Symbolic Analysis Based Pipeline for EEG Data

Oana Miruna Moisa
Computer Science
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
oana.moisa29@gmail.com

Ileana Pop
Computer Science
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
ileanapop1811@gmail.com

Eugen-Richard Ardelean
Computer Science
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
ardeleaneugenrichard@gmail.com

Vasile Vlad Moca
Transylvanian Institute of Neuroscience
Cluj-Napoca, Romania
moca@tins.ro

Raul Cristian Mureșan
Transylvanian Institute of Neuroscience
STAR-UBB Institute
Cluj-Napoca, Romania
muresan@tins.ro

Mihaela Dînșoreanu
Computer Science
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
Mihaela.Dinsoreanu@cs.utcluj.ro

***Abstract*—This paper is concerned with developing an analysis tool for electroencephalogram (EEG) data. It explores the possibility of using the self-organizing map algorithm as starting point in a symbolic analysis based approach, with the aim of reducing the challenges brought by the high dimensionality and complexity of EEG data. The solution represents a pipeline integrating different steps for accomplishing an in-depth analysis: self-organizing map training, clustering, color sequence generation, pattern specificity index, pattern triggered average and peristimulus time histogram computation.**

***Index Terms*—symbolic analysis, EEG, self-organizing maps, color sequences, pattern triggered average, peristimulus time histogram**

I. Introduction

The human brain is a remarkably complex and intricate system representing the point of interest of all research activities in the neuroscience field. By monitoring the brain activity through different techniques, we have access to a large amount of data that is continuous and unlimited. Provided with such meaningful data, the neuroscience community is on a continuous lookout for discovering highly performant analysis methods.

One method to capture electrical brain signals is given by the electroencephalogram, considered to be a 'window on the cortical brain activity' [1]. It implies acquiring brain signals using electrodes placed on standardized positions on the scalp. This method records the activity of large populations of neurons in the range of milliseconds, being characterized by its high temporal resolution capabilities.

Several steps need to be performed before being able to properly analyze EEG data, which can be integrated into a pipeline. This paper focuses on the development of such a pipeline, comprising steps from data pre-processing to visual exploration of EEG recordings. Our approach proposes a thorough analytical methodology that considers multiple perspectives such as symbolic analysis, clustering and pattern identification, which promote interpretability that may lead to meaningful findings in EEG data. The proposed approach attempts to provide a more informative perspective of the data that is easier to interpret.

II. Theoretical Foundations

A. Challenges

The main challenge of analysing EEG data comes from its high-dimensionality introduced by using multiple electrodes for recording. Regardless of its placement on the scalp, an electrode records electrical signals from multiple brain regions due to the interference of signal sources, making it difficult to distinguish data from different parts of the brain. External factors such as eye blinks, body movements or electrical noise introduce unwanted artifacts regarded as noise, hindering the analysis process. Since the electroencephalogram is a non-invasive technique, the electrical signal has to bypass several layers including the scalp and skull until it is picked up by the electrode. This results in the final signal being heavily attenuated and smeared.

EEG data were collected from healthy participants using a high-density EEG cap, specifically Biosemi ActiveTwo, with 128 electrodes recorded at a sampling rate of 1024Hz. Here, we will analyze recordings obtained during an experiment in which the subjects performed object recognition tasks. They were given a two dimensional lattice of points that is deformed to represent an object, yielding different visibility levels [2]. Presented with the stimulus, the subjects were free to visually explore the scene, the task was to identify the object and press a specific key according to what they saw. There were three possible answers: 'Nothing' (the subject was not able to identify an object), 'Certain' (the subject was able to name the object) and 'Uncertain' (the subject identified an object in the image but could not name it).

Although the experiment had multiple subjects, the proposed pipeline is shown for a single subject for consistency across all the presented analysis techniques. As it is an analysis pipeline, it can be easily applied to all subjects. For each subject, the examination consisted of seven blocks of trials, one for each visibility level (a visibility parameter is stepped

by 0.05 from 0 to 0.3). One experimental block consisted of 30 trials with a different stimulus for each trial. In the context of a trial, due to the free exploration nature of the experiment which resulted in trials of different lengths, several event timestamps are defined to allow the identification of significant moments in the recorded neuronal activity: 'stimulus on', 'key press' and 'message for verbal response'.

The conducted human experiments adhered to Directive (EU) 2016/680 and Romanian Law 190/2018. They were reviewed and approved by the Local Ethics Committee (1/CE/08.01.2018). Furthermore, all participants involved in the study provided their written informed consent.

To overcome the previously mentioned challenges, our pipeline proposes a solution having as starting point the principles of symbolic analysis. Using self-organizing maps as a symbolic analysis method, we are able to capture properties of high dimensional data through means of symbols. After obtaining this simplified data representation, we have developed several visualization methods that helped us to properly examine the particularities of the EEG data set. The pipeline was implemented using Python programming language, version 3.8, as well as the following ecosystem of libraries: NumPy, SciPy, Pandas, MatplotLib, Plotly.

