Enhancing sEMG Classification with Data-Centric Techniques

GAGE TYLEE

This project implements a classification model to determine the maximum volunatry contraction percentage (%MVC) of surface electromyography (sEMG) signals. The dataset used was collected from the biceps brachii of 12 patients at 10%, 30%, and 50% of maximum voluntary contraction. User metadata is explored through training data valuation techniques to gain insights into how an individual's characteristics impact model output. The results are used in discussing data-centric approachs to EMG collection.

A GitHub repository for this project can be found here: https://github.com/gagetylee/mvc-classification

1 INTRODUCTION

Recent breakthroughs in AI have significantly enhanced the capacity to extract meaningful insights from biomedical data. In particular, electromyography (EMG) signals hold great promise for health science domains as they contain valuable information about muscular activation patterns. Maximum Voluntary Contraction (MVC) measures the peak force a muscle can generate. This project classifies surface EMG signals into MVC percentage categories. Applications of such models could possibly extend to areas such rehabilitation, exercise optimization, prosthetic development, and more.

Machine learning and deep-learning models are found to be increasingly successful in classifying EMG signals. However both the quantity and quality of current datasets are limited by a lack of conformity to data-centric approaches. Overcoming this barrier involves developing a tailored data-collection pipeline that collects relevant subject metadata. The primary goal of this project is to utilize data valuation methods on the MVC classification model to better understand the impact of subject metadata, paving the way for more standardized pipelines.

2 RELATED WORK

The dataset used in this project was used to create a MATLAB application to better visualize HD-sEMG signals, and can be found at https://github.com/lyanet-upc/hd-emg-app [1]. This experimentation involved monitoring four tasks at varying degrees of maximum voluntary contraction. Data was collected from the biceps, triceps, and forearm. This project only uses the bicep data. HD-sEMG is effective in extracting both spatial and temporal information due to it's grid-like shape and high number of channels. Research done at Queens University explored channel selection methods such as Correlation Ratio Maximization to aid in reducing dimensionality [4]. Another study at Queens used HD-sEMG to estimate force with artificial neural networks (ANN's) [2]. The preprocessing steps in this project were kept mostly consistent, given the similarities in data used. The classification of %MVC has been effectively achieved with the use of synthetic sEMG data, as demonstrated in Hickman et al. (2014) [6]. Data valuation in this project involves retraining approaches, which were derived from Hammoudeh et al [5].

3 METHODS

3.1 Preprocessing

EMG data is highly susceptible to noise. By applying preprocessing steps, we reshape the data to be more effective for modeling. This involves several steps. Given the large amount of channels in HD-sEMG we first must determine how to select the most relevent of these channels. This is done using the Power-Correlation Ratio Maximization method [4].

Author's address: Gage Tylee.

This finds the channels with maximum output and with minimal correlation. This approach ensures the inclusion of a broader array of channels, while prioritizing those that provide the most significant information.

A bandpass filter was applied to the data with a 10 Hz low-pass and 500 Hz high-pass. This helps in eliminating movement artifacts and high frequency noise. Afterwards, the signal are rectified and smoothed. Rectification transforms all negative values into positive ones, enabling a more accurate representation of muscle contraction magnitude. A 300-point moving average filter is applied to further smooth the signals and minimize fluctuations. The resulting linear envelope represents the original signal trends in amplitude while omitting finer details [3]. This simplified representation provides a clearer understanding of muscle contraction strength.

