Final Project

Properties of Back Propagation Algorithm and its Application to A EEG-Based Classification of Subclinical Depression

Yuzuki Ishikawa

Spring 22 PSYC 489

Professor John Hummel

University of Illinois at Urbana-Champaign

Abstract

Computer-aided diagnosis (CAD) is a promising machine learning model that classifies patients with various psychiatric disorders, such as depression. Recent studies have applied non-invasive electroencephalogram (EEG) signal-based CAD and shown potential to diagnose major depressive disorder (MDD). MDD is a common psychiatric disease characterized by anhedonia, rumination of negative thoughts, and other negative cognitive and behavioral symptoms. It is estimated that 280 million people around the world have MDD. Thus, there is a need for prevention, diagnosis, and treatments. However, few studies have attempted to implement CAD for the subclinical MDD population, which may play a significant role in the early detection and prevention of MDD. Here we applied a multi-perceptron to an EEG dataset of subclinical MDD patients identified by structured clinical interviews. In our model, we used a back propagation algorithm to update weights. We also explored hyperparameters to elucidate the behavior of the model and to improve accuracy. Our model showed high classification accuracy of 63.3% for low and 67.9% for high susceptibility to MDD, respectively. This study suggests the possibility of employing CAD in the subclinical MDD population to advance the early detection of the disorder.

I. Introduction

Psychiatric disorders are forms of abnormal mental state that affect social and economic welfare of people. Liberation from those disorders is a crucial issue because there is a dramatic increase in the number of patients. WHO predicted over 264 million people have some form of mental disorder around the world, which shows that mental problems influence around 27% of the general population at some point in their lives (Yasin et al., 2021). Among those psychiatric disorders, major depressive disorder (MDD) is the most common disorder characterized by anhedonia, rumination of negative thoughts, physical and mental fatigue, sleep disturbance, and/or suicidal thoughts (Kennedy, 2008). Although 3.8% of the general population around the world have depression reported by WHO, little is known about the pathological mechanism, and there are currently no solid biological data-based criteria for the diagnosis.

Recently, computer-aided diagnosis (CAD) has demonstrated promising results in the detection of MDD. CAD is a type of artificial intelligence (AI)-based categorical classification that learns existing data and provides a diagnostic label to patients. CAD can provide an automated diagnosis of MDD, which could facilitate the prevention and detection of depression outside of the clinical services. Various types of biomarkers have been used as inputs of AI, including blood pressure, glucose level, temperature, and neuroimaging data from functional magnetic resonance imaging (fMRI), electroencephalography (EEG), near-infrared spectroscopy (NIRS), and so on. According to Yasin et al. (2021), there are more than 35 studies on EEG-based detection of MDD using machine learning, and they reported high accuracies (80~99%).

While auto-diagnosis of MDD is essential for mental health innovation, machine learning-aided prevention of disorders is also crucial. It is expected that CAD can perform well in the early detection of MDD. However, only one study has focused on the detection of subclinical MDD using EEG (Ghiasi et al., 2021). Thus, this study aims to construct a multi perceptron-based classifier for subclinical depression. Algorithms and hyperparameters are significant factors that determine the performance and runtime of the model. Here we applied a back propagation algorithm for updating weights and explored hyperparameters to improve accuracy and understand the behavior of the model.

II. Materials and Methods

Participants and Clinical Assessment. 85 Thai participants (30 men and 55 women, aged between 13 and 22) took clinical tests for depression and were asked to record the resting state EEG. All participants did not meet the criteria for MDD after a careful clinical evaluation by the Structured Clinical Interview for DSM-5. The Beck Depression Inventory-II (BDI-II) and the Patient Health Questionnaire-9 (PHQ-9) were used to assess levels of depression severity. Then, the participants were divided into three groups: 30 Minimal (BDI-II 0~13 and PHQ-9 1~6), 27 mild (BDI-II 14~19 and PHQ-9 7~13), and 28 moderate (BDI-II 20~28 and PHQ-9 14~19). The minimal and moderate groups were used for the present study. Data used for this study are available at https://doi.org/10.1186/s13104-021-05673-x (Rachamanee & Wongupparaj, 2021).