B. Self-Organizing Maps

Self-Organizing Maps (SOM) or Kohonen Maps are a particular kind of artificial neural networks which employ an unsupervised way of learning, being trained on unlabeled data. Their most important characteristic is the ability to offer another representation of high-dimensional data, reducing the dimensions but preserving the original topology so the information loss is minimal.

From a structural point of view, a SOM represents a two or three dimensional lattice of processing units, also called neurons. Each cell from the lattice contains a model vector whose length is equal to the number of features of the data set. The dimensions of the map are chosen in correspondence with the size of the input data set and can vary.

The way in which the SOM learns from the input data is by taking the samples one by one and updating itself accordingly. The number of times this process repeats is given by the number of epochs. During an epoch, each time the map learns an input sample, it completes an iteration.

For every input vector, the SOM must first find its best matching unit (BMU), which is the model vector from the lattice that best describes it or that is closest to it. To find the BMU, the Euclidean distance between the current sample and all the model vectors is computed with the intention of choosing the minimum one.

In the next phase, the best matching unit and its neighbors in the map are adjusted with the aim to make them more similar to the current input sample. This is called the updating step and the main idea is that certain neurons from the map are drawn closer to the input vector. Which neurons around the BMU are affected and how much they are updated is given by the neighborhood function and the learning rate. The neighborhood function together with the neighborhood radius determine the vicinity around the BMU to be adapted. The best matching unit will be updated the most and, as we go further away from it, still within the neighborhood radius, the neurons will be updated less and less. The most common neighborhood function is the Gaussian.

The radius defines the size of the neighborhood around the BMU and it decreases at each iteration according to an exponential decay function.

The learning rate is a parameter that determines how much the model vectors are updated at each iteration, defining the rate at which the algorithm will learn. Similar to the neighborhood radius, the learning rate decreases exponentially in time. The updating rule [3] that comprises the previously mentioned operations is given by the following equation:

$$w_i(t) = w_i(t-1)+\alpha(t-1)h_{ci}(t-1)(x(t-1)-w_i(t-1)) \quad (1)$$

, where w_i is the model vector of neuron i, t is the current iteration, $\alpha(t)$ is the learning rate, $h_{ci}(t)$ represents the Gaussian neighborhood function and x is the current input sample.

After training is completed, the resulting grid will showcase a key property of self-organizing maps, which makes them resemble the behaviour of the human brain: similar samples will be mapped on the same region on the lattice, creating 'clusters'. This mirrors the way in which neurons are grouped together based on the cognitive functions they respond to, forming brain maps.

C. Color sequences

The technique of Color Sequences was introduced in [4] as a visualization tool for obtaining representations of multiple spike-trains. As its name implies, the visual exploration of the data set is done using sequences of colors that encapsulate the characteristics of input data in a time defined context. It aims to detect meaningful patterns of the activity of neurons during the evolution of the recording.

The working principle of Color Sequences relies exclusively on a three-dimensional Self-Organising Map for localizing and obtaining the color label of each sample from the input data set. Color assignment is performed depending on the position of each model vector in the self-organised map. Each dimension of the map defines the intensities of red, green and blue channels respectively, used in the composition of final RGB values of the color label. In this way, the space of model vectors envisions an RGB subspace delimited by the size of the self-organising map. All colors composing a color sequence of a trial reside in this subspace and are found using the same process. In order to obtain the final representation of a color sequence, the colors of all samples in a trial are placed one after the other where each color band indicates a unit timestamp of 1 ms.

An individual color sequence may not be useful for obtaining meaningful information on the possible brain activity patterns. Nevertheless, groups of color sequences organized according to a specific criterion, could point out regions with similar patterns. In the case of multi-electrodes spike

trains several patterns could be identified by employing such a visualization method. Our aim is to investigate whether meaningful patterns can be observed in the EEG data as well.

D. Pattern Specificity Index

The large amount of visual information conveyed by color sequences plots of high dimensional data sets cannot be fully comprehended only by visual color inspection. Some meaningful patterns can appear by themselves multiple times throughout color sequences groups, rather than in regions of patterns at certain timestamps.

Pattern Specificity Index (PSI) can perform an in-depth color pattern analysis by ensuring the capture of all the patterns related to a predefined relevant threshold value. The index is computed for a specific pattern, p, in the context of a stimulus, s, by dividing the number of occurrences of p during trials corresponding to stimulus s, by the number of occurrences of p across all stimuli [4].

$$PSI_{p,s} = \frac{count(p|stim = s)}{\sum_j count(p|stim = j)} \tag{2}$$

PSI enables us to define a measure for the specificity of a pattern in a stimulus and represents a starting point of several analysis methods in the proposed pipeline. The numerical values, given by pattern specificity indexes of patterns (colors) in the RGB subspace delimited by the size of the self-organised map, are thresholded and color sequence plots can be regenerated containing only the remaining patterns.

