Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230324

Interpretable EEG-Based Emotion Recognition Using Fuzzy Cognitive Maps

Georgia SOVATZIDI[a] and Dimitris K. IAKOVIDIS[a,1]
[a] *Dept of Computer Science and Biomedical Informatics, Univ. of Thessaly, Greece*

Abstract. The brain is one of the most complex parts of the human body, consisting of billions of neurons and it is involved in almost all vital functions. To study the brain functionality, Electroencephalography (EEG) is used to record the electrical activity generated by the brain through electrodes placed on the scalp surface. In this paper, an auto-constructed Fuzzy Cognitive Map (FCM) model is used for interpretable emotion recognition, based on EEG signals. The introduced model constitutes the first FCM that automatically detects the cause-and-effects relations existing among brain regions and emotions induced by movies watched by volunteers. In addition, it is simple to implement and earns the trust of the user, while providing interpretable results. The effectiveness of the model over other baseline and state-of-the-art methods is examined using a publicly available dataset.

Keywords. Electroencephalography (EEG), Emotion Recognition, Fuzzy Cognitive Map (FCM), Fuzzy Logic, Interpretability

1. Introduction

Emotions are associated with many tasks in human cognition, including decision-making processes, communication, perception, and intelligence. To detect emotions, functional neuroimaging techniques, such as EEG, are used to study the electrical activity of the brain of humans, while receiving stimuli, e.g., film clips, music, pictures [1]. Machine learning techniques have been widely used to deal with the challenging task of emotion recognition [2]. However, such approaches are considered as "black boxes", as they do not provide sufficient explanations for the resulting outcome.

During the last decades, the use of fuzzy network structures has shown great potential in various fields, including medicine [3]. Specifically, FCMs are graph-based methods, which have received considerable attention from researchers, due to their simplicity, effectiveness, and high ability to deal with uncertainties [4]. However, a limitation of FCMs is that there is a need for human participation to determine the structure of the graph. Recent modified FCMs have been effectively applied to various medical problems, including Constructive Fuzzy Representation Model (CFRM) for heart disease classification [5], and Constructive FCM (CFCM) for depression severity estimation [4]. In this paper, we introduce an FCM model for interpretable emotion recognition, based on EEG signals. Unlike conventional FCMs, the proposed model contributes to automatically detect the cause-and-effects relations that exist between its concepts without the need for human intervention, from the datasets used in each

[1] Corresponding Author: Dimitris K. IAKOVIDIS, E-mail: diakovidis@uth.gr.

experiment. Regarding the emotion recognition task, the proposed FCM model has the following contributions: a) it automatically detects the relations existing among the examined brain regions and the emotions induced by movies watched by volunteers; b) it is simple to implement; and, c) it provides understandable results.

2. Materials and Methodology

2.1. Dataset

To perform emotion recognition based on the proposed FCM, the DREAMER dataset was used [1]. It is a multimodal dataset consisting of EEG and electrocardiography (ECG) signals obtained from 23 volunteers, while watching 18 film clips selected to elicit certain emotions. For our experiments, we utilize the EEG recordings (128 Hz) of the dataset, which were collected from 14 EEG channels and analyzed in terms of arousal, and valence [1]. Depending on the evoked valence and arousal values, the 2-D valence-arousal space is derived (Figure 1). Each quadrant represents one of the following possible combinations of High (H)/Low (L) Valence (V)/Arousal (A) states and includes relative emotions. For example, as can be observed from Figure 1, HAHV includes positive emotions, *e.g.,* happiness and amusement. In addition, to examine the emotional state of the brain, a segmentation of the EEG electrode positions into brain regions was performed, based on [4], and adapted to the dataset used.