EEG Data Acquisition and Preprocessing. The 64-channel resting-state EEG was recorded for 30 seconds with eyes closed. Data preprocessing includes 1-40 Hz offline filtering, baseline correction, and independent component analysis for the removal of ocular and muscle artifacts. The whole EEG data were segmented into 2-second non-overlapping Hanning-windowed epochs, smoothed using a fast Fourier transform, and averaged over five frequency bands: delta (1~4 Hz), theta (4~8 Hz), lower alpha (8~10 Hz), upper alpha (10~12 Hz), and beta (13~30 Hz). The absolute and relative EEG powers for five frequency bands were calculated.

Since Newson and Thiagarajan (2019) reported that the frontal theta power band represents depression, we used the theta absolute power obtained from 14 frontal electrodes. To increase the accuracy of the model, each data value was distributed to seven nodes using population coding Specifically, each of the seven nodes was assigned a value between 3 and 9. For example, if the input is 3.5, the inputs of the node representing 3 and the node representing 4 are 0.5, and the inputs of the other nodes are 0 for the other nodes (Figure 1). Thus, the number of nodes representing each participant was 14 channels $\times$ 7 nodes = 98.

Network Architecture and Back Propagation Algorithm. Multi-perceptron is a neural network wherein each node in each layer is fully connected to every node in the previous and following layers. Our model consists of an input layer with 98 nodes, a hidden layer with 50 nodes, and an output layer with two nodes. The net input of each node is a summation of activation * weight (1), and the net input is converted to the activation using the sigmoid function (2). One node with activation 1.0 was added to each layer except the last layer, representing a bias term.

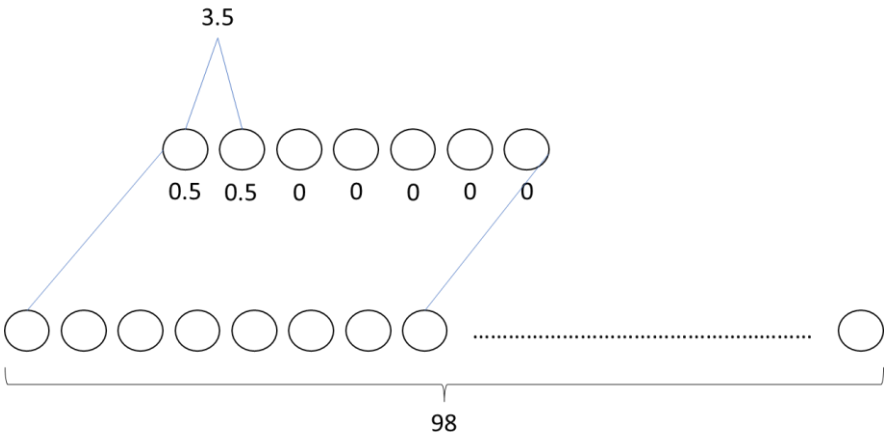

Figure 1. Population coding. Each theta band power was distributed to seven nodes that represent discrete values.

$$n_i = \sum w_{ij} a_j + b \tag{1}$$

$$a_i = \frac{1}{1 + e^{-n_i}} \tag{2}$$

Since sigmoid function converts data to the probability, outputs of the last layer can be considered probabilities of being minimal subclinical depression (low susceptibility) and moderate subclinical depression (high susceptibility). Thus, the model estimate is argmax of the two output nodes (3).

$$Diagnostic\ Label = argmax(a_i\ for\ i\ in\ nodes\ in\ the\ last\ layer) \tag{3}$$

The model continues to forward-propagate activations and back-propagate errors until the global error goes below 0.01, or the number of iterations goes beyond 1000. The raw error of the last layer is equivalent to (desired activation – activation obtained from the forward propagation), and the corrected error was calculated by (4).

$$e_i = raw\ error \times a_i(1 - a_i) \tag{4}$$

The raw error of the nodes in the following layers is calculated by (5).

$$raw\ e_j = \sum w_{ij} e_i \tag{5}$$

On each iteration, the model continuously changes its weights using (6) and (7), where $\eta =$ learning rate and $\varepsilon =$ momentum.

$$\Delta w_{ij} = \eta e_i a_j \tag{5}$$

$$\Delta w_{ij}^{n+1} = \varepsilon e_i a_j + \Delta w_{ij} \tag{6}$$

The global error (7) was calculated based on values of raw errors of the output layer to evaluate convergence. m: number of nodes in the output layer, n: number of nodes in the input layer.

$$E = \frac{\frac{\sum (d_i - a_i)^2}{m}}{n} \tag{5}$$

Initial weights are randomized, ranging from -1 to 1.