E. Pattern Triggered Average

Event Triggered Average is a commonly used measure in various domains, employing different names but having an equivalent meaning and representation. This measure has the purpose of observing a certain process around the time interval an event took place. In neuroscience, the Spike-Triggered Average [5] describes how another signal, such as the local-field potential, evolves around individual spikes. In this paper, the Pattern-Triggered Average (PTA) is used to evaluate what happens, on average, with the EEG signal around the moments when a pattern appears.

The general idea of computing the event triggered average of a repeating signal is to observe it in a time-defined interval by taking a window of values before and after the timestamp of occurrence. For every signal, these values are added in an array of size $2 \times window\_size + 1$ at the corresponding positions, which are then divided by the number of signal occurrences.

F. Peri-stimulus time histogram

Peri-stimulus time histogram (PSTH) [6] represents an analysis tool used in neuroscience in order to obtain a visual estimation of the rate at which certain patterns of neural activity occur over time. PSTH divides the time axis into small bins of width δ and then counts the frequency of each pattern within the bin. As a result, it establishes a relationship between the occurrences of a certain pattern and time, providing information on the temporal dynamics of a meaningful pattern across stimuli.

From a statistical perspective, the classical PSTH is considered a model approximating the Poisson process intensity through bins with width of duration δ. PSTH aims to provide an estimation on the number of occurrences of an event between two timestamps given by the width of the bin.

III. Solution

A. Self-Organizing Maps

Our proposed pipeline draws inspiration from the analysis methods for spike-trains introduced in paper [4]. A first question that emerges is whether such analysis methods can be applied to EEG data resulting in relevant findings. We have opted to develop our own custom implementation for a three-dimensional version of the SOM following the concepts and principles described in section IIA.

For initializing the SOM, we have implemented three methods: random, random sampling and Principal Component Analysis (PCA) based [7]. Although random initialization has the advantage of simplicity, we have decided to use the PCA based version because it ensures determinism, as each model vector is built as a linear combination of the first three principal components.

The size of the map that we have regularly used is 10 for every dimension, resulting in a lattice of 1000 neurons, which complies with the size of our data set (726164 samples with 128 features each). However, we have also experimented with sizes as small as 3, which are more suitable for some types of plots and analyses.

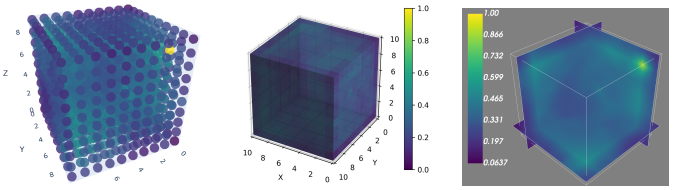
The number of epochs was chosen through the analysis of multiple values between 1 and 3, satisfactory results were obtained with the minimum value and no improvement was found when more epochs were used. The reasoning behind this is that in our implementation, each epoch has learning iterations equal to the number of samples, which is sufficient for producing accurate results. The input set is shuffled at each iteration before being passed to the map for learning, which indicates that some randomness is still present in the algorithm even when using PCA initialization. All plots presented in this article are generated using epochs=1 or 2, size=10, radius=2, learning rate=1, rectangular map topology, Euclidean distance and Gaussian neighborhood function, unless specified otherwise. We tested various learning rates for training and settled on 1, which yielded the best results.

The distance map represents a 3D array of values with the same size as the original SOM, being built after training and utilized for visualizing the results. This structure showcases the positions of the neurons in the map relative to each other. Each cell from the lattice contains a value from the interval [0,1] which indicates how close a neuron is to its neighbors. The value for a single neuron is obtained by computing the normalized sum of the Euclidean distances between itself and all its neighbors.

After the training of the SOM, the distance map can be computed. In Figure 1, we illustrate three types of 3D plotting techniques for the visualization of the distance map.

Figure 1a represents a scatter plot, where each point symbolizes a value of the distance map placed at its corresponding coordinates in the Cartesian system. Figure 1b shows a voxel plot in which the values of the distance map are $1 \times 1 \times 1$ cubic cells forming the three-dimensional volume. Figure 1c illustrates a volume slice plot, which displays the distance map as an interactive cube, whose x, y and z planes can be moved on their corresponding axes to showcase the values of the map at any coordinates.

Every sample from the data set has a corresponding best-matching unit in the map. If the BMU is close to its adjacent neurons, then the input sample is very similar to its neighbors, forming a cluster with related features. However if the BMU is distant, having values close to 1 in the distance map, this means that it becomes differentiated from its neighbors, having distinct characteristics. In the same context, yellow regions on the plots usually act as cluster separators, while the input samples are mapped on the purple areas.



(a) (b) (c)

Fig. 1. Visualization of the distance map (a) scatter plot, (b) voxel plot, (c) volume slice plot

B. Clustering

The subsequent phase in the pipeline after visualizing the results of the SOM training is clustering the data set. The distance map plots have already given us an idea about how the possible clusters are distributed in the 3D SOM space, but we cannot indicate visually exactly how many groups the data can be divided in. Finding clusters in the SOM lattice implies identifying samples which resemble each other in some aspects.