2.2. Dataset

To conduct the experiments, the EEG signals were preprocessed, based on [1]. Specifically, the signals were filtered between 4 and 48 Hz, using a FIR filter with a Hamming window of 212 samples. Moreover, an artefact rejection process was performed, *i.e*., Artefact Subspace Reconstruction (ASR) [1], utilizing the EEGLab toolbox [6]. Then, the EEG signals were separated into the following frequency bands: theta (4 Hz - 8 Hz), alpha (8 Hz - 13 Hz), and beta (13 Hz - 20 Hz). In addition, the Power Spectral Densities (PSDs) of the EEG signals in different frequency bands were calculated, given that they are significantly correlated with human emotions [1]. All the extracted features are concatenated into a final feature vector $F_{r,\lambda}^{w} = \left(f_1^w, f_2^w, \dots, f_{N_r}^w\right)$, where w corresponds to the frequency band, $r = 1, \dots, R$ are the examined segmented brain regions, N_R is the number of EEG electrodes in each r, and $\lambda = 1,2, \dots, \Lambda$ represents the class of the examined problem. In our experiments, λ corresponds to positive and negative emotion. The fuzzy set construction is performed, aiming to characterize linguistically the calculated brain electrical activity, based on [4]. Specifically, a clustering algorithm is applied to group $F_{r,\lambda}^{w}$ into a set of M clusters with $M < V_\lambda$, where V_λ corresponds to the total number of volunteers participating in the emotion recognition problem. The resulting centroids $q_m, m = 1, \dots, M$ are sorted ascendingly and their standard deviations $\sigma_m, m = 1, \dots, M$ are used to define the fuzzy sets $\Phi_r^m, m = 1, \dots, M, r = 1, \dots, R$. These fuzzy sets have triangular membership functions, where the top of the triangle is located at the q_m, and the base is extended to the range $[q_m - \sigma_m, q_m + \sigma_m]$. Each fuzzy set corresponds to a linguistic term, *e.g.*, for $M = 3$ the linguistic terms are "*Low*", "*Medium*", "*High*", which is described by a membership function $\mu_r^m(f_{v,r}^w)$, v =1,2,…, V_λ.

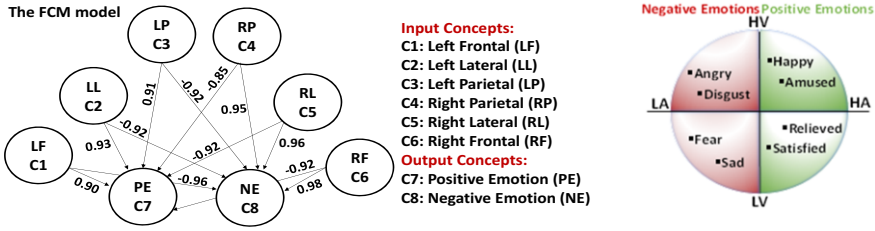

Figure 1. Proposed FCM structure (left). 2D valence-arousal emotion space -High (H) / Low (L) Arousal (A), Valence (V) (right).