Hyperparameter Validation, Training, and Testing. We used different hyperparameters (momentum and learning rate) and recorded changes in the global error to find the most efficient values (yielding the highest accuracy and shortest computing time). Then we ran the model using 5-fold cross-validation with determined hyperparameters. The overall accuracy was the average of five sets.

III. Result

Hyperparameter Validation. We tested three momentum (0.1, 0.9, 0.99) and learning rate (0.001, 0.01, 0.1) values and recorded the global error on each iteration (Figure 2). We confirmed that the model took a longer time to converge as both learning rate and momentum decreased. However, the model did not converge within 1000 iterations when we used momentum = 0.99, which implied that too large value of momentum could delay learning. Notably, every condition provided the same output labels. Based on these results, we applied momentum = 0.90 and learning rate = 0.1 for training and testing.

Training and Testing. Data were divided into 5 folds. The model was trained using four folds and tested using the remaining folds. Table 1 shows the accuracy of our model. The overall accuracy reached 63.3% for minimal and 67.9% for moderate subclinical depression.

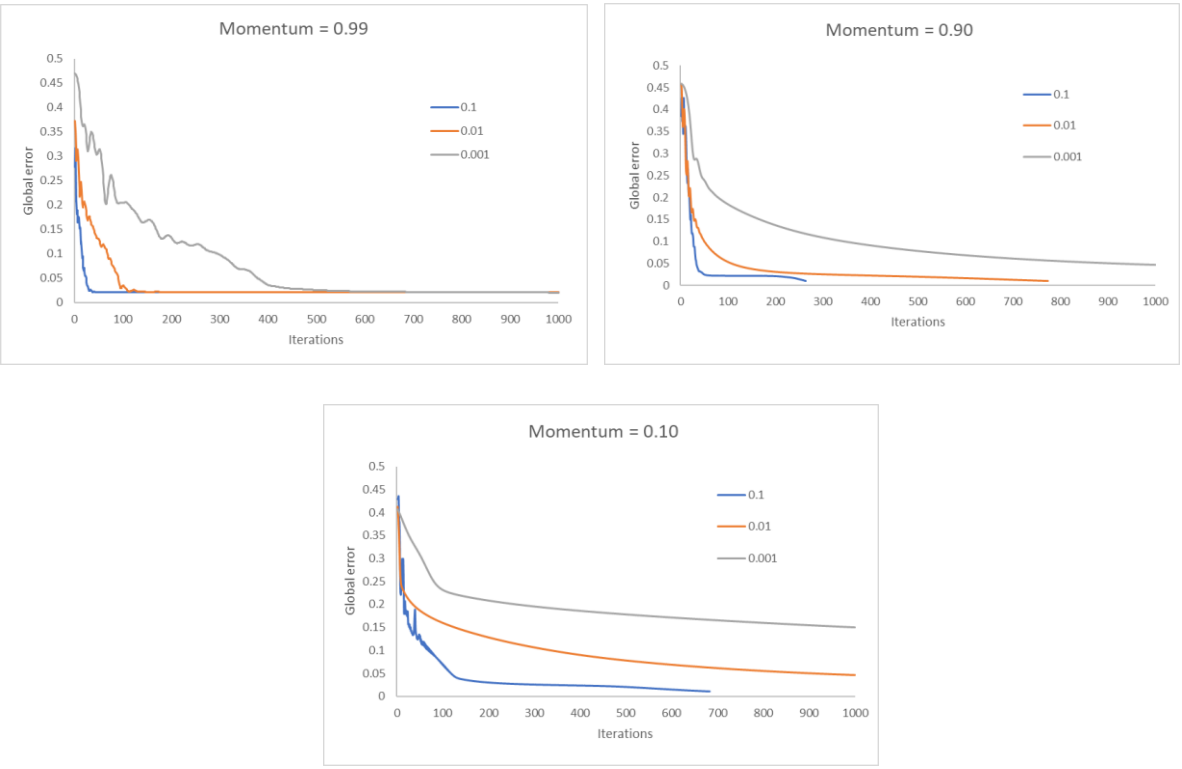

Figure 2. The global error was recorded under different values of momentum and learning rate. Different colors of the line correspond to different values of learning rate (0.1, 0.01, 0.001). Each graph shows results using different values of momentum values (0.99, 0.90, 0.10).