The method that we have implemented iterates through the samples from the data set and assigns them cluster labels one by one. Initially, we set the number of found clusters to 0. Then, we take one sample from the input data set, we find its best matching unit from the trained SOM and we assign it to a new cluster. At this point of the algorithm, the assumption is that the current sample will form a cluster on its own. Meanwhile, we also keep track of the samples and their best matching units that have already received a cluster number. The next step in the algorithm is to iterate through all the already clustered BMUs and compute the normalized distance between them and the current sample. If the minimum value from all the distances is also smaller than a predefined threshold, it means that the current sample is close enough to the cluster of that BMU to be considered part of it. This indicates that the initial assumption was incorrect and we update the cluster label of the sample accordingly. If however we don't find any distance within the threshold, the initial cluster assigned to the sample remains unchanged. If the cluster has indeed changed since the initial assignment then the number of total clusters that have been found is incremented. These steps are repeated for all input vectors from the data set.

After having labeled every sample, we assign to each individual cluster a unique combination of a marker and a color that will represent it for visualization purposes. Regarding the threshold used in the clustering method, it should be a number between 0 and 1, as all the distances are normalized in that range. Its value should be chosen depending on the data set by employing a trial and error principle. The resulting number of clusters that will be obtained using this method is influenced by the threshold as well as the values of the parameters used in training the self-organizing map.

For visualizing the clustered data we have implemented a scatter plot similar to the one used for illustrating the distance map, replacing the points with specific markers. The plot is made by taking each sample, finding its best matching unit and placing the marker corresponding to the cluster of that sample on the BMU's coordinates in the 3D lattice. In Figure 2, the plot obtained after clustering the data set using a threshold of 0.21 is displayed. The number of clusters is 7.

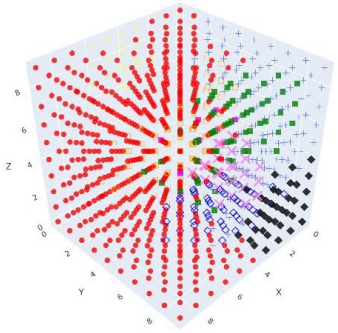


Fig. 2. Scatter plot with EEG data set clusters

C. Color Sequences

The proposed pipeline uses the technique of color sequences to build a visual representation of the brain activity acquired during EEG recording. Each trial will have its own color sequence defined by the colors of the samples in the trial. The pipeline defines two approaches for the color assignment of each sample.

In the first method, the set of colors is obtained using the corresponding colors of the best matching units from the three dimensional self-organised map of the EEG data set. The process of color labeling of a sample implies finding its model vector in the self-organised map and then mapping its spatial coordinates to corresponding RGB values in the RGB color subspace defined by the size of the map. The whole process is illustrated in Figure 3.

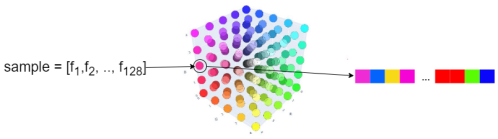


Fig. 3. RGB color assignment process

The second approach uses the clusters of the EEG data set and assigns each sample the color of the cluster it is a part of. Therefore, a clustering of the data set and a color assignment for clusters have to be performed prior to building color sequences using this method.

Since the color sequences are positioned in an adjacent and time-aligned manner, one below the other, an essential thing to consider is the grouping approach of the sequences. Different regularities and meaningful patterns can be detected according to the criteria by which the trials are organised into subsets. Based on the particularities of the EEG data set, we decided to group the trials considering three different key criteria: the response of the subject (Nothing/Certain/Uncertain), the stimulus and the visibility level of the stimulus. For example 'Nothing', 'Certain' and 'Uncertain' are refereed to as subgroups within the grouping criteria 'group by response'.

As stated in the beginning, the color sequences are visual representations of the EGG Trials. Consequently, they encapsulate the structural properties of trials, including their length. Due to the exploratory manner of the object recognition task, the subject is allowed to respond when he has reached a decision, therefore the length of each trial will be different. Moreover, this temporal variation of events lead to the misalignment of color sequences which implies that they cannot properly capture the regularities of a pattern possibly triggered by an event. The solution to this issue involved trimming the color sequences representation between the timestamp corresponding to 'stimulus on' event and the timestamp corresponding to 'stimulus off' and then performing an alignment operation using these two events, identified as left alignment (Figure 4) and right alignment, respectively (Figure 5). The former figure represents the left-alignment of color sequences of trials in which the subject responded 'Certain' and is able to capture a pattern appearing at approximately 150 ms since 'stimulus on' event. It can be observed that this pattern reappears multiple times during a trial. A right-alignment of the color sequences aims to detect regularities triggered by events happening at the end of the trials.

