Exploring Day-to-Day Variability in the Relations Between Emotion and EEG Signals

Yuan-Pin Lin(✉), Sheng-Hsiou Hsu, and Tzyy-Ping Jung

Swartz Center for Computational Neuroscience, Institute for Neural Computation, University of California, San Diego, USA
{yplin, goodshawnl2, jung}@sccn.ucsd.edu

Abstract. Electroencephalography (EEG)-based emotion classification has drawn increasing attention over the last few years and become an emerging direction in brain-computer interfaces (BCI), namely affective BCI (ABCI). Many prior studies devoted to improve emotion-classification models using the data collected within a single session or day. Less attention has been directed to the day-to-day EEG variability associated with emotional responses. This study recorded EEG signals of 12 subjects, each underwent the music-listening experiment on five different days, to assess the day-to-day variability from the perspectives of inter-day data distributions and cross-day emotion classification. The empirical results of this study demonstrated that the clusters of the same emotion across days tended to scatter wider than the clusters of different emotions within a day. Such inter-day variability poses a severe challenge for building an accurate cross-day emotion-classification model in real-life ABCI applications.

Keywords: EEG-based emotion classification · Day-to-day variability

1 Introduction

Electroencephalography (EEG)-based emotion classification has drawn increasing attention over the last few years and become an emerging direction in brain-computer interfaces (BCI), namely affective BCI (ABCI) [1]. Researchers continue to develop advanced computational frameworks by leveraging the domain knowledge of brain functions, computational neuroscience, and machine learning. Prior to the deployment of real-life ABCI applications, a practical challenge to the emotion-aware model is its ability to adapt to the EEG variability associated with emotional responses over time.

Referring to the literature, many prior studies endeavored to the analyses focusing on EEG feature extraction/selection [2], EEG electrode optimization [3], machine-learning framework [4, 5], and multidisciplinary information sources [6]. All these empirical results contributed to the understanding of how to develop an accurate emotion-aware model. However, nearly all of the previous works accessed the data collected within a single-day session and ignored the day-to-day variability. To the best of our knowledge, only few studies have examined the inter-day physiological-signal

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2015, LNAI 9183, pp. 461–469, 2015.
DOI: 10.1007/978-3-319-20816-9_44

variability related to emotional responses. Picard et al. [7] assessed inter-day variations based on a single-subject multiple-day dataset, in which the peripheral signals, such as facial muscle, blood volume pressure, skin conductance, and reparation, were measured. Their results showed that the features of the same emotion across days tended to cluster looser than those of different emotions within a single day. Our recent study [8] also reported that the inter-day variability in the EEG recordings in two subjects; each performed music-listening experiments on four different days. In principle, incorporating more training data for building an individualized model should improve the classification performance. However, pilot results only showed a very subtle improvement when the classification model was trained with data from multiple days. It is worth noting that if the inter-day variability can somehow be alleviated, a multiple-day data recording should be more desirable than a conventional single-day experiment. Presumably, multiple-day EEG data can better encompass EEG dynamics about implicit emotional responses.

This study extended our previous pilot study [8] to collect more longitudinal data from a larger population, *i.e.*, a 12-subject five-day dataset, for assessing the inter-day EEG variability. The goal of this study was to better understand the complications of inter-day variability so that an affective BCI can be appropriately modeled and/or accounted for in the design toward real-world applications.

2 Material and Method

2.1 Participants

Twelve non-musician subjects (9 males and 3 females; age: 21.83 ± 3.76 years; all right handed) participated in a 5-day music-listening experiment with an averaged time interval of 7 ± 1.13 days (min: 5 days, max: 11 days). Each subject read and signed a consent form before the experiment, which was approved by the Human Research Protections Program of University of California, San Diego.

