First published in Cardiologia Croatica. 2014;9(3-4):99. Republished with permission.

Extended abstract

Descriptive and predictive modeling

Dragan Gamberger[1]*

[1] Rudđer Bošković Institute, Zagreb, Croatia

* Corresponding author phone: +385 (1) 456-11-11, e-mail: dragan.gamberger@irb.hr

Introduction Intelligent data analysis (IDA) is an important tool for data based medical research. It is often combined with statistical techniques. The primary goal of IDA is data understanding and hypothesis creation while statistics is used for hypotheses validation. Aim: Presentation of novel approaches for pattern recognition, example clustering, and data understanding tasks in cardiological applications.

Methods A new generation of machine learning algorithms, like Random Forest [1] and Random Rules, is based on efficient and systematic construction of many independent classifiers. Besides increased predictive quality, the algorithms have some distinguished properties like inherent estimation of the probability of correct classification for each example and estimation of the similarity of pairs of examples. The probability of correct classification is measured by the difference in the number of votes for different classes while similarity of examples is measured by the percentage of classifiers that correctly predict both examples. These new options are potentially very useful for pattern recognition and example clustering results that may be applied for the predictive and descriptive analysis of medical data.

Results The problem of fetal heart rate (FHR) monitoring is known as an important and difficult problem [2]. Typically it is solved by sophisticated signal processing techniques. We demonstrate how it may be solved by transforming the four simultaneous noninvasive fetal ECG signals into a single probability sequence. From such probability sequence it is much easier to identify positions of actual fetal QRS complexes. A model based on many independent classifiers is used for this transformation. The starting point for constructing the model is a few ECG sequences on which medical experts have already identified the positions of fetal QRS complexes.

The second application is clustering of coronary heart disease (CHD) patients. The experiments start from a set of 238 patients. A classification problem is formed so that the original set of examples is used as positive examples while negative examples are obtained by randomized shuffling of attribute values of the original set. The goal of constructing a predictive model for discriminating original from randomized data is to obtain a similarity table for original examples. In the second step clusters of examples are iteratively constructed by minimization of the internal variance of the similarity table for all 238 examples. The main result is identification of relevant subgroups of CHD patients, description of properties of these subgroups, and detection of outlying examples.

Conclusion The presented methodology and its appropriate applications enable novel approaches to data analysis in cardiology. The applications that are discussed in this work are only an illustration of its potentials. A long term goal is better understanding of the methodology and its comparative evaluation with standard techniques.

Keywords Pattern recognition • Clustering of examples • Data understanding • Random forest algorithm • Example similarity table

Literature

1. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.

2. Sameni R, Clifford GD. A review of fetal ECG signal processing: issues and promising directions. Open Pacing Electrophysiol Ther J. 20103:4-20