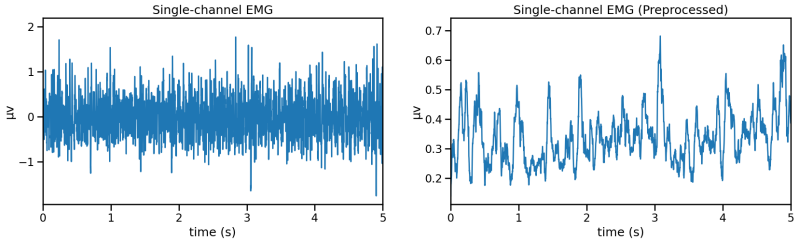

Fig. 1. Preprocessing transformation

3.2 Feature Extraction

Three features were selected from the time-domain and one from the frequency domain. The time-domain features include the maximum value, standard deviation, and mean. The mean frequency is extracted from the filtered data, before smoothing and rectification is applied. In addition to the calculated features, patient metadata was used. This includes the weight(kg), height(cm), and age. These features were used for each of the 10 selected channels of each recorded EMG signal. This resulted in 1440 rows of feature data.

3.3 Model Implementation

The model was created using the Random Forest algorithm. Rather than using regression-based approach that predicts a continuous output, the signals are classified into distinct categories. Since the dataset used in this project only had labeled signals for 10, 30, and 50% MVC, these are the target categories. In future iterations that include more data or synthetically generated data, the idea would be that each signal gets classified into bins ranging from 10-20%, 21-30%, and so on as seen in figure 2.

% of MVC	10 to 20	21 to 30	31 to 40	41 to 50	51 to 60	61 to 70	71 to 80	81 to 90	91 to 100
Output Class	1	2	3	4	5	6	7	8	9

Fig. 2. Classification outputs for a more complete dataset

4 RESULTS

The model was trained on 80% of the data using the Random Forest algorithm. This resulted in 91% accuracy for the test set, and the following confusion matrix was generated.

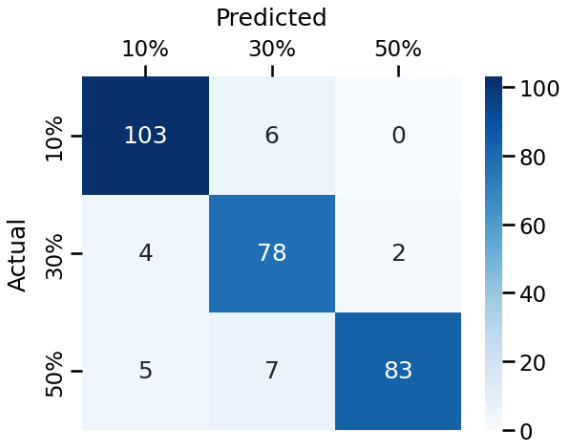

Fig. 3. Confusion Matrix of Test Set

To gain a better understanding into how subject metadata influences model effectiveness, the model was re-trained using various subsets of features. The initial experiment used the derived set of features from the EMG signals resulting in an accuracy of 86%. This demonstrated a 5% drop in performance when metadata was disregarded.

Afterwards, subject characteristics were reintroduced one-by-one, starting with weight, followed by height, and finally age. The ensuing results showed accuracy levels of 91%, 93%, and 92%, respectively. This trend suggests that optimal model performance is attained when the training set incorporates a combination of biometric characteristics (in this case height and weight) in conjunction with the EMG features.

