Processing and Analyisis of Biomedical Nonlinear Signals by Data Mining Methods

Nikola Bogunovic and Alan Jovic
University of Zagreb
Faculty of Electrical Engineering and Computing
Zagreb, Croatia
Nikola.Bogunovic@fer.hr, Alan.Jovic@fer.hr

Abstract—The paper demonstrates a nonlinear signal processing method based on an approach found in intelligent data mining. ECG signals were used as an interesting and readily available representative nonlinear domain. These signals were fed in an innovative software platform for feature extraction based on chaos theory. The resultant files were loaded into an open source machine learning software for clustering and classification analysis. The results depict 78% clustering and around 90% classification accuracy rate, which is quite impressive considering the number of features involved in the study.

Keywords-nonlinear signal processing, chaos theory, data mining, clustering, classification

I. INTRODUCTION

In the analysis of signals represented by its sampled time series data, there are basically two approaches: analyses of global integral statistical properties of the whole or analyses of significant parts of the signal. The scope of these analyses encompass some integral linear and non-linear properties of signal (such as variability, histogram distributions, various measures of information content like entropies, and measures from non-linear dynamics and chaos theory). An overview of complexity measures in signal is given in [1]. More recent direction emerged which stems from the field of intelligent data analysis (machine learning, data mining and knowledge discovery). These approaches employ methods of intelligent statistical data analysis for time series. As in static data analysis, the emphasis here is on the search for regularities or patterns and their application in predicting events, classes, or in correlating between different patterns. A good overview of these approaches is given in [2]. The most frequently used time-series transformations are embedology [3] (transforming time series into the vector space) and various symbolic transformations (e. g. [4]).

Incorporation of nonlinear dynamical measures into a multivariate discrimination provides a signal classification system that could exhibit superior traits when applied to a particular domain. We have opted for a biomedical domain, specifically electrocardiogram (ECG) signals representing Atrial fibrillation (AF), as the most common major cardiac arrhythmia. It affects several million people, with an increasing prevalence in the elderly.

The paper is organized as follows. In section II, the nonlinear ECG signals domain is further explicated. Section III gives a brief repetition of chaos theory and points out chaos based features used in the analysis. Section IV enumerates popular signal and data files taken from the Internet sites, explains feature extraction process and gives details on clustering and classification methods. The results are presented in section V. The paper ends with some discussion and final conclusion.

II. NONLINEAR SIGNALS DOMAIN

Electrocardiography (ECG) is one of the most widely used methods for the cardiac function assessment. It is extensively available and inexpensive procedure. The ECG analysis has been perfected in recent years by using ever more sophisticated instruments and powerful computer tools. Still the question, whether there is some important information contained in ECG signal that has not yet been revealed, has remained open, because the ECG feature space is indefinite. The approaches to computer based ECG analysis can be roughly divided into three groups: deterministic (frequency, wavelet), statistical (time analysis, PCA) and non-linear. While deterministic and statistic approaches have well established roles in ECG signal analysis, the efficiency and application of non-linear methods is still not refined.

One of the goals of ECG analysis in general is to determine whether a signal can be classified with respect to a heart disorder that it contains. Furthermore, if the record can be successfully classified, it is expected that a predictor model for the disorder can be constructed. Several techniques using non-linear chaos features of the signal have been proposed in order to meet the classification and/or prediction demands. For example, multiscale entropy was found to be able to distinguish between RR intervals from healthy subjects and those with a heart disorder such as atrial fibrillation [6]. Symbolic dynamics, a non-linear method, also demonstrated the advantage over deterministic and statistical methods in distinguishing ventricular tachycardia and ventricular fibrillation patients [7].

Authors [8] have found that the heart rate variability is driven by non-linear processes and that linear analysis using time and frequency only is inadequate for obtaining the complete information. The principal task of a non-linear method is to examine the possible existence of chaotic properties in the signal, i.e. the inherent unpredictability of the future despite the determinism of the underlying system. This task is almost always a complex one, because ECG records usually contain several types of background noise and can display both linear and non-linear behavior. The problem of noise can be solved in most cases by using new empirical and model-based filtering methods [5].

III. CHAOS THEORY BASED FEATURES

Chaotic system is in its base a deterministic nonlinear system. Therefore it is best used for modeling of deterministic nonlinear systems. However, not every deterministic nonlinear system exhibits chaotic behavior, although it always has the potential to become chaotic [9]. Single dimensional deterministic nonlinear system can be presented using the equation:

$$x_{n+1} = f(x_n, r) \tag{1}$$

Here, function f is called iterator and is a general nonlinear function; r is a control parameter, a constant; x_n is the present value of the system state variable and x_{n+1} is the following value of this variable. The main properties of a chaotic system are aperiodicity, determinism, confinement and sensitive dependence on initial conditions. Aperiodicity means that values of state variables exhibit no periodic patterns, i.e. their values never repeat. Determinism means that the values of system state variables can be calculated in every moment if we know their past value. Confinement signifies that the values of system variables are always constrained between some boundary values. Every chaos system has sensitive dependence on initial conditions.