2.3. FCM Construction

An FCM is a graph-based model consisted of concept nodes, $C_i = C_1, C_2, \ldots, C_K$, where K is the total number of concepts, and weighted arcs w_{ij} that determine the interconnection between C_i to C_j. The concept values of nodes represent the state vector $A^t = (A_1^t, \ldots, A_K^t)$. Thus, regarding the proposed FCM structure, the input concepts (C_1-C_6) of the graph represent the brain regions, *i.e.*, Left (L)/ Right (R) Frontal (F), Parietal (P), Lateral (L), as depicted in Figure 1 (left). The output concepts (C_7, C_8) represent the Positive and Negative Emotion (PE/ NE), which is correlated with the corresponding arousal/valence values (Figure 1). The interconnections between two concepts C_i and C_j, $i \neq j$, are defined in relation to the differences observed in brain activity in the examined brain regions, regarding positive and negative emotions. Specifically, the weight of each edge $i \rightarrow j$ is calculated based on the membership functions obtained by the respective fuzzy sets Φ_r^m. The fuzzy sets are then aggregated, using a union operation and the calculation of their center of gravity $(g_{i,j})$ follows. For example, for the calculation of w_{16}, we calculate the center of gravity of fuzzy sets that correspond to C_1 and C_6 ,*i.e.,* Φ_1^m and Φ_6^m . The calculation of interconnections between two input concepts C_i and C_j , $i \neq j$ is given by: $w'_{ij} = \frac{max_{m=1}^{M}\left(\mu_i^m(g_{i,j}),\mu_j^m(g_{i,j})\right)}{\arg\left(max_{m=1}^{M}\left(\mu_i^m(g_{i,j}),\mu_j^m(g_{i,j})\right)\right)}$. Similarly, the calculation of interconnection between an input and an output concept is given by $w_{ij} = \frac{max_{m=1}^{M}\left(\mu_i^m(g_{i,j})\right)}{\arg\left(max_{m=1}^{M}\left(\mu_i^m(g_{i,j})\right)\right)}$. In the constructed graph of the examined problem (Figure 1), all the input concepts have interconnections among them. However, for illustrative purposes, the interconnections among the input and output concepts have been selected for demonstration. The proposed FCM model iteratively calculates its states until convergence, based on $A_i^{t+1} = h\left(A_i^t + \sum_{j=1,j\neq i}^{n} w_{ij}A_j^t\right)$, where $t = 1 \ldots., T$ is the iteration number, w_{ji} is the weight matrix of C_i to C_j, and h is a transfer function. The values of the initial state vector $A^0 = (A_1^0, \ldots, A_K^0)$ are estimated based on [4].

3. Results and Discussion

3.1. Classification Outcomes

The proposed model was evaluated in terms of its decision-making performance with

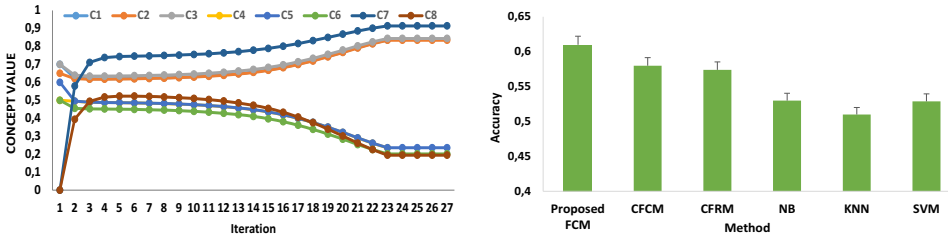

Figure 2. (left) Convergence plot of the proposed FCM model. (right) Comparisons of the average accuracy and standard deviation using the PSD features.

well-known classifiers, *i.e.,* Naïve Bayes (NB), k-Nearest Neighbor (kNN), Support Vector Machine (SVM), as well as state-of-the-art approaches, *i.e.,* CFCM , CFRM [3], [4]. As it can be observed from Figure 2 (right), the proposed model outperforms the rest methods in terms of average accuracy, using the calculated PSD features, while providing interpretable results. In addition, it reaches a steady state and finally converges, after 23 iterations. The convergence plot is depicted in Figure 2 (left).

