International Journal of Medical Informatics 108 (2017) 55–63

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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

Research Paper

Improved detection of congestive heart failure via probabilistic symbolic pattern recognition and heart rate variability metrics

Ruhi Mahajan[a,b], Teeradache Viangteeravat[b,c], Oguz Akbilgic[a,b,d,*]

[a] UTHSC-ORNL, Center for Biomedical Informatics, University of Tennessee Health Science Center, Memphis, TN 38103, USA
[b] Department of Pediatrics, University of Tennessee Health Science Center, Memphis, TN 38103, USA
[c] Children's Foundation Research Institute, Le Bonheur Children's Hospital, Memphis, TN 38103, USA
[d] Department of Preventive Medicine, University of Tennessee Health Science Center, Memphis, TN 38103, USA

ARTICLE INFO

Keywords:
Congestive heart failure
Ensemble bagged trees
Heart rate variability
Probabilistic symbolic pattern recognition
RR intervals

ABSTRACT

Objective: A timely diagnosis of congestive heart failure (CHF) is crucial to evade a life-threatening event. This paper presents a novel probabilistic symbol pattern recognition (PSPR) approach to detect CHF in subjects from their cardiac interbeat (R-R) intervals.

Method: PSPR discretizes each continuous R-R interval time series by mapping them onto an eight-symbol alphabet and then models the pattern transition behavior in the symbolic representation of the series. The PSPR-based analysis of the discretized series from 107 subjects (69 normal and 38 CHF subjects) yielded discernible features to distinguish normal subjects and subjects with CHF. In addition to PSPR features, we also extracted features using the time-domain heart rate variability measures such as average and standard deviation of R-R intervals.

Results: An ensemble of bagged decision trees was used to classify two groups resulting in a five-fold cross-validation accuracy, specificity, and sensitivity of 98.1%, 100%, and 94.7%, respectively. However, a 20% holdout validation yielded an accuracy, specificity, and sensitivity of 99.5%, 100%, and 98.57%, respectively. Results from this study suggest that features obtained with the combination of PSPR and long-term heart rate variability measures can be used in developing automated CHF diagnosis tools.

1. Introduction

Congestive heart failure (CHF), a common manifestation of the late stages of coronary diseases, is a chronic medical condition in which the heart is unable to pump enough blood for the body [1]. It is one of the prevailing major health problems, and affects about 5.7 million subjects in the United States and more than 23 million worldwide [2,3]. CHF is asymptomatic in its initial stages and is usually a progressive condition. The earlier detection of CHF provides the possibility to initiate the treatment on time to improve the pumping function of the heart, decrease the cardiovascular risk, and avoid the progression of underlying life-threatening conditions like arrhythmias, kidney and liver disease, etc. With the advent of sophisticated signal processing methods to timely diagnose CHF, the existing rate of morbidity and mortality can be reduced [4].

As a common practice, subjects at high risk of progression of heart failure are identified by the widening of their QRS complex on electrocardiography. Heart rate variability (HRV), which represents a variation of interbeat (R-R) interval, provides very useful information on the dynamics of heart rate behavior [5–8]. Over the last decade, many noninvasive techniques based on R-R intervals have been proposed to identify subjects with CHF condition. These techniques can be broadly categorized as 1) time-domain metrics based on average of R-R intervals (AVRR), standard deviation of R-R interval (SDRR), the root mean square of successive differences (RMSSD), number of adjacent R-R intervals differing by more than 50 ms (NN50), etc. [9] 2) frequency-domain metrics which reflects sympathetic and parasympathetic activity of the autonomic nervous system by evaluating spectral power in 0-0.40 Hz frequency band using discrete Fourier transform, wavelet, autoregressive models, etc. [10–12]. Apart from the aforementioned techniques, other non-linear methods using entropy, sample asymmetry, and geometric methods like Lorenz (or Poincaré) plot of RR intervals, HRV triangular index, etc. have also been proposed to distinguish healthy subjects from those with CHF condition [13–18].

Among the recent advances in this field, Costa and Healey [19] developed a novel Multiscale sample entropy (MSE) algorithm to identify subjects with CHF based on underlying complexity in the R-R interval series. Using Fisher discriminant analysis, they achieved 92%

* Corresponding author at: Le Bonheur Research Center, 50N Dunlap Avenue, Suite 490R, Memphis, TN 38103, USA.
E-mail address: oakbilg1@uthsc.edu (O. Akbilgic).

http://dx.doi.org/10.1016/j.ijmedinf.2017.09.006
Received 18 May 2017; Received in revised form 22 July 2017; Accepted 19 September 2017

accuracy on the training set but on the test data the classifier could only separate with a positive predictivity of 76% and specificity of 76%. Jing Hu et al. [20] proposed a scale-dependent Lyapunov exponent to characterize the HRV dynamics. They achieved a true positive rate of 100% and a false positive rate < 5%. Yu and Lee [21] used higher order spectra up to third-order cumulant to calculate the bispectrum features of the HRV signals. By selecting features using a genetic algorithm (GA) and a support vector machine (SVM) classifier, a classification accuracy of 98.7% was achieved. Paolo et al. [22] analyzed long-term HRV measures and used a Classification and regression tree (CART) classifier to identify CHF patients with a sensitivity and specificity of 93.3% and 63.6%, respectively. Kamath [23], however, proposed to use Dispersion entropy (DispEn) and a second-order difference plot to quantify the degree of complexity in the HRV signals. The results of this study suggest that DispEn is a promising tool for chaotic analysis of R-R intervals and can help identify subjects with CHF up to an accuracy of 92%. Acharya et al. [24] applied an empirical mode decomposition method on the HRV signals and obtained 13 nonlinear entropy-based features. Using the probabilistic neural network and SVM classifier on all features, authors achieved a classification accuracy, sensitivity, and specificity of 97.6%, 97.0%, and 98.2%, respectively. On the other hand, Tscharner and Zandiyeh [25] explored fuzzy sample entropy and sample entropy variables of R-R intervals series. These variables reflect the complexity in the phase of the R-R intervals series which helped in identifying a group of CHF subjects with a sensitivity of 87% and a specificity of 89%.

