International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 10 (2018) pp. 7267-7272

Analysis of Non-Linear Behaviour through Signal Segmentation

Miguel Salvador Gómez-Díaz, David Asael Gutiérrez-Hernández, Manuel Ornelas-Rodriguez,

Victor Zamudio, Luis Ernesto Mancilla-Espinosa

Tecnológico Nacional de México. Instituto Tecnológico de León.
División de Estudios de Posgrado e Investigación. León, Guanajuato

Abstract

The study of non-linear signals is a field that has several decades of research working, where the main issues of interest of study are the acquisition, the preprocessing to eliminate noise and outliers of the signal, as well as the method, techniques and algorithms to classify or clustered signals that are not easily to be treated to get visible patterns. This paper proposes the processing and analyzing of pure data, taken just after having been acquired, to find features and similar behaviors and to make groupings through clustering, by means of unsupervised ways and with validation methods based on techniques of probability and statistics which propose the optimal values to process the acquired data. The main objective is to make a segmentation of pure signals, without a preprocessing, which will determine the most significant groups obtained in the experiments and thus, visualize that data are represented according to a pattern behavior which will give us important information to be analyzed.

Keywords: Analysis of non-linear behaviour;Signal segmentation; k-means; Clustering; GAP Statistic; Silhouette analysis clustering; didactic model for signal segmentation.

INTRODUCTION

Today, different studies and research are carried out from the need to investigate or determine the behavior of biological signals issued by our body. It can be find a wide range of proposals in analysis and signal processing, but all finding the not linearity as a difficulty to get appropriated results that can be interpreted. This has been a well-studied field in decades, and many scientists have proposed methods and techniques to make appropriate use of the obtained signals after applying preprocessing techniques that improve the quality of the signal.

The representation of the signals issued by our body is a challenge to interpret them, e.g., modeling the behavior of an electroencephalogram signals is important for storing faithful and efficient data with the purpose of making diagnoses in clinical neurophysiology [1]. EEG signals consist of transitory fluctuations of voltage in the brain, monitored by electrodes placed on the scalp. Studies show that through analyzing EEG signals is possible to measure non-invasively wear or mental work during a task in process [2]. However, quantitative methods for analyzing interpreting biological signals from an EEG is an active area that is still in progress [1].

In this work we propose a method to analyses non-linear signals coming from a simulated prototype that represents, through light sensors, brain activity according to a well-known distribution of them, as it is shown in figure 1. Computational methods are used to learn how those signals can be grouped or segmented.

THEORICAL DESCRIPTION

A. Analysis and processing of non-linear signals

It is accurate to say that it is important to know its characteristics, the distribution of their data and their behavior, to analyze a specific phenomenon but each problem depends on a consideration and it is to determine whether the nature of the problem is linear, quadratic or any other behavior. That is the reason because of today, for the development of solutions for nonlinear problems, many research has been published by scientists with the analysis of non-linear data using and applying their own proposed methods, some of them to minimize the number of descriptors obtained at the stage of data acquisition, others use processing or functions that the initial problem leads to more comfortable space to work, i.e. to propose methodologies and computational techniques that make easier the treaty of data as in [3].

B. Clustering Techniques

Clustering techniques are a set of data as input, which are grouped according to common characteristics that they possess. These clustering methods are very useful if you don't have with a result expected, that is believed to be training or unsupervised learning [4]. Although these methods of clustering solutions that show satisfactory results, are a comprehensive task to find the number of optimal clusters or groups that offer better results than another number of k values.

C. GAP statistic

Gap statistic is well-known index of clustering validity, and plays a dominant role in some applications. Gap statistic uses the output of any clustering algorithm, comparing the change in within-cluster dispersion to that expected under an approximate null distribution reference. By maximizing a Gap statistic, the optimal number is found [5]. The derivation for the gap test assumes that there are well-separated uniform clusters. In cases where there are smaller sub-clusters within larger well-separated clusters, it can exhibit non-monotone behavior [6].

D. Silhouette Clustering

The silhouettes constructed below are useful when the proximities are on a ratio scale (as in the case of Euclidean distances) and when one is seeking compact and clearly separated clusters. Indeed, the definition makes use of average proximities as in the case of group average linkage, which is known to work best in a situation with roughly spherical clusters. To construct silhouettes, we only need two things: the partition we have obtained (by the application of some clustering technique) and the collection of all proximities between objects. For each object i we will introduce a certain value s(i), and then these numbers are combined into a plot[7].