All figures of color sequences have a horizontal time-axis attached, for which the steps are shown at every 500 ms, depending on the length of the trials. Regarding the vertical axis, each unit represents a single trial within the subgroup.

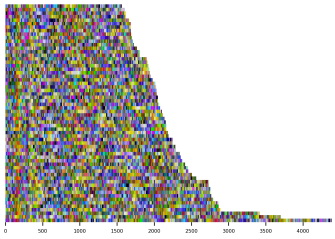


Fig. 4. Left-alignment of color sequences of the 'Certain' response of subject

Figure 6 provides the representation of color sequences built using clusters of input data set. However, this representation could not preserve the patterns that have already been identified with the previous method.

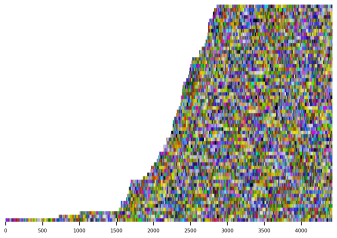


Fig. 5. Right-alignment of color sequences of the 'Certain' response of subject

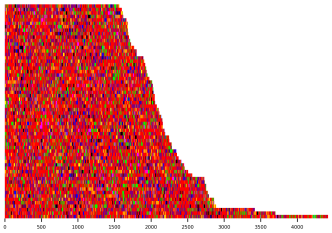
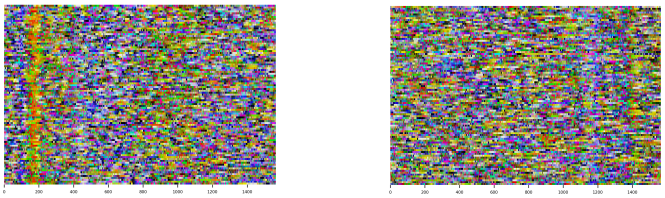


Fig. 6. Left-alignment of color sequences, based on clusters, of the 'Certain' response of subject

The subsequent step, in the process of accomplishing a thorough analysis of the information depicted by the color sequences, was to inspect the representation on a time-limited window given by the shortest duration among all trials in the group. After establishing a fixed length interval, the pipeline applies two rules for building the sequences: the former implies taking a number of samples equal to the length of the interval from the beginning of each sequence, while the latter comprises a visualization of the sequences on a same length window but aligned at the end. The result of these two approaches is illustrated by Figure 7a and Figure 7b respectively, allowing for the in-depth analysis of EEG activity on a given window. As it can be seen in both figures, the analysis on a time-limited window emphasises the presence of a pattern in the brain activity triggered by 'stimulus on' respectively by 'stimulus off'.



(a) (b)

Fig. 7. Windowed color sequences of the 'Nothing' response of subject (a) with left-alignment and (b) with right-alignment

Even though an electrode records electrical signals from multiple brain regions, which is one of the challenges of EEG data analysis, the pipeline also investigates whether a color sequence based visualization of EEG activity from individual channels could provide useful information. We have generated sequences of EEG activity recorded on electrode Oz placed in the occipital region and sequences presenting the state of parietal channels acquiring EEG data predominantly from motor function areas that record information about invocation of voluntary muscles. However, this analysis did not bring any significant changes in the discovered patterns.

D. Pattern Specificity Index

Equation (2) gives the base formula for computing the PSI value of a pattern p for a stimulus s. Our pipeline employs alternative methods of grouping that do not solely rely on the stimulus of each trial, but rather it uses conditions specific to the EEG data set structure. Consequently, the PSI of a pattern belonging to a subgroup defined by the subject's response (Nothing/Certain/Uncertain), should be computed in relation with its number of occurrences in the other subgroups as well. Yet, we still need to consider other aspects that prevent the formula from providing accurate results due to the structural particularities of the EEG trials.

Dealing with groups of color sequences containing trials of different lengths and an unequal number of trials, the PSI computation will be unbalanced if we use equation (2). More specifically, grouping the sequences according to the subject's response will result in having 95 trials for response 'Nothing', 62 trials for response 'Certain' and 53 cases in which the response was 'Uncertain'. Besides this, the result is even more influenced by the fact that some trials corresponding to response 'Nothing', are significantly longer, i.e. contain a larger number of patterns, than subgroups of response 'Certain' and 'Uncertain'.

The pipeline integrates a method named "Weighted PSI" that presents itself as a solution when working with an imbalanced distribution of structural properties of groups of color sequences. Its purpose is to weight the contribution of PSI values from each subgroup within a condition with the intention of ensuring that each PSI has an equal impact on the final value. The final PSI formula for a pattern within a subgroup becomes a derivation of (2) as it implies a multiplication of the initial PSI value with a weight, W_{sub_group}, corresponding to the subgroup it belongs to.

$$Weighted_{PSI_{p,s}} = W_{sub\_group} * \frac{count(p|sub\_group = s)}{\sum_j count(p|sub\_group = j)} \tag{3}$$

The PSI weights of a group of trials, $W_{sub\_group1}, W_{sub\_group2}, \cdots W_{sub\_groupN}$, are the unknowns of a linear system of equations given by (4) :

$$\begin{cases} W_{sub\_group1} * P_{sub\_group1} = \cdots = W_{sub\_groupN} * P_{sub\_groupN} \\ W_{sub\_group1} + W_{sub\_group2} + \cdots + W_{sub\_groupN} = 1 \end{cases} \tag{4}$$

The first equation involves equalizing the products between the unknown weights and the probabilities of choosing a sample from a subgroup within a condition. The latter equation constraints the sum of weighs to keep PSI values in the same range of values [0,1] as before.