2.2 Music-Listening Experiment Setup

This study selected 24 music excerpts from an emotion-tagged music database [9] to induce two emotions (happiness and sadness). Each of the selected excerpts was associated with a consensus label of happiness (001 ∼ 012.wav from Set 1) or sadness (031 ∼ 042.wav from Set 1). Interested readers are referred to [9] for more details about the music excerpts and the collection of consensus labels. All selected excerpts were repeated once to make each trial last around 37 s. The 24 excerpts were randomly separated into three four-block sessions for each subject and for each day. Each block comprised both happy and sad trials in random order. Each trial began with a 15-s rest and ended with a self-emotion-assessment task, in which the subjects required to report one of the target emotions (happiness or sadness) or neutral (no feelings) based on what they had experienced. The experiment is completely self-paced so that each subject can decide the time lapse of rest (typically a few tens of seconds) before proceeding to next trial. The protocol was designed to avoid possible auditory fatigue and/or distraction.

The music-listening experiment took place in a dimly lit room. All subjects wore an in-ear earphone and were instructed to keep seated with eyes closed during music listening.

2.3 EEG Data Acquisition

This study used a 14-channel Emotiv EEG headset (Emotiv, Inc.) to sample EEG signals at 128 Hz and in a bandwidth of 0.16 and 43 Hz. Twelve channels (AF3, AF4, F3, F4, FC5, FC6, F7, F8, P7, P8, O1, and O2) were included for measuring EEG signals referentially against left and right mastoids in accordance to the international 10-20 system. The electrode impedance was checked before each session via the Emotiv Control Panel. During rest, to ensure the recording of best-quality EEG signals for analysis, a syringe was used to inject a small amount of saline solution to the sponge of poorly contacted electrodes (if any) without uncapping the headset.

2.4 EEG Feature Extraction, Selection, and Classification

Each subject performed the music-listening experiment on five different days and each experiment comprised 24 EEG trials. Each of the trails corresponded to a self-reported emotion label (either happiness, sadness or neutral). The recorded EEG data were first processed by a 1-Hz high-pass finite impulse response filter to remove low-frequency drift. The short-time Fourier transform with a 1-second non-overlapping Hamming window was then employed to estimate the power spectral density in five frequency bands for each of the 12 channels, including delta (1-3 Hz), theta (4-7 Hz), alpha (8-13 Hz), beta (14-30 Hz), and gamma (31-43 Hz).

This study adopted an EEG feature type namely differential laterality (DLAT) [6] to associate the EEG spatio-spectral patterns with implicit emotional states. The DLAT calculates the differential band power asymmetry between left-right symmetric electrode pairs in order to reflect the extent of hemispheric lateralization over the scalp. Previous studies have demonstrated the efficacy of DLAT for emotion classification [3, 6]. Upon the 12-channel montage, the DLAT formed a feature dimension of 30 for six left-right electrodes pairs (AF3-AF4, F7-F8, F3-F4, FC5-FC6, P7-P8, and O1-O2) across five frequency bands.

In addition, this study adopted a computationally efficient feature-selection method, namely F-score, to exclude the features that were relatively irrelevant to the emotion classification from the entire DLAT space. The F-score first calculates the ratio of between-class and within-class variance upon the data distribution formed by each feature individually, and then to sort the contributions of the features according to the F-score values. The larger the F-score value is, the more contributions the feature makes. The F-score index has been proven effective in improving emotion-classification performance by selecting emotion-relevant features [3, 6].

Lastly, a Gaussian Naïve Bayes (GNB) classifier was employed to model the data distributions associated with the emotional states in the F-score trimmed feature space for each subject. Note that this study discarded the EEG trials labeled as neutral and performed a two-class emotion (happiness versus sadness) classification task.

