameters located in 0.9-1.4, it represents mild fatigue state. When feature of energy parameters located in 1.4-1.75, it represents moderate fatigue state. When feature of energy parameters located in 1.75-2.3, it represents severe fatigue state. When feature of energy parameters located in 2.3-2.8, it represents extreme fatigue state. The conclusion in compliance with experimenters behavior.

Analyzing the result of BDPCA model trained the EEG data which collecting by EMOTIV when experimenters operate virtual driving equipment. The result in compliance with behavior data and experimenters’ subjective feelings. The experiment results verified the feasibility of BDPCA model analyzing EEG data. Due to fatigue state needs takes label by subjective feeling, it hardly to observe BDPCA accuracy with other cluster algorithm. To solved this problem, we used SEED to observe different cluster algorithms performance.

5.2. BDPCA in identifying different moods
In this current paper, the SEED was selected to observed different cluster algorithm performance. The EEG data set was collected by 15 subjects (7 males and 8 females). 15 movie fragments which can generate positive, neutral and negative moods were chosen as stimuli used in the experiments. The 15 Chinese film fragments was clipped by 6 movies. The movies name and induced emotions shown as Table 2.

| Movies Name                  | Induce Emotion |
|------------------------------|----------------|
| Tangshan Earthquake          | negative       |
| Back to 1942                 | negative       |
| Lost in Thailand             | positive       |
| Flirting Scholar             | positive       |
| Just Another Pandora’s Box   | positive       |
| World Heritage in China      | neutral        |

There is a 5s hint before each clip, 45s for self-assessment and 15s for rest after each clip in one session. The order of presentation is arranged so that two film clips targeting the same emotion are not shown consecutively. In this current paper, The EEG data was divide different trails, select the EEG trails when movies were in climax. The EEG trails were taken a preprocessing step which introduced in chapter 3.1 and extracted the feature vectors by using the step of chapter 3.2. Finally, the energy parameters were extracted which select beta band and gamma band of channel Fp1 and Fp2. The feature vectors was trained by BDPCA model. The decision graph shown as Figure 3.

![Figure 3. The decision graphs generated by the SEED](image-url)
According to the decision graph, the number of suspected cluster center points was 2-6, and calculated BIC parameters when the number of cluster center points was 2-6. The calculating results shown as Table 3.

**Table 3.** Bic parameters with different K values

| K=2   | K=3   | K=4   | K=5   | K=6   |
|-------|-------|-------|-------|-------|
| -418.88 | -169.04 | -258.40 | -268.46 | -319.18 |

The value of BIC parameter determined that the optimal number of clusters was 3. The EEG data-set consisted of three different type moods which was sad, happy and neutral.

The analysis results of BDPCA model in SEED shown as Table 4.

**Table 4.** SEED data clustering results

| Type of moods | Actual Number | Dividing Number | Accuracy  |
|---------------|---------------|-----------------|-----------|
| Happy         | 206           | 135             | 65.53%    |
| Neutral       | 238           | 193             | 81.09%    |
| Sad           | 235           | 177             | 75.34%    |

In addition to BDPCA, other cluster algorithms was adopt to train the SEED. Finally, adopted GMM cluster algorithm, Hierarchical cluster algorithm and K-means cluster algorithm trained SEED. The accuracy of different cluster algorithms shown as table 5.

**Table 5.** Cluster Algorithms accuracy

| Type of moods | Happy | Neutral | Sad  |
|---------------|-------|---------|------|
| K-means       | 62.98%| 63.31%  | 67.63%|
| Hierarchical  | 60.92%| 81.70%  | 64.73%|
| GMM           | 45.38%| 50.21%  | 65.05%|
| BDPCA         | 65.53%| 81.09%  | 75.34%|

To observe different cluster algorithm performance more intuitive. The histogram of cluster algorithms accuracy was shown as Figure 4.

