Pattern Recognition 42 (2009) 2979 – 2988

Contents lists available at ScienceDirect

Pattern Recognition

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

A two-stage mechanism for registration and classification of ECG using Gaussian mixture model

Roshan Joy Martis[a], Chandan Chakraborty[a,*], Ajoy K. Ray[a,b]

[a]*School of Medical Science and Technology, I.I.T. Kharagpur, India*
[b]*Department of Electronics and Electrical Communication Engineering, I.I.T. Kharagpur, India*

ARTICLE INFO

Article history:
Received 30 August 2008
Received in revised form 31 December 2008
Accepted 16 February 2009

Keywords:
ECG
Pan Tompkins algorithm
Linear prediction
PCA
GMM
Chernoff bound
Bhattacharya bound

ABSTRACT

An automatic classifier for electrocardiogram (ECG) based cardiac abnormality detection using Gaussian mixture model (GMM) is presented here. In first stage, pre-processing that includes re-sampling, QRS detection, linear prediction (LP) model estimation, residual error signal computation and principal component analysis (PCA) has been used for registration of linearly independent ECG features. GMM is here used for classification based on the registered features in a two-class pattern classification problem using 730 ECG segments from MIT-BIH Arrhythmia and European ST-T Ischemia datasets. A set of 12 features explaining 99.7% of the data variability is obtained using PCA from residual error signals for GMM based classification. Sixty percent of the data is used for training the classifier and 40% for validating. It is observed that the overall accuracy of the proposed strategy is 94.29%. As an advantage, it is also verified that Chernoff bound and Bhattacharya bounds lead to minimum error for GMM based classifier. In addition, a comparative study is done with the standard classification techniques with respect to its overall accuracy.

1. Introduction

Electrocardiogram (ECG), generated due to the action potentials developed at the cardiac muscles of the heart, is acquired at the surface of human body. The cardiologist looks into different waves, their amplitude, polarity, etc., in the ECG to infer the underlying disease. The technique gives a guide to prognosis, appropriate care and prevention for the suspected future likely illness. Pattern identification of ECG is gaining importance, since a robust classification algorithm enhances the accuracy in diagnostics.

Several studies have been reported on ECG data analysis based on a number of features and different classification algorithms. The features include R–R interval [2], frequency features [12], time–frequency features [3,4] and higher order cumulants [5]. The classification techniques include discriminant analysis [6], backpropagation neural networks [7] and self-organizing feature maps using neural networks [8], probabilistic neural networks [9], support vector machines [11] and independent component analysis (ICA) [10]. Most of the techniques used in the ECG classification use one or the other clustering algorithm for classification. These clustering algorithms can be divided into partitioned approaches and hierarchical approaches [14]. A partitioned method partitions data objects into a pre-specified number of groups using some optimization algorithm. A hierarchical method creates a nested set of groupings and thus a structured view of data objects.

Commonly used distance measures are Euclidian distance and Mahalanobis distance for data that can be represented in a vector space [14]. For complex data types such as varied length sequences defining a very good similarity measure is very much data dependent. Parametric model based approaches attempt to learn generative models from the statistical distribution of the data with each model corresponding to one particular cluster. In addition to the computational advantage model based clustering provides several other benefits [13]. First each cluster is described by a representative model, which provides the probabilistic interpretation of the clusters. Second online algorithms can be easily constructed for model based clustering using competitive learning techniques. Finally probabilistic models can capture the common underlying dynamics within a cluster for complex data types.

In the proposed methodology, ECG data are subjected to preprocessing and classification. In pre-processing stage, re-sampling, QRS extraction, LP model estimation, segmentation based on R point, residual error signal (RES) computation and dimensionality reduction by principal component analysis (PCA) are performed. In the second phase, the ECG signal is classified by Gaussian mixture model (GMM) based classifier. In addition, the algorithm can explain the common underlying dynamics of the system responsible for data

---

[*] Corresponding author. Tel.: +91 3222 28880; fax: +91 3222 28881.
E-mail address: chandanc@smst.iitkgp.ernet.in (C. Chakraborty).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.02.008

generation. GMM based technique provides higher accuracy compared to K-means and fuzzy c-means clustering, which is substantiated in this paper by using error bound called Chernoff bound (Duda et al., 2007) and Bhattacharya bounds.

2. ECG dataset description

ECG [1] corresponds to a single heart beat consisting of P wave, QRS complex and T wave (Fig. 1), which are approximately 0.2, 1 and 0.1–0.3 mV in amplitude, respectively. The short duration and high amplitude of the QRS complex become more important features [2] for registration. The normal heart beat is called the normal sinus rhythm. Out of several cardiac problems, arrhythmia and ischemia have become frequent worldwide. In fact, 'arrhythmia' refers to an irregularity in the heart rhythm reflected specially in the R–R interval (Fig. 2). And 'cardiac ischemia' (Fig. 3) is another type of cardiac disease leading to the elevation of ST-T segment in the ECG.

In this study, MIT-BIH Arrhythmia and European ST-T datasets (www.physionet.org) on ECG signals are used. The datasets are briefly described here in order to set a two-class disease classification problem using GMM based classifier.