2.5 EEG Feature Validation and Visualization

This study aimed to explore the day-to-day EEG variability associated with emotional responses and its impacts on multiple-day, *i.e.*, cross-day, emotion-classification accuracy. To this end, two validation methods were performed on the five-day dataset for each subject individually: leave-day-out (LDO) and leave-trial-out (LTO) validation methods. First, the LDO method used data from D training days (D = 1, 2, 3, and 4) to build an emotion classifier and to test it against data from a separate day to evaluate the multiple-day, *i.e.*, cross-day, classification accuracy. Specifically, the EEG data from D training day(s) were concatenated to explore the optimal feature set, *i.e.*, features with high F-score values, to train the GNB model, and then to test against the data from a separate day. Given D = 4, for example, the LDO returned a subject-dependent 5-day classification performance by averaging the results of five training-testing day pairs with each of five days being the testing data once. Such validation accounting for heuristic training-testing day pairs presumably led to a fairly reliable cross-day classification result. Second, the LTO method calculated single-day, *i.e.*, within-day, classification accuracy based on the data from each single day, which was not affected by inter-day variability and considered to be the benchmark for the multiple-day scenario. The LTO calculated the averaged performance of T classification repetitions (T = 24, trials in this study) with each of T trials being equally tested using T-1 training trials. Note that the LTO performed F-score feature selection on the entire data of a single day rather than on the training trials in order to comply with the LDO condition.

In addition, this study explored the distributions of EEG features across different emotions and days to better interpret the empirical cross-day classification results. The linear discriminative analysis (LDA) was employed to reduce the dimension of the DLAT feature space (30) to a comprehensible space spanned by first two LDA components (corresponding to the two largest eigenvalues) with the maximization of discriminatory information between classes. Prior to the LDA projection, the features from each day were normalized, making the input samples varied between 0 and 1. Note the LDA was not involved in the aforementioned classification framework. In addition to the 2D informative feature space, this study further superimposed a linear decision boundary that was calculated from the data distributions of D training days. The normal vector and intercept of the decision boundary were defined as $\vec{v} = \vec{u_1} - \vec{u_2}$ and $\vec{c} = (N_1\vec{u_1} + N_2\vec{u_2})/(N_1 + N_2)$ respectively, where $\vec{u_1}$ and $\vec{u_2}$ are the means of the class distributions, and N_1 and N_2 are the numbers of samples in each class. It is worth mentioning that although such boundary was not the high-dimensional hyperplane appeared in the actual GNB classifier, it remained effective to conceptually reveal the interrelationship among training-testing data distributions and their corresponding actual classification accuracy.

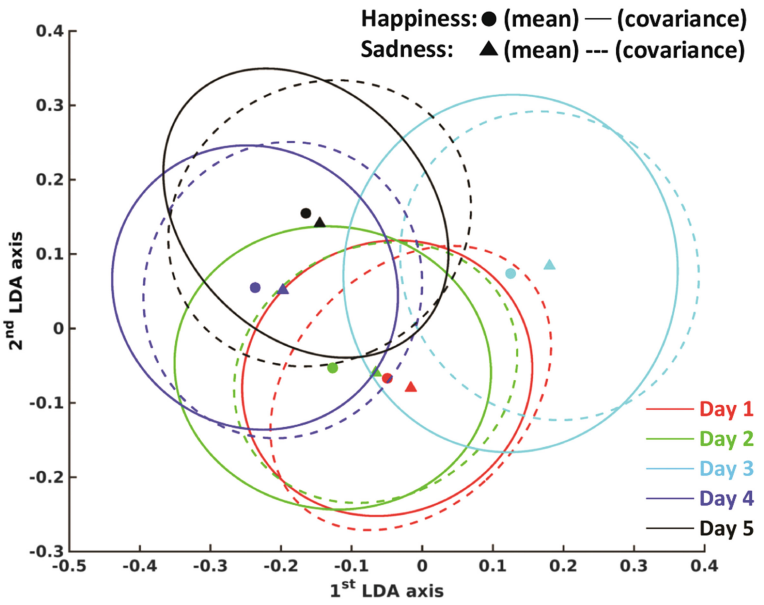

Fig. 1. EEG features' distributions of happiness and sadness across five days from a representative subject. X and Y axes represent the first two LDA components corresponding to the largest eigenvalues of the data variance in the original 30-dimensional DLAT feature space. Filled circles and triangles represent the mean values of the data distributions of happiness and sadness, respectively, whereas solid and dotted lines indicate their covariance ellipses.

