Random Forest-Based Classification of Heart Rate Variability Signals by Using Combinations of Linear and Nonlinear Features

Alan Jovic[1], Nikola Bogunovic[1]

[1]Faculty of Electrical Engineering and Computing, University of Zagreb/Department of Electronics, Microelectronics, Computer and Intelligent Systems, Zagreb, Croatia

Abstract— **The goal of this paper is to assess various combinations of heart rate variability (HRV) features in successful classification of four different cardiac rhythms. The rhythms include: normal, congestive heart failure, supraventricular arrhythmia, and any arrhythmia. We approach the problem of automatic cardiac rhythm classification from HRV by employing several features' schemes. The schemes are evaluated using the random forest classifier. We extracted a total of 78 linear and nonlinear features. Highest results were achieved for normal/supraventricular arrhythmia classification (93%). A feature scheme consisting of: time domain (SDNN, RMSSD, pNN20, pNN50, HTI), frequency domain (Total PSD, VLF, LF, HF, LF/HF), SD1/SD2 ratio, Fano factor, and Allan factor features demonstrated very high classification accuracy, comparable to the results for all extracted features. Results show that nonlinear features have only minor influence on overall classification accuracy.**

Keywords— **heart rate variability, ECG, linear features, nonlinear features, random forest**

I. INTRODUCTION

Heart rate variability (HRV) analysis examines fluctuations in the sequence of cardiac interbeat (RR) intervals, usually obtained from electrocardiogram (ECG) recordings. It allows us to assess how the fluctuations can be employed in detecting presence of cardiovascular diseases [1]. Decrease in HRV has been associated with old age as a result of progressive autonomic system dysfunction. Cardiac dysfunction is often manifested by systematic changes in the variability of the RR interval sequence relative to that of a normal rhythm [2].

HRV is analyzed by using both linear and nonlinear features. Linear features are mostly oriented on time and frequency characterization of the RR interval series [3]. The field of nonlinear analysis of biological rhythms is a relatively new area of scientific exploration. A pioneer work done by [4] introduced the concept of nonlinear dynamics into the field of cardiology. He showed that healthy physiological systems have fractal complexity whereas unhealthy biological systems lack the nonlinear properties and are marked by periodical dynamics and loss of the ability to adapt. Several authors later demonstrated the existence of nonlinear components in HRV [5, 6]. Author [6] pointed out that linear analysis using time and frequency features is inadequate for obtaining complete information about HRV. Regarding the nature of HRV series, author [2] showed that HRV series is nonlinear and stochastic. Nevertheless, authors continue to successfully utilize features stemming from nonlinear dynamics that rely on the assumption of underlying determinism. Nonlinear features are mostly used in combination with linear features [7]. For short-term analysis of cardiac rhythms, wavelet transform, a specific type of time-frequency localization, gives satisfying results [2, 8].

It is the purpose of this work to demonstrate the efficacy of several different schemes of features in a difficult classification setting. We want to examine how much predictive potential the linear and nonlinear features possess in the case when classification of a large number of different patients' rhythms is required. Our purpose is to determine the classification potential of these combinations of features.

II. METHODS AND MATERIAL

A. Cardiac records

We collected several hundred patient records from PhysioBank, a web database collection of biological signals [9]. In Table 1, we list the analyzed records. We decided to extract features for the following cardiac rhythms: normal, any arrhythmia, supraventricular arrhythmia (SVA), and congestive heart failure (CHF). The primary reason why these cardiac rhythms were analyzed, and not some others, is due to sufficient number of the records to be able to establish valid conclusions. We analyzed 500 RR intervals at the time, which constitutes to about five minutes of recording. An overlapping window was used that covers half of RR intervals from the preceding window, i.e. intervals 1-500, 251-750, 501-1000... were analyzed. In this way, we extracted a large number of feature vectors. There were a few nonexistent or invalid records within some of the databases listed in Table 1 that were omitted

P.D. Bamidis and N. Pallikarakis (Eds.): MEDICON 2010, IFMBE Proceedings 29, pp. 29–32, 2010.
www.springerlink.com

