Feature Extraction for ECG Time-Series Mining Based on Chaos Theory

Alan Jovic[*], Nikola Bogunovic[†]
[*] *Rudjer Boskovic Institute, Laboratory of Informational Systems, ajovic@irb.hr*
[†] *Faculty of Electrical Engineering and Computing, University of Zagreb, nikola.bogunovic@fer.hr*

Abstract. *Chaos theory applied to ECG feature extraction is presented in this article. Several chaos methods, including phase space and attractors, correlation dimension, spatial filling index, central tendency measure and approximate entropy are explained in detail. A new feature extraction environment called ECG Chaos Extractor has been created in order to apply these chaos methods. System model and program functions are presented. Some of the obtained results are listed. Future work in this field of research is discussed.*

Keywords. chaos theory, feature extraction, ECG analysis

1. Introduction

Human heart is an uttermost complex biological system. There exists no model that can take into consideration all of the psycho-physiological factors that collaborate, thus allowing the proper heart functioning. The most common accepted procedure used in monitoring heart rhythms is electrocardiogram (ECG). ECG measures weak currents that are the result of electrochemical signal conduction within heart muscle. A graph of this voltage measurement is called electrocardiograph. It contains a lot of information about heart functioning, including the length of contraction and relaxation periods, peak voltage, conduction time, etc. Medical doctors usually analyze ECG in search for anomalies that could indicate a heart disease. Recently, many automated and semi-automated methods for ECG analysis have been explored and implemented, extracting features from ECG. The goal of feature extraction is to find as few properties as possible within ECG signal that would allow successful anomaly detection and efficient prognosis. Theoretically, the number of features held in ECG signal is infinite. Therefore, methods have been devised in order to restrain the number of dimensions of the problem to a calculable level. Typically, not more than twenty features of ECG are observed, and usually less than ten. The methods can be categorized into several classes depending on mathematical approach. These are deterministic methods, statistic methods and chaos analysis. The main method of deterministic frequency investigation is Fourier analysis. It is a fast and reliable method that uses Fourier transform in order to determine significant amplitudes in the frequency domain [8]. Recently, wavelet analysis, an advanced type of time-frequency analysis has been successfully used [6]. Statistical methods include time analysis [5] and principal component analysis (PCA) [4]. It is also possible to have systems that combine several types of extraction methods in order to improve the results. Chaos methods propose an entirely different approach to signal analysis. They are based on non-linear dynamics of the system. The role of chaos analysis applied to ECG is still not clear and is a subject of ongoing research. A new semi-automatic program for ECG feature extraction has been implemented and is presented in this article. Section 2 shows the applied chaos theory and section 3 describes a chaos-based extractor program ECE which is used to extract the ECG features.

2. Chaos theory applied to feature extraction

There is no agreeable definition of chaos. When one speaks or thinks about chaos, usually a deterministic chaos is implied. It denotes the irregular motion which is generated by non-linear systems, whose dynamical laws uniquely determine the time evolution of a state of the system from knowledge of its previous history [7]. Single dimensional deterministic non-linear system can be presented using the equation:

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

Here, function *f* is called iterator and is a general non-linear 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 a state variable exhibit no apparent periodic pattern, i.e. the values of the state variable never repeat. Determinism means that the values of system state variables can be calculated in every moment if we know their past value (1). Confinement signifies that the values of system variables are always constrained between some boundary values. Every chaotic system has sensitive dependence on initial conditions. Already a small difference in seventh decimal place may lead to a large trajectory divergence of the variable after a number of cycles of a system variable *x* have passed (Fig.1). [http://www.vanderbilt.edu/AnS/psychology/cogsci/chaos/workshop/Sensitivity.html].

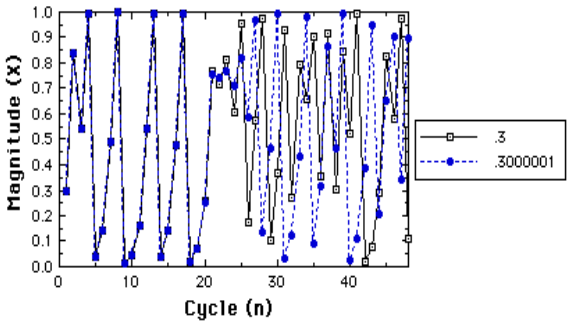

Figure 1. Sensitive dependence on initial conditions

2.1. Confirming the non-linearity and determinism of a system

Successful feature extraction using chaos analysis can only be conducted if a system is shown to be deterministic and non-linear. Although this statement can not be exactly proven, logic dictates that it is true. Consequently, if a system is, for instance, linear and deterministic, chaos analysis has no meaning. Because human heart diverges slightly in its anatomy and other properties, it can not be stated that every heart is in every moment a non-linear 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 disprove that the system is linear and stochastic. This can be efficiently conducted by showing that system is not in accordance with the null hypothesis. This means that the time series of a system variable does not behave in a way that 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 the frequency domain in interval $[0,2\pi]$. Thereafter, an inverse Fourier transform is made, thus returning the shuffled values back to the 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 the null hypothesis is disproved and the system is not both linear and stochastic [3]. Three options are left. The system can still be:

