Pattern Recognition Letters 70 (2016) 45–51

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Cardiac arrhythmia classification using statistical and mixture modeling features of ECG signals ☆

Rashid Ghorbani Afkhami*, Ghanbar Azarnia, Mohammad Ali Tinati

Faculty of Electrical Engineering, University of Tabriz, Tabriz, Iran

ARTICLE INFO

Article history:
Received 18 March 2015
Available online 2 December 2015

Keywords:
Decision tree
Expectation maximization
Gaussian mixture model
Heartbeat classification
Higher order statistics

ABSTRACT

In this paper we propose a novel method for accurate classification of cardiac arrhythmias. Morphological and statistical features of individual heartbeats are used to train a classifier. Two RR interval features as the exemplars of time-domain information are utilized in this study. Gaussian mixture modeling (GMM) with an enhanced expectation maximization (EM) solution is used to fit the probability density function of heartbeats. Parameters of GMM together with shape parameters such as skewness, kurtosis and 5th moment are also included in feature vector. These features are then used to train an ensemble of decision trees. MIT-BIH arrhythmia database containing various types of common arrhythmias is employed to test the algorithm. The overall accuracy of 99.70% in "class-oriented" scheme and 96.15% in "subject-oriented" scheme is achieved. Both cases express a significant improvement of accuracy compared to other methods.

1. Introduction

The analysis of the electrocardiographic (ECG) signal provides detailed information on the patient's cardiac health status. Cardiac arrhythmias are group of conditions in which the electrical activity of the heart is irregular, faster or slower than normal. While some types of arrhythmias impose an imminent threat on patient's life (e.g. ventricular fibrillation and tachycardia) other types represent long-term threats, which will still need special care. Cardiac arrhythmias are also one of the most common causes of death and as a result, their accurate detection has been of great interest in biomedical studies.

ECG is a simple and specific diagnostic test for assessment of heart rhythms. In most cases arrhythmias don't show any consistent changes in ECG signal, therefore a Holter monitor is needed. Holter monitor is an ECG recorded over a longer period time (i.e. longer than 30 mins), which helps the detection of dysrhythmias that may happen briefly and unpredictably throughout that period. While analyzing such a long signal by physicians is extremely time demanding, signal processing techniques can offer rapid, real-time categorization with acceptable accuracies. Our goal in this paper is to introduce a legitimate method for automatic detection of cardiac arrhythmias based on novel techniques in signal processing.

Electrocardiographic based classification of cardiac arrhythmias has been investigated in numerous papers. These methods can usually differ in three main aspects, features, classifiers and evaluation schemes.

Different features used in papers are Hermit coefficients [12,14,20], higher order statistical features [7,20], morphological features [6,8,27], independent component analysis and wavelet features [10,13,24,25].

Classifiers such as self-organizing map (SOM) [14], support vector machine (SVM) [13,20,25], artificial neural network (ANN) [10,12], conditional random field (CRF) [7], linear discrimination analysis (LDA) [6,15], and ensemble methods [27] are considered in different papers.

Two different evaluation schemes, namely "class-oriented" and "subject-oriented", are used in literature. Heartbeat segmentation of extensive ECG signals such as MIT-BIH arrhythmia records, results in more than a thousand heartbeats for an individual record. As a result, it is highly probable that in random selection of training set for a supervised classifier there will be samples of the same patients used in both training and testing sets. This method is referred to as "class-oriented" and it is criticized for poor generalization performance. While class-oriented method is used in many papers (e.g., [8,13,14,20,21,22]), having particularly correlated samples in training and testing sets may cause the overfitting problem and lead to promising results that might not be reachable in practice. In contrast, subject-oriented method, which was proposed by [6], creates a patient-based division of dataset into training and testing sets, i.e. prior to heartbeat segmentation procedures. This method is believed to be more realistic and is adopted by [7,15,25].

Although promising results have already been seen in cardiac arrhythmia classification, there is a long way to go before applicable

☆ This paper has been recommended for acceptance by Jie Zou

* Corresponding author. Tel. +9 8413 339 3721.
E-mail address: r.ghorbani91@ms.tabrizu.ac.ir, ghorbani.rashid@gmail.com (R. Ghorbani Afkhami).

http://dx.doi.org/10.1016/j.patrec.2015.11.018
0167-8655/

methods and totally automatic classifications can be employed in practice. This means more accurate and stable algorithms should be developed.

The paper is organized as follows: Section 2 describes the dataset, Section 3 provides background materials and methods and Section 4 represents the proposed method; in Section 5 we share the simulation results and compare it with previous works and Section 6 has the conclusion.

2. Background

2.1. Higher order statistics

Probability distribution moments are the generalization of concept of the expected value and can be used to define the characteristics of probability density function [9].