For For a given cluster, Xj (j = 1, . . ., c), the silhouette technique assigns to the ith sample of Xj a quality measure, s(i) (i = 1, ..., m), known as the silhouette width. This value is a confidence indicator on the membership of the ith sample in cluster Xj and it is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{u(i), b(i)\}} \qquad (1)$$

where a(i) is the average distance between the ith sample and all of the samples included in Xj; and b(i) is the minimum average distance between the ith sample and all of the samples clustered in Xk (k = 1, . . ., c; k # j). From this formula it follows that s(i) has a value between -1 and 1 [8][9].

Silhouette ranges from −1 and 1, and it is a measure of how appropriately the data has been clustered (the greater the value, the better the partition). In particular, singleton clusters (for which the silhouette is always equal to one) are not included in the average computation across clusters because they inflate incorrectly this value[10][11][12].

E. Silhouette Clustering

K-means is a method of clustering that has as objective the partition of a set of n observations on k groups in which each observation belongs to the group whose average value is closest.

Given a set of observations (x1, x2,…, xn), where each observation is a real vector of d dimensions k-means builds a partition of the square within each group (WCSS): S = {S1,S2,…,Sk}, as is it described below:

$$\arg\min \sum_{i=1}^{k} \sum_{x=S_i} ||x - \mu||^2 \qquad (2)$$

Where μ_i is the average of points in S_i.[13]

However, this algorithm can only be applied to numerical attributes and the outliers can affect the performance and produce strange behavior [14]. The pseudocode of k-means is shown in Algorithm 1 has was be taken from [14].

K-means pseudocode.

1. Select an initial partition with k clusters; repeat steps and 3 until cluster membership stabilizes.
2. Generate a new partition by assigning each pattern to its closest cluster center.
3. Compute new cluster center.

EXPERMENTAL SETUP AND RESULTS

This paper aims to propose a teaching model to acquire, process and analyses non-linear signals through unsupervised clustering, this with the purpose of finding features similar [3] [15] in the performance of tasks wherever It is possible to measure the significant variations, wear, work and/or areas of execution, etc. [2]. The focus of acquisition, processing and grouping of signals, are based on an emulation of brain activity as measured by a voltage measurement system as mentioned in the description of the theoretical description and functioning of the EEG.

a) Acquisition data

It intends to collect data based on voltage along time. Voltages were collected through 16 analog inputs of an Arduino Mega 2560. Across the analog inputs is counted variation in voltages by 16 light sensors, called photo-resistances, which ranged in the values of voltage when were stimulated by a white color LED light between the values of 0 and 5 volts.

The photo-resistances were placed so strategic in a dummy head with the distribution at the interface of brain machine Brain-Computer-Interface (BCI), this experiment used the EMOTIV EPOC headset 14 channels of distribution. The photo-resistances vary from voltage to be stimulated or not by Ray of light, in the case of be stimulated with light voltage value increases and lacking stimulus of light this decrease in Figure 1 shows the distribution of the BCI EMOTIV EPOC. Table 1 shows the number of the analog input with respect to the EMOTIV EPOC sensor.

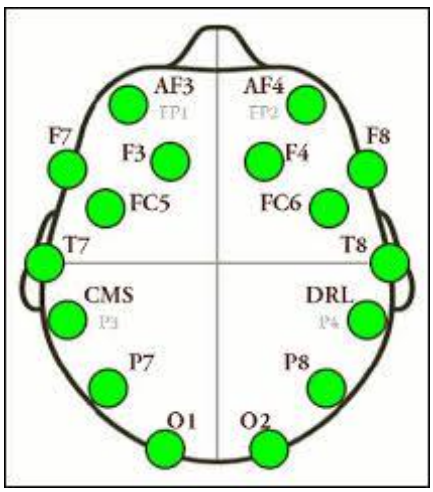

Figure 1. Distribution of sensors of EMOTIV EPOC 14 chanels.

Table 1. Analogic in Arduino - sensor EMOTIV EPOC

Analogic in	Sensor EMOTIV EPOC
A0	F4
A1	AF4
A2	F8
A3	TC6
A4	T8
A5	DRL
A6	P8
A7	O2
A8	O1
A9	P7
A10	CMS
A11	T7
A12	FC5
A13	F7
A15	AF3

For this experiment were used two different light stimuli; the first experiment consisted of white light and did a sweep that was executed manually with pendulum movement; the second experiment had variations on the pulses emitted by the led, it was programmed with an SOS Morse code signal and sweep was like the previous.

