Towards An Unified Analysis Framework for EEG-based Emotion Recognition

some people

ABSTRACT

EEG-based emotion classification research is characterized by a sequence of steps, where prepossessing and feature engineering have a preponderant role, given the richness and complexity of the data under study. This has led to a a high number proposed methodologies in recent years. Although this is positive, it naturally makes it difficult to easy visualize and compare, as each method usually uses its own dataset and evaluation metric and without providing enough detail (or code) for replication. In an attempt to formalize the analysis and evaluation methodologies, in this paper it is presented the initial step towards a unified analysis framework. We performed a replication study on two of the most representative approaches for classification and evaluate them using a unique metric on the DEAP dataset.

As the feature extraction represents one of the most heterogeneous and also laborious steps, we propose an alternative based the use of a Convolutional Neural Network, with the aim to study if a setting where the model automatically learn to represent the data can be competitive in comparison with manually crafted features. Results of our empirical study show that...TODO

Keywords

EEG, Machine Learning, Emotion Recognition

1. INTRODUCTION

Emotion is critical aspect of the human behavior that plays a significant role in activities such as communication and learning []. Historically, researchers have focused on psycho physiological variables, such as posture, language (voice intonation), facial expression and gestures to identify and classify emotions and relate them to behavior patterns or decision making processes []. The use of electroencephalography (EEG) to study the emotion recognition has allowed the support of disciplines such as Human Computer Interaction (HCI) and Brain Computer Interfaces (BCI) []. In this context, the vision proposed is that in the future, machines should have the ability to identify the emotional state of humans in order to provide an effective assistance. This could represent a relevant advance in fields such as health care and education [].

EEG-based emotion recognition usually follows a sequence of steps, such as stimuli selection, feature selection and engineering, classification and evaluation []. Literature reports a high level of heterogeneity on each of these steps, given the inherent richness of the data (usually captured through a sophisticated configuration of electrodes), the multimodality of the stimuli and the wide range of analysis tools. Although this could be seen as a beneficial element, the recent increment in reported approaches makes it difficult to perform a reliable comparison to assess the real impact, applicability and level of generalization that the current studies. Our vision is that, based on the collaborative work with other members of the community, in the future it will be possible to have a solid benchmark that could increase the performance and pace of the research in this area, similarly as benchmarks such as [13] and [5] have boosted the progress in the areas of image recognition and visual saliency, respectively.

In this paper, our goal is to firstly study the quality of the available datasets for EEG-based recognition, in terms of their level of generalization. After choosing one, select the current state of the art in terms of feature extraction and try to replicate the their results keeping in mind the unification as a primary goal. Additionally, a evaluation framework is proposed, in order to formalize and normalize the results.

After setting up a baseline based on the study of related work, we explored the use of a new paradigm based on Representation Learning: the use of Convolutional Neural Networks (CNN) for EEG-based emotion recognition. Our main motivation is try to generate an alternative to the feature engineering process present in EEG-based emotion recognition, which usually turns out to be difficult and excessively laborious. In that sense, we are interested in analyzing if an automatic feature extraction could be competitive compared to hand crafted features. Our main hypothesis is that the ability of CNN to generate hierarchical representations of the features could lead to an effective understanding of the relationship between signal patterns and the valence of emotions.

We performed an initial empirical study to compare the collected emotion classification methods based on an unified way. Our results show that ... ******* TODO

The rest of the paper is structured as follows. Section

BrainKDD '15 Sydney, Australia

2 explores the publicly available EEG datasets for emotion recognition. Section 3 describes what we consider the most representative methods for emotion classification, along with a deep learning alternative. Section 4 sets a formal way to unify and compare the performance of the methods. Section 5 discusses the results and the limitations of the analysis. Finally, Section 6 outlines the conclusions and the future work.

2. AVAILABLE DATASETS DESCRIPTION

Literature shows several attempts to provide datasets for EEG-based emotion recognition. These approaches differ in many factors, such as nature of the stimuli and number and type of participants. In this section, a detailed summary is provided, followed by a qualitative and quantitative comparison. The goal is to highlight the strengths and limitations in order to facilitate the selection, given a specific goal. The following list is based on information available online by May 2015.

