Biomedical Signal Processing and Control 19 (2015) 130–136

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Genetic algorithm-based method for mitigating label noise issue in ECG signal classification

Edoardo Pasolli[a,*], Farid Melgani[b]

[a] School of Civil Engineering, Purdue University, 47907 West Lafayette, IN, United States
[b] Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy

ARTICLE INFO

Article history:
Received 14 March 2014
Received in revised form
24 September 2014
Accepted 28 October 2014
Available online 27 November 2014

Keywords:
ECG signal classification
Genetic algorithms
Mislabeling issue
Multiobjective optimization
Training sample validation

ABSTRACT

Classification of electrocardiographic (ECG) signals can be deteriorated by the presence in the training set of mislabeled samples. To alleviate this issue we propose a new approach that aims at assisting the human user (cardiologist) in his/her work of labeling by removing in an automatic way the training samples with potential mislabeling problems. The proposed method is based on a genetic optimization process, in which each chromosome represents a candidate solution for validating/invalidating the training samples. Moreover, the optimization process consists of optimizing jointly two different criteria, which are the maximization of the statistical separability among classes and the minimization of the number of invalidated samples. Experimental results obtained on real ECG signals extracted from the MIT-BIH arrhythmia database confirm the effectiveness of the proposed solution.

1. Introduction

In the last decades, growing attention has been given in the biomedical engineering community to the problem of automatic analysis of electrocardiographic (ECG) signals. The great interest for ECG analysis derives from its role as an efficient and noninvasive tool for detecting and monitoring cardiac diseases. In particular, significant effort has been spent in the development of efficient and robust systems for ECG signal classification in order to detect automatically heartbeat abnormalities.

For such purpose, different solutions based on pattern recognition approaches have been proposed in the literature. Most of the attention has been given on improving the accuracy of the classification process by acting mainly at two different levels: (1) signal representation and (2) optimization of the discriminant function. In terms of signal representation different types of features have been extracted from the acquired ECG signals in order to have a better discrimination among the classes. Some examples of features are given by high-order statistics [1], morphological features [2], temporal intervals [2–4], wavelet transform coefficients [3–5], frequency domain features [6], and statistical features [7]. Moreover, given the high number of features that is usually involved, some feature reduction techniques have been applied to project the features into a lower dimensional feature space, such as principal component analysis [4,8] and independent component analysis [8]. The problem of discriminant function optimization has been addressed by considering different classification approaches. Although linear models have shown good results [2], in the last few years more attention has been given to nonlinear approaches. In particular, different works have focused on neural networks [3,4,8,9] and kernel methods such as support vector machines (SVMs) [1,5,7,8]. Moreover, classification improvements have been obtained by combining classifiers with optimization processes, such as those based on particle swarm optimization (PSO) [10,11].

Although these works have demonstrated their effectiveness, they are based on an essential assumption that is the samples used to train the classifier are statistically representatives of the classification problem to solve. Therefore the quality and the quantity of such samples are very important, because they have a strong impact on the performance of the classifier. However, the process of training sample collection is not trivial since it is based on a human user (cardiologist) intervention and so it is subject to errors and costs both in terms of time and money. In general, scarce attention has been given to this problem in the literature. Only in the last few years there has been a growing interest in developing semi-automatic strategies for the problem of training set construction.

* Corresponding author. Tel.: +1 7654097937.
E-mail addresses: epasolli@purdue.edu (E. Pasolli), melgani@disi.unitn.it (F. Melgani).

http://dx.doi.org/10.1016/j.bspc.2014.10.013
1746-8094/© 2014 Elsevier Ltd. All rights reserved.