The duration of data extraction time was 60 seconds with a collection of values of voltage every two milliseconds per sensor, this was controlled with a program made in MATLAB. The program also had the ability to display signals in real time on a computer through the serial port, generate the log of the 16 sensors and segmenting them into a vector with length of 1 x 300 for each sensor. Figure 2 shows the behavior of the voltages in real time.

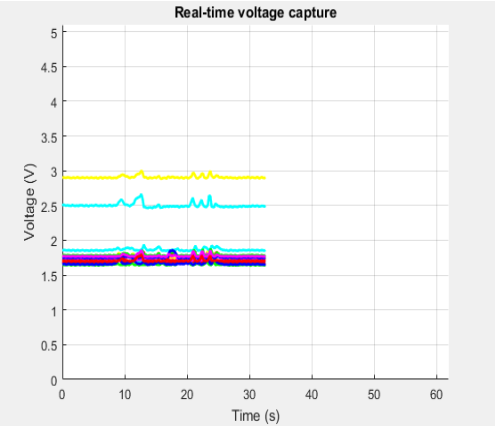

Figure 2. Real-time voltage capture

The experiment was repeated 35 times with SOS flashing light and another time with fixed light having total 70 samples to analyze. Figure 3 shows the final design of the experiment in data acquisition.

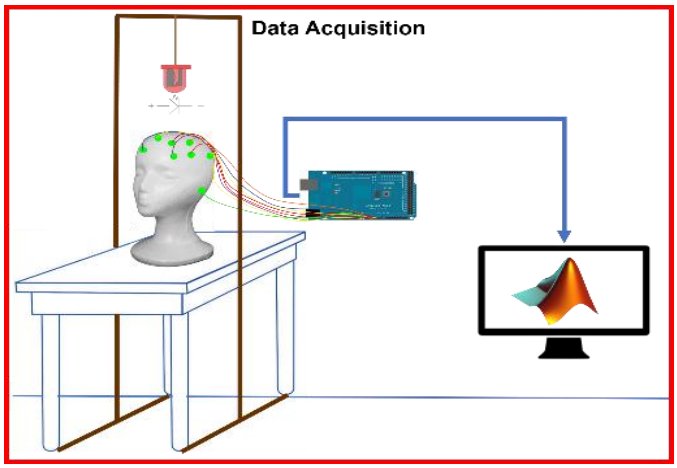

Figure 3. Data acquisition design

b) Characteristics and behavior of the samples

With the total number of samples taken were created two databases the first named dataC (data with continuous light) and the second dataSOS (light S.O.S. data) each with dimensions of 560 x 300 where 560 are the resulting number of rows of 16 test voltage ratings and multiplied by 35 that is the number of repetitions per experiment and the number of columns (300) corresponds to the voltage values collected every 2 milliseconds for 60 seconds. Figure 4 shows the behavior of the types of experiment.

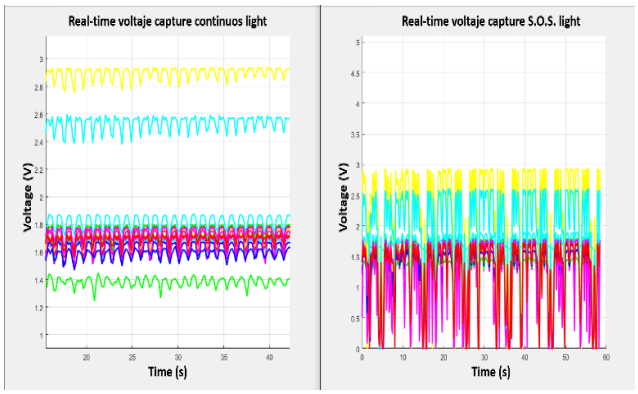

Figure 4. Continuous Light and S.O.S. Behaviors

To review the final data sets detected that the last 8 values, belonging to each vector voltages had values assigned at 0's, after repeating a couple of times the experiment in isolation. was completed that processing speed between the computer and the Arduino board were not equal and had a desfasamento of time. Therefore, we conclude that they gathered information at right time but each one with out-of-date speed, that's that when the computer processor core i5 (2.30 GHz) would finish register, the Arduino Mega from lower speed (16 MHz) processor about how many milliseconds it took to finish. For

this reason, were the first 291 values of voltage for all samples taken and the dataC databases and dataSOS were finally left with dimensions of 560 x 291. Figure 5 shows the process of acquisition and grouping of tests (create database).