1. Non-linear and deterministic
2. Non-linear and stochastic
3. Linear and deterministic

Next, an assumption is made for ECG time series and heart in general. It is assumed that the human heart is not both 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 the heart exhibits deterministic or stochastic non-linear behavior, a method using attractor [7] 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 too 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 system.

2.2. Phase space reconstruction

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 $\left\{\vec{X}(n)\right\}$ of the system, where each point is given by:

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

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. If a phase space has more than three dimensions (three state variables), there is a problem with its graphical presentation.

Phase space reconstruction is a standard procedure when analyzing chaotic systems. It shows the trajectory of the system in time. An example of the attractor is given in Fig. 2. This figure was obtained using ECG Chaos Extractor program presented in section 3. It is a phase space reconstruction of normal heart rhythm, taken from MIT-BIH Normal Sinus Rhythm Database, record number 16272, first N=500 points of the record, parameter T=1 point. [http://www.physionet.org/physiobank/database/nsrdb/]. Two numbers on the right side of the attractor are spatial filling index η (see section 2.3), multiplied by the factor of 10^3 and correlation dimension D_2. More on D_2 can be found in [7].

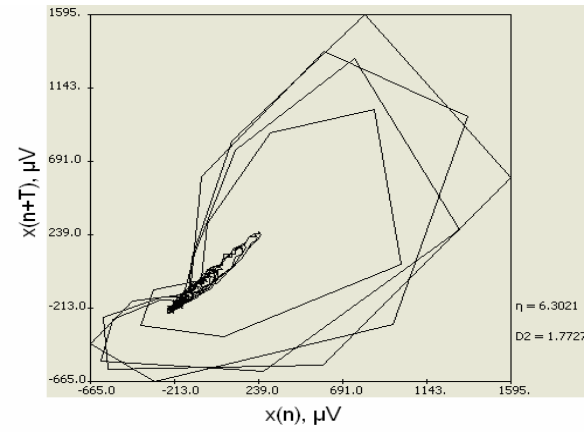

Figure 2. A chaos attractor of a normal heart rhythm

2.3. Spatial filling index

Spatial filling index is a quantitative description of the density of points of an attractor [1]. Let the dynamic behavior of a system be determined by the trajectory $\vec{X}$ of points in phase space. A point in d dimensional phase space is given by (2). A matrix A_d is formed such that

$$A_d = \begin{bmatrix} \vec{X}(1) \\ \vec{X}(2) \\ \vdots \\ \vec{X}(M) \end{bmatrix} \quad (3),$$

where $\vec{X}(n)$ is a point, and $M = N - (d-1)T$, N is a total number of points in phase space. If $d = 2$ is assumed, then $M = N - T$, thus giving $A_2 = \begin{bmatrix} x(1) & x(1+T) \\ x(2) & x(2+T) \\ \ldots & \\ x(M) & x(N) \end{bmatrix}$.

Next, matrix B_d is formed from A_d by dividing elements of A_d with $x_{\max} = \max|x(k)|$, $1 \le k \le N$. B_2 is consequently

$$B_2 = \frac{A_2}{x_{\max}} = \begin{bmatrix} x(1)/x_{\max} & x(1+T)/x_{\max} \\ x(2)/x_{\max} & x(2+T)/x_{\max} \\ \ldots & \\ x(M)/x_{\max} & x(N)/x_{\max} \end{bmatrix}.$$

It is obvious that the elements of the matrix B_d are $-1 \le b_{ij} \le 1$. Phase space is divided into $n \times n$ squares, each of them of size $R \times R$, $R \in \Re$, $\frac{2}{R} \in \mathrm{N}$. Matrix C is constructed with elements c_{ij} and dimensions $n \times n$ such that c_{ij} is the number of points that fall into square $g(i,j)$, $i, j \in \{1, \ldots, n\}$. The matrix C is called the matrix of phase space. It is generally a d-dimensional field and is limited to a matrix in two dimensional phase space. 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, $q_{ij} = p_{ij}^2$. A sum $s = \sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij}$ is determined. Spatial filling index is defined by expression

$$\eta = \frac{s}{n^2} \quad (4).$$

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.