3 Results

Figure 1 shows the EEG features' distributions of two emotion classes for five different days from a representative subject. The data distributions of emotional classes (solid vs. dotted ellipses) projected onto the two most discriminative LDA axes exhibited considerable overlap between the covariance ellipses within any given day. These class distributions scattered widely across different days (ellipses in different colors). These results provided insights into the day-to-day variability in EEG, which makes inter-day EEG-based emotion classification extremely difficult. It is also worth mentioning that even though the EEG features varied widely, the EEG features in some of days had relatively comparable distributions, *e.g.*, Days 4 and 5.

Figure 2 shows the results of the within- and cross-day two-class (happiness versus sadness) emotion classification. As shown in Fig. 2(A), the within-day classification obtained by the LTO validation returned the averaged accuracy of 65.49 ± 5.20 %, which exceeded chance level (50 %). There was no trend on the classification accuracy versus the recording days. Figure 2(B) shows the cross-day classification results using two scenarios namely 'heuristic-day-pair' and 'optimal-day-pair'. The heuristic-day-pair (black dashed line) profile showed the averages of all training-testing day pairs, which presumably suffered more from inter-day variability than the optimal- day-pair (red solid line) profile that reflected the best performance among all the training-day

pairs for each test day. The heuristic-day-pair result showed that the accuracy did not monotonically increase but remained around 55 ∼ 56 % with the increasing number of training days. Furthermore, regardless of the number of training days involved, all cross-day emotion-classification accuracies were worse than those of within-day classification by ∼ 10 %. Unlike the heuristic results, the optimal-day-pair result exhibited a noticeable jump in classification accuracy to 60.43 ± 2.55 % and 60.58 ± 3.05 % while using training data from one and two day(s), respectively. However, the accuracy profile started declining when more training data (days) were included (three days: 58.75 ± 2.80 %, four days: 55.76 ± 3.81 %).

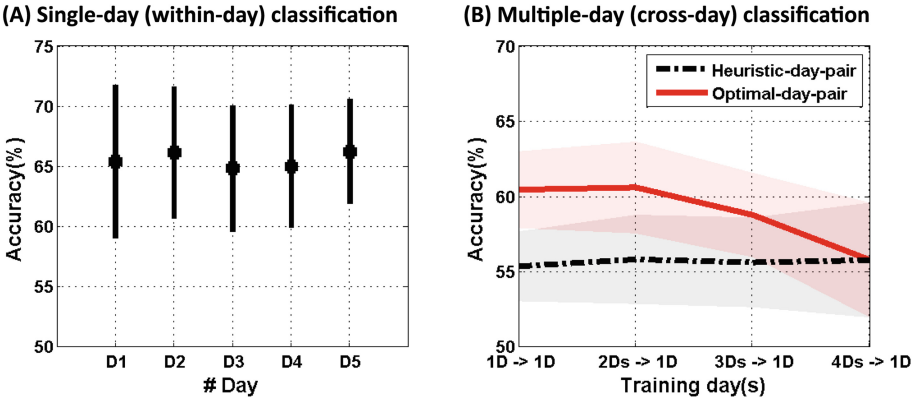

Fig. 2. Averaged results of within- and cross-day two-class emotion classification in this study. (A) Within-day classification obtained by the LTO validation was regarded as a benchmark for evaluating (B) the cross-day scenario obtained by LDO validation that accounted for inter-day EEG variability. In the cross-day plot, the heuristic-day-pair (black dashed line) shows the averaged results of all combinations of training-testing day pairs, whereas the optimal-day-pair result (red solid line) shows the best performance among all D training-day conditions (D = 1, 2, 3, and 4) for each of testing days. The shaded areas indicate the standard deviation (Color figure online).