2.1 DEAP

Koelstra et al. [7] released the Database for Emotion Analysis using Physiological Signals (DEAP). This dataset contains electroencephalogram and peripheral physiological signals of 32 healthy participants (50% female), aged between 19 and 37 (mean age 26.9), while they were watching 40 one-minute long excepts of music videos. For each video, the dataset has a label for valence, arousal, dominance and liking levels according a process that combined Last.fm application and subjective annotation of subjects. The data is available include 48 channels (32 EEG channels, 12 peripheral channels, 3 unused channels and 1 status channel) at a sample rate of 512Hz. Due to different revision of the hardware, there are some minor differences in the format, mainly regarding to the order of the channels.

2.2 MAHNOB-HCI

In 2012, Soleymani et al. [15] published the MAHNOB-HCI[1], a Multimodal Database for Affect Recognition and Implicit Tagging. Face videos, audio signals, eye gaze data and peripheral/central nervous system physiological signals are available for researchers in those mentioned fields. The characteristics of the database include a total of 27 subjects (11 male and 16 female) with ages between 19 and 40 years old (M=26.06; SD=4.39), and the following recordings: 32-channel EEG (256 Hz); peripheral physiological signals(256 HZ); face and body videos using 6 cameras (60 f/s); eye gaze (60 Hz) and audio (44.1 Hz). This work includes two experiments, the first one was related to the emotional responses to visual stimuli (videos), for which 20 videos were selected for subjects to self-assess using emotional keywords, arousal, valence, dominance and predictability. The second experiment was related to implicit tagging and subjects had to report (dis)agreement to the displayed tag for 28 images and 14 videos.

2.3 eNTERFACE'06

In the context of Enterface 2006, the project Emotion Detection in the Loop from Brain Signals and Facial Images by Savran et al. [14] considered the generation of an affective assessment database[2]. EEG, fNIRS (functional Near Infrared Spectroscopy) and peripheral signals, namely galvanic skin response (GSR), respiration and blood volume pressure were recorded in order to prove the feasibility of a multimodal approach for emotion recognition. Two experiments were conducted for collecting the data. The first one included 5 subjects who were asked to self assess the emotions elicited by diverse IAPS[3] images of three classes (calm, exciting positive and exciting negative), while the mentioned responses were being recorded at 1024 Hz. The second experiment considered the display of three kind of emotions, neutral, happiness and disgust. In experiment 2, 16 subjects (10 male and 6 female; M=25 years old), where shown videos from the DaFEx database[4] while face video and fNIRS where recorded.

2.4 Selection

Based on the data sets described, we consider DEAP as the most complete source for analysis. Firstly, it provides the highest number of participants and the highest number of instances, in this case, one minute length music video clips (also available online). Secondly, it provides the most reliable two way emotion tagging data, making it more flexible to study the labels in the classification problem. Lastly, in terms of accessibility and ease of use, DEAP provides the most structured schema.

3. EEG-BASED EMOTION CLASSIFICATION

Based on the work of Jenke et al.[4], where a comprehensive survey on the features and models for emotion classification is presented, we performed a selection of the most representative in terms of the treatment of the data. These methods heavily rely on manually generated features, which basically represents the state of the art in the field []. We provide a detailed description of the steps involved and the replication efforts we carried out.

Additionally, as it is well known that the performance of any classification method depends on the representation of the data [1], we propose a way to adapt a Convolutional Neural Network approach to classify emotions, in order to compare manually crafted and automatic feature generation.

3.1 Multi-Modal Bio-Potential Signals

In [17], Takahashi et al. proposed a method for classification of 5 emotions (joy, anger, sadness, fear, and relax) using multi-modal biopotential signals such as brain activity, pulse and skin conductance. The classification is based on SVM and NN, reaching accuracies of 41.7% 66.7% for 5 and 3 emotions, respectively.

As this study represents one of the first attempts to implement machine learning methods on biological signal features for emotion recognition, we selected as a baseline for our empirical study.

3.1.1 Data Collection

The data collection was performed through a controlled experiment using 3 different devices for each type of signal. Firstly, a brain-computer interface was used for to measure

[1] http://mahnob-db.eu

[2] http://www.enterface.net/results/

[3] http://csea.phhp.ufl.edu/media.html

[4] https://i3.fbk.eu/resources/dafex-database-kinetic-facial-expressions

the brain activity. This device included three dry electrodes that have to be located on the subject's forehead. Secondly, a pulseoxymeter was used to capture pulse signal. It consists of a sensor clip and an amplifier, which is located on the subject's earlobe. Lastly, a skin conductance meter composed by two electrodes and an amplifier.

The experiment was applied to 12 male subjects and consisted of stimulation based on commercial films presented on a screen. These films were previously evaluated and manually labeled. This methodology resembles partially the experimental setting performed for DEAP, such as the use of video, however, the magnitude and complexity of the data captured is drastically lower.

