Graph-Based Data Augmentation Approach for Electroencephalogram Analysis

Yuchou Chang

Computer Science and Engineering Technology Department, University of Houston - Downtown, Houston, Texas, United States 77002

Abstract: Brain state classification of Electroencephalogram (EEG) data acquired by Brain Computer Interface (BCI) is a new dimension in an immersive human-computer interaction. However, labeled EEG data is limited due to long time of data acquisition and expensive data post-processing. For this reason, a preliminary study on brain state classification with limited EEG labeled data is proposed. A small amount of labeled data and a large amount of unlabelled data are learned together via a graph-based semi-supervised learning approach for classifying brain states. Our method produces promising results on EEG signal classification and shows the effective use of unlabelled brain state data.

Keywords: Semi-Supervised Learning, Graph, Electroencephalogram (EEG), Brain State Classification, and Data Augmentation

Introduction

Electroencephalogram (EEG) signals [1, 2] are produced from brain activity, when many neurons have synchronized occurrence of synaptic potential. It records the brain activity when the brain wave changes [3, 4]. It reflects electrophysiological activity of brain cells in the cerebral cortex or scalp surfaces. Brain waves originate from the postsynaptic potentials of pyramidal neurons. Brain waves are the basic theory of neuroscience research and brain wave monitoring is widely used in clinical practice. Brain waves’ frequency ranges from 1 to 30 Hz, which can be divided into four bands, namely δ (1-3Hz), θ (4-7Hz), α (8-13Hz), and β (14-30Hz). In addition, in the awakening and focus on a thing, often see a higher frequency γ wave with the frequency of 30 ~ 80Hz, which is higher than the β wave.

Collecting EEG data and labels of a subject requires setting up an experimental environment. The basic experimental procedures contain online stimulus presentation, scalp EEG amplification, analog-to-digital conversion (in data acquisition), and offline data analysis (like ocular artifact reduction, epoch, filter, and so forth) after the end of the experiment [5]. However, this data acquisition and data labeling process needs a long time and high cost, and therefore requiring expensive human labor. It is impossible to collect and label almost all types of data from all subjects. EEG records are limited by scarcity of meaningful data labels.

On the other hand, traditional machine learning techniques are divided into two categories: unsupervised learning and supervised learning. Unsupervised learning uses only unlabelled data sets, while supervised learning uses a set of labeled data for learning. But in many practical problems, there is only a small amount of labelled data, because the cost of labelling the data is sometimes high [6]. As an example, biologists may spend many years labeling data on structural
analysis of a protein or functional identification [7]. However, a lot of unlabelled data is easy to get. This has led to the rapid development of semi-supervised learning techniques that simultaneously use a small amount of labeled data and a large amount of unlabelled data [8-10].

In the paper, EEG brain wave state classification with a graph-based semi-supervised learning is preliminarily studied. It uses only a small amount of labeled data of brain states for classification. For a subject, the amount of time for data acquisition and processing can be reduced, but brain state classification accuracy can also be kept relatively high in comparison to classification when more data is labeled. The first part of the paper presents the introduction. Background of the work is introduced in the second section. The proposed method and experimental results are shown in the third and fourth sections, respectively. A conclusion is provided in the final part.

**Background**

The Brain-Computer Interface (BCI) establishes direct communication and control between the human brain and computers or other electronic devices without relying on the normal output pathways of the brain (peripheral nerves and muscle tissue). BCI technology can help deaf patients to provide new information exchange channels, improve patients' quality of life, and has great practical value in the fields of medical field, cognitive science, psychology, military field, entertainment, and wearable smart equipment. In the classification of Electroencephalogram (EEG), traditional supervised learning requires the collection of a large number of labeled EEG data to train classifier, but the cost of obtaining a large number of labeled samples is expensive because multiple tests are performed. It takes a lot of time and effort, which is easy to cause fatigue in the subject and hinders the development of the BCI system. Moreover, the EEG state also changes due to its nonstationary characteristics, which makes the classification difficult. However, unlabeled EEG data is readily available. If this data is not used, a large amount of unlabeled EEG data will be wasted. For unsupervised learning, classifier is trained using unlabeled EEG data. However the lack of information on the marked EEG data decreases generalization ability of the model. Therefore, it is reasonable to apply semi-supervised learning to classification of EEG data. Semi-supervised learning only needs to collect a small number of labeled EEG data for training classifiers, and then it uses a large number of unlabeled EEG data to train classifier. It not only shorten training time of the collected samples of subjects and improve the classification performance, but also semi-supervised learning is also an adaptive process, which is able to promote the enhancement of BCI adaptability.

In traditional supervised learning, the learning system learns a large number of labelled data, and establishes a model for predicting the label of unknown data. Here, the label corresponds to the output of the data, and is used to represent the target concept to be learned. Semi-supervised learning can automatically use unlabelled data to learn models that have strong
generalization ability across the entire data distribution. The learning process without human intervention is based entirely on the learning machine itself to achieve the use of unlabelled data. The key to semi-supervised learning is the consistency assumption that the basic requirements of the classification function are sufficiently smooth for the internal structure revealed by a large number of labelled and unlabelled data. Zhou et al proposed a simple graph-based algorithm to get such a smooth solution to the internal structure collected by the known and unmarked points [10].