2.4. Central tendency measure

Central tendency measure (CTM) is a quantitative measure of variability for second – order difference plot [2]. It also shows the concentration of points in a 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. A point in second – order difference plot in E – dimensions is given by

$$\vec{X}(t)=[x(t+T)-x(t),\ x(t+2T)-x(t+T),\ldots \\ \ldots,x(t+ET)-x(t+(E-1)T)] \quad (5).$$

Usually only two dimensions are observed. CTM is defined as

$$CTM=\sum_{t=1}^{N-E}\delta(d(t)) \quad (6),$$

where

$$\delta(d(t))=\begin{cases}1, & for\ \sqrt{(x(t+T)-x(t))^2+\ldots} \\ \sqrt{+\,(x(t+ET)-x(t+(E-1)T))^2} & \\ <r & \\ 0, & otherwise\end{cases} \quad (7).$$

Here, N is a number of points in time series, T is time period between two consecutive points and r is central area radius, dependent on data.

2.5. Approximate entropy

Approximate entropy (ApEn) is a statistical measure used to quantify the regularities in data without a priori knowledge of the problem [2]. It adds a real number to a series. The greater this number the higher is the 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

$$\vec{X}(i)=[x(i),x(i+1),\ldots,x(i+m-1)], \\ i=1,\ldots,N-m+1 \quad (8).$$

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 highest norm, such that

$$d\left[\vec{X}(i),\vec{X}(j)\right]=\max_{k=1,2,\ldots,m}\left|x(i+k-1)-x(j+k-1)\right| \quad (9).$$

3. For each $\vec{X}(i)$, the number of $\vec{X}(j)$ $(j=1,\ldots,N-m+1,\ j\neq i)$ is counted, such that $d\left[\vec{X}(i),\vec{X}(j)\right]\leq r$ 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 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:

$$ApEn(m,r,N)=\phi^m(r)-\phi^{m+1}(r) \quad (10).$$

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.

3. ECG Chaos Extractor program

ECG Chaos Extractor program (ECE) has been created as a feature extraction environment for chaos methods presented in section 2. It is written in Java 1.5 language and is platform independent. Graphical interface is used to specify ECG files employed in the extraction procedure as well as for method selection and results saving. The program extracts features from ECG files, as given in Table 1. The extracted features are bolded, while the extraction parameters are in italics. Block diagram of the system of which ECE program is an essential part is given in Fig. 3. The idea is to

Table 1. Features extracted by program ECG Chaos Extractor

Feature and parameter name	Domain	Explanation
Correlation dimension D_2	$[0, \infty)$, in reality: $[0,10)$	D_2 is extracted from the attractor using Grassberger – Procaccia method.
Spatial filling index η	$(0, 0.01)$	η is determined from the attractor using expression (4). It has been multiplied by the factor 10^3 for more appropriate presentation in ECE program.
Central tendency measure CTM	$(0, 1]$	CTM is determined from second – order difference plot using equation (6).
Approximate entropy ApEn	$[0, \infty)$, in reality $(0.5,5)$	ApEn is determined from expression (10) using four r values: [0.1, 0.15, 0.2, 0.25] σ, where σ is time series' standard deviation
Number of points	$[100, 15000]$	Number of consecutive points in ECG record or number of consecutive annotations analyzed.
Dimension	$[2, 100]$	Reconstruction dimension of the attractor.
Interval T	[1, number of points - 1]	Interval between observed consecutive points, given in points and not in seconds.
Starting point	[0, total number of points – number of points]	Starting point within the ECG record from which the analysis is started.
m factor	$[1,\infty)$, in reality: 1 or 2	Factor that determines the size of the vector for ApEn.

extract features either from an ECG signal file or from an ECG annotation file. The annotation file contains heart beat times and beat type, whereas signal file contains the whole sampled signal. The analysis is basically the same for both file types. Also, a 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. ECE program has a separate window that presents the relation between correlation dimension D_2 and the reconstruction dimension d. User has to confirm if D_2 comes into saturation for some dimension d as well as to check the significance of difference for the real and surrogate series (section 2.1). If determinism and non-linearity is satisfied, then other feature extraction methods can be pursued. Feature extraction is done in a separate window. After the required features have been extracted, results can be stored in a file in Arff format. This file has to be created first and then later it is filled with the extracted data. Even though multiple ECG signals and annotations can be loaded into the program, feature extraction should be conducted on one signal at a time. The result file in Arff format can later be loaded into Weka system for a classification process [10]. 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. Some results obtained for normal and arrhythmia files