Table 1. Patient records

Heart rhythm (total no. of feature vectors)	PhysioBank database	ECG annotations records	RR intervals analyzed
Normal heart rhythm (665)	MIT-BIH Normal Sinus Rhythm Database, Normal Sinus Rhythm RR Interval Database	MIT-BIH: 16265-19830 NSR: nsr001-nsr054	1-500, 251-750, 501-1000, 751-1250, 1001-1500, 1251-1750, 1501-2000, 1751-2250, 2001-2500
Any arrhythmia (492)	MIT-BIH Arrhythmia Database, CAST RR Interval Sub-Study Database	MIT-BIH: 100-234 CAST: e001a-e130a, f001a-f130a	1-500, 501-1000
Supraventricular arrhythmia (312)	MIT-BIH Supraventricular Arrhythmia Database	800-894	1-500, 251-750, 501-1000, 751-1250
Congestive heart failure (747)	BIDMC Congestive Heart Failure Database, Congestive Heart Failure RR Interval Database	BIDMC: chf01-chf15 CHF RR: chf201-chf219	1-500, 251-750, 501-1000, ... , 3751-4250, 4001-4500

from the analysis. A total of 2216 feature vectors were extracted from the patients' annotated records.

B. Features

We implemented many linear and nonlinear features for HRV described in literature. Full list is comprehensive (78 features) and is given in Table 2. References to implementation details and partition in schemes are shown.

Advanced sequential trend analysis (ASTA) is not covered in literature. It is an extension of the idea to describe RR interval prolongation and shortening [18] with a more detailed specification of the degree of pace change. In ASTA, two out of possible four quadrants are analyzed in detail: prolongation / prolongation (+/+) and shortening / shortening (-/-). The features include: no change in RR interval length, small change, medium change, large change and very large change (nine features in total). Additionally, total number of points in each of the two quadrants is taken (two additional features). Number of RR interval changes falling in each of these subsections divided by the number of RR interval changes in all four quadrants is represented by each feature in ASTA.

Carnap 1D entropy has not been previously applied to HRV or ECG analysis. We implemented the algorithm proposed by [15] for time series analysis and allowed for Carnap entropy extraction on multiple scales, similar to sample entropy [14].

Nonlinear chaos attractor features possess a parameter named interval T (lag) that shows which pairs of RR intervals are used in the calculation (e.g. if T=2, an RR interval between two RR intervals is skipped). Authors [17] showed that if multiple intervals are taken into consideration, the classification accuracy improves.

Table 2. Feature schemes

Scheme number	Features in scheme	Number of features	Description	Comment
1	SDNN [3], pNN20 [3, 10], pNN50 [3, 10], RMSSD [3], HTI [3]	5	Linear, time domain	
2	(PSD, VLF, LF, HF, LF/HF) [3]	5	Linear, frequency domain	
3	Linear (time domain), linear (frequency domain)	10	Linear	
4	Linear, SD1/SD2 ratio [11], Fano factor [2], Allan factor [2]	13	Linear + nonlinear	
5	(Spatial filling index (SFI), Correlation dimension (D_2), Central tendency measure (CTM)) [12]	3	Nonlinear chaos attractor features	Time interval (lag) , T={1,2,5, 10, and 20}, reconstruction dimension d=2
6	Approximate entropy (ApEn1-ApEn4) [12], Maximum approximate entropy (MaxApEn) [13], r for MaxApEn, Multiscale sample entropy (SampEn1-SampEn20) [14], Multiscale Carnap 1D entropy(Carnap1-Carnap20) [15]	46	Entropies	Dimension m=2 for ApEn and SampEn
7	Advanced sequential trend analysis (ASTA): ASTA1-ASTA11	11	ASTA	
8	Detrended fluctuation analysis (DFA): DFA 5, DFA 7, DFA 10, DFA 15, DFA 20 [16]	5	DFA	
9	(SFI, D_2, CTM, ApEn1-ApEn4, SDNN, pNN20, RMSSD, HTI) [17]	11	Features combination	T=1, d=2, m=2
10	All features	78	Advanced linear + nonlinear chaos attractor features (T=1) + entropies + ASTA + DFA	

Therefore, for the analysis of scheme number 5 from Table 2, we extracted five times the amount of feature vectors: $T = \{1, 2, 5, 10, \text{and } 20\}$, 11.080 feature vectors in total.

Most of the feature extraction algorithms were implemented in a Java-based platform, ECG Chaos Extractor [12]. The only exceptions are spectral (frequency) features, which were extracted from RR interval series in Matlab using autoregressive (AR) model of order 12.

C. Classification procedure

In order to classify feature vectors with high accuracy, we used a state-of-the-art classifier named random forest (RF) [19]. Random forest is composed of a large number of decision trees that choose their splitting features from a random subset of *k* features at each internal node. Best split based on Gini index is taken among these randomly chosen features and the trees are built without pruning. Feature vectors are sampled using the bootstrap procedure. RF ensures at the same time the smallest obtainable bias and very low data variance, which often gives excellent classification results.