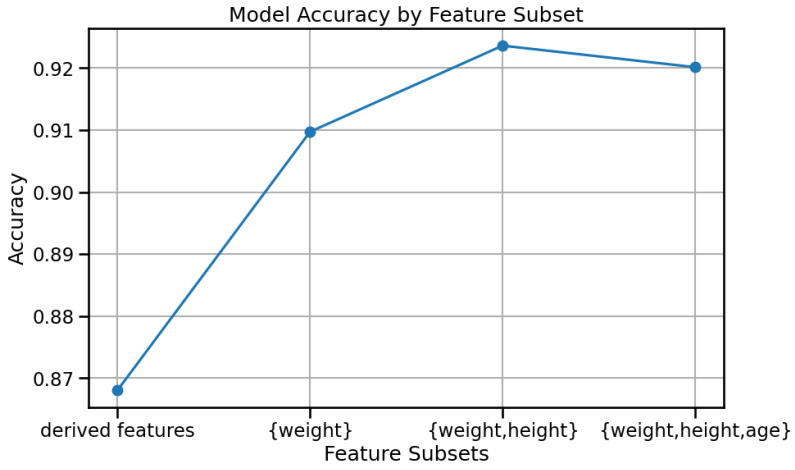

Fig. 4. Preprocessing transformation

5 DISCUSSION

Historically, most EMG studies tend to neglect subject metadata, as analysis is primarily focused on the raw signals. The strength of AI however, lies in its ability to infer relationships with large pools of information. Adopting a more data-centric approach requires a few steps.

5.1 Training Data Valuation

In this experiment, results of the training data valuation showed that incorporating patient metadata increases the model accuracy. Access to additional data would allow for a more extensive review of this claim, but it still highlights the importance of adopting a more robust framework for EMG collection that can store this extra information. Access to additional data would enable a more thorough evaluation of this claim. However, this still underlines the need for a more comprehensive EMG collection framework capable of preserving this additional information. While it didn't end up getting used in the project code, Shapley Values (SV) would be effective in measuring the contribution of each feature for the model [5]. This involves retraining the model using the power set of the features.

As AI continues to gain prominence, there will be increased demand for abundances of high-quality data.

5.2 Data-Centric EMG Collection

To design such a system for EMG data, we can refer to the training data development tasks outlined in the Zha et al. survey literature.

The goal of training data development is to collect and produce rich and high quality training data to support the training of machine learning models [7]. With an ever growing size of available data, it's essential that automated algorithms be developed to streamline the collection process. This facilitates consistency across a multitude of sessions, and reduces likelihood of human error. This could come in the form of a software application which prompts for physical attributes, such as the subjects age, height, weight, as well as the exercise they are performing.

This streamlined system would be beneficial to achieve the collection sub-goal, but it also adds increased time to an already lengthy process. Most forms of EMG collection requires a group of voluntary participants to perform various tasks in a controlled environment, which does not always produce enough results to sufficiently train a model. In such cases, data augmentation should be considered to synthesize additional EMG signals. One approach commonly used for time-series data like EMG, is by using a Generative Adversarial Network (GAN) [7]. GAN's are leveraged to produce new data based on the statistical property of previously seen data and randomly added noise.

REFERENCES

[1] [n. d.]. dataset. https://github.com/lyanet-upc/hd-emg-app

[2] [n. d.]. EMG-based Force Estimation using Artificial Neural Networks. 42, 18 ([n. d.]). https://proceedings.cmbes.ca/index.php/proceedings/article/view/876

[3] Ouriel Barzilay and Alon Wolf. [n. d.]. A fast implementation for EMG signal linear envelope computation. 21, 4 ([n. d.]), 678–682. https://doi.org/10.1016/j.jelekin.2011.04.004

[4] Gelareh Hajian, Ali Etemad, and Evelyn Morin. [n. d.]. Automated Channel Selection in High-Density sEMG for Improved Force Estimation. 20, 17 ([n. d.]), 4858. https://doi.org/10.3390/s20174858 Number: 17 Publisher: Multidisciplinary Digital Publishing Institute.

[5] Zayd Hammoudeh and Daniel Lowd. [n. d.]. Training Data Influence Analysis and Estimation: A Survey. arXiv:2212.04612 [cs] http://arxiv.org/abs/2212.04612

[6] Stephen Hickman, Rocio Alba-Flores, and Mohammad Ahad. [n. d.]. EMG based classification of percentage of maximum voluntary contraction using artificial neural networks. In *2014 IEEE Dallas Circuits and Systems Conference (DCAS)* (Richardson, TX, USA, 2014-10). IEEE, 1–4. https://doi.org/10.1109/DCAS.2014.6965337

[7] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. [n. d.]. Data-centric Artificial Intelligence: A Survey. arXiv:2303.10158 [cs] http://arxiv.org/abs/2303.10158