Project report
-
Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers

BERGER Clément - DESBLANCS Dorian
Binôme 33

March 2021

Abstract

In this project we studied the work done by John Atkinson and Daniel Campos in their article *Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers* [1]. Their paper presents a method to recognize emotions using EEG signals. More precisely, they aim to classify arousal and valence. Their method is split into two sections. The first is centered around feature extraction, and makes use of the minimum-Redundancy-Maximum-Relevance (mRMR) technique [7]. The second is centered around emotional classification, and makes use of a Support Vector Machine (SVM) classifier. While the paper demonstrates the method to be competitive against other state-of-the-art models, our implementation of their pipeline failed to generate the same results. Most notably, they achieved more than 60% accuracy on the DEAP data set [3] for each emotion using 8-fold classification. We were unable to break 50%.

1 Introduction

Emotion classification plays a crucial role in the field of human-computer interaction (HCI). While some emotional changes can be successfully recognized based on electroencephalogram (EEG) signals, most automatic methods to do so are restricted to only a few classes. *Brain-computer interfaces* (BCIs) are communication systems based on sensors capable of tracking brain activity. They are designed such that the user can control computers using only their brain activity. Over the past few decades, the deciphering of EEG signals has played a key role for such tasks.

In their paper *Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers*, Atkinson and Campos proposed a full pipeline for emotion recognition for EEG-based BCIs. Their goal was to extend the range of emotion classes that can automatically be recognized. Here an emotion is represented as a combination of two dimensions called Arousal and Valence. The former refers to levels of emotional activation, and usually serves to scale one's calmness or excitement. The latter refers to positive and negative emotional intensity, and usually serves to scale one's happiness or sadness. Their method is centered around a feature extraction and selection step, followed by an emotional classification step that leverages the use of Support Vector Machines (SVM). Their methodology was tested on the DEAP data set [3]. This data set regroups the EEG measurements of 32 patients across 40 trials using a 32-channel EEG-recording system. For each trial, patients were asked to rate their levels of arousal and valence between 1 and 9. In this report, we will first explore some of the work related to this article. We will then explore the methodology presented in Atkinson and Campos' paper. Finally, the results we obtained on the DEAP data set using their methodology will be presented. Our most important finding is that we were unable to reproduce their results.

2 Related work

As mentioned previously, the task of automatic emotional recognition has been popular for a long time. This is especially true since the publication of the DEAP data set [3] in 2012. This section will present some of the key methods that have been used in the field.

Before Atkinson and Campos published their paper in 2016, emotional classification methods were centered around astute pre-processing and feature selection steps. This is notably the case in Huang et al.'s 2008 work [9]. They proposed a method for emotional classification based on spectral density features, extracted from a variety of EEG frequency bands. From there, other metrics such as Hjorth parameters [9], Logarithmic Band Power (Log BP) [2], and time-frequency transformations (most notably the wavelet transform [6]) were proposed for successful emotional classification. These metrics clearly influenced the methodology proposed by Atkinson and Campos.

Since 2016, however, deep neural networks have taken over the emotion recognition problem. The use of multimodal networks was most notably successfully leveraged in the the following two works [10] [4]. The pre-processing and feature selection steps conducted are however quite limited, and most often centered around splitting the input EEG signals in frequency bands. The results obtained in these two works are state-of-the-art on the DEAP data set.

3 Feature Extraction and Selection

3.1 EEG Signal Pre-Processing

Before extracting any feature, the signals provided needed to be transformed using common EEG pre-processing techniques. These included downsampling the data to 128 Hz and removing electrooculography (EOG) artifacts. These steps were already performed on DEAP data we were given access to.