Most of these studies rely on using a set of complex features which are difficult to interpret by clinicians. Also, their capability in real-time applications is limited. Most of the frequency-domain techniques are implemented on a time window (usually two to three minutes) with an underlying assumption of stationarity during that period. This study, therefore, aims to develop a simple, yet robust model amenable to accurately detect CHF from the given R-R interval data. In our previous work, we implemented a novel probabilistic symbolic pattern recognition (PSPR) approach to screen subjects with paroxysmal atrial fibrillation (PAF) with an accuracy of 92% [26]. In this study, we address the greater challenge of identifying subjects with CHF and propose to use PSPR along with other HRV metrics (e.g., AVRR and SDRR) to improve the CHF diagnostic. The dynamics of R-R interval time series are evaluated by learning the pattern transition behavior in the symbolically discretized R-R interval series using PSPR. Further, the diagnosis efficacy is enhanced by calculating the variability in R-R intervals. A supervised classifier using the ensemble of decision trees is thereby proposed to automatically classify subjects into two classes − class 0 (normal subjects) and 1 (subjects with CHF). The motivation to use the proposed PSPR is summarized below:

1. R-R interval series of different lengths can be compared and analyzed.
2. It learns the dynamics of the series by modeling the underlying morphology so it can be used to analyze any chaotic or deterministic series.
3. It is independent of the stationarity of the series, unlike conventional frequency-domain measures.
4. Easy implementation and fast computation make it suitable for future embedded computing applications.

The rest of the manuscript is organized as follows. Section 2 gives conceptual details of the PSPR approach and highlights its novel features. Section 3 explains the database used for analysis and the procedure followed to extract discernible features for classification. In Section 4, we describe key results for classification, validation, and compare our results with related studies. We mention the study limitations in Section 5 and summarize the key findings of this study in the concluding Section 6.

Table 1
Eight-symbol Discretization rules for R-R interval data used in this study. Symbols *c* through *f* are typically associated with R-R intervals observed during normal sinus rhythm.

Symbol	Range
a	[0 0.30)
b	[0.30 0.60)
c	[0.60 0.75)
d	[0.75 0.90)
e	[0.90 1.05)
f	[1.05 1.20)
g	[1.20 1.50)
h	[1.50 ∞)

2. Probabilistic symbolic pattern recognition

The PSPR technique aims to model the pattern transition behavior in sequential series represented by n_s unique symbols [26] similar to n-grams [27] for natural language processing using Shannon's information theory. When applied to a stationary time series, the series is first discretized using user defined thresholds generally based on the distribution of the series (i.e. normal, skewed, etc.) or domain knowledge depending on the problem at hand. The commonly used Piecewise Average Approximation (PAA) [28,29], Symbolic Aggregate Approximation (SAX) [30], Extended SAX [31], etc. based techniques can be also applied for discretizing a given series.

Table 1 depicts the discretization rules used in this study based on the uniform intervals over a nominal R-R interval range. For illustration purposes, Fig. 1 shows an example of a symbolically discretized one second Electrocardiography (ECG) excerpt using quantile-based discretization rules.

After discretizing the series, the PSPR method finds the joint occurrence of observed patterns of up to length n_p that is followed by a single symbol. For example, a two-symbol pattern transition is the pattern *ca* followed by *b* or a three-symbol pattern transition can be the pattern *cdd* followed by *c*. By observing the frequency of occurrence of such pattern transitions, pattern transition probabilities (PTP) are calculated. Consider an example of discretized series, S = {*ababbbaccbbabcabacc*} defined using three-symbol {*a, b, c*} i.e., $n_s = 3$. Table 2 shows two-symbol pattern transition probabilities (PTP_2) in lexicographical order. According to PTP values, if we observe either *ca* or *cb*, there is a 100% probability that the next symbol will be *b*. However, if we observe *ba*, symbols *b* and *c* are equiprobable, as each of them has a 50% chance. Therefore, the pattern transition behavior of the series S, PTP_S, is represented by PTP matrices $PTP_S = \{PTP_1, PTP_2\ldots, PTPn_p\}$. In other words, PTP_S is a probabilistic model of how the series S evolves.

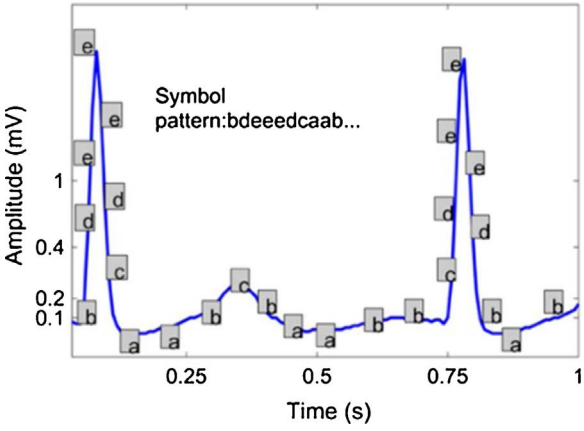

Fig. 1. An example of a five-symbol {a,b,c,d,e} quantile-based discretization of a sample ECG recording.