Journal of Electronic Research and Application

Research Article

Classify mental states from EEG signal using XGBoost algorithm

Yihao Wang

Shandong Experimental High School, Jinan, Shandong Provinces, China

Abstract: Brain-computer interface (BCI) is a leading-edge technique which allows the brain communicates with external devices. It has been applied in several fields, such as medical rehabilitation, virtual reality and so on. This invention introduces a technique that can be applied in education field to monitor and analyze users' electroencephalogram (EEG) so that the mental states could be identified. The algorithm of classifier used XGBoost which combined Bayes, KNN and SVM in it and its accuracy could reach to 80%. By using this technique, teacher could obtain the concentration status of students in real time and adjust his or her teaching method or remind the student who is wandering.

Publication date: December, 2019

Publication online: 31 December, 2019

****Corresponding author:*** Yihao Wang, guoliming@xyzrgroup.com

1 Introduction

Brain-computer interface (BCI), also called direct neural interface or brain-machine interface, is a technique providing a directly communication channel built between the brain and external devices independent of peripheral nervous system. A common function of BCI equipment is to monitor and analyze wearers' brainwaves in order to get detailed information about the functional activity condition on different brain areas and then transform those data into specific form. Based on this function, the technique has been applied in several fields, including medical rehabilitation, video games etc. Furthermore, with the development of EEG technology and wireless transmission technology, BCI equipment has been designed to be wireless and became smaller and lighter, so it can be wearable and used in more fields. Thus, the public is attaching more expectations to the application of BCI in the education field which is supposes to create a revolutionary change in education field. As a matter of fact, a prevailing problem of most education methods is to get real-time learning states of each student, and obviously, attention monitoring based on BCI can probably play an important role. So creating an attention monitoring method becomes a prior choice.

An attention monitoring system requires a set of BCI equipment to detect the EEG signals of the wearer and then process these signals through an algorithm which can extract effective data and transform those data into specific statements of attention concentration. This invention used XGBoost algorithm to process data. XGBoost initially started as a research project by Tianqi Chen as part of the Distributed (Deep) Machine Learning Community (DMLC) group. It became well known in the ML competition circles after its use in the winning solution of the Higgs Machine Learning Challenge. A main advantage of this method is its conciseness and immediateness which allows real-time monitoring to be achieved. Furthermore, by using this algorithm, the accuracy of identifying concentration state could reach up to 80%.

2 Experimental verification

2.1 Bayes

Naïve Bayes is a one of the classifiers that make predictions by predicting probability. It is based on Bayes' theorem which describes the probability of an event, based on prior knowledge of conditions that shown in the datasets. One of the advantages of naive Bayes classifier is that it doesn't need a quite huge

number of data for classification while it works well in some complex situations.

2.2 KNN

K-nearest neighbor (KNN) is a basic classification and regression method. Its basic method is: given a test case, based on some distance measure to find the nearest K instance points in the training set, and then based on the k nearest neighbor information to predict.

Distance measures, selection of K value and classification decision rule are three basic elements of k-nearest neighbor method. According to the selected distance measure such as Manhattan distance or Euclidean distance, the distance between the test case and each instance point in the training set can be calculated, and K nearest neighbors can be selected according to the K value. Finally, the test cases are classified according to the classification decision rules.

2.3 SVM

As an algorithm which has been widely used in the biological field, support-vector machines (SVMs, also support-vector networks) prove to perform well for data classification. The principle of SVM is to separate categories by a clear gap and that means to find a hyperplane in an N-dimensional space so that the distance from it to the nearest data point on each side is maximized. For example, if N=2, then the hyperplane is just a line and if N=3, the hyperplane is a plane that separates the features.

In this experiment, support-vector machine is employed to classify the four degrees of one's attention. Given that there are several features in our research, different kernel functions are taken into account to realize the conversion from two-dimensional space to multi-dimensional space.

2.4 XGBoost

Extreme gradient boosting (XGBoost) is one of the most powerful approaches for most regression and classification problems. In this algorithm, many classifiers whose performance was poor in experiment were integrated together to form a strong classifier. The algorithms mentioned before were considered as tree models and the XGBoost is a combination of those tree models and form an ascending tree model which was based on CART regression tree model. The objective function of XGBoost is consisted of two parts: train loss and complexity of the tree.