A. Nonlinearity and Determinism Test

Successful feature extraction using chaos analysis can only be conducted if a system is shown to be deterministic and nonlinear. Consequently, if a system is, for instance, linear and deterministic, chaos analysis has no sense. Because human hearts diverse slightly in their anatomy and other properties, it can not be stated that every heart is in every moment a nonlinear and deterministic system. In order to confirm the non-linearity and determinism, two conditions have to be fulfilled and also an assumption has to be made. First, one must prove that the system is not linear and stochastic. This can be efficiently conducted by showing that system is not in accordance with null hypothesis. This means that the time series of system variable does not behave in a way as Gauss noise does. In order to prove this, a Fourier transform of original time series has to be made. Then, its phases are randomized in frequency domain in interval $[0,2\pi]$. Thereafter, an inverse Fourier transform is made thus returning the shuffled values back to time domain. If the values of the original time series correspond significantly to this, so called, surrogate time series, that signifies that the original time series is linear and stochastic. Chaos analysis can not be applied in this case. However, if their values differ significantly (usually one order of magnitude) then null hypothesis is disproved and system is not linear and stochastic [10]. The system can still be (i) Nonlinear and deterministic, (ii) Nonlinear and stochastic, (iii) Linear and deterministic.

It is further assumed that human heart is not linear and deterministic. Although this assumption is generally true, some hearts may exhibit linear and deterministic behavior, usually those that have very low heart rate variability. Finally, to determine if heart exhibits deterministic or stochastic nonlinear behavior, a method using attractor reconstruction dimension d and correlation dimension D_2 is used. If in some d dimensional description of the system attractor correlation dimension D_2 comes into saturation, then the system can be considered deterministic. If to much noise exists in ECG, then it is possible that the attractor is "masked" and so its correlation dimension never saturates, thus making ECG a stochastic process.

B. Phase Space Reconstruction Method

Phase space reconstruction is a standard procedure when analyzing chaotic systems. It shows the trajectory of the system in time. A set to which a dynamic system evolves after a long enough time is called attractor. Phase space or phase diagram is such a space in which every point describes two or more states of a system variable. The number of states that can be displayed in phase space is called phase space dimension or reconstruction dimension. It is usually symbolized by letter d or E. Phase space in d dimensions will display a number of points of the system, where each point is given by:

$$\vec{X}(n) = [x(n), x(n+T), \ldots, x(n+(d-1)T)] \tag{2}$$

Here, n is a moment in time of a system variable, and T is a period between two consecutive measurements of the variable. The trajectory in d – dimensional space is a set of k consecutive points, where $n = t_0, t_0 + T, \ldots, t_0 + (k-1)T$, t_0 is the starting time of observation. When phase space has been created, other methods such as spatial filling index and correlation dimension can be used in order to numerically describe the attractor.

C. Spatia Filling Index

Spatial filling index is a quantitative description of the density of points of an attractor. Let the dynamic behavior of a system be determined by the trajectory of points in phase space. A point in d dimensional phase space is given by (2). Spatial filling index is defined by the expression

$$\eta = \frac{s}{n^2} \tag{3}$$

The phase space is divided into $n \times n$ squares. A matrix C of phase space is constructed with elements c_{ij}

and dimensions $n \times n$ such that c_{ij} is the number of points that fall into square. A matrix P is next formed, with elements p_{ij}, such that $p_{ij} = \frac{c_{ij}}{m}$, $m = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}$. Finally, Q matrix is formed that contains the squared elements of P, $q_{ij} = p_{ij}^2$, and $s = \sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij}$ is determined. It can be shown that the order of magnitude for η is 10^{-3} and it rises with greater concentration of points in the attractor.

D. CentralTtendency Measure

Central tendency measure (CTM) is a quantitative measure of variability for second – order difference plot. It also shows the concentration of points in the plot; however it is not used on the phase space, but on the second – order difference plot, i.e. $x(n+2)-x(n+1) / x(n+1)-x(n)$ diagram. This type of plot also gives an accurate description of chaotic behavior of the system.

E. Approximate Entropy

Approximate entropy (ApEn) is a statistical measure used to quantify the regularities in data. It adds a real number to a time series. The greater ApEn shows the higher complexity and irregularity of the series. The algorithm for determining ApEn can be divided into several steps:

1. $N-m$ vectors of dimension *m* are formed such that

$$\vec{X}(i) = [x(i), x(i+1), \ldots, x(i+m-1)], \quad i = 1, \ldots, N-m+1 \tag{4}$$

These vectors represent *m* consecutive *x* signal values, starting with *i*.