Proposed Methods

Due to insufficiency of labelled EEG data, we propose to apply the graph-based semi-supervised learning [10] for classifying brain states via EEG data. The brain wave learning machine is trained by both labelled and unlabelled EEG data as shown in Figure 1. EEG data is acquired from multiple channels and labelled with positive and negative tags. A partial set of labelled data is selected and combined with unlabelled data to construct a data set \(\{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l), (x_{l+1}, y_{l+1}), \ldots, (x_{l+u}, y_{l+u})\}\), where \(x\) is data and \(y\) is data labels. The \(l\) labelled data is only a small portion of the unlabelled data \((l << u-l)\). An affinity matrix is constructed based on \(x\) data and follow the steps proposed in the reference [10]. In the affinity matrix, every point (as a graph node) is similar to its local neighbourhood points. A smoothness constraint is added to the regularization framework with iterative solutions. Local and global consistency of EEG data is learned for predicting labels of unknown EEG data. The learning machine is then applied on testing data for classifying brain states, where data doesn’t have any label information. The wrongly classified instances are divided by total instances related to that specific brain state data to serve as an indicator of classification performance.

We use P300 data [11] for brain state classification. The specific brain waves for the P300 can be used as a marker to identify a place, object, or other detail that a person sees and identifies in daily life. Studies of EEG recorded using brain activity have shown that P300 brain waves become larger when a person identifies a meaningful item from a series of meaningless items. Using the P300, a hidden information test (CIT) can performed on the subjects to determine their identification of information related to crime or other events. Since distance-based classification needs to normalize features, all EEG data is normalized to the range \([0, 1]\) before training process. The P300 data was sampled with 1024 Hz frequency with 32 channels, it is also down sampled to 32 points for each channel. 7 subject’s EEG data were used for training model. Each subject has 6 brain states with observing 6 scenes in the data acquisition, respectively.
Figure 1. Demonstration of a small amount of labelled EEG data with a graph-based semi-supervised learning.

Results

A small amount of labelled data are selected to be combined with unlabelled data for training the graph-based semi-supervised learning model. The numbers of labelled data selected are 3, 6, 8, 10, 12, 15, 16, 20, 30, 50, and 60, respectively. They are a portion of the all labelled state for one subject with a specific brain state. For example, they are 3.33%, 5.56%, 6.67%, 8.89%, 11.11%, 13.33, 16.67%, 17.78%, 22.22%, 33.33%, 55.56%, and 66.67% of the total 90 labelled
instances for the subject 1 with one brain state. Two-class classification is accomplished between the target state and other 5 states for each subject.

As shown in the Figure 2, classification of three brain states for one subject is not accurate when the numbers of labelled instances selected are 3 and 5. But when the number of labelled instances is around 12, the error rate is reduced much from around 30% to 17%. When the number of labelled instances is increased further (e.g. 20, 30, 40, 50, and 60 labelled instances), error rate doesn’t have obvious improvement. In addition, classification results of one same brain state with three different subjects are listed in the Figure 3. It is seen that the classification performance across three subjects are not very different. The error rates can be reduced when 8 labelled instances are incorporated into the training model. For the 6th brain state, all 7 subjects have been selected with 3, 6, 10, 15, 20, and 30 labelled instances for classification performance comparison. As shown in the Table 1, error rates can be reduced from classification by using 3 labelled data to 30 labelled data for each subject. It is seen that if 10 labelled data (10% ~ 20% of the total labelled data for the 6th brain state across 7 subjects), error rates can be reduced below 20%. Furthermore, using more than 15 labelled data cannot improve error rates obviously.
Figure 2. Three brain state classification error rates for one subject with three different brain states.
Figure 3. One brain state classification error rates across three different subjects.

Greedy Max-Cut based semi-supervised learning [12] is also used for testing EEG data classification. As shown in the Figure 4, Greedy Max-Cut algorithm outperforms other two methods: Gaussian fields and harmonic functions (GFHF) method [13] and LGC method [14].
Figure 4. One brain state classification error rates using the numbers of labeled data: 6, 8, 10, 12, 14, 16, 18, and 20.

<table>
<thead>
<tr>
<th>S</th>
<th>L3</th>
<th>L6</th>
<th>L10</th>
<th>L15</th>
<th>L20</th>
<th>L30</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.33%</td>
<td>6.67%</td>
<td>11.11%</td>
<td>16.67%</td>
<td>22.2%</td>
<td>33.33%</td>
</tr>
<tr>
<td></td>
<td>29.05%</td>
<td>24.99%</td>
<td>19.85%</td>
<td>17.22%</td>
<td>21.0%</td>
<td>31.58%</td>
</tr>
<tr>
<td>2</td>
<td>3.16%</td>
<td>6.32%</td>
<td>10.53%</td>
<td>17.69%</td>
<td>21.0%</td>
<td>31.58%</td>
</tr>
<tr>
<td></td>
<td>33.22%</td>
<td>25.18%</td>
<td>17.72%</td>
<td>17.56%</td>
<td>16.6%</td>
<td>16.60%</td>
</tr>
<tr>
<td>3</td>
<td>3.23%</td>
<td>6.45%</td>
<td>10.75%</td>
<td>16.13%</td>
<td>21.5%</td>
<td>32.26%</td>
</tr>
</tbody>
</table>
Table 1. Classification error rates of 7 subjects (represented as “S”) with the 6th brain state is listed. For each row of a subject, the first row represents the percentage of labelled instances in the total labelled instances of that subject for the 6th brain state. The second row represents error rates.

Conclusion

A preliminary study on EEG based brain state classification via a small portion of labelled data is proposed. Graph-based semi-supervised learning is applied on EEG data for brain state classification. High classification accuracy can be achieved with a small portion of labelled data. More number of labelled data over a threshold cannot improve classification error rate obviously. Since noise of EEG data spreads label information to neighbours and deteriorates global stable state. Future work will focus on noise data removal under semi-supervised learning.

Reference