Table 1. Accuracy of each training/test set and overall accuracy (%). Minimal: Participants who scored 0~13 in BDI-II and 1~6 in PHQ-9. Moderate: Participants who scored 20~28 in BDI-II and 14~19 in PHQ-9.

set	Minimal	Moderate
1.0	66.6	60.0
2.0	50.0	66.6
3.0	33.3	66.6
4.0	83.3	50.0
5.0	83.3	100.0
Overall	63.3	67.9

IV. Discussion

Here we constructed a three-layer multi perceptron with a back propagation algorithm and obtained classification accuracy of 63.3% for low and 67.9% for high susceptibility to MDD, respectively. First, we confirmed a tendency that as the leaning rate increased, the model takes less time to converge. We also confirmed that the momentum value of 0.9 would be best for

our model. This result is in line with fact that many back propagation-based deep learning packages set the value of momentum as 0.9 (Wolfe, 2021).

Secondary, there are several hyperparameters we should consider other than momentum and learning rate: The number of nodes in the hidden layer, the depth of the network, and the settling criterion. Although there is no method to determine the size and depth of the network, greater depth seems to result in better classification accuracy (Goodfellow et al., 2016). However, it takes more time to compute as the network becomes bigger. Thus, there exists a trade-off between computation time and accuracy. The settling criterion is significant if there is a risk to be stuck in the local global error minimum because the model could provide underfitting in such a situation. Our model took more than 1000 iterations (maximum number of iterations) in most cases, which suggests that a higher value for the settling criterion would yield the same results with a shorter computation time.

Finally, the activation function also largely affects the results. The activation function in the back propagation algorithm must be differentiable because the derivative of the activation function is used for updating weights. It should also be nonlinear because multiple layers with a linear activation function should be equal to one layer, which yields trivial results. According to Cao et al. (2018), the rectified linear unit (ReLU) has shown superior performance to the sigmoid function and is currently the most popular activation function in deep learning. Thus, we should consider the application of ReLU for further studies.

References

Cao, C., et al. (2018). Deep Learning and Its Applications in Biomedicine. Genomics Proteomics Bioinformatics, 16(1), 17-31.

DOI: https://dx.doi.org/10.1016%2Fj.gpb.2017.07.003

Ghiasi, S., et al. (2021). Classifying Subclinical Depression Using EEG Spectral and Connectivity Measures. Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2050-2053. DOI: https://doi.org/10.1109/embc46164.2021.9630044

Goodfellow, I., et al. (2016) Deep Learning (Adaptive Computation and Machine Learning Series). The MIT Press.

Kennedy, S. (2008). Core Symptoms of Major Depressive Disorder: Relevance to Diagnosis and Treatment. Dialogues in Clinical Neuroscience, 10(3), 271-277. DOI: https://dx.doi.org/10.31887%2FDCNS.2008.10.3%2Fshkennedy

Newson, J. & Thiagarajan, T. (2019) EEG Frequency Bands in Psychiatric Disorders: A Review of Resting State Studies. Frontiers in Human Neuroscience, 09.

DOI: https://doi.org/10.3389/fnhum.2018.00521

Rachamanee, S. & Wongpparaj, P. (2021). Resting-state EEG Datasets of Adolescents with Mild, Minimal, and Moderate Depression. BMC Research Notes, 14, 256. DOI: https://doi.org/10.1186/s13104-021-05673-x

Wolfe, C. (2021). Why 0.9? Towards Better Momentum Strategies in Deep Learning. Available at: https://towardsdatascience.com/why-0-9-towards-better-momentum-strategies-in-deep-learning-827408503650

World Health Organization. Depression. Available at: https://www.who.int/news-room/fact-sheets/detail/depression

Yasin, S., et al. (2021). EEG Based Major Depressive Disorder and Bipolar Disorder Detection Using Neural Networks: A Review. Computer Methods and Programs in Biomedicine. DOI: https://doi.org/10.1016/j.cmpb.2021.106007