3.2 Feature Extraction

We then extracted the features listed in Atkinson and Campos' paper. We notably extracted statistical features, band power for different frequencies, Hjorth parameters, and fractal dimension for each channel. In their work, Atkinson and Campos only mention the use of the signals' means, standard deviations, and kurtosis as statistical features. They however suggest that many more were extracted without explicitly mentioning them. Inspired by work done in our Time Series course, we took the liberty to also extract each signal's maximum, minimum, skew, and 25th, 50th, and 75th percentiles. Note that we ended up only using our signals' means, standard deviations, and kurtosis in our final notebook as the other features were found to lead to worst performances. We then extracted the band power of the following frequency bands: *theta* (4-8 Hz), *low alpha* (8-10 Hz), *alpha* (8-12 Hz), *beta* (12-30 Hz), and *gamma* (30-45 Hz).

Once all the features were extracted, we discretized the data according to the following formula:

$$f(x) = \begin{cases} 1 & \text{if } x \geq \mu + \frac{\sigma}{2} \\ 0 & \text{if } \mu - \frac{\sigma}{2} \leq x < \mu + \frac{\sigma}{2} \\ -1 & \text{if } x < \mu - \frac{\sigma}{2} \end{cases} \tag{1}$$

Note that μ is the mean of a subject's feature values, and σ its standard deviation.

We extracted the features from all 32 channels. However, in their paper, Atkinson and Campos mention conducting their experiments on 14 relevant channels. They did not indicate how these channels were selected for relevance. We therefore tried to use the 14 relevant channels listed in this paper [8] for our feature computation. These were selected using GA-SVM, and are most likely also those selected by Atkinson and Campos (both papers use the DEAP data set).

Finally, our labels were split into three classes for each emotion. Arousal and valence ratings between 1 and 3.66 were given a label of 0, ratings between 3.66 and 6.33 were given a 1, and the rest were given a value of 2.

3.3 Feature Selection

Once all the features were computed and discretized, we used the minimum-Redundancy-Maximum-Relevance (mRMR) [7] algorithm to select a subset of features for classification. The mRMR algorithm selects the features that correlate the strongest with the classification variable. In our case, either valence or arousal labels of 0, 1, or 2. The mRMR method for feature selection reduces the information redundancy between bad and good information. It does so using mutual information, which is defined for two random variables X and Y as:

$$I(X, Y) = H(Y) - H(Y|X) \tag{2}$$

($H(Y)$ corresponds to Y's entropy; $H(Y|X)$ is known as the conditional entropy between both variables). Mutual information allows us to quantify how much information we have about X using just Y (and vice-versa). As such, by optimizing the following, known as the minimum redundancy condition:

$$\min W_1, W_1 = \frac{1}{|S|^2} \sum_{f_i, f_j \in S} I(f_i, f_j) \tag{3}$$

we can create a feature set whose elements are as uncorrelated as possible. Note that f_i and f_j are the features in the set S, and $|S| = n$ is the number of features from the set.

On the other hand, the maximum relevance condition implies that the features selected must be as correlated as possible with the class labels. Mathematically, it is expressed as:

$$\max V_1, V_1 = \frac{1}{|S|} \sum_{f_i \in S} I(C, f_i) \tag{4}$$

where C denotes the class label.

By optimizing both conditions, the mRMR algorithm is capable of selecting features that are optimal for future classification. The number of features selected is controlled by the user.

4 Emotions Classification

4.1 Methodology

Once useful features were adequately selected, we had to train a classifier to classify our labels. As in the paper, we used a *Support Vector Machine* (SVM). We denote our data by $(x_i)_{1 \leq i \leq n}$ where x_i is a vector of features, and our labels by $(y_i)_{1 \leq i \leq n}$. We train separate classifiers for arousal and valence. Hence, let us consider y_i as a one dimensional vector, taking three different values. For binary classification, an SVM aims to find a hyperplane separating $\mathbb{R}^N$ - where N is the feature dimension - into two sets, one for each value. To classify between k different classes (here $k = 3$), $k(k-1)/2$ different SVMs are trained, one for each pair of classes. This is the principle of a *One-Versus-One* mechanism. The global classification is based on a voting mechanism, where we take the class that maximizes the number of correct classifications among the different SVMs.