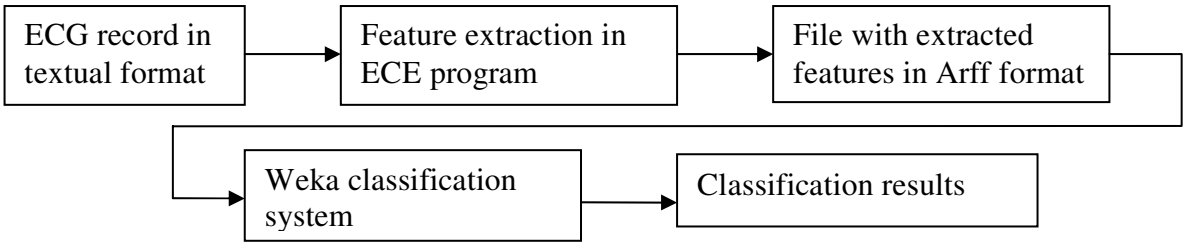

Figure 3. ECE program in ECG analysis process

Table 2. Results obtained by feature extraction using ECE program

ECG class	ECG record		$\eta * 10^3$	D_2	CTM	ApEn (annotations only)
Test record	aami3a		3.862	1.634	0.718	N/A
	aami3b		3.491	1.583	0.524	N/A
Normal heart rhythm	16265	1.trail	7.322	1.583	0.844	[1.56, 0.97, 0.97, 0.60]
		2.trail	2.905	2.681	0.891	[1.56, 0.97, 0.97, 0.60]
	16272	1.trail	5.361	2.053	0.943	[1.66, 1.38, 1.21, 1.12]
		2.trail	5.276	1.904	0.928	[1.66, 1.38, 1.21, 1.12]
Supraventricular arrhythmia	800	1.trail	6.580	1.789	0.946	[1.23, 1.03, 0.90, 0.77]
		2.trail	7.047	2.663	0.870	[1.23, 1.03, 0.90, 0.77]
MIT-BIH arrhythmia database	100	1.trail	6.775	1.654	0.941	[2.48, 1.97, 1.65, 1.45]
		2.trail	5.830	1.810	0.939	[2.48, 1.97, 1.65, 1.45]
	101	1.trail	4.116	1.780	0.957	[2.00, 1.58, 1,29, 1.06]
		2.trail	4.351	2.628	0.573	[2.00, 1.58, 1,29, 1.06]

are given in Table 2. Phase space dimension d=2, first N=1000 points, period T=1 point. Results obtained by the program ECE show some diversity in features obtained from different patient records. This is unfortunately also true for patients with the same disorder. No significant pattern can be recognized on this small sample.

4. Discussion and conclusion

ECG signal often demonstrates the characteristics of non-linearity and can be unpredictable. Nevertheless, the existence of chaos within ECG can not be easily determined. Surely, the transition to chaos can be observed in tachycardia – fibrillation transition [9], however it is not clear whether normal heart or heart with other disorders exhibits significant chaotic behavior. Program ECE has been tested only on a small sample of ECG records and further work should be pursued using a larger number of records under different extraction parameters. Further study could point to existent patterns in features for the same disorder type.

5. References

[1] Faust, O., Acharya U.R., Krishnan, SM., Min, L.C., "Analysis of Cardiac Signals Using Spatial Filling Index and Time-Frequency Domain", *BioMedical Engineering OnLine*, September 2004

[2] Hornero, R., Abásolo, D. *et al.*, "Variability, Regularity, and Complexity of Time Series Generated by Schizophrenic Patients and Control Subjects", *IEEE Transactions on Biomedical Engineering*, Vol. 53, No. 2, February 2006

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

[4] Lux, R.L., "Principal Component Analysis: An Old but Powerful Tool for ECG Analysis", *International Journal of Bioelectromagnetism*, Vol. 5, No. 1, 2003

[5] Malik, M., Kulakowski P., Hnatkova, K., Staunton, A., Camm, A.J., "Spectral turbulence analysis versus time-domain analysis of the signal-averaged ECG in survivors of acute myocardial infarction", J Electrocardiol., 27, Suppl: 227-32, 1994

[6] Morlet, D., et al. "Time-scale analysis of high-resolution signal-averaged surface ECG using wavelet transformation." *Proceedings of Computers in Cardiology Conference*, September 1991

[7] Schuster, H. G., "Deterministic Chaos: An Introduction", *Physik-Verlag GmbH*, 1984.

[8] Strohmenger, H.-U., Lindner, K.H., Brown, C.G., "Analysis of the ventricular fibrillation ECG signal amplitude and frequency parameters as predictors of countershock success in humans", CHEST, March 1997

[9] Weiss, J.N., Garfinkel, A. *et al.* "Chaos and the Transition to Ventricular Fibrillation: A New Approach to Antiarrhythmic Drug Evaluation", UCLA Cardiovascular Research Laboratory, 1999

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