3.2. Classification Outcomes

To better understand the interpretable emotion recognition based on the proposed FCM model, the following indicative example is included in this section. Let us consider a randomly selected volunteer from the dataset, for automatically detecting and interpreting his emotional state. The initial state vector $A^0 = (0.70, 0.65, 0.60, 0.60, 0.60, 0.50, 0, 0\ )$ is calculated as described in the methodology and is inserted into the FCM to start the reasoning process. After $k = 23$ iterations, the FCM converges into a steady state, which results in $A^{23} = (0.86, 0\ 85, 0.86, 0.21, 0.21, 0.18, 0.92, 0.17)$. Specifically, regarding A^0 and A^{23}, the first three numbers correspond to the concept values of C_1, C_2, C_3, which represent the Left Hemisphere of the Brain, *i.e.*, Left Frontal Lateral, and Temporal regions, while C_4, C_5, C_6 represent the Right Hemisphere, respectively (Figure 1). The last two values of A^0 and A^{23} represent the output concept values, *i.e.*, the values of C_7 and C_8 that correspond to the positive and negative emotions (Figure 1). To interpret the outcome in a way compatible to human logic, the calculated PSDs are then characterized linguistically, using fuzzy sets. Specifically, the PSDs are described using the following four linguistic terms: "*Low*" (L)=[0, 0.4], "*Medium*"(M)=[0.2, 0.6], "*High*" (H)=[0.4, 0.8], "*Very High*" (VH)=[0.6, 1]. Consequently, looking at A^{23} the examined volunteer has a total higher electrical activity in the left-brain regions compared to the right ones. Specifically, Left Frontal Lateral and Parietal Regions resulted in "*Very High*" PSDs, with corresponding calculated concept values $C_1 = 0.86, C_2 = 0.85, C_3 = 0.86$. Regarding the Right Frontal Lateral and Parietal Regions, they are characterized by "*Medium*" PSDs, as the respective concept values are $C_4 = 0.21, C_5 = 0.21, C_6 = 0.18$. The values of the outcome concepts, which determine the final results, are $C_7 = 0.92$ and $C_8 = 0.17$, which is linguistically described as "*Very High*" Positive Emotion, "Low" Negative Emotion. An interpretation of the proposed FCM model (Figure 1 (left)) considering the calculated interconnections, regarding the examined case, follows:

- Left-brain regions have positive relations ($w > 0$) with the positive feelings, such as joy or happiness, and negative ($w < 0$) with the negative feelings, *e.g.,* fear or disgust. Consequently, an increase in the electric activity of the left-brain region

Figure 3. Scalp maps that show the scalp distribution of power at 6 Hz on the left-brain region (left), and right brain region (right).

evokes an increase (decrease) in the degree of the positive (negative) emotional state.

- Right-brain regions have positive relations ($w > 0$) with the negative emotional states, and negative ($w < 0$) with the positive feelings. Thus, an increase in the activity of the right-brain region evokes an increase (decrease) in the degree of the negative (positive) emotional state.

Figure 3 illustrates the scalp distribution of power, at 6 Hz. The warm (cold) color in scalp topography indicates high (low) energy, respectively.

4. Conclusions

In this paper, an auto-constructed FCM model was proposed for interpretable emotion recognition, based on EEG signals. The presented graph-based model automatically detects the cause-and-effects relations between the examined brain regions and the emotions induced to volunteers, after having watched movie clips. In addition, it is simple to implement and earns the trust of the user, while providing interpretable results. Future work includes further investigation of the proposed framework on various domains, using different types of membership functions and datasets.

Acknowledgment: We acknowledge support of this work by the project “Smart Tourist” (MIS 5047243) which is implemented under the Action “Reinforcement of the Research and Innovation Infrastructure”, funded by the Operational Program "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

References

[1] Katsigiannis S, Ramzan N. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. IEEE journal of biomedical and health informatics. 2017 Mar 27;22(1):98-107.

[2] Zhang J, Yin Z, Chen P, Nichele S. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. Information Fusion. 2020 Jul 1;59:103-26.

[3] Vlamou E, Papadopoulos B. Fuzzy logic systems and medical applications. AIMS neuroscience. 2019;6(4):266.

[4] Sovatzidi G, Vasilakakis M, Iakovidis DK. Constructive Fuzzy Cognitive Map for Depression Severity Estimation. Studies in health technology and informatics. 2022 May 25;294:485-9.

[5] Vasilakakis MD, Iakovidis DK, Koulaouzidis G. A Constructive Fuzzy Representation Model for Heart Data Classification. InMIE 2021 May 1 (pp. 13-17).

[6] Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of neuroscience methods. 2004 Mar 15;134(1):9-21.