While in a linear SVM, the classification is made by considering the dot product between the normal vector of the hyperplane and the feature vector, we can achieve non linear classification by replacing the dot product by a function called the kernel function $k(.,.)$. Here *radial basis functions* (RBF) are used. This family of functions is defined by:

$$\Big(k(x_1, x_2) \mapsto \exp(-\gamma ||x_1 - x_2||^2)\Big)_{\gamma > 0} \tag{5}$$

We varied the γ parameter. The results of these variations will be displayed in the Results section.

4.2 Testing Procedure

As the dataset is quite small, *k-fold cross-validation* was used to both test and train our models. The full dataset was divided into k different folds. Then, we trained k different models. These were each trained on $k-1$ different folds of data and tested on the remaining fold. We then calculated the mean and standard deviation of the accuracy obtained for each model.

Note that before training our models, we randomized our dataset. We used multiple seeds for each one of our trials. This allowed us to avoid the possibility of biased training or testing sets, and hence biased results. We used a value of $k = 8$, as did Atkinson and Campos in their original paper.

5 Results

As explained in Section 3.2, the Atkinson and Campos' paper did not indicate the channels that were selected for training. Therefore, we first attempted to use all 32 channels, and let our selection process select a good subset of training features for each emotion.

Concerning our SVM classifier, we began with $\gamma = 0.05$, as recommended by the paper. We started with 75 selected features and report both the total accuracy and per-patient accuracy in Figure 1. Surprisingly, the average accuracy did not exceed 45% for both Arousal and Valence. The variance between test set results was also found to be quite large. This is even more striking when we look at the accuracy computed for each patient. While some of the patients' results are comparable to those found in Atkinson and Campos' paper, for others, the model is no better than a random baseline. There is even a patient for whom the variance is bigger than the actual mean.

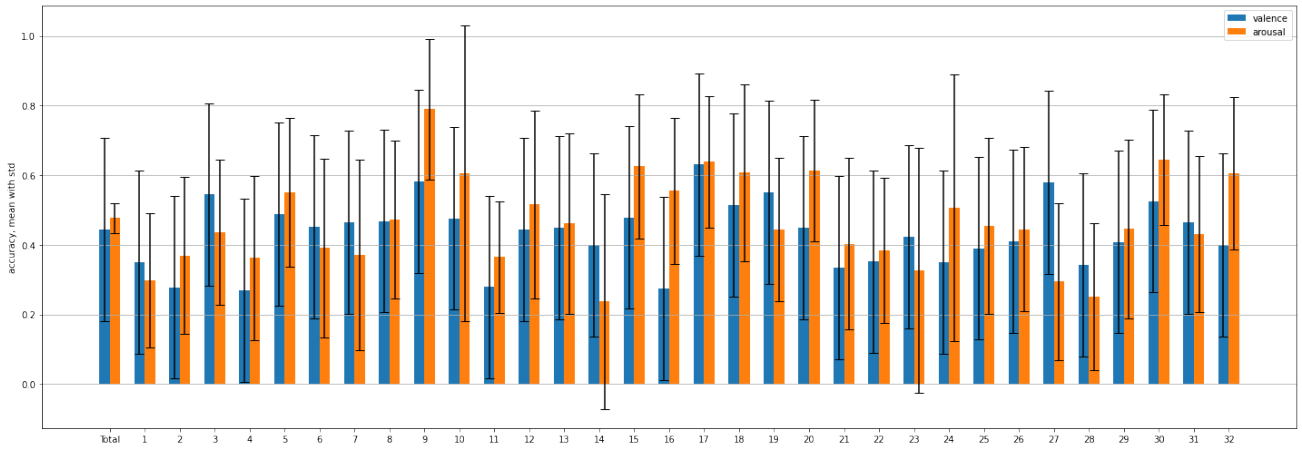

Figure 1: Mean and standard deviation of the total and per-patient accuracy for both valence and arousal. Here we consider 75 features from the features computed on all 32 channels. We set $\gamma = 0.05$.

