https://doi.org/10.52326/ic-ecco.2022/CS.14

Kolmogorov-Chaitin Algorithmic Complexity for EEG Analysis

Victor Iapascurta [1,2], ORCID: 0000-0002-4540-7045
Ion Fiodorov [2], ORCID: 0000-0003-0938-3442
[1] N. Testemitanu State University of Medicine and Pharmacy, 165, Stefan cel Mare si Sfant Blvd., Chisinau, MD – 2004, Republic of Moldova, victor.iapascurta@doctorat.utm.md
[2] Technical University of Moldova, 168, Stefan cel Mare si Sfant Blvd., Chisinau, MD – 2004, Republic of Moldova, ion.fiodorov@ati.utm.md

Abstract— **Electroencephalography as a generally accepted method of monitoring the electrical activity of brain neurons is widely used both in diseases and in healthy conditions. The recorded electrical signal is usually obtained from several electrodes located on the scalp. While EEG recording techniques are largely standardized, the interpretation of some aspects is still an open question. There is hardly questionable progress in detecting abnormal EEG signals known as seizures.**

A less explored field is the detection and classification of non-pathological conditions such as emotional and other functional states of the brain. This requires special approaches and techniques that have been widely developed over the past decade.

The current paper describes an attempt to use algorithmic complexity concepts and tools for EEG transformation making it possible to combine this approach and machine learning for classification purposes.

Keywords—Electroencephalography; EEG analysis; algorithmic complexity; block decomposition method; machine learning

I. INTRODUCTION

The goals of using EEG as a monitoring method can be summarized as: (a) to help researchers gain a better understanding of the brain; (b) to assist physicians in diagnosis and treatment choices; (c) to boost brain-computer interface (BCI) technology [1].

There are many ways to roughly categorize EEG analysis methods. As shown in a review article [2] most EEG analysis methods can be divided into four categories: (1) time domain, (2) frequency domain, (3) time-frequency domain, and (4) non-linear methods. There are also more recent methods, including machine learning (ML). As for specific mathematical signal analysis methods, there is a multitude of approaches in every of de domains listed above: linear prediction (LP) and independent component analysis (ICA), fast Fourier transform (FFT), autoregressive (AR) methods, short-time Fourier transform (STFT), wavelet transform (WT), etc.

Since the EEG signal is far from stationarity and may contain much noise, linear methods of analysis were thought not the best choice, at least in some situations. Nonlinear dynamical analysis has been a powerful approach to understanding these physiological signals. It has been observed that nonlinear dynamics theory will be a better approach than traditional time domain and frequency domain methods in analyzing and characterizing the EEG signals. The collection of non-linear methods also looks impressive: higher order spectra (HOS) techniques, phase space plot (PPS) methods, correlation dimension (CD) and fractional dimension (FD) methods, largest Lyapunov exponent (LLE), entropy estimators, etc.

Among the non-linear methods, there is a group of Entropy estimators (e.g., Spectral entropy (SEn), Approximate Entropy (ApEn), Sample entropy (SampEn), etc.). Most of them are based on Shannon's entropy, which is also presented as a measure of algorithmic complexity (AC) [2,4].

However, recent researches question the use of Shannon's entropy as the best (and sometimes even appropriate) estimation for algorithmic complexity (AC) and the Kolmogorov-Chaitin definition of AC is used instead [3,4].

The current research is trying to use the algorithmic complexity (by Kolmogorov-Chaitin approach) as a metric and data representation method for processing the data before they are fed to a machine learning algorithm.

II. EEG Data and Processing Methods

A. Data

The dataset [5] on which this research is based was originally collected to study the EEG correlates of mental activity during an intense cognitive task (mental arithmetic task—serial subtraction). The arithmetic tasks in this study involved the serial subtraction of two numbers. Each trial started with the oral communication of the 4-digit (minuend) and 2-digit (subtrahend) numbers (e.g., 4753 and 17, 3141 and 42, etc.).

In this experiment all subjects were divided into two groups: (a) group "G" (or "good counters") - performing good quality count (mean number of operations per 4 minutes = 21, standard deviation (SD) = 7.4) and (b) group "B" (or "bad counters") - performing bad quality count (mean number of operations per 4 minutes = 7, SD = 3.6).

Table 1 and Figure 1 show the general characteristics and appearance of the data.

