Comparison of Classification Methods for EEG-based Emotion Recognition

Wei-Long Zheng[1,2], Roberto Santana[3] and Bao-Liang Lu[1,2]

[1] Center for Brain-Like Computing and Machine Intelligence, Department of Computer Science and Engineering
[2] Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering,
Shanghai Jiao Tong University, Shanghai, China
[3] Intelligent Systems Group, University of the Basque Country (UPV/EHU), San Sebastian, Spain

Abstract— **In this paper, we review different classification methods for emotion recognition from EEG and perform a detailed comparison of these methods on a relatively larger dataset of 45 experiments. We propose to combine the classifiers using stacking to improve the emotion recognition accuracies. Experimental results show that the combination of classifiers using stacking can achieve higher average accuracies than that without stacking methods. The weights derived from the classifiers are investigated to extract the relevant features and present their biological interpretation as critical brain areas and critical frequency bands.**

Keywords— **emotion recognition, EEG, classification methods, stacking**

I. Introduction

Emotion plays an important role in our daily life, especially for human communication. However, the current human computer interactions (HCI) lacks of emotional intelligence, which has an ability to detect and response to users' emotional states. Although the emotion research can be dated back to the 19th century [1], the introduction of emotion into HCI has only happened in recent decades [2]. This trend tries to narrow down the emotional gap between human and computers. One of the key challenges is emotion recognition. In the past decades, many approaches to emotion recognition have been proposed based on different modalities (e.g., facial expression, gesture, voice, physiological signals) [3]. Among these approaches, EEG-based emotion recognition, which allows direct assessment of cognitive states of users, has gained more and more attention recently [4, 5]. Emotional brain-computer interfaces can integrate brain computer interfaces with emotion factors.

In general, there are two primary aims we focus in emotion recognition from EEG. One is to improve the accuracies of emotion classification and have a good generalization on the new data, and the other is to interpret critical features with respect to the cognitive processes under study. So far, many EEG-based emotion recognition methods have been proposed. For example, Wang *et al.* [6] compared three types of EEG features (power spectrum features, wavelet features, nonlinear dynamical features) with SVM classifiers and proposed an approach to track the trajectory of emotion changes with manifold learning. Lin *et al.* [4] extracted power spectrum density and asymmetry features of five frequency bands and employed two classifiers, multilayer perceptron and SVM, to classify four emotional states. However, a major limitation of these methods is that only a handful of features and classifiers have been compared in each study. Additionally, most studies evaluate their methods on different dataset, so the results of these studies cannot be compared directly due to different setups of experiments. It is difficult to judge which types of features and classifiers are most appropriate for EEG-based emotion recognition. Although recently Jenke *et al.* [7] have presented a complete review of feature extraction methods and performed a systematic comparison of features on one database, there is still a lack of detailed comparison of classification methods for emotion recognition from EEG.

In this work, we aim for a systematic comparison of classification methods used for EEG-based emotion recognition and perform a qualitative evaluation of different classifiers on a relatively large EEG dataset of 45 experiments. Therefore, the performance of different classifiers can be compared in a unified framework. Since different classifiers may have different discriminative power for emotion classification, we further propose an approach to combine the classifiers using stacking [8, 9] to improve accuracy and have a robust generalization. Moreover, we perform a contrastive analysis of the weights derived from the classifiers as a way to detect and extract the most relevant features for classification and their biological interpretation.

II. Method

A. Feature Extraction

The raw EEG data was firstly down-sampled to 200Hz. In order to filter the noise and remove the artifacts, trials that contain large amplitude are removed and the EEG data are processed with a bandpass filter between 0.3Hz to 75 Hz.

Then we extract each segment corresponding to each trial of the experiments. We further extract features from the preprocessed EEG segments.

According to our previous studies [5, 10], differential entropy (DE) features outperform other conventional EEG features such as power spectral density. Since our focus is to evaluate the performance of different classifiers, here, we employ the DE features as the input of the classifiers. Duan *et al.* [10] have proven that, for a fixed length EEG segment, DE is equivalent to the logarithm energy spectrum in a certain frequency bands. So DE feature can be computed in five frequency bands (delta: 1-3Hz, theta: 4-7Hz, alpha: 8-13Hz, beta: 14-30Hz, gamma: 31-50Hz) with a non-overlapped Hanning window of one second. Each frequency band signal has 62 channels, so we can extract 310-dimension DE features for each sample.

