Analysis of ECG Records using ECG Chaos Extractor Platform and Weka System

Alan Jovic, Nikola Bogunovic

Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, *HR-10000 Zagreb, Croatia*, *alan.jovic@fer.hr and nikola.bogunovic@fer.hr*

Abstract. *Clustering and classification of ECG records for four patient classes from the internet databases by using the Weka system. Patient classes include normal, atrial arrhythmia, supraventricular arrhythmia and CHF. Chaos features are extracted automatically by using the ECG Chaos Extractor platform and recorded in Arff files. The list of features includes: correlation dimension, central tendency measure, spatial filling index and approximate entropy. Both ECG signal files and ECG annotations files are analyzed. The results show that chaos features can successfully cluster and classify the ECG annotations records by using standard and efficient algorithms such as EM and C4.5.*

Keywords. chaos features, ECG analysis, clustering methods, classification methods

1. Introduction and related work

Electrocardiography (ECG) is one of the most used methods for the cardiac function assessment. It is a widely available and not an expensive 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. 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 [2].

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 [3]. Symbolic dynamics, a non-linear method, also demonstrated the advantage over deterministic and statistical methods in distinguishing ventricular tachycardia and ventricular fibrillation patients [8]. Authors [1] 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.

In our previous work, we presented a platform, called ECG Chaos Extractor (ECE) [5]. We have also explained the chaos features and their corresponding algorithms that have been implemented in the platform. These features include: spatial filling index (*SFI*), correlation dimension (D_2), central tendency measure (*CTM*) and approximate entropy (*ApEn*). *ApEn* contains four measures *ApEn1-ApEn4*, each for a different range inclusion, dependent of the data standard deviation. We present in the current article the results of using the ECE platform in the classification and clustering of four patient classes. The goal of this work is to determine if it is possible to successfully classify the patient record with respect to the disorder by using only the previously mentioned chaos features.

The structure of this paper is as follows: In section 2, we present the methodology of our work, i.e. which ECG records are used and how is the platform configured in order to obtain the desired features. In section 3, we give a short overview of the methods that we use for the clustering and classification process and in section 4 we present the obtained results. The discussion of the results is given in section 5 and the conclusion in section 6.

2. Methodology

ECG signal records have been obtained for four different classes of patients (Table 1). 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. The starting internet page is [6]. 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, as specified in the table.

Both the signal files and the annotations files are input files to the ECE platform. We have not performed any filtering of these records, since they had already been filtered. An automation of the feature extraction process has been performed on the ECE platform. It allows the user to specify a list of input files and the extraction parameters as the input arguments for the platform. The platform then performs the extraction of the specified chaos features and stores them in an output Arff Weka file, ready to be analyzed. The schema of the analysis process is given in Fig. 1. First, the ECG files are downloaded and prepared. The exact file format is specified in [5]. Next, a user starts the ECE platform with specified parameters, including: the name of the Weka file to write the results to, the starting point of the analysis, the number of points to be analyzed, the time interval between two consecutive points, the ECG trail number, the *m* factor for approximate entropy analysis, the list of features to be extracted and finally, the list of ECG files to extract the features from, given by their file path.

As the input parameters for the extraction process, we have taken first 500 samples from the ECG signal files and first 500 beats from the annotations files. The 500 signal files' samples correspond to a period from one up to four beats of the signal, depending on a record's sampling frequency. We have empirically taken five different intervals between two consecutive points: {1, 2, 5, 10, 20} for each record in the databases. For the signal files, we have obtained the extracted data both for trail 1 and trail 2. For the annotations files, since the beat time is equal in both trails, we have specified the *m* factor: {1,2} for *ApEn* evaluation. For the signal files, the platform has been requested to extract *SFI*, D_2 and *CTM* for each of the trails and for the five intervals. For the annotations files, *SFI*, D_2, *CTM* and *ApEn* features have been extracted for each of the *m* factors and for the five intervals.

