Asociación Argentina
de Mecánica Computacional

Mecánica Computacional Vol XXXVI, págs. 1807-1814 (artículo completo)
José G. Etse, Bibiana M. Luccioni, Martín A. Pucheta, Mario A. Storti (Eds.)
San Miguel de Tucumán, 6-9 Noviembre 2018

MODEL BASED FEATURE EXTRACTION METHOD FOR MYOCARDIAL INFARCTION DETECTION

Sergio Liberczuk[a,b] and Lorena Bergamini[b]

[a]*Instituto de Ingeniería y Agronomía (IIyA), Universidad Nacional Arturo Jauretche UNAJ, , Av. Calchaquí 6200 (1888) Florencio Varela, Buenos Aires, Argentina, , https://www.unaj.edu.ar/*

[b]*Centro de Altos Estudios en Tecnología Informática (CAETI), Universidad Abierta Interamericana UAI, , Av. Montes de Oca 745, (C1270AAH) CABA, Argentina, , http://caeti.uai.edu.ar/inicio.aspx*

Keywords: ECG, SVM, Myocardial Infarction, McSharry ECG Model

Abstract. The electrical activity of the heart represented by the electrocardiogram (ECG) has been widely used for the detection of heart diseases. Long-term records require the automatic detection of cardiac events. In this work, the detection of myocardial infarction (MI) is performed by means of novel ECG features based on a synthesis ECG model previously described in the literature. The model consists of a sum of five Gaussians centered on each wave of the ECG (P, Q, R, S and T). Each Gaussian is fully specified by three parameters; location in time, amplitude and broadness. By fitting this set of Gaussians, and performing numerical and nonlinear optimization procedures in the resulting 15-dimensional space, we get this set of 15 parameters as features for classification. Although the model was widely used previously with different purposes, its parameters had never been used as features for heartbeat classification even though they reflect the morphology of the ECG in an accurate manner. Physikalisch-Technische-Bundesanstalt (PTB) database was used to validate training and testing algorithms. Data was obtained from 48 healthy subjects (HS) and 95 patients with MI and was split into two datasets. The first dataset contains 190 beats from 26 HS, and 140 beats from 60 patients with MI and was used to train a support vector machine (SVM) classifier with linear kernel. The second dataset contains 88 beats from 22 HS, and 70 beats from 35 subjects with MI and was used to provide a detection performance assessment of the previously trained SVM. This assessment yielded an overall accuracy above 93%. The results show the feasibility of performing the separation between infarcted beats and physiological beats based on the new model-based features proposed. The simplicity of the linear kernel used in the SVM classifier shows the power of the proposed features for classification tasks.

1 INTRODUCTION

According to data from the World Health Organization (WHO) and the Ministry of Health from Argentina, just in Argentina, over 40% of all registered deaths were due to Cardiovascular Diseases (CVD) in 2013 Ministerio de Salud (11-07-2018). The most prevalent form of heart disease is myocardial infarction resulting from a thrombus that obstructs blood flow in one or more coronary arteries. The sooner thrombolytic medication such as tissue plasminogen activator or urokinase is placed into the patient's bloodstream after the occurrence of a myocardial infarction, the sooner an obstructive thrombus will be dissolved and some perfusion of the myocardium can occur. The damage to the myocardium is strongly dependent on the length of time that occurs prior to restoration of some blood flow to the heart muscle. This makes early detection have a strong impact on the quality of life of thousands of people locally and worldwide.

The analysis of biomedical data to determine patterns describing physiological and pathological behaviors is crucial to achieve this goal. Specific algorithms must be developed to perform the analysis and processing of the data and thus to obtain rich and useful information to transfer. This information will allow the semi-automation of earlier and more accurate diagnoses, supported by specific devices designed for that purpose. Over the years the ECG signal has been used to assess the cardiovascular condition of humans recording electrical activity of the heart. Several electrodes, arranged conveniently on the surface of the thorax, acquire the temporal variation of the electrical potential that cardiomyocites produce. The morphology of this record and its interpretation from the detection of its characteristic waves (so-called fiducial points that comprise the P, Q, R, S, T waves) as well as various calculations that arise from the detection of such waves (ST segment, QT interval, PR interval and others) allow the diagnosis of various pathologies such as different cardiac arrhythmias, ischemic heart disease or conduction abnormalities. Therefore this type of non-invasive and low-cost analysis continues being a fundamental tool for the cardiovascular evaluation of patients who arrive by spontaneous demand to the emergency rooms of any health center.

