Subject Identification Using an EEG-based Biometric System

Steven Cao

ABSTRACT

Electroencephalography (EEG) signals present major advantages over other biometrics modalities (i.e. iris, face, fingerprint), such as being resilient to physical injuries, extremely hard to reproduce, and cannot be furtively captured at a distance. However, the field of EEG-based biometrics is new, and there are still many challenges that need to be solved for its real-world application, including improvement of accuracy, stability, and robustness.

The objective of this project is to create a machine learning (ML) pipeline to examine the feasibility of EEG-based subject identification for long periods of time. For better practical use, the ML pipeline we built here focused on the tuning of different modified ML models, including K-Mean, Naïve Bayes, 4-layer neural network, XGBoost, ResNet, Inception and EEGNet. We also enhanced the EEG signals based on their statistical information and showed that data preprocessing is critical in tuning different ML models.

Results show that, compared to XGBoost, the most used model in literature, which had an accuracy of 57.82%, our version of EEGNet model achieved the best result with an accuracy of 86.74%, and Inception ranked the second with an accuracy of 70.18%.

INTRODUCTION

Traditional access control approaches require individuals to remember or possess some information or item that must be presented to the access system. However, there are two main problems: information can be forged or stolen, and information can be hard to remember.

Biometrics is an alternative to other access methods. Biometrics captures the biological characteristics of the subject for identification and classification using a pre-built database. Electroencephalography (EEG) signals present major advantages over other biometrics modalities (i.e. iris, face, fingerprint) – resilient to physical injuries, extremely hard to reproduce, and cannot be furtively captured at a distance.

However, the field of EEG-based biometrics is new and there are many problems that will have to be solved: improvement of accuracy, stability of performance, and robustness.

BACKGROUND RESEARCH

An effective person identification system must be able to recognize subjects, even when they return days, weeks, or years later.

However: 1) Few studies have examined the stability of their systems over long spans of time. (Chan, Kuo, Cheng, and Chen, 2018). 2) Some show that the error rate performance of an identification system could increase within days (Marcel and Millan, 2007; Hu et al., 2011). 3) Others demonstrate a drop in performance when the training and test sessions of EEG data get further apart (Pozo-Banos, etc., 2014).

MATERIALS AND METHODS

Materials

I used the BED: Biometric EEG Dataset and data from Hero Lab. We used the consumer-grade Emotiv EPOC+ headset at a sampling rate of 256 Hz for EEG acquisition.
Our data contains 21 subject recordings, and each subject recording contains 14 channels: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4. In addition, our dataset contains three chronologically disjointed sessions one week apart.

For the software, I used PyCharm IDE for my coding and used Anaconda: Python 3.8 to set up my virtual environment. My project required the use of several Python modules, which are listed below: Torch, Pandas, Numpy, Tensorflow, Tensorboard, Keras, Sklearn, Scipy, Seaborn, Pyprep, Mne, XGBoost, Lazy Predict, Librosa.

Methods

The first two sessions of data were used for training and the third session was used for testing.

I used two methods for artifact removal, which I designed and implemented: Central Limit Theorem (CLT)-based data enhancement and Epoch-based Least-Squares Regression (EBLSR).

For the Central Limit Theorem (CLT)-based data enhancement, good segments of EEG signals are selected and split into epochs, then statistical information based on STD (standard deviation) of epochs is computed, and finally, a threshold is determined based on Central Limit Theorem to clean the data.
For the Epoch-based Least-Squares Regression (EBLSR), we first calculate the least-squares regression line as the trend for each epoch. Then we remove the trend for each epoch.

Each EEG subject recording is first split into epochs. Then each epoch will go to the Epoch-based Least-Squares Regression (EBLSR) to get its trend removed. After that, then it will go to the Central Limit Theorem (CLT)-based data enhancement, where its std will be calculated and compared to the threshold, which if the std is greater than the threshold, then it will be

identified as a bad epoch and if the std is less than the threshold, then it will be identified as a good epoch. This is done for each subject recording.

Then based on the result, we will also remove the bad subjects.

Then the preprocessed data get trained by our seven modified ML models: K-Mean, Naive Bayesian, XGBoost, 4-layer neural network, ResNet, Inception, and EEGNet. These models will be evaluated using the confusion matrix by evaluating the accuracy, precision, recall, and f1 score of each model. We also run a case-by-case investigation to examine the model performance.

ANALYSIS AND RESULTS

Results from the data preprocessing show that both of my implemented methods successfully removed artifacts from the EEG signals and enhanced their quality. Results from model training show that EEGNet achieved the best accuracy with 86.74%, followed by Inception with 70.18%, then ResNet with 63.21%, while XGBoost came in fourth with 57.82%.

The high accuracy from EEGNet of 86.74% shows a promising future for EEG-based subject identification over long periods of time.

Our model is deep learning-based, so the layer structure of deep learning models allows an optimal prediction-oriented feature representation of input EEG signals, which could outperform the manually engineered features used in XGBoost. The accuracy of the EEGNet model is 50.02% higher than the traditional best model from XGBoost, which only has an accuracy of 57.82%.

Data preprocessing is critical in deep learning model solutions. We showed that without proper data cleaning or enhancement, the accuracy could drop from 86.74% to 47.69%, a drop of 39.05%, for the EEGNet model. Thus proper data preprocessing for EEG signals is very important to achieve optimal performance for deep-learning-based solutions.

REFERENCES CITED

1. https://ieeexplore.ieee.org/document/9361690
2. https://www.hindawi.com/journals/cin/2018/5483921/
3. https://www.frontiersin.org/articles/10.3389/fninf.2018.00066/full
4. https://arxiv.org/abs/1909.04939
5. https://arxiv.org/pdf/1512.03385v1.pdf
6. https://ieeexplore.ieee.org/document/6313536
7. https://arxiv.org/abs/2106.03253
8. https://arxiv.org/abs/1611.08024#:~:text=version%2C%20v4
9. https://www.sciencedirect.com/science/article/abs/pii/S0957417414002930?via%3Dihub
10. https://arxiv.org/abs/1611.06455
11. https://braininformatics.springeropen.com/articles/10.1186/s40708-021-00142-4