Altogether, we have extracted 590 feature vectors from 59 ECG signal files and 590 vectors from 59 ECG annotation files. Total number of vectors per patient class is also given in Table 1. Number of vectors taken from the two trails is the same, as well as the number of vectors with the same *m* factor: half of total count per class.

Since the automatic procedure disregards the real absence or presence of abnormal beats in the extracted time period, we have performed an additional manual extraction for the two disorder

Table 1. Database ECG files used for feature extraction

Patient class / vector count	Internet database	Sampling frequency, Hz	ECG signal and annotations records
Normal / 120	MIT-BIH Normal Sinus Rhythm Database	125	16265, 16272, 16273, 16420, 16483, 16539, 16773, 16786, 16795, 17052, 17453, 18177
Atrial arrhythmia / 200	MIT-BIH Arrhythmia Database	360	100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 111, 112, 113, 114, 115, 116, 117, 118, 119, 121
Supraventricular arrhythmia / 150	MIT-BIH Supraventricular Arrhythmia Database	125	800, 801, 802, 803, 804, 805, 806, 807, 808, 809, 810, 811, 812, 820, 821
Congestive heart failure (CHF) / 120	BIDMC Congestive Heart Failure Database	250	chf1,chf2, chf3, chf4, chf5, chf6, chf7, chf8, chf9, chf10, chf11, chf12

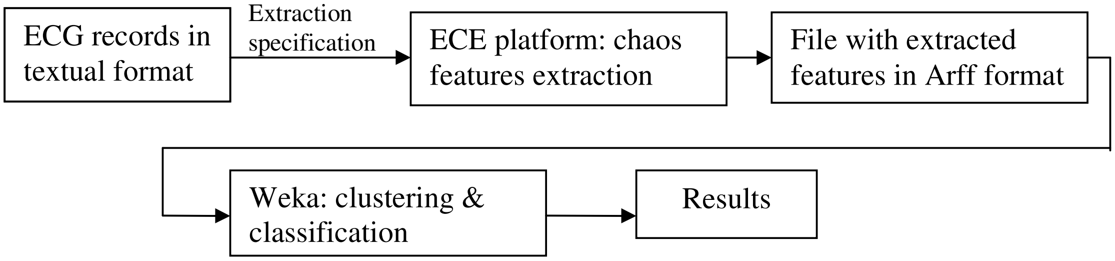
ECG records in textual format
Extraction specification
ECE platform: chaos features extraction
File with extracted features in Arff format
Weka: clustering & classification
Results

Figure 1. ECG analysis process

patient classes, atrial arrhythmia and supraventricular arrhythmia. We have examined the annotations files for the corresponding disorder beats. For example, a file in the MIT-BIH Arrhythmia Database may or may not have an atrial arrhythmia beat (“a” or “A”). If this file contains no such beat, we have disregarded it. For those files that do contain such a beat, we have determined empirically an interval of samples around the time of the occurrence of the disorder for which we have extracted the chaos features. We have performed this kind of manual overview and extraction on fewer files (only 7), with their list and vector count given in Table 2. Only annotations files have been treated in this way, since they show better classification results. We have taken first seven of the normal patients and of congestive heart failure patients. This set of files is named annotations specialized files.

Table 2. Set of the ECG annotations specialized files

Patient class / vector count	ECG files	Points interval
Atrial arrhythmia / 70	100, 101, 113	0 – 499
	103, 112	1000 – 1499
	108	220 – 719
	114	600 – 1099
Supravent. arrhythmia / 70	800, 801, 806, 807, 808	0 – 499
	809	1100 – 1599
	810	1400 – 1899
Normal / 70	16265 – 16773	0 – 499
CHF / 70	chf1 – chf7	0 – 499

When we obtained the Arff files intended for clustering and classification, it became apparent that additional preparation would have to be performed. Namely, it would be better not to base the analysis on redundant information such as the number of the ECG trail or the *m* factor, since these are the same for a great number of feature vectors. We have therefore either removed these attributes prior to the analysis process from an Arff file, or specified to the Weka system that these attributes should be disregarded. In this way, only the chaos features are the analyzed attributes.