2.1. MIT-BIH Arrhythmia dataset

The MIT-BIH Arrhythmia database contains 48 half hour recordings of two channel ambulatory ECG, obtained from 47 subjects in the year 1975 and 1979 by the Beth-Israel Hospital Arrhythmia Laboratory at Boston. Twenty-four hour ambulatory ECG recordings were collected from a mixed population of size 4000 having inpatients (around 60%) and outpatients (around 40%). The recordings were digitized at 360 samples per second per channel with 11-bit resolution over a 10 mV range. Two or more cardiologists independently annotated each record; consensus was made to obtain the computer-readable reference annotations for each beat.

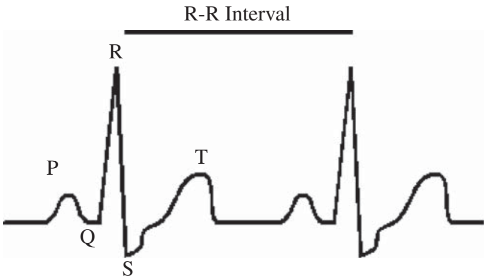

Fig. 1. Structural components of an ECG signal.

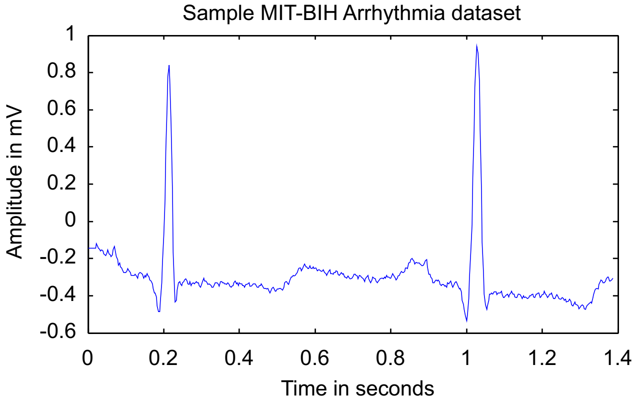

Fig. 2. Sample MIT-BIH Arrhythmia dataset.

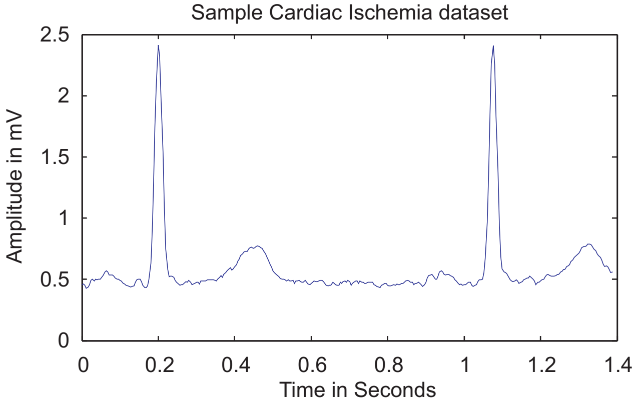

Fig. 3. Sample European ST-T dataset.

2.2. European ST-T dataset (cardiac ischemia)

European society of cardiology has provided a standard ST-T database [22] consisting of 90 annotated samples of ambulatory ECG recordings from 79 subjects having myocardial ischemia disease. The subjects were 70 men aged from 30 to 84 years, and some women aged from 55 to 71 years. Additional selection criteria were established in order to obtain a representative selection of ECG abnormalities in the database, including baseline ST segment displacement resulting from conditions such as hypertension, ventricular dyskinesia, and effects of medication. Each record is of 2 h duration and contains two signals. Each is sampled at 250 samples per second with 12-bit resolution over a nominal 20 mV input range.

3. Methodology

The proposed methodology consists of pre-processing followed by classification. Practically, pre-processing consists of re-sampling by FFT based decimator, QRS middle point extraction by extended Pan Tompkin's method, segmentation based on R point, linear predictive model estimation, RES computation and dimensionality reduction by PCA. Thereafter, the principal components (PCs) extracted as features are to be used for classification using GMM classifier. Finally, diagnostic performance of the proposed strategy is evaluated by using overall accuracy and probability of classification error using Chernoff and Bhattacharya bounds. The methodology is described below using a schematic diagram (see Fig. 4).

3.1. Pre-processing

Before actual classification of ECG a few pre-processing steps are carried out. They are re-sampling, QRS detection, segmentation based on R point, LP model estimation, residual signal estimation, and PCA.

3.1.1. Re-sampling

In the databases, initially ECG signal is sampled at 360 Hz for MIT-BIH Arrhythmia, whereas the same is sampled at 250 Hz for European ST-T database. In order to maintain the uniformity, the European ST-T database is re-sampled at 360 Hz by using FFT based interpolation method [20,21]. The re-sampling process is briefly described below:

(a) Take 250 point FFT of the ECG signal, $x(n)$ which is given by

$$X(k)=\sum_{n=0}^{N-1} x(n)e^{-j2\pi kn/N} \tag{1}$$

where $N=250$ points and $x(n)$ is the signal and $X(k)$ is its discrete Fourier transform.