TABLE I. General Characteristics of Data

Data source	36 healthy volunteers performing an arithmetic task
Data type	Multimodal multivariate time series: EEG and ECG, with 500 Hz sampling rate
The volume of the set and format	175 MB ".edf"
The volume of a subset and format	1285-3883 KB ".edf"
Parameters present in data	EEG signals from 20 electrodes and one-lead ECG
Data set peculiarities	EEG clip duration equal to 60-180 seconds
The task to be solved with the data	Two class classification: (a) good counters and (b) bad counters

The appearance of the data is shown in Fig.1.

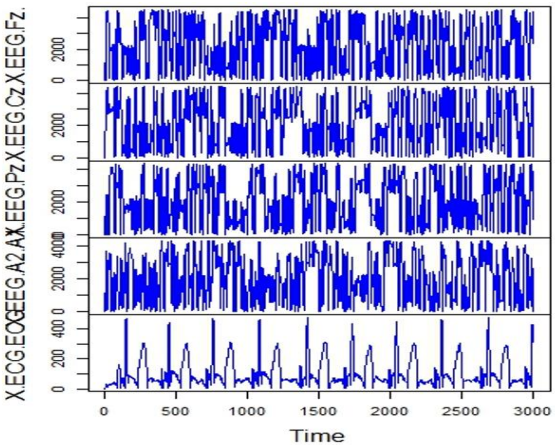

Figure 1. Appearance of EEG and ECG signals.

Figure 2.

B. Methods

A central role in data processing flow in this research is assigned to the estimation of *Algorithmic (Kolmogorov-Chaitin) Complexity* performed using the *Block Decomposition Method* which comes from the field of *Algorithmic Information Dynamics* [4, 6].

Of primary importance here is the definition of algorithmic (Kolmogorov–Chaitin or program-size) complexity (Kolmogorov, 1965; Chaitin, 1969) [6]:

$$K_T(s) = min\{|p|, T(p) = s\}, \quad (1)$$

that is, the length of the shortest program p that outputs the string s running on a universal Turing machine T.

Algorithmic Information Dynamics (AID) is an emerging field of complexity science based on algorithmic information theory, which comprises the literature based on the concept of Kolmogorov–Chaitin complexity and related concepts such as algorithmic probability, compression, optimal inference, the universal distribution, Levin's semi-measure, and others.

AID strives to search for solutions to fundamental questions about causality: why a particular set of circumstances leads to a particular outcome. In this aspect, it essentially differs from traditional statistics. As an applied science, AID is a new type of discrete calculus based on computer programming and aimed at studying causation by generating mechanistic models to help find the first principles of physical phenomena, building up the next generation of machine learning [6].

In the AID toolkit, there is a special tool for providing reliable estimations to uncomputable functions, namely the online algorithmic complexity calculator (OACC) [7], which provides estimations of algorithmic complexity and algorithmic probability for short and long strings and for two-dimensional arrays better than any other traditional tool, none of which can capture any algorithmic content beyond simple statistical patterns. The OACC uses the BDM method [3,7,8], which is based upon algorithmic probability defined by the coding theorem method (CTM) :

$$BDM = \sum_{i=1}^{n} CTM(block_i) + log_2(|block_i|). \quad (2)$$

The OACC is available as an online version as well as standalone packages in R and a number of other languages and it is used for respective calculations for the scope of the current work.

III. Data Processing Steps and Their Results

Each file (subset) in the original data is a ".edf" file describing the EEG signal voltage variations for "n channels" for 60 to 180 seconds duration. The subset is unfolded and a matrix with columns representing "n channels/electrodes" and rows denoting observations of EEG signal variation over time corresponding to particular channels is generated. The resulting matrix is split into a series of 20 x 20 (depending on the number of channels/electrodes) matrices, keeping the tie with the electrodes and time. Table 2 shows the appearance of a fragment of such a matrix (4 channels and 4 observations only).

TABLE II. Appearance of an EEG Fragment (4 channels)

Ch - 1 (μV)	Ch - 2 (μV)	Ch - 3 (μV)	Ch - 4 (μV)
4.476	-2.741	-2.502	0.095
1.208	-3.309	-4.418	-0.529
-2.546	-3.709	-6.411	-1.003
-6.187	-3.681	-8.03	-1.103

A. Calculating algorithmic complexity

These small (i.e., 20 x 20) matrices are binarized (using "BASCA" method, "Binarize" package in R) [8]. Table 3 shows what the fragment of an original matrix above looks like after binarization, with the respective thresholds and p-values.