3.1.2 Feature Extraction

To characterize the data this work used the next statistical features:

1. Mean: $\mu_X = \frac{1}{N}\sum_{n=1}^{N} X(n)$
2. Standard Deviation: $\sigma_X = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(X(n) - \mu_X)^2}$
3. Mean of absolute differences: $\delta_X = \frac{1}{N-1}\sum_{n=1}^{N-1} |X(n+1) - X(n)|$
4. Mean of normalized absolute differences: $\overline{\delta_X} = \frac{\delta_X}{\sigma_X}$
5. Mean of absolute second differences: $\gamma_X = \frac{1}{N-2}\sum_{n=1}^{N-2} |X(n+2) - X(n)|$
6. Mean of normalized absolute second differences: $\overline{\gamma_X} = \frac{\gamma_X}{\sigma_X}$

Once these features were obtained, a feature vector was created according the following expression:

$$x^T = [\mu_e \; \sigma_e \; \delta_e \; \overline{\delta_e} \; \gamma_e \; \mu_p \; \sigma_p \; \delta_p \; \overline{\delta_p} \; \gamma_p \; \mu_s \; \sigma_s \; \delta_s \; \overline{\delta_s} \; \gamma_s \qquad (1)$$

Where e indicates EEG signals, p pulse, and s skin conductance.

Following the same method of feature extraction for DEAP dataset, we could obtain a similar feature vector, however as we mentioned before, features related with pulse could not be added. As a result we generate a feature vector with 192 columns related to EEG (32 channel x 6 statistical features) plus 6 columns related to skin conductance.

3.1.3 Emotion Recognition Method

This study used two algorithms to classify emotions given bio-potentials signals. First, a Support Vector Machine (SVM) using a multiclass classification is designed under a one-vs-all method, in other words, for each emotion was created a SVM, and the best classification results indicated the corresponding class. A Gaussian function was used as a kernel function. Second, an Artificial Neural Network using three layers and trained by Levenberg-Marquart method. A sigmoid function was used as an activation function of the nodes. As same as the first algorithm, for each emotion a ANN was created using a similar process to classify.

To evaluate the results a leave-one-out cross-validation method was carried out. For the SVM method with 5 emotions the results was 41.7% and for 3 emotions was 66.7%. In the case of ANN method, for 5 emotions the results was 31.7% and for 3 emotions was 63.9%.

To build a similar process using DEAP dataset, we used A SVM algorithm using a one-vs-all method and a Gaussian function is used as a kernel function.

3.1.4 Related Work

As this work is one of the first in using statistical variables from bio-potentials signals over machine learning algorithms, there are other works that have tried to replicated it.

Wang et. al. in [18] made a similar experiment and used the same statistical features for the analysis, but also added frequency domain features such as Fast Fourier Transform. His results are better comparing when frequency domain features are not used.

Liu et. al. in [9]

3.2 Wavelet Decomposition Method

Murugappan et al. have quite an interesting story related to the emotion assessment through EEG analysis. Starting from 2008, several studies could be found in literature corresponding to this author, using different combinations of features, EEG channels and algorithms in order to classify discrete emotions such as disgust, happy, surprise, sad and anger [12, 10].

For the present benchmark, we are centered in one of the mentioned Murugappan's works, where an experiment was conducted for collecting EEG data of 20 subjects, for the classification of five emotions, namely disgust, happy, surprise, fear and neutral. A wavelet-based approach was performed, three different features were proposed for the analysis and two classifiers were used, obtaining a maximimum average classification rate of 83.26% with KNN and 75,21% with LDA [11].

Since Murugappan is an active actor in the emotion recognition field, turns to be important to include this study in benchmark and see how those proposed features behave with the DEAP dataset.

3.2.1 Data Collection

The data acquisition process considered a randomly ordered presentation of emotionally selected video clips with different time durations. The protocol followed consisted of presenting natural scene pictures and a "soothing music" for several seconds before the experimental session, which made subjects feel calm and mind relaxed. Later, 5 trials for disgust, happy and surprised emotions were followed by 4 trials of fear and neutral emotions. EEG data of 64 channels at a 256 HZ sampling rate were collected with the Nevus EEG device, whose electrodes were placed according to the International 10-10 system. A total number of 20 subjects took the experiment (3 females and 17 males with ages between 21 and 39 years old).

3.2.2 Feature Extraction

Murugappan's approach takes into account the utilization of two sorts of features: Statistical and Wavelet-based. Based in previous section (Takahashi method), we are not performing statistical feature analysis, thus we focus on the proposed wavelet features.