Meaningful patterns, those appearing the most during the color sequences of each group, are identified by thresholding the PSI values of all patterns. We identify a pattern as a meaningful or significant if its corresponding pattern specificity index value satisfies the condition:

$$PSI_{p,group} - \mu > coeff * \sigma \tag{5}$$

, where μ is defined as the mean of all PSI values of patterns in a given grouping criterion, and σ expresses the standard deviation of the values.

The value of term *coeff* defines the degree of restrictiveness that decides the significance or relevance of a pattern in a given group and it is closely connected to the size of the self-organised map. For example, by using *coeff*=3 on the color sequences of Figure 4, we obtained the meaningful patterns of Figure 8.

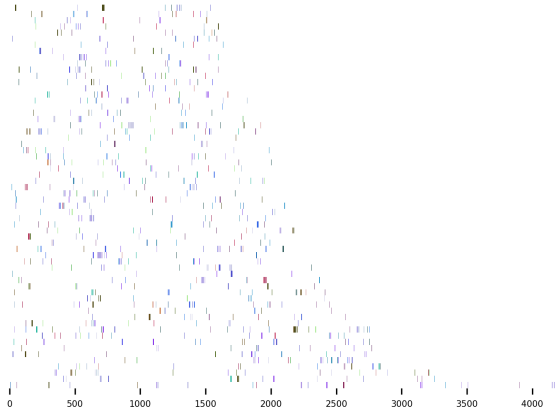


Fig. 8. PSI left-alignment for subject's response: 'Certain'

E. Pattern Triggered Average

The computation of PTA starts from the PSI that has been determined in the previous step from the pipeline. We are interested in visualizing how the signal looks like only for those patterns that are specific for a group of trials.

We will carry out the PTA computation for every meaningful subgroup of trials (e.g. 'Nothing') that we have defined when constructing the color sequences for a grouping criterion. For every pattern identified by PSI in such a group, the outcome of this process will be a figure containing a subplot for every trial in the subgroup. One such subplot will showcase the average of the EEG signal on a specific channel, offering a glimpse into how the signal behaves when that pattern occurs in the trial. If the pattern doesn't appear for all trials in the subgroup, those subplots are left empty.

Regarding the channels suitable for this analysis, we have chosen four electrodes that record electric signals from different parts of the brain (Biosemi 128 cap): A23 (positioned on the occipital lobe), B26 (positioned on the temporal lobe) , C17 (positioned on the frontal lobe) and D23 (positioned on the temporal lobe).

The computation of PTA starts with saving the timestamps at which each meaningful pattern appears within a trial. At the end of this step, each trial will be represented by a dictionary, having as keys the patterns and as values lists of corresponding positions at which the pattern appeared. This process is repeated for each subgroup generated by the grouping criterion used in the construction of the color sequences.

The next step is computing the PTA vector for each pattern in a trial. The PTA vector is obtained by averaging the signal

values for each occurrence of the pattern in the trial. The values considered in the computation of the PTA vector are the values of the signal in a time interval centered at the point of occurrence of the pattern. The limits of the interval are computed by defining a fixed portion of time, referred to as window. The default window is of 100 ms. For each pattern, we generate an image displaying the shape of the signal in every trial from the subgroup. All signal values considered in this computation correspond to a chosen channel.

Besides this way of PTA visualization, we have also generated an average signal that encapsulates the overall behavior of all the signals in a subgroup for a specific pattern. Figure 9 displays the behaviour of the signal on channel A23 for every trial in which pattern (0.4, 0.4, 0.8) appears. The average of the values of the signal on all trials is presented in Figure 10. The shape of PTAs of high specificity patterns (found through PSI) in the occipital part of the brain feature a negative deflection, indicating the presence of synchronized activity of a large number of neurons for the visual processing of the stimuli. Due to the nature of EEG data, this method is rendered potent by its ability to capture such effects.