3. Clustering and classification methods

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 in 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, as specified in Table 1. 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 [7]. 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 [9]. 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 is the landmark decision tree algorithm developed in 1993 by Quinlan [9].

C4.5 was used with reduced error pruning and a minimum amount 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) [4].

For classification purposes, a 10*4-fold cross-validation technique has been used in order to randomize the input samples and obtain a representative classification error. 4-fold was used instead of standard 10-fold because the number of vectors was relatively small. In this way, three fourths of random samples have been used for training and one fourth for testing.

4. Results

In Table 3, the results of the ECG signal files analysis are given. In Table 4, the results of the ECG annotations files are presented and in Table 5 the results of the annotations specialized files are given (see Table 2 for their complete list). Table 6 shows the results of clustering and classification for the ECG annotations

Table 3. Results for the ECG signal files, four classes

Clustered or classified samples, classification accuracy, %		Data volume		
		All data	1.trail	2.trail
Clustering	KMeans	37.1	37.3	35.6
	EM	39.9	38.0	36.9
Classification	C4.5	40.7	41.0	36.9
	BayesNet	37.6	38.6	35.6

Table 4. Results for the ECG annotations files, four classes

Clustered or classified samples, classification accuracy, %		Data volume		
		All data	$m = 1$	$m = 2$
Clustering	KMeans	42.7	46.4	45.8
	EM	46.8	46.8	52.5
Classification	C4.5	77.5	81.4	81.0
	BayesNet	78.1	87.8	70.8

Table 5. Results for the ECG annotations specialized files, four classes

Clustered or classified samples, classification accuracy, %		Data volume		
		All data	$m = 1$	$m = 2$
Clustering	KMeans	53.2	53.6	59.3
	EM	57.2	57.9	56.4
Classification	C4.5	81.8	75.0	80.7
	BayesNet	90.4	91.4	85.7

Table 6. Results for the ECG annotations specialized files, two classes

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

specialized files for two patient classes, one containing a normal heart rhythm and the other one containing a disorder. Classification accuracy performance measure is calculated for four classes' case. 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.

The frequency of feature usage for two clustering methods (KMeans and EM) is given in Fig 2. *ApEn* has been used only in the clustering of the annotations files.

5. Discussion

In order to demonstrate the problem that the clustering methods and classifiers had to face, we present in Fig. 3 the sample space in 2D, containing the samples from all the annotations files of the four mentioned patient classes, with features D_2 on the *x* axis and *CTM* on the *y* axis. The figure has been obtained using the Weka visualization package and it contains all the annotations sample vectors. It is obvious from Fig. 3 that it is difficult to perform efficient clustering.

From Table 3, we can perceive that it is not advisable to cluster or classify the ECG signal files, because the corresponding probabilities of success are too low, around 40%. However, the analysis performed on the annotations files shows some promising results. The classification and clustering of the annotations specialized files is somewhat better than of all the annotations files. This is expected, because a part of a record that contains a specific disorder beat is always easier to classify successfully than a part of a