TABLE III. Binarized Matrix and Statistics

Binarized matrix				Threshold	p-value
1	0	0	0	2.2855	0.001
1	0	0	1	-1.9190	0.001
1	1	0	1	-5.0600	0.001
0	1	0	1	-4.9340	0.001

The algorithmic complexity (by BDM) of this matrix equals 32.7241 bits. Based on (2), for a 20 x 20 matrix the AC value will be much higher. The BDM value for each such matrix is calculated with BDM values arranged over time axis obtaining time series that describe BDM value variation over time (Fig. 2).

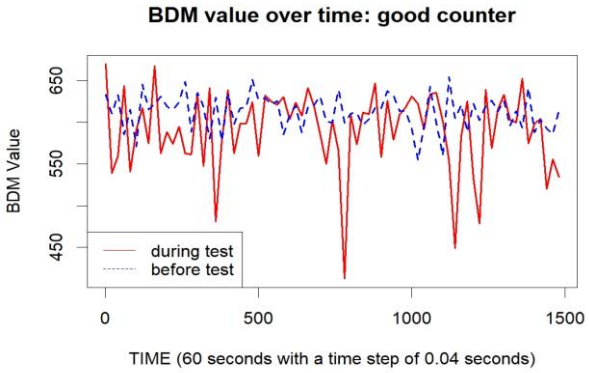

Figure 3. BDM value over time in good counter (before – blue line and during the test – red line)

Since the volume of data BDM value is to be calculated on is large, neither online nor the regular stand-alone version of the OACC is suitable. For the purpose of this research, the "core" of the R version of OACC was extracted and integrated into the data processing flow.

The binarization and BDM calculation on these large data are quite computationally expensive. To address this the E2C Amazon Web Service is being used. Considering the sampling rate of 500 Hz and the dimension of a small matrix (e.g., 20 x 20) described above, 1500 matrices are generated with each 60 seconds subset (i.e., 500Hz*60seconds/20).

B. Plotting algorithmic complexity

After calculating the BDM value, it can be plotted along the time axis with a total number of steps equal to 1500, which represents the BDM value over time for a particular subject. A detailed explanation of the data processing steps is provided in [8].

The BDM values are aggregated (using the average of ten observations/time steps) to better capture possible underlying patterns. As can be seen from Fig.2, the AC during the test fluctuates in a much larger range compared to the before-the-test AC.

In order to identify additional patterns that would help discriminate EEG before and during the test, EEG clips were randomly sampled from groups of good and bad counters (10 files from each group). After estimating the BDM value for each file ("before-the test" and "during-the-test"), the BDM value density distribution was estimated on mean per-time-step BDM values for each group. Figure 3 shows the situation in the good-counters group.

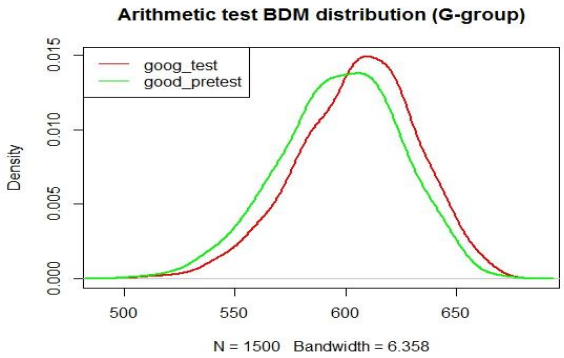

Figure 4. BDM values distribution in good counters

According to the plot above, there is a shift of mean BDM to the right (or towards increased complexity).

In the bad-counters group, the pattern seems to differ (Fig. 4). The BDM value distribution curves look quite similar. This seems to imply that AC of the brain functioning is activity agnostic in this group, or does not change depending on the type of brain activity (e.g., mental counting, as in this research).

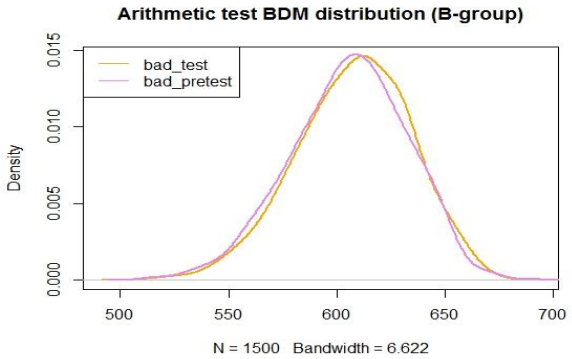

Figure 5. BDM values distribution in bad counters

IV. Discussions and Conclusions

Nonlinear dynamics has been used in neurophysiology to understand complex brain activity from EEG signals. Although linear methods have been the most commonly used in EEG analysis, non-linear approaches have expanded their presence as they reveal aspects that cannot be measured with linear approaches. However, published works in this scientific field are still rare.