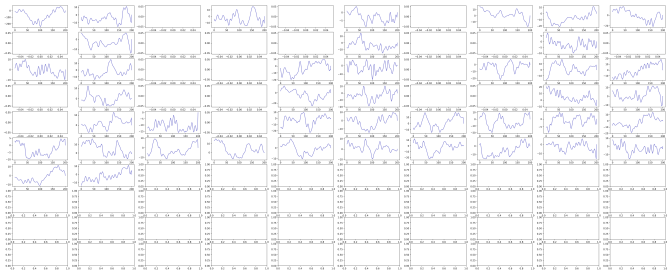


Fig. 9. PTA for each trial in the subgroup 'Certain', pattern: (0.4, 0.4, 0.8), channel A23

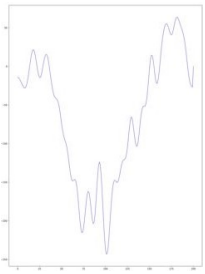


Fig. 10. Average PTA for subgroup 'Certain', pattern: (0.4, 0.4, 0.8), Channel A23

F. Peristimulus Time Histogram

The pipeline generates the PSTH of each meaningful pattern discovered after PSI computation performed for each group of trials. The PSTH computation process is carried out during the reconstruction of color sequences, which are time aligned on the stimulus onset, and it involves counting the occurrences of each color pattern at a particular timestamp across all trials in a specific group. As a result, the histogram will provide information on the distribution of the meaningful pattern across each group of trials.

Considering the length of trials, the default bin size is 50 ms and each bin contains the sum of occurrences of a pattern appearing over a time period of 50 timestamps.

PSTH is a type of analysis that can only be performed on a time-limited window as it presents the distribution of a meaningful pattern across all trials in a subgroup. Figure 11 presents the relevant patterns found for subgroup 'Certain'.

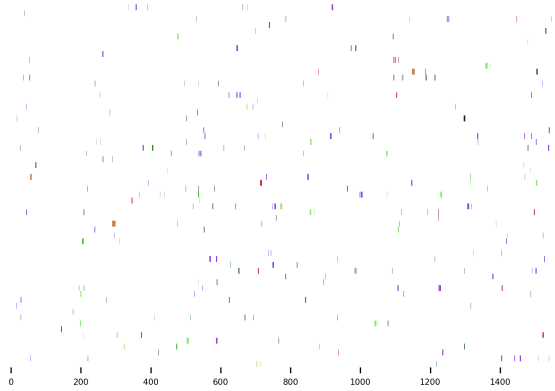


Fig. 11. PSI window for subject's response: 'Certain'

Based on the patterns from the previous figure, we generated the PSTH of pattern (0.5, 0.3, 0.9) in Figure 12 and pattern (0.6, 0.7, 0.8) in Figure 13. The histograms prove that these patterns are indeed meaningful as they are recurrent in the brain activity recorded during the experiment.

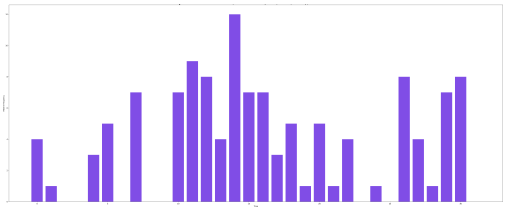


Fig. 12. PSTH pattern(0.5, 0.3, 0.9), subject's response: 'Certain'

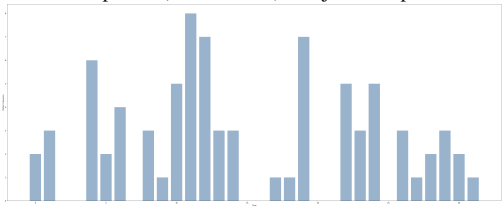


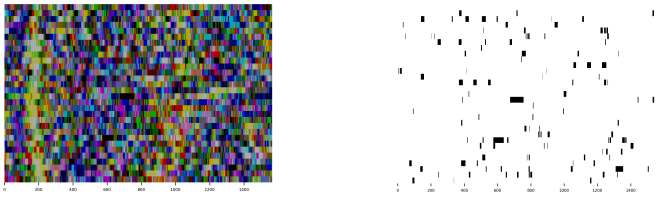
Fig. 13. PSTH pattern(0.6, 0.7, 0.8), subject's response: 'Certain'

IV. Other Relevant Results

In this section we will present other relevant results that stood out when running the pipeline with a smaller-sized SOM. The main advantage of using a map with a smaller number of neurons is that the RGB subspace defined by the size of SOM contains fewer colors and therefore, the possible patterns are visually more defined and easier to see.

Having a color sequence plot using a smaller range of colors resulted in finding significantly less meaningful patterns, which is to be expected. In the case of the sequence plot presented in Figure 14, the number of occurrences of only one pattern satisfied the PSI equation.

Moreover, it is important to mention that the pattern triggered by 'stimulus on' event, which is represented in shades of yellow, is excluded from the final PSI plot. This outcome is



(a) (b)

Fig. 14. Windowed color sequences of stimulus visibility 0.0 for a SOM size=3 (a) without PSI thresholding and (b) with PSI thresholding

anticipated because that pattern appears for color sequences of each visibility level at the timestamp corresponding to stimulus onset event. Therefore the pattern is not considered specific for a certain visibility.