As mentioned before the study of the ECG signal provides substantial information of the heart function so modelling ECG signal becomes very useful for different purposes such as characterization, compression or classification, all of them, problems of concern for the biomedical engineering community. Among the works that have dealt with the idea of modeling the wave sequence in an ECG, to extract and recognize patterns, we can mention the articles Sornmo et al. (1981); Lagerholm et al. (2000) which proposes a model and classification method for the QRS complex (formed by the Q wave, the R wave and the S wave) using an orthonormal basis of Hermite functions. Baali et al. (2014) propose a parametric model based on orthogonal transformations, that involves the mapping of the ECG in the domain of singular values, whereas Philips and De Jonghe (1992) apply a polynomial approximation for the compression of ECG data. Suppappola et al. (1997) focus on the modeling of ECG waves with Gaussian pulses. Thus, an ECG cycle results in a sum of such Gaussian pulses. Each Gaussian is characterized by its location, its amplitude and its width. The mentioned work Suppappola et al. (1997) presents an iterative algorithm to approximate a given ECG by means of this model, estimating the necessary parameters. One of the parameters to be determined is the number of pulses that are needed to achieve a good representation for a given real ECG.

The work of Clifford et al. (2005) uses 5 Gaussian functions, one for each characteristic wave and McSharry et al. (2003) proposes a dynamical model whose solution trajectories reproduce realistic synthetic ECG waves. The model generates a trajectory in the space of states (x, y, z). The approximate quasi-periodicity is reflected by the movement of the path around an attractive boundary cycle, a unit circumference in the (x, y) plane. Each beat is represented with a revolution around this limit cycle. Thus, the model remains dependent on 3M morphological parameters, where M is the number of Gaussian functions involved. In Clifford and McSharry (2005) a method to find the parameters that best reproduce a real given beat is proposed, thus achieving compression with loss. The parameter adjustment is carried out using non-linear optimization (gradient descendent method) to minimize the Euclidean distance between the data and the simulated model (least squared error methodology). This would allow, according to the mentioned authors, to predict the performance of the model in a segment of an ECG and facilitate the rejection of beats for a specific study. The deviation of these parameters with respect to the physiological parameters would thus indicate a change in the morphology of the ECG constitutive waves, showing alterations that point out or suggest certain pathologies. The model then allows the 3M-dimensional representation of any ECG, physiological or pathological, which can then be used not only in filtering schemes that require a model but in compression, clustering and / or pattern classification applications to ECG signals in the mentioned space Clifford et al. (2005); Clifford and McSharry (2005).

In the present work we extract features of normal and pathological ECG beats in a parametric way fitting the heartbeat with a sum of Gaussian curves using two techniques: the classical Least Squares Method and our new method based on Monte Carlo simulation ideas Liberczuk and Bergamini (2017). Finally we use McSharry model parameters (very known in the literature before but never with this purposes) in order to represent and classify ECG signals coming from 48 healthy subjects and 95 patients with MI. The objective is to classify and separate physiological from mycardial infarction beats using a linear kernel SVM (Support Vector Machine).

2 MATERIALS AND METHODS

2.1 Parameter estimation

The features we will use to classify heartbeats are model parameters, that arise from a heartbeat model proposed by Clifford and Mc Sharry Clifford et al. (2005); McSharry et al. (2003), which is widely utilized in the literature. The model assumes that each heartbeat in a ECG is modeled by a set of Gaussian waves, characterized by their amplitude, position and width.