Random forest was constructed with 40 trees for each feature scheme. A 10x10-fold stratified cross-validation testing procedure was used in order to obtain representative classification accuracy. Analysis was performed in Weka system, version 3.6.1 [20].

Four distinct classification tasks were pursued: simultaneous classification of all four examined cardiac rhythms; classification between normal rhythm and any arrhythmia; classification between normal rhythm and supraventricular arrhythmia; classification between normal rhythm and congestive heart failure.

III. RESULTS

Classification results are presented in Fig. 1. Scheme numbers 4, 9, and 10 give the best results. Linear + nonlinear features analyzed by scheme number 4 show almost as good classification accuracy as do all the features collectively (scheme number 10). Also, nonlinear features in feature scheme 4 have only minor influence on classification accuracy (when compared to scheme 3).

A combination of linear and nonlinear features recently proposed by authors [17] (scheme number 9) can be, for all practical purposes, replaced by a linear combination of features, i.e. scheme number 3. Even the simple feature scheme 1, which contains only time-domain linear measures, is comparable to scheme number 9.

Spectral features from scheme number 2 also demonstrate high classification capacity, comparable to time-domain features. Also, nonlinear chaos attractor and entropy features failed to achieve high classification rates, probably due to inspection of only a single dimension (d=2 and m=2).

DFA shows the worst results for all classification tasks and is not suitable for classification of examined rhythms. The results for ASTA are fair considering the fact that it was the only method in scheme 7.

Although scheme number 10 provides us with the most accurate solution, the combination of 78 features is highly impractical regarding the description of the underlying

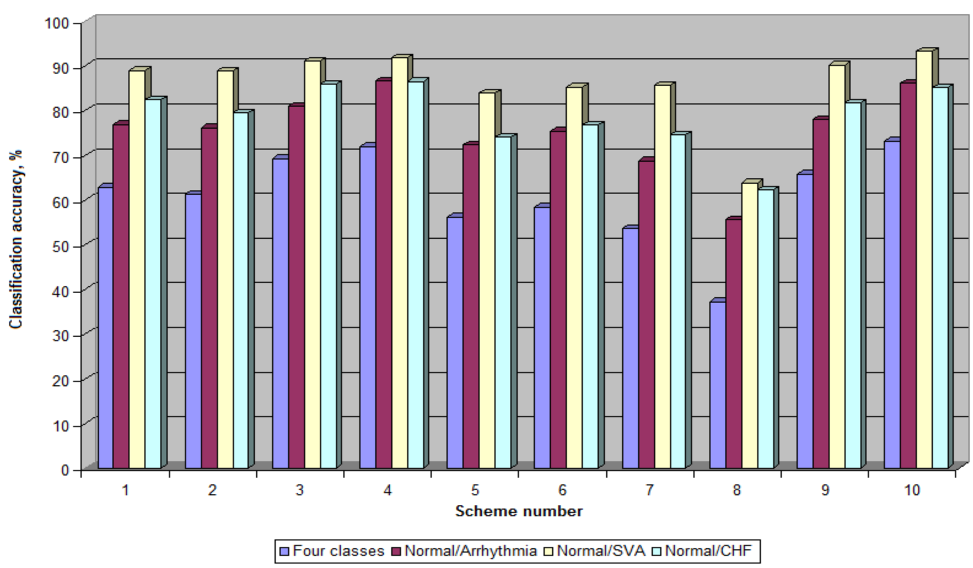

Fig. 1.Classification of HRV records using features' schemes

rhythm. The results show that the most accurate classification is achieved for discerning normal rhythms from supraventricular arrhythmia (around 93% for scheme 10), which indicates that normal heart rhythm and supraventricular ectopic beats differ significantly.

IV DISCUSSION

One of major problems in classification of HRV signals is the small number of abnormal heart beats present in most records. This fact severely limits the application of many analytical methods, because an abnormal rhythm seldom differs significantly from a normal one. In this work, we analyzed the records disregarding the actual number of abnormal beats in each record. Further work should concentrate on finding a minimal number of abnormal beats in records to be able to successfully apply the classification schemes.

Results of ASTA should be investigated further. We plan to extend the trend change algorithm with a procedure that would take into account not only two consecutive changes in RR interval duration, but three. In this way, more information about abnormal beats might prove useful for automatic classification tasks.