In Figure 14a, it can observed that the pattern specific to 'stimulus on' appears several times after the event as well. The reappearance of the pattern after a period of suppression can be due to the subject's eye movements such as saccades and fixations, which would be a possible direction for future work.

For a SOM size of 2, the pipeline succeeded in generating color sequences based on clusters, as shown in Figure 15, in which the pattern triggered by the 'stimulus on' event is more prominent than in the case of a SOM size of 10.

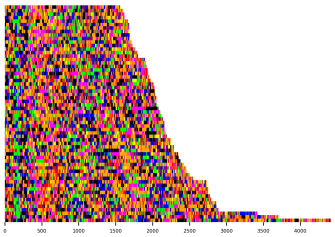


Fig. 15. Color sequence given by clusters for a SOM size=2

Figure 16 shows a clear depiction of the negative deflection that appears on channel A23 for specific patterns. Regardless the visibility level, the shape of the signal remains the same.

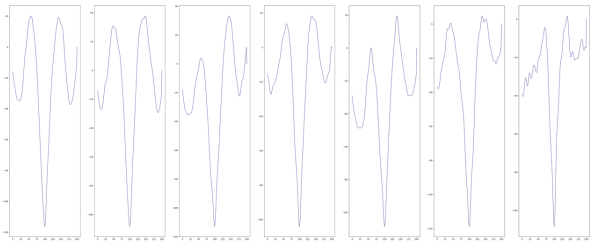


Fig. 16. Average PTA for visibility levels ranging from 0.0 to 0.3 with increment of 0.05, pattern (0.333,0.333, 0.667), SOM size = 3

V. Conclusions

The findings presented in this article allow us to point out some observations regarding the study of EEG data through a symbolic analysis approach based on the self-organizing map.

The proposed approach is capable of identifying the onset of visual processing that ensues when the subject is presented with visual stimuli, as well as the disappearance of the stimuli. The ability to identify these phenomena, clearly associated with visual processing through their timings, indicates that the proposed approach is suitable for the task.

Several patterns with a fascinating behaviour have been identified in trials in which the response of the subject has been uncertain or nothing. It seems that some of the patterns found appear at the presentation of stimuli followed by a period of inactivity after which they emerge again. The reappearance of the pattern after its suppression can be certainly motivated by occular movements, such as saccades and fixations. Nevertheless, this behaviour indicates a dualistic processing system for hard-to-identify visual stimuli. Immediately after the presentation of the stimulus there is a first attempt at identification, during which the pattern is suppressed. The reemergence could indicate another attempt of the subject to identify the stimulus and the pattern reappears as the search begins through fixations and saccades on different parts of the screen.

In conclusion, the proposed approach manages to identify a variety of patterns with specific properties that can be characterized by different analysis tools such as color-based representation, PTA or PSTH, offering us a diverse array of directions for further explorations.

Acknowledgements

The research leading to these results has received funding from: NO (Norway) Grants 2014-2021, under Project contract number 20/2020 (RO-NO-2019-0504), four grants from the Romanian National Authority for Scientific Research and Innovation, CNCS-UEFISCDI (codes PN-III-P1-1_1-TE-2021-0709, PN-III-P3-3.6-H2020-2020-0109, ERA-NET-FLAG-ERA-ModelDXConsciousness, and ERANET-NEURON-Unscrambly), and a H2020 grant funded by the European Commission (grant agreement 952096, NEUROTWIN).

References

[1] I. Constant and N. Sabourdin, "The eeg signal: a window on the cortical brain activity," Pediatric Anesthesia, vol. 22, no. 6, pp. 539–552, 2012.

[2] V. V. Moca, I. Tincaș, L. Melloni, and R. C. Mureșan, "Visual exploration and object recognition by lattice deformation," PloS one, vol. 6, no. 7, p. e22831, 2011.

[3] U. Asan and S. Ercan, "An introduction to self-organizing maps," in Computational Intelligence Systems in Industrial Engineering: with Recent Theory and Applications. Springer, 2012, pp. 295–315.

[4] O. F. Jurjuț, D. Nikolić, G. Pipa, W. Singer, D. Metzler, and R. C. Mureșan, "A color-based visualization technique for multielectrode spike trains," Journal of neurophysiology, vol. 102, no. 6, pp. 3766–3778, 2009.

[5] E. Simoncelli, J. Paninski, J. Pillow, and O. Schwartz, "Characterization of neural responses with stochastic stimuli," The cognitive neurosciences, vol. 3, 01 2004.

[6] M. Abeles, "Quantification, smoothing, and confidence limits for single-units' histograms," Journal of Neuroscience Methods, vol. 5, no. 4, pp. 317– 325, 1982. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ 0165027082900024.

[7] S. P. Mishra, U. Sarkar, S. Taraphder, S. Datta, D. Swain, R. Saikhom, S. Panda, and M. Laishram, "Multivariate statistical data analysis-principal component analysis (pca)," International Journal of Livestock Research, vol. 7, no. 5, pp. 60–78, 2017.