2. A distance between $\vec{X}(i)$ and $\vec{X}(j)$ is defined as:

$$d\left[\vec{X}(i), \vec{X}(j)\right] = \max_{k=1,2,\ldots,m} \left|x(i+k-1) - x(j+k-1)\right| \tag{5}$$

3. For each $\vec{X}(i)$, the number of $\vec{X}(j)$ is counted, such that $d\left[\vec{X}(i), \vec{X}(j)\right] \le r$ is satisfied, where *r* is so called tolerance frame parameter. This number is designated $N^m(i)$. Next, $C_r^m(i)$ coefficients are found by the expression $C_r^m(i) = \frac{N^m(i)}{N-m+1}$.

4. Natural logarithms are calculated for each $C_r^m(i)$ and their mean value is found, giving $\phi^m(r) = \frac{1}{N-m+1}\sum_{i=1}^{N-m+1} \ln C_r^m(i)$.

5. Dimension is increased to *m*+1 and steps 1 – 4 are repeated. Thus, $C_r^{m+1}(i)$ and $\phi^{m+1}(r)$ are obtained.

6. Approximate entropy is found using expression:

$$ApEn(m, r, N) = \phi^m(r) - \phi^{m+1}(r) \tag{6}$$

Parameters *m* and *r* are determined based on specific problem. Usually, starting *m* is 1 and four values for *r* are taken. This range of *r* shows differences in series complexity when more points are included.

IV. ECG Files and Data Mining Methods

ECG signal records have been obtained for four different classes of patients (Table I). The data have been collected from internet databases as specified in the table, using *rdsamp* and *rdann* programs for displaying signal data and annotations data, respectively. We have taken the first minute of signal data and the first half an hour of annotations data. The annotations files contain the exact time and type of the heart beats and the signal files contain samples taken at various sampling frequencies.

TABLE I. ECG Signal Records

Patient class / vector count	ECG files
	Database
Normal / 120	MIT-BIH Normal Sinus Rhythm Database
Atrial arrhythmia / 200	MIT-BIH Arrhythmia Database
Supraventricular arrhythmia / 150	MIT-BIH Supraventricular Arrhythmia Database
Congestive heart failure (CHF) / 120	BIDMC Congestive Heart Failure Database

Both the signal files and the annotations files are input files to the innovative software platform for chaotic features extraction (ECE), presented in more details in [11]. Graphical user interface is employed to specify the ECG files used in extraction procedure as well as for method selection and results saving. Also, input file can contain one or two ECG trails. Signal and annotations can be presented visually in a special window for signal visualization. After an ECG trail has been selected, test for determinism should first be pursued. If determinism and nonlinearity is satisfied, then other feature extraction methods can be activated. Feature extraction is done in a separate window. After the required features have been extracted, results can be stored in a file. Multiple ECG signals and annotations can be loaded into the program. However, feature extraction should be conducted on one signal at a time. The resultant file can later be loaded into Weka system for classification [12]. ECE program allows the classification based on the type of the disorder in ECG signal, including atrial, ventricular, supraventricular arrhythmia, fusion beats, nodal premature beats etc.

Two clustering analysis methods and three classification methods have been used to examine the efficiency of the chaos features. For clustering, we have used SimpleKMeans (in text: KMeans) and EM (Expectation Maximization) algorithms, both of them supported by Weka system. KMeans is a popular and efficient method for clustering. The number of data clusters is usually suspected and specified in advance for the algorithm. In our case, four clusters have been specified, each one of them corresponding to a patient class. KMeans starts with four random clusters and moves the objects between these clusters in such a way that it minimizes the variability within a cluster and

maximizes the variability between clusters with their corresponding mean values being as different as possible across all dimensions. One obtains ideally all the samples of a particular patient class in a cluster. In reality, such an ideal distinction is an exception.

EM algorithm is an extension of the KMeans algorithm. It does not assign the samples to particular classes by maximizing their mean differences, but rather by computing one or more probability distributions. It then maximizes the overall probability of the samples belonging to a certain cluster In the case of both KMeans and EM, we have specified four clusters and 100 iterations of the algorithms.

For classification purposes, C4.5 (J48 in Weka) and Bayesian network algorithms have been used. C4.5 was used with reduced error pruning and a minimum of three instances per leaf. Three instances per leaf are used instead of the standard two in order to ensure that only relevant leaves are taken into consideration.

Bayesian network is a known probabilistic graphical model classifier based on the Bayesian theorem and its implications. The network is constructed using several parameters, including the type of estimator (simple estimator based on maximum likelihood has been used) and search method (hill climbing has been used).