B. Classifiers

We use ten types of classifiers that differ according to their functioning principles, search strategies, and efficiency considerations. The classifiers selected, as implemented in the scikit-learn software [11] programmed in Python language, are:

- Regularized logistic regression with (Ll1) [12]
- Linear discriminant analysis (LDA)
- Quadratic discriminant analysis (QDA)
- k-nearest neighbor classifier (KNN) algorithm using the Euclidean distance
- Support Vector Machine (SVM) [13]
- Gaussian naive Bayes classifier (GNB)
- Decision tree (DT)
- Regularized linear models with stochastic gradient descent (SGD) learning [14].
- Random forests (RF) [15].
- Gradient boosting (GB) [16].

Different setups where used for the classifiers. The parameters determining the differences between the variants of the classifiers are described in Table 1. When no information about the parameters is provided in Table 1, the classifiers were applied with their defaults parameters in scikit-learn[1].

The classifiers investigated cover the methods most commonly applied to BCI implementations [17]. Some of these classifiers consider interactions between the features, some others incorporate regularization techniques, or take into account similarity metrics between the data.

Table 1: Parameters used by the classifiers where k is the number of neighbors, *loss* is the loss function to be used,*md* is the maximum depth, and *ne* is the number of estimators.

	Class.	*Params*		Class.	*Params*
1	*LR*	$Norml1$	14	*SGD*	$loss = hinge$
2		$Norml2$	15	*RF*	$md = 5, ne = 50$
3	*LDA*		16		$md = 7, ne = 80$
4	*QDA*		17		$md = 7, ne = 100$
5	*KNN*	$k = 5$	18		$md = 7, ne = 120$
6	*SVM*	*linear*	19		$md = 9, ne = 150$
7		*poly.*	20		$md = 5, ne = 10$
8		*rbf.*	21		$md = 5, ne = 15$
9	*GNB*		22		$md = 9, ne = 10$
10	*DT*	$md = No.$	23	*GB*	$ne = 25$
11		$md = 5$	24		$ne = 30$
12	*SGD*	$loss = log$	25		$ne = 40$

C. Stacking

Several methods have been proposed for the combination of classifiers [18]. One of these methods is stacking [8, 9], where classifiers are organized in two different layers. There is a first level layer comprising classifiers that are learned from the data, and a second layer where another classifier is learned from the output of the first-level classifiers. The classifiers in the first layer are usually called first-level learners, while the classifier in the second layer is called second-level learner, or meta learner [18]. The detailed procedure of stacking algorithms is shown in Algorithm 1. It should be noticed that the predicted value of a given trial should be the output of classifiers not trained on that trial through cross-validation in order to avoid overfitting.

Algorithm 1: **Stacking method**

```
Divide the current training data into two different data sets $Tr_1$ and $Tr_2$
For each of the $m$ first-learners
    Train the classifier using $Tr_1$ and $Tr_2$, respectively
    Output the predictions $Pre\_Tr_1$ and $Pre\_Tr_2$ of the classifier for $Tr_1$ and $Tr_2$, respectively
    Train the classifier using the whole training data
    Output the predictions $Pre\_Te$ of the classifier for the test data
Create the second-level dataset with the first-level predictions $Pre\_Tr_1$, $Pre\_Tr_2$, $Pre\_Te$
Learn the metalearner using the second-level dataset $Pre\_Tr_1$ and $Pre\_Tr_2$
Output the predictions of the test data as represented $Pre\_Te$.
```

[1] See `http://scikit-learn.org/stable/index.html` for more details on the code.

In this paper, we are interested in the investigation of the following two different problems: 1) Whether the combination of classifiers can improve the accuracy of single classifiers. 2) Whether some classifiers are more accurate when used as metalearners than used as single classifiers. There are two ways to implement stacking method. Metalearners can be trained with the output label or output probability of the first-level learners, which are represented as *Stack_Lab* and *Stack_Pro*, respectively in this paper. With this end, we evaluated the behavior of all classifiers as metalearners.

III. Experiment Results

A. Experimental framework

We evaluate the performance of different classification methods on an emotion EEG dataset proposed in our previous studies [5, 10]. In total, fifteen subjects (7 males and 8 females; MEAN:23.27, STD: 2.37) participated in the experiment. For each subject, three sessions are repeated with an interval of about a week between them. Therefore, the dataset contains 45 experiments of EEG recordings for 3 different emotions (positive, neutral and negative) while subjects are watching emotional movie clips. The EEG data are recorded using NeuroScan System at a sampling rate of 1000Hz from 62-channel electrode cap according to the international 10-20 system. A more detailed description of the dataset is given in our previous work [5, 10].

After feature extraction from the raw EEG data, we further train different classification methods with the training and test data from different trials of the same experiment. The training data contains 9 trials of data while the test data contains the rest 6 trials of data from the same experiment.

B. Performance of different approaches

We firstly compare the performance of different single classifiers for EEG-based emotion recognition. The column "*No_Stack*" in Table 2 shows the average accuracies of different single classifiers. The results shows that LR with norm *l*2, SVM and RF outperform other classifiers with the mean accuracies of 0.8126, 0,7997 and 0.7874, respectively. We observe that the performance of SVM with linear kernel is slightly better than SVM with polynomial and RBF kernel.