We then worked on the number of selected features, since the paper reports a wide array of results depending on the size of the feature subset selected. We varied the number of selected features from 3 to 179. The results are reported in Figure 2. The best accuracy found for Arousal is 48.5%, using 85 features. For Valence, the best accuracy found is 46.3%, using 10 features. Our first observation still holds: the results are very different compared to those reported by the original paper (more than 60%). We also observe that the number of features required for Arousal is far greater for us than for Atkinson and Campos, who achieved their best results using merely 35 features. The opposite holds for Valence.

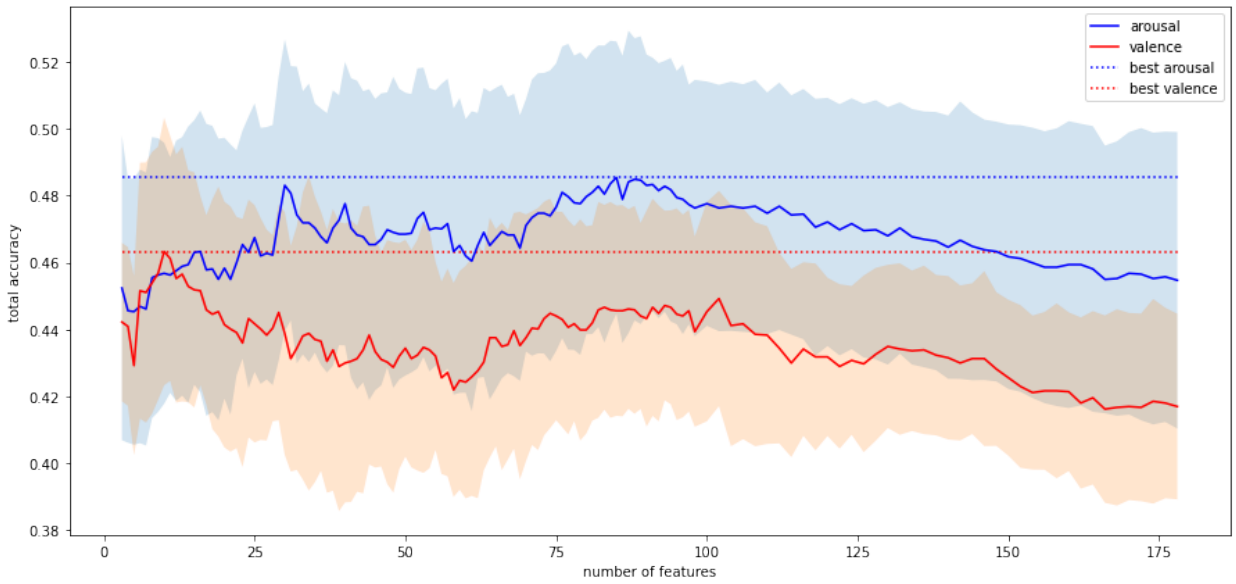

Figure 2: Total accuracy for different feature subset sizes. Here we consider all 32 channels and $\gamma = 0.05$.

As discussed in Section 3.2, we believe that the list of 14 channels used by Atkinson and Campos is the same as the one found in [8]. We therefore conducted the same experiments as before using these channels. Unfortunately, the results are similar, and even seem to be a bit worse. The maximum accuracy found for Arousal is 47.5% using 118 features. It is 46.1 % using 6 features for Valence. These results are displayed in Figure 3.

Finally, we also studied the impact of the γ parameter in our SVM classifier. We focused our study on the number of features yielding our previous best results. We chose values of $\gamma = 0.1$ and $\gamma = 0.5$. Note that we also tried $\gamma = 0.01$, but had to cut our experiment short due to computational constraints. These results can be found in the Appendix (Figures 4