2.5 ANN

Artificial neural networks are computer systems coming from biological neural network and there are a collection of connected units called artificial neurons that imitate human brains.

where X is input, Y is output and w is the weight of each input. What's more, b means Bias which provides every node with a trainable constant value. The function of f is for nonlinear data while linear neural network was employed in our experiment.

Since neural network is the most widespread algorithm nowadays, different types of artificial neural network have emerged like convolutional neural networks. Inspired by the way biological neural networks work in the human brain, artificial neural network does make computers analyze data and make predictions just like humans.

However, due to time and other factors, linear neural network was the only neural network that employed in this experiment.

2.6 Fine-Turning

In order to get a better performance classifier, the parameters in all models such as parameter K in KNN, parameter C in SVM and series parameters in XGBoost need to be adjusted. The parameter that has the best performance was different from others in each pre-processing algorithm. To test the Bayes, KNN and SVM algorithm, the data was first divided into test set and train set with proportion 3/7 and then recorded the highest accuracy. Then, for further performance test, the group used 5-fold cross validation and recorded the average accuracy. As for the XGBoost algorithm, since there were lots of parameters, one parameter was changed at one time with the rest fixed to reduce the complexity. While the ANN algorithm needs to be trained for many times to improve its performance.

3 Results

The features of absolute sum of all set were visualized. The red lines segment each chart into four parts which correspond to four types of status. From the chart it can be see that because some extreme high values exist in the data file, the differences between each sample in one type could be very large and increase the difficulty for computer to classify the sample.

In comparison, the performance of Naïve Bayes algorithm was the worst. The average accuracy of

Bayes was about 40%. The Normalizer was the best pre-processing algorithm in Bayes, the average accuracy reached about 46%. In this case, the classifier could easily distinguish state 1 and state 2 while it was confused other two types.

To fine the highest accuracy, the program traversed the value of K in range 1 to 30 in KNN algorithm and drew the line chart. Figure 5 shows the accuracy varies with parameter K. There are three charts and from left to right is StandScaler, MinMaxScaler and Normalizer data pre-processing algorithm respectively.

In KNN algorithm, the best accuracy could reach 70% when the quotient of test set and train set is 3/7. However, the best average accuracy is 57.58% when using StandScaler pre-processing algorithm and parameter K is 13.

The two charts from left to right is StandScaler and Normalizer data pre-processing algorithm respectively.

In the three basic models, the performance of SVM algorithm was the best. The accuracy was about 75% in 3/7 division and the average accuracy was 62.66% when using StandScaler.

The performances of all the basic models were not good as a classifier. However, once combined them together using XGBoost algorithm, the accuracy will boost. Table 4 shows the test results of XGBoost algorithm. The average accuracy was between 75% and 81% when using different data pre-processing algorithm. Same as KNN and SVM, StandScale also has the best performance in XGBoost.

The accuracy is depending on the train time. As the train time increase, the accuracy of train set and test set will increase. The accuracy for the test set will increase to 74% when the train time is 1000. However, the group did not continue the experiment by increasing the train time because it will take a long time to process the training and there is the possibility of over-fitting after thousands time of training.

4 Conclusion

To improve the immediateness and accuracy of the existing methods in classifying different states of attention, the research proposes a real-time attention monitoring method that is based on the machine learning method of XGboost. Through combining machine learning methods with significant features extracted from the FFT power spectrum of 64 channels of EEG signal, the proposed method notably reduces the processing time and increases the accuracy up to 81.02% when using 5-fold cross-validation, which demonstrates the model's advantage of good stability and powerful classification ability.

This attention monitoring method for EEG signal includes EEG signal database, feature extraction module, feature pre-processing module and different types of classifiers. The power spectrum of the EEG database is imported into the feature extraction module and extract features from the original data. Then, the representative characteristics will be standardized or normalized in the feature pre-processing module. At last, based on these features and 4 kinds of labels, the classifier can identify 4 states of attention, including concentration, meditation wandering, fatigue, and sleepiness.

On the whole, compared with other machine learning methods such as KNN or SVM, the XGBoost algorithm obtains the advantage of the integration of many methods, making its accuracy higher under the same condition. Compared with deep learning method such as ANN which requires plenty of training time, this method has a higher rate of convergence while maintaining acceptable accuracy.