The "db4" wavelet function is used to separate the EEG signals into 5 levels of decomposition, and three frequency bands (alpha, beta and gamma) are considered for deriving the following features:

1. *Recoursing Energy Efficiency:* $REE(\gamma - 3b) = \frac{E_\gamma}{E(total - 3b)}$
2. *Logarithmic REE:* $LREE = \log(REE)$
3. *Absolute Logarithmic REE:* $ALREE = |LREE|$

Where the total energy of the three bands is: $E(total - 3b) = E_\alpha + E_\beta + E_\gamma$

3.2.3 Emotion Recognition Method

The process of emotion recognition used by Murugappan consisted in the utilization of two machine learning algorithms, namely K-Nearest Neighbors (KNN) and Linear Discriminant Analysis (LDA), along with a 5-fold cross validation. The same process was applied in this research for classifiying three states of emotion in the valence dimension, such as *positive*, *neutral* and *negative*.

Since the best results obtained by Murugappan were those related with the wavelet energy-based features and the maximum amount of EEG channels (64), we include the same features and our maximum quantity of electrodes (32). We used a 5-Fold Cross Validation method for evaluation.

Contrary to the original results, in our case KNN performed worse than LDA, with accuracies of 41,56% and 63,28% respectively with the best behaved feature, ALREE. In [4], a Quadratic Discriminant Analysis (QDA) classifier is used for emotion recognition with different features, thus for keeping on comparing performances, QDA was applied in the same for, with 70,31% accuracy for LREE feature as the best result.

3.2.4 Related Approaches

Wang et al. [**?**] used diverse features for classifying emotional states from EEG data analysis. With a wavelet-based focus, authors obtained an accuracy of 78.41% for the wavelet Entropy feature, with a linear SVM model.

3.3 Convolutional Neural Network Method

The recent renaissance in Artificial Neural Network research through the so-called *deep learning*, has led us explore the use of these sets of techniques for EEG-based emotion recognition. In that sense, we interested in testing if the ability to automatically learn hierarchical feature representations can improve the classification accuracy in this domain.

As stated in [16], EEG can be can be modeled just a waveform, using a one dimensional representation, or as a frequency spectrum, using two dimensional representation. For simplicity, we chose a one dimensional representation.

In general terms, we modeled the problem as a multivariate time series classification, performing automatic feature extraction locally on each individual channel and then combining them into an aggregated input vector for a multilayer perceptron. In that sense, our approach resembles the work performed by Zheng et al. in [20].

3.3.1 Network Configuration

Filter Layer: The raw signal is presented to the this layer where a convolution filter *f* is applied. The size of this filters is known to influence the performance of the classification [8], therefore it is necessary to test several configurations.

Activation Layer: This layer is the main responsible for providing non linearity to the overall mapping process with the aim to learn more complex relationships [19]. Although *sigmoid* function has been historically preferred, we additionally studied the performance of $RELU$ function.

Pooling Layer: This layer performs a subsampling of the signal

3.3.2 Implementation Details

In order to make use of GPU capabilities, we implemented this approach using the Torch 7 Machine Learning Library [2]

4. EMPIRICAL STUDY

In this section, we present an exploratory analysis of the selected approaches based on a standarized evaluation process. Our main goal is to provide the basis of a initial benchmark that eventually could be used in future studies.

4.1 Evaluation Design

The first step is to define a clear way to compare the studied approaches. We are dealing with a multi class classification problem, where each subject performed a defined set of trials. As each trial encapsulates the visualization of a video clip in the DEAP repository, we consider this level as the most appropriate to evaluate. It allows to perform further analysis across different dimensions, for example, it is possible to obtain an aggregated performance score for each trial (video) across all users. Similarly, it is possible to obtain the average accuracy for one user across all trials.

Given the above, for the specific case of DEAP, it provides 32 subjects, each with 40 instances. Therefore, it is expected as output for any classification method in this study, a 40 x 32 matrix O in which each element o_{ij} represents the predicted emotion class associated to the instance i for the user j.

Another relevant aspect is to choose the representation of the class itself. DEAP provides several values that can be used to represent the polarity of an emotion . For simplicity, we chose to used the pre-computed valence score and discretize to three classes, encoding the classes from three equal segments (reported scores range from 1 to 9).

For the evaluation we adopted the *leave-one-out* cross validation technique (LOOCV)[3, 6] TODO

4.2 Results and Analysis

We performed the emotion classification task for the three approaches presented using the DEAP dataset.