The model can be stated as

$$z(k) = \sum_{i=P,Q,R,S,T} a_i e^{\frac{-(k-\theta_i)^2}{2b_i^2}} \tag{1}$$

where θ_i is the position of the corresponding wave peak, a_i corresponds to the wave amplitude and b_i corresponds to wave width. Given a real signal $s(k)$, we want to obtain the set of parameters in Eq. 1 that best represents that signal. They could be calculated as those that minimize the squared error between the mentioned signal $s(k)$ and the parametric model $z(k)$.

$$\underset{a_i,b_i,\theta_i}{argmin} \left\| z(k) - s(k) \right\|^2 \tag{2}$$

The problem stated in Eq. 2 is nonlinear, and it is highly likely to have many suboptimal solutions. It can be solved by a local optimization algorithm, but a good initial point should be provided such that the solution be reliable. Otherwise, a global optimization algorithm may be applied to solve Eq. 2. In our previous work Liberczuk and Bergamini (2017) we propose an heuristic global optimization, based on a Monte Carlo search. This type of approach avoids the problem of getting stuck in local minimum, since it permits random exit of suboptimal neighborhood.

2.2 Data Base

Real signals were taken from internationally validated databases such as the Physikalisch-Technische Bundesanstalt (PTB) ECG Database available in Physio-Bank Goldberger et al. (2000 (June 13); PTB Database (22-03-2018). The National Metrology Institute of Germany has provided this compilation of digitized ECGs for research, algorithmic benchmarking or teaching purposes to the users of PhysioNet. The ECGs were collected from healthy volunteers and patients with different heart diseases by Professor Michael Oeff, M.D., at the Department of Cardiology of University Clinic Benjamin Franklin in Berlin, Germany. It contains records of 52 healthy subjects and 148 patients with myocardial infarction and also provides some patients with other pathologies like Cardiomyopathy, Bundle branch block, Dysrhythmia, Myocardial hypertrophy and Valvular heart disease. The ECGs are digitized at 1Khz, with 16 bits resolution over a range of 16,384mV. Each record includes the 12 simultaneous leads and the orthogonal leads of Franz. The patient's medical history is available. We have selected 278 heartbeats from healthy patients and 210 heartbeats from patients who have suffered anterior myocardial infarction.

2.3 Parameter collection

Single lead ECG Data (Lead 2) was extracted from the PTB Database described in the previous section. This records were all preprocessed with 5th-order Butterworth highpass filter (Fc=0,5 Hz) for baseline wander rejection. A peak-detection algorithm was then applied to each signal, to isolate beats. R-peaks were detected in the array with Pan-Tomkins algorithm for QRS detection Pan and Tompkins (1985). From each record, a set of beats were selected and the optimization process indicated in Eq 2 was solved for each selected beat. The beats were normalized (uniformly scaled) to have all the same length. Thus, position parameters actually represents the relative peak position inside the beat. The 15-dimensional parameter array $X = [a_P, \ldots, a_T, \theta_P, \ldots \theta_T, b_P, \ldots, b_T]$ that represents each beat was first saved in a parameter vector and then appended to a file. Beat classification was carried on applying a Support Vector Machine (SVM) to the set of features acquired. In this first attempt, we worked on a two-class classification scheme that will be explained in the next section.

2.4 Classification

We have implemented a SVM algorithm for binary classification to classify the set of parameters that we collected from the heartbeats of the diferent subjects described in the previous section. The selected kernel was a linear kernel because it gave excellent results so it was no necessary to increase the complexity.

SVM is a supervised machine learning algorithm, well suited for classification or regression problems Awad and Khanna (2015). The algorithm consists of 2 stages, the training and the testing stage.

In the training phase of this algorithm, each feature vector is considered as a point in a N-dimensional space (N=15 in our case), labeled with its corresponding class label (HS healthy subject beat or Myocardial Infarction MI beat). The algorithm finds the hyper-plane that best separates these two classes, by maximizing a margin function that accounts for the separation between the points in both classes. In the testing phase, new vectors are evaluated, to determine in which side of the space they lie, and thus assigning the corresponding class label.

In some cases, the data is not completely separable. For these cases, a penalty cost is added to the margin function, penalizing points that lie in the wrong half-space. There are also cases when data is not linearly separable in the original feature space. For those cases, separation is made in a higher dimensional space. It is performed after projecting data vectors into a new space, using a kernel function. In the new space, the points are separated with a hyper-plane that represents a nonlinear separation frontier in the original feature space. In our problem, we found that SVM was able to linearly separate data in the 15-dimensional space, as results show in the next section.