We suppose that nonlinear chaos attractor features and entropy measure do not demonstrate high classification accuracy due to the calculation of only a single, low dimension (m=2 and d=2). Researches performed by other authors almost always included feature calculations over a range of dimensions.

V. CONCLUSION

We have assessed the classification capabilities of several combinations of HRV features on a large sample of cardiac records for four different cardiac rhythms. The results show that the combination of time and frequency domain linear features and several nonlinear features: SD1/SD2, Fano factor, and Allan factor gives high classification accuracy. Other examined nonlinear features have very little influence on classification accuracy. Overall results suggest that linear features carry the most weight in all four classification tasks, with only a minor improvement obtained by adding some of the nonlinear features to the feature set.

Further work has to conclude which nonlinear features should be used together with standard time and frequency domain linear features in HRV analysis in order to obtain the best results.

REFERENCES

1. Kitney R, Linkens D, Selman A et al. (1982) The interaction between heart rate and respiration: part II – nonlinear analysis based on computer modeling. Automedica 4:141–153
2. Teich MC, Lowen SB, Jost BM et al. (2001) Heart Rate Variability: Measures and Models. Nonlinear Biomed Sig Proc Vol. II, Dynamic Analysis and Modeling, IEEE Press, New York, 159–213
3. Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology (1996) Heart rate variability guidelines: Standards of measurement, physiological interpretation, and clinical use. European Heart Journal 17:354–381
4. Goldberger AL (1996) Non-linear dynamics for clinicians: chaos theory, fractals, and complexity at the bedside. Lancet 11:312–1314
5. Iyengar N, Peng CK, Goldberger AL et al. (1996) Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics. Am J Physiol 271:1078–1084
6. Braun C, Kowallik P, Freking A et al. (1998) Demonstration of Nonlinear Components in Heart Rate Variability of Healthy Persons. Am J Physiol Heart Circ Physiol 275(5)1577–1584
7. Asl BM, Setarehdan SK, Mohebbi M (2008) Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal. Artif Intell Med 44(1):51–64
8. Chen SW (2002) A wavelet-based heart rate variability analysis for the study of nonsustained ventricular tachycardia. IEEE Trans. Biomed. Eng. 49(7):736–742
9. PhysioBank, at http://www.physionet.org
10. Hutchinson TP (2003) Statistics and graphs for heart rate variability: pNN50 or pNN20? Physiol Meas 24(3)N9–N14
11. Kitlas A, Oczeretko E, Kowalewski M et al. (2005) Nonlinear dynamics methods in the analysis of the heart rate variability. Roczniki Akademii Medycznej w Białymstoku, Annales Academiae Medicae Bialostocensis 50(Suppl. 2)
12. Jovic A, Bogunovic N (2007) Feature Extraction for ECG Time-Series Mining Based on Chaos Theory. Proc. 29th Int. Conf. Inf. Tech. Interfaces, ITI 2007, Cavtat, Croatia, 2007, pp. 63–68
13. Chon KH, Scully CG, Lu S (2009) Approximate Entropy for All Signals. IEEE Eng. Med. & Biol. Mag. 28(6):18–23
14. Costa M, Goldberger AL, Peng CK (2005) Multiscale entropy analysis of biological signals. Phys Rev E 71:021906
15. Jovic F, Krmpotic D, Jovic A et al. (2008) Information Content of Process Signals in Quality Control. IPSI BgD Transactions on Internet Research 4(2):10–16
16. Acharya RU, Kannathal N, Krishnan SM (2004) Comprehensive analysis of cardiac health using heart rate signals. Physiol Meas 25(5):1139–51
17. Jović A, Bogunović N (2009) Feature Set Extension for Heart Rate Variability Analysis by Using Non-linear, Statistical and Geometric Measures. Proc. 31st Int. Conf. Inf. Tech. Interfaces, ITI 2009, Cavtat, Croatia, 2009, pp. 35–40
18. de Carvalho JLA, Rocha AF, Nascimento FA et al. (2002) Development of a Matlab software for analysis of heart rate variability. Proc. 6th Int. Conf. on Signal Processing, Beijing, China, 2002, vol. 2, pp. 1488–91
19. Breiman L (2001) Random forests. Mach Learn 45:5–32
20. Hall M, Frank E, Holmes G, Pfahringer B et al. (2009) The WEKA Data Mining Software: An Update. SIGKDD Explor. 11(1):10–18.

Author: Alan Jovic
Institute: Faculty of Electrical Engineering and Computing
Street: Unska 3
City: Zagreb, HR-10000
Country: Croatia
Email: alan.jovic@fer.hr