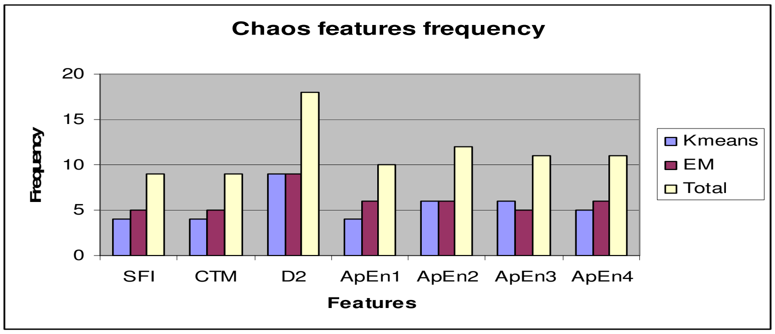

Figure 2. The frequency of chaos features occurring in optimal clustering

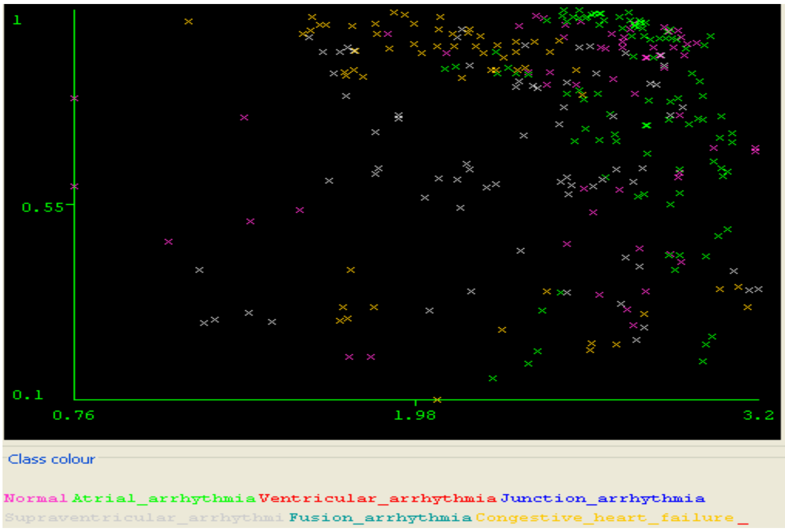

Figure 3. Annotation samples in 2D; features D_2 on y-axis and *CTM* on x-axis

record where there is no apparent disorder present. An annotations specialized file with four classes can be positively classified in about 84% of cases. We have used the annotations specialized files with only two patient classes present in order to find out if the chaos features are more 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 the 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 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 D_2 has been found the most useful feature in the clustering process in general. In several cases, it was the only feature used to optimally discern the patient classes. It is also the most established chaos feature in literature, together with Lyapunov exponent [2], with the latter not being the part of the platform at this time. *ApEn*, especially *ApEn*2 feature which determines *ApEn* trend, has been found the most useful in clustering of the annotations specialized files. We stress out that on several occasions the best clustering results have been obtained only when all four of the *ApEn* features have been used, either with other features or alone. It is also interesting to notice that *ApEn* is more often included in clustering of the annotations specialized files than of all the annotations files, probably because of the greater complexity and irregularity in the specialized files. Consequently, they are more appropriate for the *ApEn* analysis [5]. Further work should include the estimation of the efficiency of additional chaos features, such as the Lyapunov exponents and the Hurst exponent. It would be also interesting to try to select the optimal period between two consecutive points. Certainly, a successful predictor model for a disorder is the ultimate goal of the ECG chaos analysis.

6. Conclusion

The results show that it is not feasible to classify a patient using chaos features if the entire ECG signal file is analyzed, sample by sample. However, if we analyze the ECG annotations files, which contain only the heart beats (RR intervals), it is possible to classify them into the correct classes. The probability for the correct classification rises if only the parts of the annotations files that contain the disorder are analyzed and if the number of possible patient classes is reduced to two, thus discerning between normal and abnormal ECG record. EM clustering method has been found somewhat better the KMeans algorithm. Correlation dimension and approximate entropy have been found to be the most efficient features for the successful clustering of the ECG files. Additional work is needed to assert the application of the platform in clinical practice.

7. References

[1] 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.

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

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

[4] Friedman N., Geiger D., Goldszmidt, M., “Bayesian Network Classifiers”, Machine Learning 1997; 29(2-3): 131-163.

[5] 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.

[6] Chart-O-Matic, Physionet ECG database tool, http://www.physionet.org/ [1/2/2008]

[7] StatSoft, Inc., “Electronic Statistics Textbook”, Tulsa, OK, USA: StatSoft; 2007. http://www.statsoft.com/textbook/stathome.html [28/1/2008]

[8] 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.

[9] Witten, I.H., Frank, E., “Data Mining: Practical Machine Learning Tools and Techniques”, San Francisco: Morgan Kaufmann, SE, 2005.