XXVI Congresso Brasileiro de Engenharia Biomédica

Armação de Búzios – RJ – Brasil

October 21st to 25th, 2018

Symbolic dynamics of RR intervals to identify coronary heart disease via supervised learning model: Preliminary results.

C B N Freitas[1*], L dos Santos[2]**

[1] Telefonaktiebolaget L. M. Ericsson, São José dos Campos, Brasil

[2]Instituto Científico e Tecnológico da Universidade Brasil, São Paulo, Brasil

*cbnfreitas@gmail.com

**laurita.santos@universidadebrasil.edu.br

Background, Motivation and Objective. One of the simplest noninvasive ways to analyze Heart Rate Variability (HRV) is to employ RR intervals time series, where each RR represents the difference between two R waves on the electrocardiogram. It is known from medical literature that momentarily stressful or pathological conditions alter Autonomic Nervous System (ANS). In this context, the distinction between stress and pathology is very difficult to detect and classify using only HRV. However, they differ since momentary situation of stress can be reversed and normalized at ANS level. Symbolic dynamics may be a method to differentiate these groups in terms of complexity. This method is widely applied to quantify the non-linear dynamics of the time series. In terms of RR intervals time series, it is expected that HRV of non-healthy groups decreased compared with healthy group. In this work, we train supervised leaning models to distinguish individuals under three groups of heart diagnostics.

Methods. We analyzed in this work a data set of RR interval time series, whose entries are pre-labeled according to individual's heart health condition: healthy (H), pathological conditions (P) and healthy but stressful conditions (S). The groups have 54, 29 and 62 members, respectively; and each time series was pruned to 2290 points, which corresponds to approximately 1 hour. We employ RR symbolic dynamics in association with Shannon Entropy (SE) to determine the features for the machine learning models. This measure may be related with the repetition patterns in RR intervals time series. More specifically, given a time series $\Delta RR = (\beta RR_1, \beta RR_2, \ldots, \beta RR_N)$, where $\beta RR_i = x_{i+1} - x_i$ with $l = 1, 2, \ldots, N-1$, the new time series $S = \{s_1, s_2, \ldots, s_N\}$ was elaborated. Three symbols were used, as following: $s_i = 0$ if $|\beta RR_i| \leq \alpha$; $s_i = 1$ if $\beta RR_i > \alpha$; $s_i = 2$ if $\beta RR_i < -\alpha$, where α is an empirical difference value between two RR intervals. This parameter represents the difference between two consecutive heartbeats, according to the demand of the body for homeostasis balance. After this step, it was obtained the S sequence into a W series, $W = \{w_1, w_2, \ldots, w_{N-(l-1)}\}$, where l is word's size. For this experiment, we evaluated SE considering α over {1, 2, ..., 19, 20, 30, ..., 140, 150} and l over {2, 3, …, 9, 10}, that is, 297 features. We train five supervised learning algorithms (Python/Scikit-learn v0.19.1), with the following parameters:

- RandomForest (RF): max_depth=10
- DecisionTree (DT): criterion = "entropy", max_depth = 3, min_samples_leaf = 5
- SVM (SVM): kernel = 'linear', C = 1
- LogisticRegression (LR): default
- MLP (MLP): hidden_layer_sizes = (20,20,20), max_iter = 0,500,

(all remaining parameters are the default ones). All SE values are employed as features to predict the corresponding label/class (H, P or S). As recommend by some authors, we considered 5 folds cross-validation, including stratification by label. So, model accuracy in our context means the average accuracy regarding all test folds.

Results. We report a reasonable model accuracy for all models to classify heart health condition. The accuracy of each model was: DT: 0.75, RF: 0.77, MLP: 0.81, SVM: 0.85, LR: 0.85. The confusion matrix of LR can be found at the table below. Notice that, confusion matrix displays good classification of the S group from others, but for the C group there are misleading classifications with control group (H). This result suggested that there is discrimination between pathological conditions group from stressful conditions group using SE method.

Tables I. Confusion Matrix of one the model (Logistic Regression - LR), considering three groups healthy (H), pathological conditions (P) and healthy but stressful conditions (S).

Predicted	H	C	S
Actual			
H	48	5	1
C	12	16	1
S	1	1	60

Discussion and Conclusions. Our study shows a fair model accuracy for the 5 supervised learning models. This strongly indicates that Shannon Entropy of Symbolic Dynamics (SE) is a relevant feature to predict heart health condition. However, model accuracy analysis alone can be misleading, as shown by the Confusion Matrix at Table I, mostly due to class imbalance. In fact, accuracy was higher because models were able to correctly identify the class S in almost all cases, which was the biggest class in our data set. As future research, new features could to be introduced to enhance the distinction classes H and C.

Keywords. heart rate variability, noninvasive methods, symbolic dynamics, machine learning, supervised learning models.