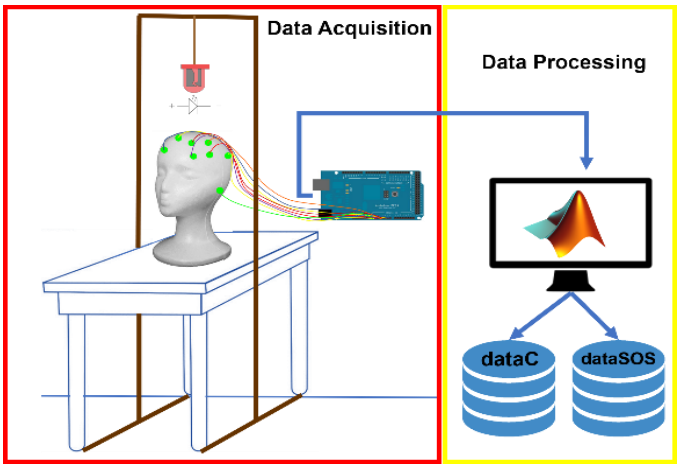

Figure 5. Databases and data processing

c) stage of clustering and adjustment of parameters

Whereas the databases that we have so far, it will undergo a process of grouping to find significant characteristics in common to be able to relate behavior and/or contain parameters that are of particular or general [3] for each instance of the database with their respective values of voltage over time.

The following methodology is proposed for this grouping process.

APPLY EVALUATION GAP TO DATABASE

The search for the best parameter will be k - optimal for best results in the clustering. The data will be processed in raw, because when adjusting our signal will lose important data and us we care about full signal to have a wide view over the behavior of the data and the proposed experiment. You choose a range of 1-16 for the evaluation of the best k, this for the two databases, you choose to evaluate k 1-16 hoping that the values of each of the sensors can be separated or grouped them into areas. Other purposes achieve a grouping of signal and see if behaviors reflecting on a sensor can reflect on another with equal or similar behavior. Figure 6 shows the graph of the evaluation in this range of k, for the continuous pulse database. (dataC) with a KOptimo = 6. Figure 6 shows the graph of the evaluation in this range of k, for the intermittent pulses S.O.S. database. (dataSOS) with a KOptimo = 16.

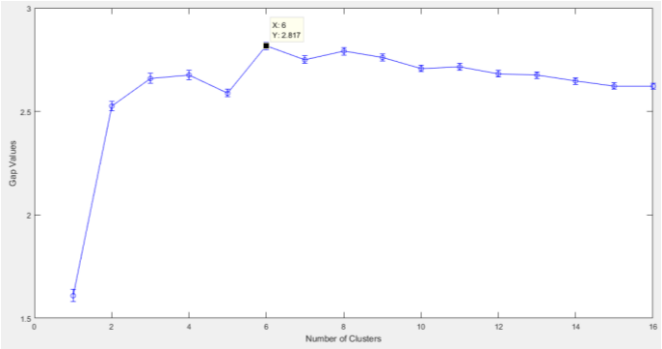

Figure 6. GAP for database “dataC” with continuous light

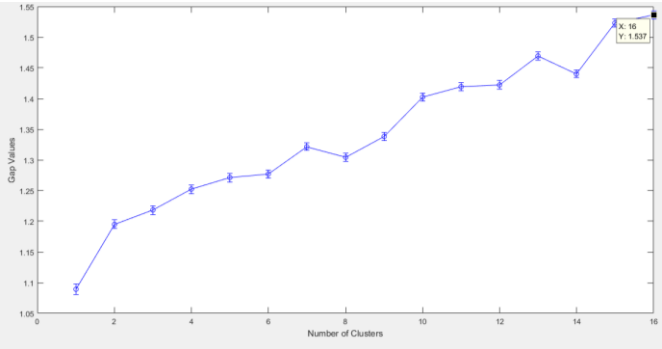

Figure 7. GAP for database “dataSOS" with signal S.O.S.