Figure 3 further illustrates the effects of the selection of training days on the classification accuracies from the same sample subject shown in Fig. 1. In general, given the inter-day variability, the trained decision boundary worked well to a certain extent on estimating emotional responses in a separate day whose EEG feature distributions were compatible to those of the training day(s). Otherwise, it could fail and only return accuracy close to chance level. As shown in Fig. 1, Days 4 and 5 exhibited most comparable class distributions among five different days. The model used to test against Day 5 provided better classification performance as long as the data from Day 4 were included in the training data (see the left column of Fig. 3). In contrast, the class distributions of Days 1, 2, and 3 largely differed from those of Day 4, and thereby built a useless decision boundary for estimating emotional responses in Day 4 (see the right column of Fig. 3). Furthermore, as comparing the best 1 ∼ 4-day models, the classification accuracy did not improve progressively as adding more data from more days.

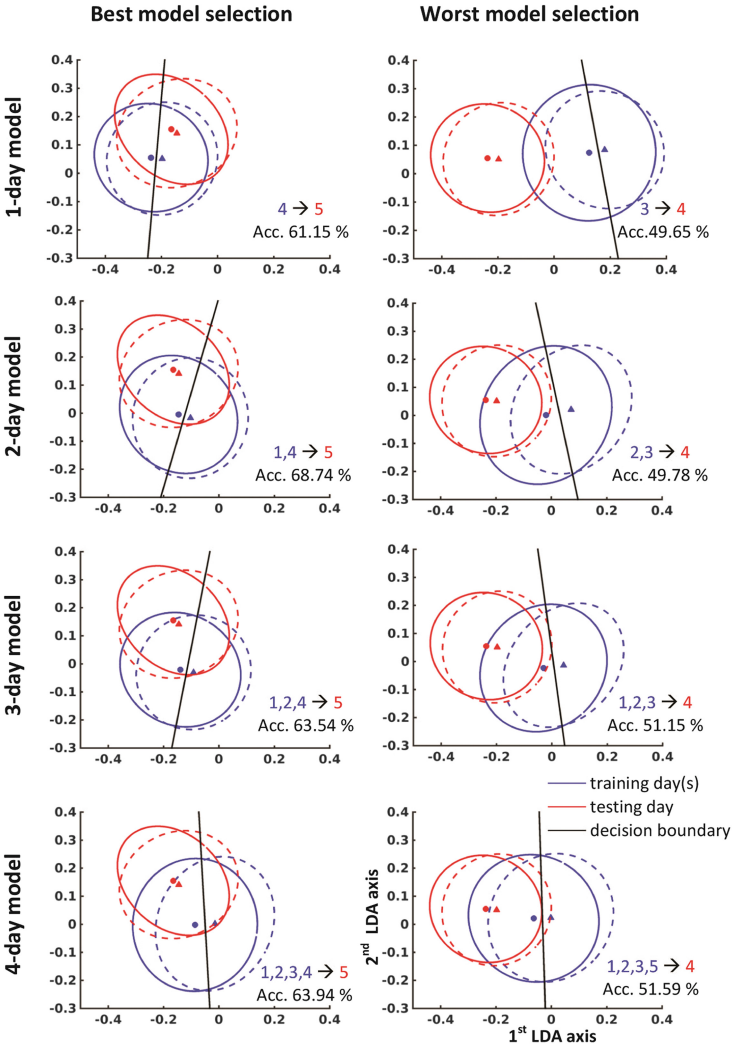

Fig. 3. The impacts of the selection of training days on the emotion-classification performance from the same sample subject shown in Fig. 1. Blue and red ellipses show the distributions of training and testing data along the first two discriminative LDA-transformed DLAT space, respectively, whereas black lines represent the linear decision boundaries calculated from the training-data distribution. Filled circles and triangles show the mean values of the data distributions of happiness and sadness, respectively, whereas solid and dotted lines indicate their covariance ellipses. Numbers in blue and red represent the day(s) for training and testing, respectively, whereas Acc. represents the classification accuracy (Color figure online).