The training dataset was built with 190 beats from 26 HS, and 140 beats from 60 patients with MI. The testing dataset contained 88 beats from 22 HS, and 70 beats from 35 subjects with MI and was used to provide a detection performance assessment of the previously trained SVM.

3 RESULTS

We load the parameter file that resulted from the parameter collection process described in the previous section and run 10 times using different set of beats randomly taken for training, and leaving the rest for testing. Then we classified the test beats obtaining an average classification rate of 93% over the 10 trials.

To evaluate the classification performance, we report the statistical measures in Table 1.

For the only purposes of possible graphing and visualization we have selected three 2-dimensional graphs to show the relative position of the parameters in each class.

Figure 1 shows the classification results for the testing data in the two dimensional parameter plane a_P vs. a_R. These parameters represent amplitudes in the P-wave and the R-wave respectively. Figure 2 shows the classification of testing data in the two dimensional parameter plane b_P vs. b_S. These parameters represent widths in the P-wave and the S-wave respectively. Figure 3 shows the testing classification in the space b_P vs. b_T. These parameters represents widths in the P-wave and the T-wave respectively.

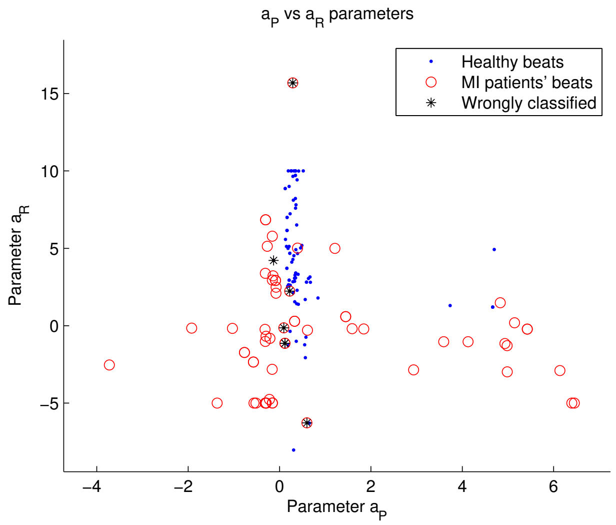

Figure 1: Testing results: a_P vs. a_R parameter space

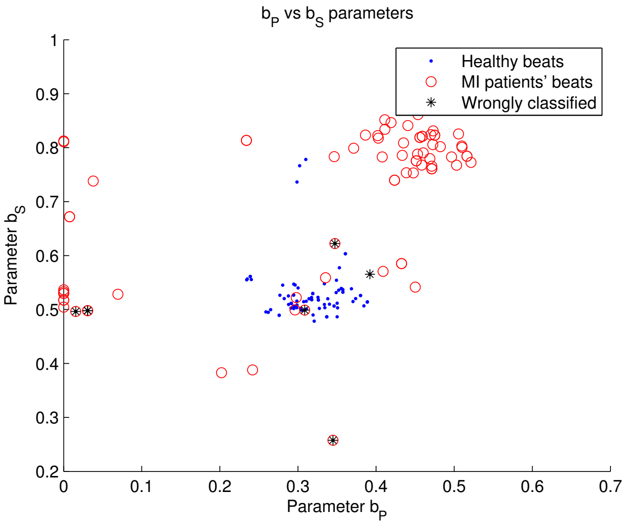

Figure 2: Testing results: b_P vs. b_S parameter space

Table 1: Statistical Measures

Sensitivity (true positive rate - TPR)	0.957
Specificity(true negative rate - TNR)	0.909
Precision (positive predicted value - PPV)	0.893
Negative predicted value (NPV)	0.964
Accuracy (ACC)	0.93