**Figure 4.** Different algorithms accuracy histogram.

By observing Table 5 and Figure 4. It easy to found that in positive emotion identification, BDPCA accuracy raised more than 5% comprehensive than other algorithms; in neutral emotion identification, BDPCA accuracy was close to the highest accuracy in other algorithms; in negative emotion identification, BDPCA accuracy both raised about 10% than other algorithms. It can conclude that accuracy of BDPCA was higher and BDPCA performance was steadier in identifying different emotions than other cluster algorithms.
6. Conclusion
In this current paper, to improve the problem that comprehending limitation of EEG, proposed a improved cluster algorithm which based on DPCA into EEG analyzing. The experiment of fatigue level dividing in lab was successful, BDPCA model achieved different fatigue level. The result of experiment verified that BDPCA analyzed EEG data was feasible. Due to the character of dividing fatigue level that different fatigue level transition is not clearly and easy to influenced by subjective feeling, It’s hardly verified BDPCA performance was better than other cluster algorithm. Thus, we adapt SEED to identify different emotions. By identifying there emotions, found that BDPCA accuracy comprehensively improved about 5% than other cluster algorithms. In identifying different emotion, BDPCA performance was steady, other algorithms accuracy existed relatively large differences. It could say that BDPCA model has better robustness than other cluster algorithms. The result of experiment verified that BDPCA don’t need to take subjective label to EEG data in advance, can automatically divide EEG data with high accuracy and have robustness. Therefore, BDPCA has a good application prospect in using EEG data to detect diseases and other body abnormalities.

7. References
[1] Christoph M, Michel, Micah M and Murray 2012 Towards the utilization of EEG as a brain imaging tool J. Neuro Image. 2012, 61(2).
[2] Cichocki A and Sanei S 2007 EEG/MEG signal processing J. Computational intelligence and neuroscience. 2007.
[3] Pavlo A, Fred P, Roland G and Tamara G 2010 Using Electroencephalography to Measure Cognitive Load J. Educational Psychology Review. 2010, 22(4).
[4] Leuthardt Eric C, Schalk G, Wolpaw Jonathan R, Ojemann Jeffrey G and Moran Daniel W 2004 A brain-computer interface using electrocorticographic signals in humans J. Journal of neural engineering. 2004, 1(2).
[5] Jianfeng H. 2018 An approach to EEG-based gender recognition using entropy measurement methods J. Knowledge-Based Systems. 2018,140.
[6] Saeedi M, Saeedi A and Maghsoudi Arash 2020 Major depressive disorder assessment via enhanced k-nearest neighbor method and EEG signals J. Australasian Physical & Engineering Sciences in Medicine.2020,43(3).
[7] Waili T, Sidek K.A,Mohd N.S.H,Yaacob H and Othman M 2019 EEG Based Biometric Identification Using Correlation and MLPNN Models J International Journal of Online and Biomedical Engineering.2019, 15(10).
[8] Bouallegue G, Djemal R and Belwafi K.2021 Artificial EEG signal generated by a network of neurons with one and two dendrites J. Results in Physics.2021,20.
[9] Lei S, Xin-Yi G and Shou-Yan W.2020. Epileptic States Recognition using Transfer Learning and Dilated CNN J. Chinese Journal of Biomedical Engineering.2020,39(06):700-710.
[10] Zhang-Jin H, Chen C, Jin Z, Jia-Qi Z. 2020. Lie detection based on wavelet coherence analysis of multichannel EEG signals J. Journal of Electronic Measurement and Instrumentation. 2020,34(10):156-162.
[11] Da-Feng Q,2020 A fatigue detection method based on multimodal fusion neural network J. Modern computer 2020(35):32-36+42
[12] Feng-Jie F, Yang B,Hui-Fang J.2021 Research on EEG denoising algorithm based on EEMD-ICA J. Journal of measurement.2021,42(03):395-400.
[13] Zhen P,Ming Y,Wei-Jia H,Guo-Jun M.2021 An EMG artifact removal method based on PCA and adaptive step size IVA J. Modern electronic technology.2021,44(01):39-44.
[14] Cichocki A and Sanei S 2007 EEG/MEG signal processing J. Computational intelligence and neuroscience 2007.
[15] Wei-Long Z and Bao-Liang L 2015 *IEEE Transactions on Autonomous Mental Development (IEEE TAMD)*. 2015.
[16] Alex R and Alessandro L.2014 Clustering by fast search and find of density peaks *J. Science*.2014.344.
[17] Tsunenori ISHIOKA 2000 Extended K-means with an Efficient Estimation of the Number of Clusters *J. Tsunenori ISHIOKA*.2000.29(3).
[18] ZHANG X, LI J and YU H 2011 Local density adaptive similarity measurement for spectral clustering. *J.Pattern Recognition Letters*.2011,32(2):352-358.
[19] Chao L, Cheng G and Yun-Qing L.2020 Aeroengine fault data tag using on improved BDPCA algorithm *J. Journal of Vibration and Shock*.2020,39(09):35-41.