Considering the diversity and robustness of the evaluated 24 variants of classifiers, we propose to select some efficient variants as the first-level learners for stacking instead of all the variants according to the performance of single ones. The selected variants of classifiers as first-level learners are listed in bold in Table 2. We select 2 variants of LR, 3 variants of SVM, 1 KNN, 3 variants of GB and 3 variants of RF as first-level learners. For metalearners, we evaluated the behavior of all variants as metalearners. The results of stacking are shown in Table 2. We can see that almost all classifiers with stacking can achieve higher average accuracies than that without stacking methods, which shows its efficiency for emotion classification. Stacking with output label of first-level learners achieves slightly better performance than that with output probability. The best average accuracies of single classifiers, stacking with label and stacking with probability are 0.8126, 0.8112 and 0.8218, respectively.

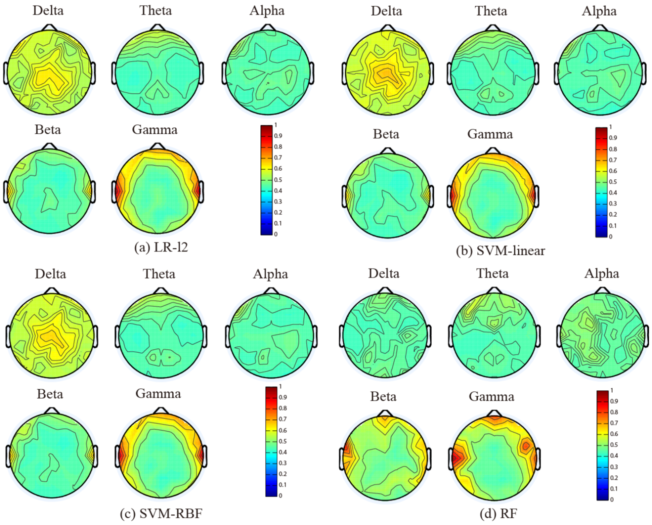
Table 2: The average classification accuracies of 45 experiments for different classification methods, where *md* is the maximum depth, and *ne* is the number of estimators. The variants of classifiers in bold indicate the first-level learners we select for stacking

Classifiers	No_Stack	Stack_Lab	Stack_Pro
LR-l1	0.7811	0.7693	**0.7948**
LR-l2	0.8126	0.7550	**0.8218**
QDA	0.5937	0.3172	**0.7110**
LDA	0.7274	**0.7392**	0.7193
KNN-5	0.7103	0.7737	**0.7822**
SVM-linear	0.7997	0.6075	**0.8067**
SVM-poly	**0.7880**	0.5910	0.5666
SVM-rbf	0.7963	0.7858	**0.7972**
GNB	0.6625	**0.7149**	0.5327
DT_md=no	0.6447	0.7318	**0.7320**
DT_md=5	0.6566	**0.7369**	0.7317
SGD_log	0.3795	0.5594	**0.8020**
SGD_hinge	0.3968	0.6792	**0.8099**
RF_md=5_ne=50	0.7781	0.8092	**0.8156**
RF_md=7_ne=80	0.7866	**0.8165**	0.8163
RF_md=7_ne=100	0.7872	**0.8205**	0.8157
RF_md=7_ne=120	0.7874	0.8089	**0.8130**
RF_md=9_ne=150	0.7806	0.8106	**0.8151**
RF_md=5_ne=10	0.7817	0.8112	**0.8160**
RF_md=5_ne=15	0.7741	0.8088	**0.8160**
RF_md=9_ne=10	0.7581	**0.8068**	0.7936
GB_ne=25	0.7060	**0.7919**	0.7510
GB_ne=30	0.7027	**0.7968**	0.7740
GB_ne=40	0.7044	**0.8029**	0.7603

C. Feature relevance analysis from the classifiers

In this section, we aim to investigate the critical features associated with emotion recognition on the interpretation of weight vectors of the classifiers. We choose logistic regression with the norm *l*2, SVM with linear kernel, SVM with RBF and random forest, since these classifiers achieve comparatively high accuracies. To clearly explore the weight distributions of the classifiers, we project the absolute average weight vectors of the classifiers to the brain scalp. The topographic plots of weights from LR-*l*2, SVM-linear, SVM-RBF and RF are shown in Figure 1.

From the weight distributions, we can extract the relevant features in emotion recognition. We find that the relevant features of different classifiers are very similar. These results show that most relevant channels locate on the lateral temporal and prefrontal brain areas and the critical frequency bands are beta and gamma bands. These findings are consistent with our previous studies [5, 10, 19]. Additionally, the weights of LR-$l2$, SVM-linear and SVM-RBF indicate some relevant features in delta bands except for RF.