The 2-day model that leveraged Days 1 and 4 showed the highest accuracy of 68.74 % in predicting emotions in Day 5, especially as compared to the 4-day model (63.94 %).

4 Discussion and Future Work

This study explored the day-to-day variability between emotion and EEG signals based on the dataset of 12 subjects, each underwent the music-listening experiment on five different days. The inter-day variability was assessed by inter-day data distributions and cross-day emotion-classification performance. The empirical results of this study demonstrated that the data distributions associated with emotional responses (happiness and sadness in this study) in the LDA-transformed DLAT feature space (*c.f.* Fig. 1) varied widely across different days, which was consistent with the findings in [7]. This can explain why the performance of the 'heuristic' multiple-day LDO classification was worse than that of the within-day LTO classification by around 10 % (*c.f.* Fig. 2). Furthermore, due to the inter-day EEG variability, the model trained with the data from more or all available days might not necessarily lead to better classification performance. The reason may be that the aggregation of more days which had very distinct data distributions presumably resulted in an inaccurate decision boundary. Figures 2B and 3 further provide evidence that combining informative training days (often having comparable data distributions in the explored feature space) would lead to better classification performance than naively pooling all available data together.

To conclude, large inter-day EEG variability poses severe challenges to cross-day emotion classification. How to develop a consistent and accurate emotion-classification model for each subject based on a longitudinal dataset is an emerging research direction in the EEG-based emotion classification community. The empirical results of this study may shed light on future research efforts, including: (1) a larger longitudinal database on a larger subject population is imperative for a systematic evaluation, (2) prior to pooling data from different days, an advanced normalization algorithm might be required to mitigate the salient inter-day variability, (3) other EEG feature types should be explored with emphasis on a common set of emotion-relevant signatures that are less sensitive to inter-day variability, and (4) a procedure that measures the similarity of inter-day data distributions of emotion classes should be incorporated into the data-pooling framework to better leverage the big longitudinal database.

Acknowledgement. This work was support in part by Army Research Laboratory under Cooperative Agreement Number W911NF-10-2-0022.

References

1. Mühl, C., Allison, B., Nijholt, A., Chanel, G.: A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. Brain-Comput. Interfaces **1**, 66–84 (2014)
2. Jenke, R., Peer, A., Buss, M.: Feature extraction and selection for emotion recognition from EEG. IEEE Trans. Affect. Comput. **5**, 327–339 (2014)
3. Lin, Y.P., Wang, C.H., Jung, T.P., Wu, T.L., Jeng, S.K., Duann, J.R., Chen, J.H.: EEG-based emotion recognition in music listening. IEEE Trans. Bio-Med. Eng. **57**, 1798–1806 (2010)
4. Koelstra, S., Patras, I.: Fusion of facial expressions and EEG for implicit affective tagging. Image Vis. Comput. **31**, 164–174 (2013)

5. Chanel, G., Kierkels, J.J.M., Soleymani, M., Pun, T.: Short-term emotion assessment in a recall paradigm. In.t J. Hum. Comput. Stud. **67**, 607–627 (2009)
6. Lin, Y.P., Yang, Y.H., Jung, T.P.: Fusion of electroencephalogram dynamics and musical contents for estimating emotional responses in music listening. Front. Neurosci. **8**, 94 (2014)
7. Picard, R.W., Vyzas, E., Healey, J.: Toward machine emotional intelligence: analysis of affective physiological state. IEEE Trans. Pattern Anal. **23**, 1175–1191 (2001)
8. Lin, Y.P., Jung, T.P.: Exploring day-to-day variability in EEG-based emotion classification. In: IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 2226–2229 (2014)
9. Eerola, T., Vuoskoski, J.K.: A comparison of the discrete and dimensional models of emotion in music. Psychol. Music **39**, 18–49 (2011)