1. Apply k-means and print the silhouettes of the clusters assigned by the k-means algorithm:

Shall be a method of grouping k-means with the intention of verifying if the number of clusters proposed by GAP offers satisfactory results. Be subsequently displayed the silhouettes graphs resulting from applying k-means with a k-optimal¨, if the silhouettes of each cluster are approaching the 1, then will conclude that the grouping is good with the k value. Otherwise if the silhouette has a negative or very distant 1 [9] [10] [11], will conclude that the evaluation by the GAP and k-means clustering algorithm is not good for the type of problem, the amount of data or be better try another number of k-optimal.

In Figure 8 you can see that the cluster 5 proposed by the GAP are grouping positively and the other does not have the certainty of being in the correct cluster.

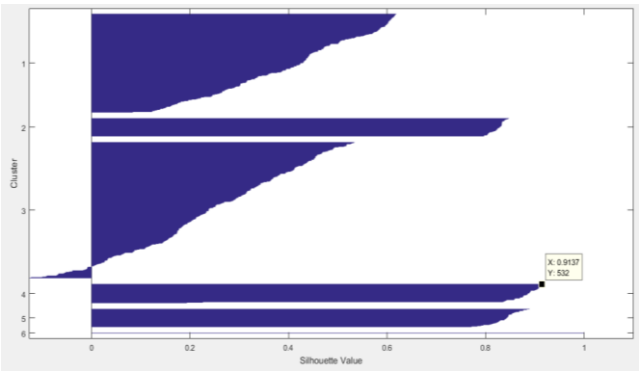

Figure 8. Silhoutte for dataC with 6 clusters

In Figure 9 you can see 12 clusters proposed by the GAP are grouping positively and the other 4 do not have the certainty of being in the correct cluster. In this case the clusters that are closer in his characteristic group are above 0.6.

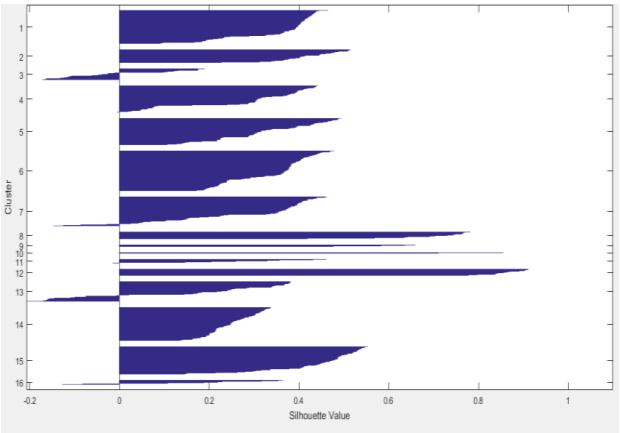

Figure 9. Silhoutte for dataSOS with 16 clusters

ANALYSIS OF RESULTS

The results obtained with the techniques of clustering and validation of the k "optimal" methods of GAP statistic and Silhouette, will be compared with information referred to, i.e., the placement of the sensors, the type of lighting and the trajectory of the pendulum help us to interpret the results.

According to the results shown in figures 8 and 9, we find a significant difference in terms of the number of clusters formed in each dataset.

The first corresponding to the figure 8 experiment and a k = 6, shows fewer cluster by the type of lighting, to which he was subjected during the first experiment. In addition, two areas marked in the cluster number 5 and 6 can be identified. The same two sensors in the original space placed on dummy are far apart from the 12 remaining sensors. Figure 10 shows the original arrangement of sensors with better separation in the original space.

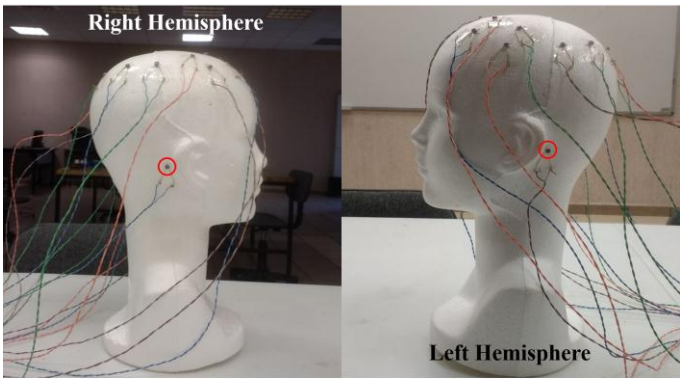

Figure 10. Sensors farther away in manikin

The second experiment denotes a greater number of k = 16, this is due to the behavior of light applied with variations in time and different spaces of light projected onto the mannequin, for this reason each higher number of groupings can be removed. Similarly, you can find 2 sites with similar behavior in the cluster"8" and "cluster 12" determined by sensors farthest from the top of the mannequin., same which you can see in Figure 10.