(a) LR-l2

(b) SVM-linear

(c) SVM-RBF

(d) RF

Fig. 1: The weight distribution of LR-l2, SVM-linear, SVM-RBF and RF

IV. Conclusions

In this paper, different classification methods for emotion recognition from EEG have been evaluated on a relatively large dataset. We have presented a quantitative analysis comparing a wide range of classification methods that use machine learning techniques. We have also investigated the combination of the classifiers by stacking. The experimental results show that the stacking approach can improve the performance with respect to single classifiers. Moreover, we have proposed how to identify the relevant features from the weight vectors of the classifiers, which indicate the critical brain areas and critical frequency bands.

Acknowledgments

WL.Zheng and BL.Lu has been partially supported by the grants from the National Natural Science Foundation of China (Grant No. 61272248), the National Basic Research Program of China (Grant No. 2013CB329401), the Science and Technology Commission of Shanghai Municipality (Grant No.13511500200) and the NICaiA Project PIRSES-GA-2009-247619 (European Commission). R. Santana has been partially supported by the Saiotek and Research Groups 2013-2018 (IT-609-13) programs (Basque Government) and TIN2013-41272P (Ministry of Science and Technology of Spain).

References

1. James William. What is an emotion? *Mind.* 1884:188–205.
2. Picard Rosalind W. *Affective computing.* MIT press 2000.
3. Calvo Rafael A, D'Mello Sidney. Affect detection: An interdisciplinary review of models, methods, and their applications, *IEEE Transactions on Affective Computing.* 2010;1:18–37.
4. Lin Yuan-Pin, Wang Chi-Hong, Jung Tzyy-Ping, et al. EEG-based emotion recognition in music listening, *IEEE Transactions on Biomedical Engineering.* 2010;57:1798–1806.
5. Zheng Wei-Long, Zhu Jia-Yi, Peng Yong, Lu Bao-Liang. EEG-based emotion classification using deep belief networks, in *IEEE International Conference on Multimedia and Expo*:1-6 2014.
6. Wang Xiao-Wei, Nie Dan, Lu Bao-Liang. Emotional state classification from EEG data using machine learning approach, *Neurocomputing.* 2014;129:94–106.
7. Jenke Robert, Peer Angelika, Buss Martin. Feature Extraction and Selection for Emotion Recognition from EEG, *IEEE Transactions on Affective Computing.* 2014;5:327-339.
8. Smyth Padhraic, Wolpert David. Stacked density estimation, *Advances in Neural Information Processing Systems.* 1998:668–674.
9. Wolpert David H. Stacked generalization, *Neural networks.* 1992;5:241–259.
10. Duan Ruo-Nan, Zhu Jia-Yi, Lu Bao-Liang. Differential entropy feature for EEG-based emotion classification, in *6th International IEEE/EMBS Conference on Neural Engineering*:81–84 2013.
11. Pedregosa F., Varoquaux G., Gramfort A., et al. Scikit-learn: Machine learning in Python, *The Journal of Machine Learning Research.* 2011;12:2825–2830.
12. Yu Hsiang-Fu, Huang Fang-Lan, Lin Chih-Jen. Dual coordinate descent methods for logistic regression and maximum entropy models, *Machine Learning.* 2011;85:41–75.
13. Vapnik V.N.. *The Nature of Statistical Learning Theory.* Springer-Verlag New York Inc 2000.
14. Zadrozny Bianca, Elkan Charles. Transforming classifier scores into accurate multiclass probability estimates, in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*:694–699ACM 2002.
15. Breiman L.. Random forests, *Machine learning.* 2001;45:5–32.
16. Friedman Jerome H. Greedy function approximation: a gradient boosting machine, *Annals of Statistics.* 2001;29:1189–1232.
17. Lotte F., Congedo M., Lecuyer A., Lamarche F., Arnaldi B.. A Review of Classification Algorithms for EEG-Based Brain–Computer Interfaces, *Journal of Neural Engineering.* 2007;4:R1-R13.
18. Zhou Zhi-Hua. *Ensemble methods: foundations and algorithms.* CRC Press 2012.
19. Nie Dan, Wang Xiao-Wei, Shi Li-Chen, Lu Bao-Liang. EEG-based emotion recognition during watching movies, in *5th International IEEE/EMBS Conference on Neural Engineering*:667–670 2011.

Address of the corresponding author:

Author: Bao-Liang Lu
Institute: Shanghai Jiao Tong University
Street: 800 Dong Chuan Road
City: Shanghai
Country: China
Email: bllu@sjtu.edu.cn