Skewness (3rd order statistics) and kurtosis (4th order statistics) of random variable, $\boldsymbol{x}$, are defined respectively as follows,

$$\gamma_3 = \frac{E\left\{[x - E(x)]^3\right\}}{\left(E\left\{[x - E(x)]^2\right\}\right)^{3/2}}, \tag{1}$$

$$\gamma_4 = \frac{E\left\{[x - E(x)]^4\right\}}{\left(E\left\{[x - E(x)]^2\right\}\right)^{2}} - 3. \tag{2}$$

in which E denotes the expected value. While skewness gives a measure of the lopsidedness of the distribution, kurtosis gives a comparative measurement of the signal's distribution with normal distribution of the same variance. Estimates of the two are given respectively by,

$$\hat{\gamma}_3 = \frac{\sum_{i=1}^{N}\left(x_i - \hat{m}\right)^3}{(N-1)\hat{\sigma}^3}, \tag{3}$$

$$\hat{\gamma}_4 = \frac{\sum_{i=1}^{N}\left(x_i - \hat{m}\right)^4}{(N-1)\hat{\sigma}^4} - 3, \tag{4}$$

where x_i is a realization of the random variable x. $\widehat{m}$ and $\widehat{\sigma}$ are estimates of mean and variance for the observed signal with N samples.

2.2. Gaussian mixture model

A Gaussian mixture model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs with many degrees of freedom allow arbitrary density modeling. In this model, data $\{x_1, x_2, \ldots, x_n\}$ in R^d are assumed to arise from a random vector with density,

$$f(\boldsymbol{x}) = \sum_{i=1}^{K} p_i \phi(\boldsymbol{x}|\mu_i, \Sigma_i). \tag{5}$$

in which, K is the number of mixture components, p_i is the mixture weight, i.e. prior probability of a certain component and $\phi(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$ is the normal distribution with mean vector and covariance matrix of $\boldsymbol{\mu}$ and Σ, respectively. Generally, the mixing parameters $\theta = (p_1, \ldots, p_k, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k)$ are estimated by maximizing the log-likelihood,

$$L(\theta|x_1, \ldots, x_n) = \sum_{j=1}^{n} ln\left[\sum_{i=1}^{K} p_i \phi\left(x_j|\mu_i, \Sigma_i\right)\right]. \tag{6}$$

A widely used approach for maximum likelihood solution is the Expectation Maximization (EM) algorithm. EM is an iterative algorithm that starts from an initial parameter θ^0. The E step computes conditional probabilities of $\hat{p}_i(x_j)$ where $(1 \le i \le K, 1 \le j \le n)$ and x_j emerges from the ith mixture component and the M step updates $\hat{p}_i, \hat{\mu}_i, \hat{\Sigma}_i$ with maximum likelihood estimation using $\hat{p}_i(x_j)$ as conditional mixing weights [4].

2.3. Decision trees

A Decision tree (DT) is a classifier expressed as a recursive partition of the instance space. Discernibility, ability to handle missing attributes, having characteristics of non-parametric classifiers and the implementation simplicity are the main advantages of DTs. A decision tree is made of nodes and edges. The tree starts from a node called *root node*, which has no incoming edge. Nodes with outgoing edges are called *internal nodes* or *test nodes* and all the other nodes are called *leaves* or *terminal nodes*. Internal nodes are a split of their input space into two or more subspaces according to a certain discrete function of the input attributes values. Moreover, each leaf node may hold a probability vector indicating the probability of the target attribute having a certain value. According to the decisions made by internal nodes, instances are navigated from the root of the tree down to a leaf and then classified. This process faces the task of recursive partitioning the input space. The input space is generally represented by a training set like as

$$\mathcal{L} = \{(x_n, y_n)|n = 1, \ldots, N\}. \tag{7}$$

This set includes N instances which are represented by a feature vector x and its associated class, y. After training the tree with the set $\mathcal{L}$, the new instances propagate through the tree and assign to the class which the leaf belongs [16–26].

2.4. Ensemble learners

Ensemble is a supervised learning algorithm that combines several *weak learners* (also called *base learners*) which are slightly better than random guessing to construct one *strong learner*. Integration of these weak learners can be done by various methods, like as majority voting or weighted aggregation of individual results. It is evident that combining several identical classifiers provides no gain, so this method can be useful only if there is dissimilarity among learners. Thus, the diversity of weak learners is crucial in the efficiency of ensemble learners. Diversity can be achieved by using different presentations of the input data, as in bagging. Variations in learner design and adding a penalty to the output are other methods to encourage diversity [16,19,26].

Bagging was proposed by [5] and it is the most well-known *bootstrap* ensemble method that processes samples concurrently. Bagging generates individuals for its ensemble by training each inducer on a random reconstruction of the training set. The training set for a particular classifier is generated by randomly drawing N samples with replacement, where N is the size of original training set. As a result, each individual learner in the ensemble is configured with a different random sampling of the training set. This means that the initial training samples might appear zero or multiple times in a certain training set. Bagging combines these learners by majority voting, i.e. the most voted class is selected. The obtained strong learner will have the misclassification error lower than a particular inducer because the dissimilarity among these learners can compensate for the increase in error rate of each single learner. Breiman notes that bagging is more effective on *unstable* learners because it can remove their uncertainty. In this context, a learner is considered unstable if perturbing the training set can result in large changes in predictions. He claimed that decision trees are example of unstable learning algorithms.

3. Method

This section introduces the methodology used in the paper. Firstly, we talk about the dataset and then we follow the overall processing steps illustrated in Fig. 1. After preprocessing, which is baseline wandering removal and beat segmentation, we extract three sets of features; higher order statistical features, RR interval features and