XXVI Congresso Brasileiro de Engenharia Biomédica

Armação de Búzios – RJ – Brasil

October 21st to 25th, 2018

Cardiac Arrhythmia Diagnosis Applying Machine Learning Techniques

A L Pereira[1*], L S Sá[1], L A Pinto[1]

[1]Federal Institute of Espírito Santo, Serra, Brasil

*mandspereira@gmail.com

Background, Motivation and Objective. Cardiac arrhythmia is a health problem that affects a large number of people around the world. Such problem consists of alterations in the normal sequence of electrical impulses that control the heartbeats, causing abnormal rhythms of functioning. Under the arrhythmia condition, the heart may present very fast beats (tachycardia), very slow beats (bradycardia), or even completely irregular beats that can oscillate between fast and slow in short time intervals. The precise determination of the type of arrhythmia is an important condition in specifying the most appropriate treatment. However, preparing the diagnosis may not be a simple task, even for the most experienced experts. In many cases, the pathology does not present apparent symptoms. In addition, the diversity of types is another factor that can make diagnosis difficult. To make an accurate diagnosis, experts analyze the outcomes of medical exams, such as, echocardiogram, stress test, holter and, mainly, the electrocardiogram. Such tests investigate problems related to the functioning and the heart anatomy. In addition, lifestyle related factors are considered to be associated with episodes of arrhythmias. In view of the difficult in making good diagnosis as previously stated, this work, still in progress, investigates the cardiac arrhythmia diagnosis applying machine learning techniques.

Methods. The development of this research consists of the following steps: *(1) Data Acquisition:* The experimental phase is accomplished using the *dataset* "Arrhythmia", which can be freely obtained on the UCI-Machine Learning Repository website. The dataset consists of a fusion between outcomes of medical exams and information related to the patient's lifestyle. In its original form it consists of 452 samples and 279 attributes (206 numerical and 73 nominal), distributed in 16 classes. Class 1 consists of data from healthy patients, and the remaining 15 are associated with patients with different types of arrhythmias. *(2) Pre-processing:* To fix problems in the dataset structure, in all tests the classes 11, 12 and 13 are removed because they lack samples. In the same way, column 14 is removed because it has 84% of unknown values among the 452 attributes. *(3) Modelling:* In the experimental stage the following tests are performed: *(3.1) Two class problem:* The initial tests investigate the capability of the models to identify the occurrence of cardiac arrhythmias. For this stage, classes 7, 8, 14 and 15 are also removed because they have a reduced number of samples, generally less than 9, which would make it impossible to model these classes. With the samples removed, the dataset samples are reorganized into two classes, the Class 1 corresponding to healthy individuals and the Class 2 corresponding to individuals with arrhythmia. Since some samples have unknown attribute values, it is decided to investigate the use of Principal Component Analysis to estimate such values. *(3.2) Multiclass problem:* The multiclass problem consists in identifying the occurrence and the type of the arrhythmia. On the first test, Principal Component Analysis is used to estimate unknown values of attributes, as well as to eliminate variables that are poorly correlated with the sources of variability associated with the phenomena to be modelled. For this case, the dataset consists of 452 samples and 59 attributes, distributed in 13 classes. In the second test, unknown values and classes with reduced number of samples (7, 8, 14 and 15) are eliminated and Principal Component Analysis is applied to

reduce the size of the dataset. The dataset used in the modelling step has 412 samples and 53 attributes, distributed in 9 classes. *(3.3) Classification by sex:* the third phase investigates the occurrence of arrhythmias considering the gender of the patient. Samples of female and male patients are organized in separate *datasets*, which corresponds to a total of 234 and 178 samples, respectively. For these tests the attributes with unknown values and classes with reduced amounts of samples are eliminated. *(4) Classification:* all models are constructed using k-Nearest-Neighbour (kNN) and Support Vector Machine (SVM) algorithms. In each case 70% of the samples are used to train the models and 30% in the test phase. The performance of the models is measured by using the correct classification rate. All tests are accomplished using the Matlab functions.

Results. *Two class problem:* samples from classes 7, 8, 14 and 15 are removed in all tests. Using the PCA to estimate the unknown attribute values, the best result obtained with the k-NN and SVM classifiers are 84.88% and 76.69%, respectively. When attributes with unknown values are removed, the correct classification rate for k-NN and SVM are, respectively, 84.47% and 73.38%. In both cases, the k-NN classifier settings are 3 neighbours, Euclidean distance and uses the exhaustive search method. *Multiclass problem:* keeping the samples of classes 7, 8, 14 and 15 and using the PCA to estimate unknown attributes values, the best result with the k-NN and SVM classifiers are, respectively, 74.12 of 67.4%. When the samples of classes 7, 8, 14 and 15 and the attributes with unknown values are removed, the correct classification rate obtained with the classifiers k-NN and SVM are, respectively, 82.52% and 66.40%. In both cases the k-NN classifier settings are 2 neighbours, Euclidean distance and uses the exhaustive search method. *Classification by sex:* when the analysis considers the gender of the patient, the best result for the k-NN classifier is, for males 78.65% and for female 94.03%. In turn, the results with SVM are, 60.34% and 70.0% for the male and female sex, respectively.

Discussion and Conclusions. This work investigates the use of pattern recognition techniques to assist specialists in identifying the type of arrhythmia in cardiac patients. The classification models are implemented using the k-NN and SVM classifiers, and for the experiments the dataset "Arrhythmia" is used, which gathers information from male and female patients obtained from medical examinations, as well as information related to the lifestyle of the patients. It can be observed from the results that in all tests scenarios the best performance is obtained with k-NN classifier, and that among the scenarios tested, the best result is obtained with the two-class problem. We can justify this result by arguing that as the used dataset presents a high level of imbalance, in addition, some classes have reduced quantities of samples. Such conditions are not favourable for good performance of the SVM classifier, which considers the distribution of probabilities of the samples. On the other hand, the performance of the k-NN classifier, whose decision is based on the distance between samples, is influenced a bit by the imbalance of the classes. As expected, the average performance of the classification improves slightly when the analysis is done by sex, which confirms the theoretical knowledge that cardiac diseases affect men and women differently. Subsequently, this paper will examine the application of Partial Least Square and Deep Learning techniques to classify the types of cardiac arrhythmias on the Arrhythmia dataset.

Keywords. Cardiac arrhythmia, machine learning, support vector machine, nearest neighbour.