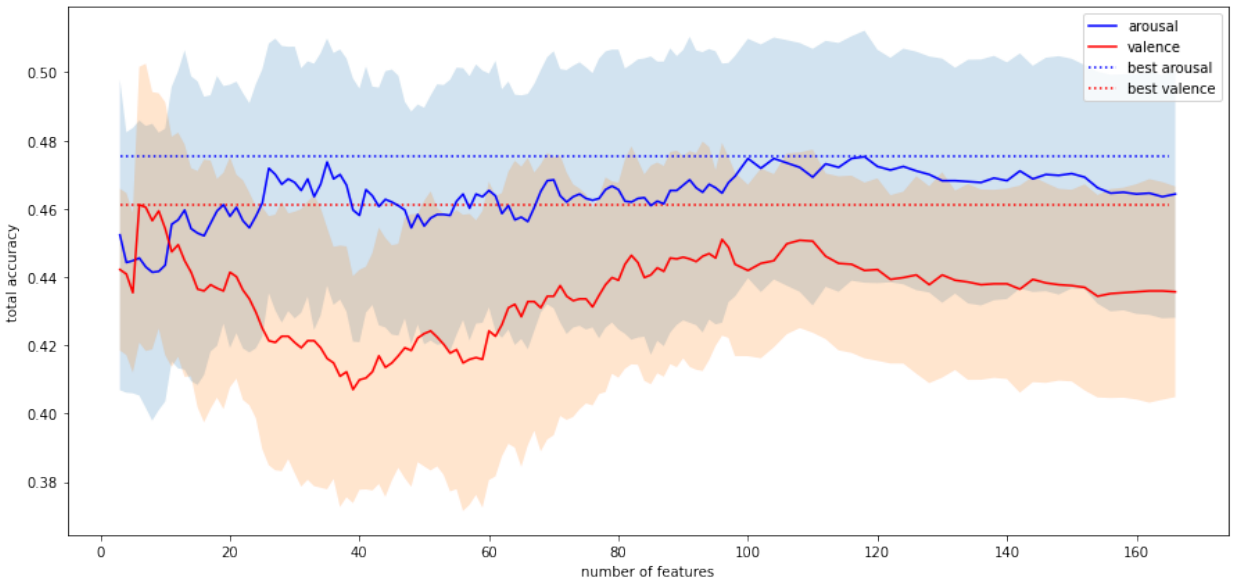

Figure 3: Total accuracy for different feature subset sizes. Here we consider only 14 channels and $\gamma = 0.05$.

and 5). A γ value of 0.05 is undeniably better for our model's performance, for both valence and arousal. It contributes to both a 5% and 3% gain in accuracy for each emotion. This observation is similar to that made by Atkinson and Campos.

6 Conclusion

In this project, we implemented the methodology found in the paper *Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers* by Atkinson and Campos [1]. This paper is centered around the task of automatic emotion recognition using EEG signals. The emotions that are to be recognized are arousal and valence. The methodology used can be split into two parts: a feature extraction and selection step, and a classification step that leverages the use of an SVM classifier. The results we obtained using their methodology are quite disappointing. This could be due to the fact that we are missing some of the features they computed, or selecting the wrong channels for classification. Whereas they indicate accuracies superior to 60% on each emotion, we were unable to break 50%. We hope to find the reason behind this difference in results in the future. On a separate note, this project enabled us to learn about time series classification applied to brain signals. We look forward to expanding our knowledge about the field in the future, and thank Professor Laurent Oudre for the opportunity to explore this field.

Acknowledgements

Statement of Contribution:

- Clément: emotional classification, results, report
- Dorian: feature extraction and selection, report

For citation purposes, we used the following library: scikit-learn [5].

References

[1] John Atkinson and Daniel Campos. Improving bci-based emotion recognition by combining eeg feature selection and kernel classifiers. *Expert Systems with Applications*, 47:35–41, 2016.

[2] Clemens Brunner, Martin Billinger, Carmen Vidaurre, and Christa Neuper. A comparison of univariate, vector, bilinear autoregressive, and band power features for brain–computer interfaces. *Medical & biological engineering & computing*, 49(11):1337–1346, 2011.

[3] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[4] Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. Emotion recognition using multimodal deep learning. In *International conference on neural information processing*, pages 521–529. Springer, 2016.

[5] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[6] Panagiotis C Petrantonakis and Leontios J Hadjileontiadis. Emotion recognition from brain signals using hybrid adaptive filtering and higher order crossings analysis. *IEEE Transactions on affective computing*, 1(2):81–97, 2010.