4.3 Remarks

The code used in this empirical study is available at :

5. DISCUSSION

The reported results show interesting aspects worth to discuss, specially in relation to the replication and the level of generalization.

5.1 Threats to Validity

The proposed study has a main goal the formalization of the analysis into a unified analysis framework that could allow researchers to compare and visualize and improvements in more clear way. As this is quite a challenging goals, there are several threats to the validity of the results.

Internal validity: ****hablar sobre la metrica

External validity: In this study, we chose only three approaches for comparison.

Construct validity:

5.2 Feasibility

6. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed the first step towards the unification TODO

Our future work is aimed in two directions. Firstly, as a broad goal, our idea is to continue replicating and storing EEG-based classification methods reported in the literature. With this, we hope to generate a public repository where the community of researchers could carry out studies and visualize improvements in a easier way. With this we hope to contribute to field by providing a benchmark framework for EEG-based emotion classification.

Secondly, given the promising results obtained by CNN, we are interested on studying how the hyperparameter setup affects the performance of the classification. To this end, we are currently testing the use of Bayesian optimization as well as testing unsupervised methods for pre-training.

7. REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, 2013.

[2] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, number EPFL-CONF-192376, 2011.

[3] A. Elisseeff, M. Pontil, et al. Leave-one-out error and stability of learning algorithms with applications. *NATO science series sub series iii computer and systems sciences*, 190:111–130, 2003.

[4] R. Jenke, A. Peer, and M. Buss. Feature extraction and selection for emotion recognition from eeg. *Affective Computing, IEEE Transactions on*, 5(3):327–339, 2014.

[5] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012.

[6] M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999.

[7] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis ;using physiological signals. *Affective Computing, IEEE Transactions on*, 3(1):18–31, Jan 2012.

[8] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361:310, 1995.

[9] Y. Liu and O. Sourina. Eeg databases for emotion recognition. In *Cyberworlds (CW), 2013 International Conference on*, pages 302–309, Oct 2013.

[10] M. Murugappan, R. Nagarajan, and S. Yaacob. Comparison of different wavelet features from eeg signals for classifying human emotions. In *Industrial Electronics Applications, 2009. ISIEA 2009. IEEE Symposium on*, volume 2, pages 836–841, Oct 2009.

[11] M. Murugappan, N. Ramachandran, Y. Sazali, et al. Classification of human emotion from eeg using discrete wavelet transform. *Journal of Biomedical Science and Engineering*, 3(04):390, 2010.

[12] M. Murugappan, M. Rizon, R. Nagarajan, S. Yaacob, D. Hazry, and I. Zunaidi. Time-frequency analysis of eeg signals for human emotion detection. In N. Abu Osman, F. Ibrahim, W. Wan Abas, H. Abdul Rahman, and H.-N. Ting, editors, *4th Kuala Lumpur International Conference on Biomedical Engineering 2008*, volume 21 of *IFMBE Proceedings*, pages 262–265. Springer Berlin Heidelberg, 2008.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015.

[14] A. Savran, K. Ciftci, G. Chanel, J. C. Mota, L. H. Viet, B. Sankur, L. Akarun, A. Caplier, and M. Rombaut. Emotion detection in the loop from brain signals and facial images. pages 60–80, 2006.

[15] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *Affective Computing, IEEE Transactions on*, 3(1):42–55, 2012.

[16] S. Stober, D. J. Cameron, and J. A. Grahn. Using convolutional neural networks to recognize rhythmï£¡j stimuli from electroencephalography recordings. In *Advances in Neural Information Processing Systems*, pages 1449–1457, 2014.

[17] K. Takahashi. Remarks on emotion recognition from multi-modal bio-potential signals. In *Industrial Technology, 2004. IEEE ICIT '04. 2004 IEEE International Conference on*, volume 3, pages 1138–1143 Vol. 3, Dec 2004.

[18] X.-W. Wang, D. Nie, and B.-L. Lu. Eeg-based emotion recognition using frequency domain features and support vector machines. In B.-L. Lu, L. Zhang, and J. Kwok, editors, *Neural Information Processing*, volume 7062 of *Lecture Notes in Computer Science*, pages 734–743. Springer Berlin Heidelberg, 2011.

[19] Z. Xing, J. Pei, and E. Keogh. A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter*, 12(1):40–48, 2010.

[20] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao. Time series classification using multi-channels deep convolutional neural networks. In *Web-Age Information Management*, pages 298–310. Springer, 2014.