CONCLUSIONS

After ending the experiment described in the "experimental setup and results" section, we can reach the following conclusions.

According to the figures 8 and 9, you can see that with the use of unsupervised clustering techniques (k-means) and the use of clustering techniques based on probability and statistics, we find significant and similar behaviors in the analysis of non-linear signals. With these methods acceptable approximations are reached when behavior is not known, or when you have a slight idea of how you can react the phenomenon being studied.

It is important to mention that different light pulses applied in experiments and the evaluation of k optimum were important parameters to infer that areas or regions have greater number of features in common and which have a high activity of coincidence with each other which makes it difficult to analyze information in a visible way. Despite the large amount of data that was used for these experiments, superior results were obtained by not altering the nature of these same transformations, dimensions and standards. These if single in its original form shows the true nature with actual values and their own noise values that give the original meaning of the phenomenon under study.

As future work, this experiment is intended to replicate to work with biological signals extracted from clinical character measuring equipment to find the behavior that you must perform some task specific and if it is possible to find patterns to detect some or several pathologies and thus be able to characterize these behaviors to provide efficient and low-cost analysis for the people who most need it.

ACKNOWLEDGMENT

Author whould like to acknowledge CONACYT, TecNM and Instituto Tecnológico de León and friends for their unconditional support to this research. This work was supported in grand part for GIIA (Grupo Interdisciplinario de Investigación Aplicada) of DEPI (División de Estudios de Posgrado).

REFERENCES

[1] R. B. Pachori and P. Sircar, “EEG signal analysis using FB expansion and second-order linear TVAR process,” *Signal Processing*, vol. 88, no. 2, pp. 415–420, 2008.

[2] C. Ling, H. Goins, a Ntuen, and R. Li, "EEG signal analysis for human workload classification," *IEEE SoutheastCon 2001, Mar 30-Apr 1 2001*, pp. 123–130, 2001.

[3] U. Orhan, M. Hekim, and M. Ozer, "EEG signals classification using the K-means clustering and a multilayer perceptron neural network model," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 13475–13481, 2011.

[4] M. T. G. González *et al.*, "Analysis of pupillary response after a stimulus of light to generate characteristical groups," in *2017 International Conference on Electronics, Communications and Computers, CONIELECOMP 2017*, 2017.

[5] Z. Hongwei, Z. Xuehua, and Z. Bao'an, "System Dynamics Approach to Urban Water Demand Forecasting," *Trans. Tianjin Univ.*, vol. 15, no. 1, pp. 70–74, 2009.

[6] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, vol. 63, no. 2, pp. 411–423, 2001.

[7] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, 1987.

[8] N. Bolshakova and F. Azuaje, "Improving expression data mining through cluster validation," *Proc. IEEE/EMBS Reg. 8 Int. Conf. Inf. Technol. Appl. Biomed. ITAB*, vol. 2003–Janua, no. i, pp. 19–22, 2003.

[9] H. Ghasemzadeh and R. Jafari, "Coordination analysis of human movements with body sensor networks: A signal processing model to evaluate baseball swings," *Sensors Journal, IEEE*, vol. 11, no. 3, pp. 603–610, 2011.

[10] S. Spadone, F. de Pasquale, D. Mantini, and S. Della Penna, "A K-means multivariate approach for clustering independent components from magnetoencephalographic data," *Neuroimage*, vol. 62, no. 3, pp. 1912–1923, 2012.

[11] H. Liang, Z. Wang, A. Maier, and N. K. Logothetis, "Single-Trial classification of bistable perception by integrating empirical mode decomposition, clustering, and support vector machine," *EURASIP J. Adv. Signal Process.*, vol. 2008, 2008.

[12] C. Arizmendi *et al.*, "Data mining of patients on weaning trials from mechanical ventilation using cluster analysis and neural networks," *2009 Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, pp. 4343–4346, 2009.

[13] A. Vattani, "k-means Requires Exponentially Many Iterations Even in the Plane," *Discret. Comput. Geom.*, vol. 45, no. 4, pp. 596–616, 2011.

[14] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.

[15] S. U. López *et al.*, "Identification of parameters for the study of diabetes from light reflex with controlled stimulus," in *2017 International Conference on Electronics, Communications and Computers, CONIELECOMP 2017*, 2017.