[7] Alper Unler, Alper Murat, and Ratna Babu Chinnam. mr2pso: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. *Information Sciences*, 181(20):4625–4641, 2011.

[8] Zhongmin Wang, Zhaoping Zhang, and Wenlang Wang. Emotion recognition based on framework of badeba-svm. *Mathematical Problems in Engineering*, 2019, 2019.

[9] Aihua Zhang, Bin Yang, and Ling Huang. Feature extraction of eeg signals using power spectral entropy. In *2008 international conference on BioMedical engineering and informatics*, volume 2, pages 435–439. IEEE, 2008.

[10] Yuxuan Zhao, Xinyan Cao, Jinlong Lin, Dunshan Yu, and Xixin Cao. Multimodal emotion recognition model using physiological signals. *arXiv preprint arXiv:1911.12918*, 2019.

Appendix

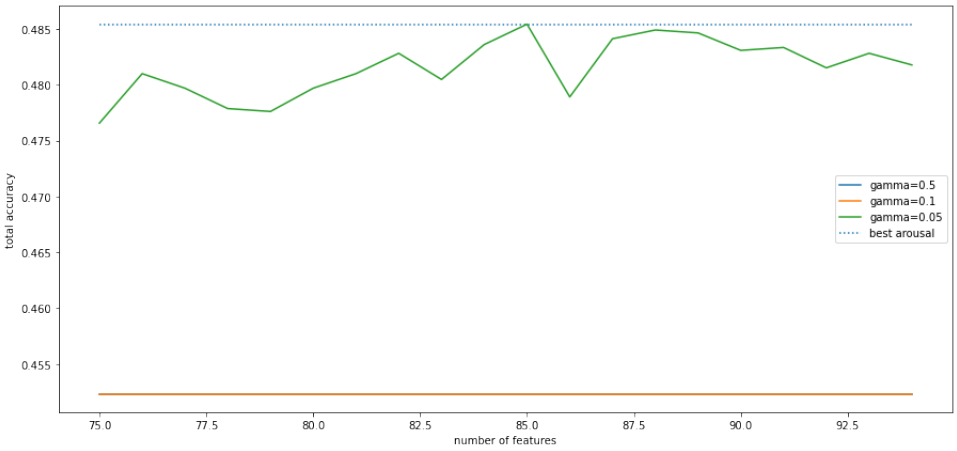

Figure 4: Mean arousal accuracy for various feature subset sizes and γ

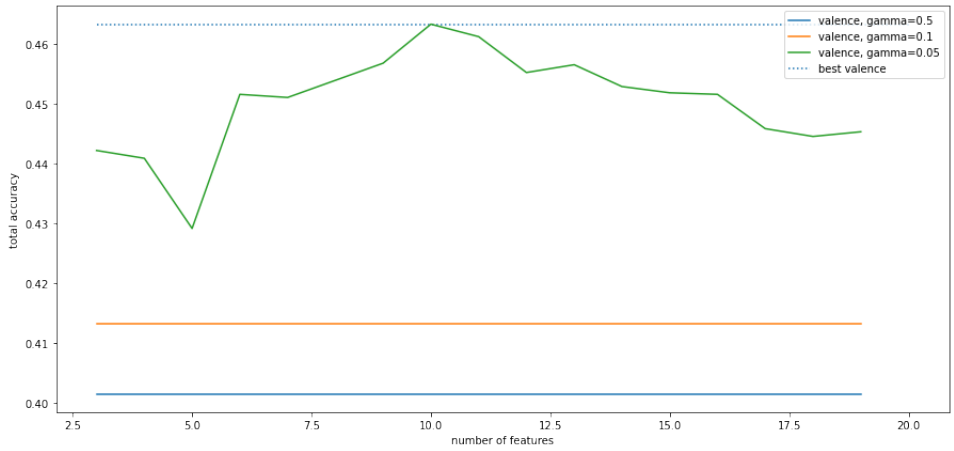

Figure 5: Mean valence accuracy for various feature subset sizes and γ