For classification purposes, a 10*4-fold cross-validation technique has been used in order to randomize the input samples.

V. RESULTS

The classification results of the ECG annotation files for two classes are shown in Table II. One class contains a normal heart rhythm and the other one contains a disorder. Total classification accuracy together with sensitivity and specificity for normal patients are used to evaluate two classes case. For each clustering method a selection of features has been performed such that the optimal clustering results are achieved. First, a recommended set of features for the corresponding samples set has been obtained using the Weka's Attribute Evaluator *CfsSubsetEval*, with best first search method. After the features were determined, we used them to obtain clustering results. Next, we have manually tried to find a better set of features by either omitting some of the recommended features or adding them to the set.

TABLE II. CLUSTERING AND CLASSIFICATION RESULTS

Classification accuracy(total)/ sensitivity(Normal)/ Specificity(Normal), %		**Classes type**		
		Normal – Atrial arrhythmia	***Normal – Supraventricular arrhythmia***	***Normal – Congestive heart failure***
Clustering	**KMeans**	71.4	81.4	80.0
	EM	78.6	78.6	78.6
Classification	**C4.5**	90.0/85.7/94.3	92.9/90.0/95.7	92.9/90.0/95.7
	BayesNet	82.1/84.3/80.0	90.0/94.3/85.7	94.3/95.7/92.9

VI. DISCUSION AND CONCLUSION

We have used the annotation specialized files with only two patient classes present in order to find out how chaos features are successful in discerning a healthy patient from a patient with a disorder.

The results are promising. We have obtained a 78% clustering and around 90% classification accuracy rate, which is an impressive result considering small number of features involved in the study. There was no apparent difference in efficiency between C4.5 and Bayesian Network algorithms. EM clustering algorithm has been found more efficient than KMeans in most cases. Also, no significant difference was perceived between ECG trails 1 and 2 or between *m* factors 1 and 2. The conclusion is that any of the two trails and any of the *m* factors can be used. Correlation dimension and approximate entropy have been found to be the most efficient features for the successful clustering of the ECG files.

REFERENCES

[1] J. Kurths, U. Schwarz A. Witt R. Th. Krampe and M. Abel, "Measures of complexity in signal analysis", in Chaotic, Fractal and Nonlinear Signal Processing, AIP Conference Proceedings, pp.33-54.

[2] J. F.Roddick, M. Spiliopoulou, "A survey of temporal knowledge discovery paradigms and methoda", IEEE Trans. On Knowledge and Data Eng., Vol. 14, No. 4, (2002), pp.750-767..

[3] T. Sauer, J.A. York,and M.Casdagli, "Embedology", Journal of of Statistical Physics, Vol. 65, No. 3, (1992), pp. 579-616.

[4] J. Kurths, A. Voss, P. Saparin, A. Witt, H.J. Kleiner and N. Wessel, Quantitative analysis of heart rate variability", Chaos, Vol. 5, No.1, (1995), pp.88-94.

[5] Clifford, GD, Azuaje, F., McSharry, P.E. editors, "Advanced Methods and Tools for ECG Data Analysis", Norwood MA, USA: Artech House; 2006.

[6] Costa, M., Goldberger, A.L., Peng, C.K., "Multiscale Entropy Analysis of Complex Physiologic time Series", Phys. Rev. Lett 2002; 89, No. 6.

[7] Wessel, N. et al., "Short-Term Forecasting of Life-Threatening Cardiac Arrhythmias Based on Symbolic Dynamics and Finite-Time Growth Rules", Phys. Rev. E 2000; 61(1): 733-739.

[8] Braun, C. et al., "Demonstration of Nonlinear Components in Heart Rate Variability of Healthy Persons", Am. J. Physiol. Heart Circ. Physiol. 1998; 275: H1577-H1584.

[9] Sharma, S., "An Exploratory Study of Chaos in Human-Machine System Dynamics", *IEEE Transactions on Systems, Man and Cybernetics – Part A: Systems and Humans*, Vol. 36, No. 2, March 2006

[10] Kaplan, D., Glass, L., "Understanding Nonlinear Dynamics", Springer-Verlag, 1995

[11] Jović, A., Bogunović, N., "Feature Extraction for ECG Time-Series Mining Based on Chaos Theory", Proceedings of the 29th International Conference on Information Technology Interfaces; 2007 June 25-29; Cavtat, Croatia. Zagreb: SRCE University Computing Centre, University of Zagreb; 2007. p. 63-68.

[12] Witten, I.H., Frank, E., "Data mining: Practical Machine Learning Tools and Techniques with Java Implementations", *Morgan Kaufmann Publishers*, 2000