The EEG signals reflect the electrical activity of the brain. They are considered to be highly random in nature and may contain useful information about the brain state. However, it is difficult to get useful information from these signals just by observing them. They are basically non-linear and nonstationary in nature. Hence, important features can be extracted using advanced signal-processing techniques. This paper describes the effect of a mental arithmetic task on the EEG signal using a less traditional processing method, namely algorithmic complexity estimation. This method allows the extraction of hidden information from the signal.

The main steps for extracting this information are presented based on the block decomposition method, which is a tool from the newly emerging field of algorithmic information dynamics. Although the Kolmogorov-Chaitin complexity (as the core of the method) apparently resembles the Shannon entropy approach as a measure of complexity, they are different, and this is explained in detail in [3, 4, 7].

According to the results presented in this paper, it seems possible to use AC of the brain functioning to gain insights into the brain state and use it to potentially classify the brain functional state (e.g., relaxed/background functioning vs performing a mental arithmetic task).

The EEG signals obtained from two groups of human subjects performing mental arithmetic tasks (good and bad counters) were split into two groups: (a) before the test and (b) during the test, processed and finally analyzed for differences using the Kolmogorov-Smirnov test.

The two-sample Kolmogorov-Smirnov (KS) test for good counters provides the following statistics: D = 0.08, p-value = 0.0001355. Since the p-value is much less than 0.05, it can be inferred that the distribution of BDM values as a measure of the algorithmic complexity of brain functioning differs before and during the arithmetic test activity in good counters.

The statistics for the two-sample KS test in bad counters are: D = 0.046, p-value = 0.08367. Since the p-value is higher than 0.05, it can be concluded that the algorithmic complexity of brain functioning in this group does not differ regardless of mental activity.

Thus AC by BDM can be used as a metric that can help distinguish between these two groups (i.e. good counters vs bad counters). The distance (D) between paired (i.e., before and during the test) states for good counters is almost twice that for bad counters. But given the small value of D in both groups, special care will be required when using this approach for specific tasks such as machine learning.

References

[1] J. Simeral et al., ”Home Use of a Percutaneous Wireless Intracortical Brain-Computer Interface by Individuals with Tetraplegia,” *IEEE Transactions on Biomedical Engineering*, 68(7), 2021, pp. 2313–2325. doi:10.1109/TBME.2021.3069119.

[2] D. Puthankattil Subha , P. K. Joseph , Rajendra Acharya U and Choo Min Lim, “EEG Signal Analysis: A Survey”, *Journal of Medical Systems*, vol. 34(), 2010, pp.195–212

[3] H. Zenil, "A Review of Methods for Estimating Algorithmic Complexity: Options, Challenges, and New Directions," *Entropy* vol. *22*(6), 2020, 612. doi.org/10.3390/e22060612.

[4] H. Zenil, **"**Towards Demystifying Shannon Entropy, Lossless Compression, and Approaches to Statistical Machine Learning," *Proceedings of the International Society for Information Studies 2019 summit*, University of California, Berkeley, 2020, 47, 24; doi:10.3390/proceedings2020047024

[5] I. Zyma et al. "Electroencephalograms during Mental Arithmetic Task Performance," *Data*, vol. 4(1), 14, 2019

[6] H. Zenil and N. Kiani, instrs. "Algorithmic Information Dynamics: A Computational Approach to Causality and Living Systems from Networks to Cells" *MOOC by Complexity Explorer*, Santa Fe Institute, Santa Fe, NM (Jun 12, 2018 to Oct 13, 2018). www.complexityexplorer.org/courses/63-algorithmic-informationdynamics-a-computational-approach-to-causality-and-livingsystems-from-networks-to-cells-2018

[7] H. Zenil, S. Hernández-Orozco, N. A. Kiani, F. Soler-Toscano, A. Rueda-Toicen and J. Tegnér, "A Decomposition Method for Global Evaluation of Shannon Entropy and Local Estimations of Algorithmic Complexity," Entropy, vol. 20(8), 2018 p. 605.

[8] V. Iapascurta, "Block Decomposition Method and Traditional Machine Learning for Epileptic Seizure Prediction,", *26th IFIP WG 1.5 International Workshop AUTOMATA 2020, Special Session on Algorithmic Information Dynamics, August 10-12*, Stockholm, Sweden 2020, https://www.automata2020.com/videos--material.html