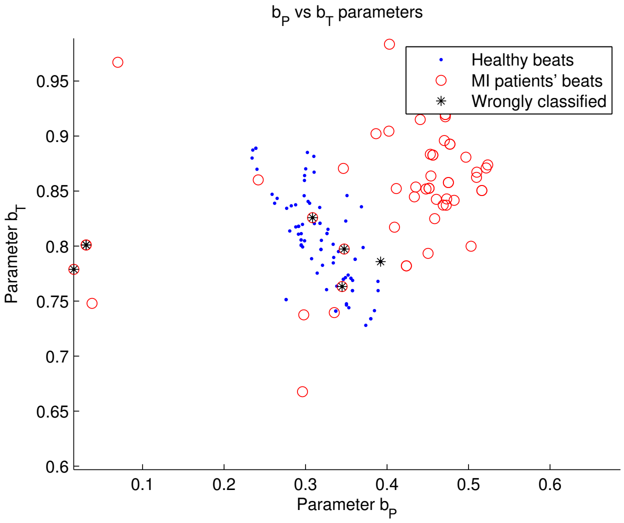

Figure 3: Testing results: b_P vs. b_T parameter space

4 DISCUSSION

We found that SVM with a linear kernel could very accurately classify the heartbeat characteristics of healthy subjects from the characteristics of non-physiological heartbeats.

In the figures you can see that the points in both classes were grouped into two regions. The regions do not seem to be completely defined, but we must bear in mind that they are two-dimensional projections of the space of original parameters. It can also be seen that the parameters that were erroneously classified (both false positives and false negatives, a 7 % of the total test parameters), are located at the edge of the grouping regions. It is reasonable to think that the points near the separation hyperplane do not show a clear pattern for the classes considered, and then the SVM algorithm could confuse them. These poorly coded patterns could be due to factors in the generation stage of the characteristics (parameter estimation), that is, suboptimal representation of the heartbeat.

In Table 1 we can see that the True Positive Rate (TPR) is almost 96 %, this rate is higher than the True Negative Rate (TNR) which is almost 91 %. This is an important point because the algorithm better detects the cases that require attention i.e. people with myocardial infarction. In contrast, 91 % of TNR means that for 9 % of healthy subjects, the SVM would detect it erroneously as infarction, but this case would not be so serious because those subjects could be observed with other studies and later determined they were in good condition.

5 CONCLUSIONS AND FUTURE WORK

We have used novel features for classification of ECG beats with Anterior Myocardial Infarction (AMI) from ECG beats coming from healthy subjects. The results have shown the feasibility of performing the separation between infarcted beats and physiological beats based on these new model-based features proposed. The simplicity of the linear kernel used in the SVM classifier shows the power of the proposed features for classification tasks. In the future we will incorporate more pathologies such as Bundle Branch Block or Myocardial Hypertrophy as well as other types of infarction, and the possibility to work with more than two classes in the classification task.

ACKNOWLEDGEMENTS

The authors acknowledges to Universidad Abierta Interamericana and Universidad Nacional Arturo Jauretche for the financial support that allowed the realization of the present paper.

REFERENCES

Awad M. and Khanna R. *Support Vector Machines for Classification*, pages 39–66. Apress, Berkeley, CA, 2015. ISBN 978-1-4302-5990-9. doi:10.1007/978-1-4302-5990-9_3.

Baali H., Akmeliawati R., Salami M.J.E., Khorshidtalab A., and Lim E. Ecg parametric modeling based on signal dependent orthogonal transform. *IEEE Signal Processing Letters*, 21(10):1293–1297, 2014.

Clifford G. and McSharry P. Method to filter ecgs and evaluate clinical parameter distortion using realistic ecg model parameter fitting. In *Computers in Cardiology, 2005*, pages 715–718. IEEE, 2005.

Clifford G., Shoeb A., McSharry P., and Janz B. Model-based filtering, compression and classification of the ecg. *International Journal of Bioelectromagnetism*, 7(1):158–161, 2005.

Goldberger A.L., Amaral L.A.N., Glass L., Hausdorff J.M., Ivanov P.C., Mark R.G., Mietus J.E., Moody G.B., Peng C.K., and Stanley H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000 (June 13). Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

Lagerholm M., Peterson C., Braccini G., Edenbrandt L., and Sornmo L. Clustering ecg complexes using hermite functions and self-organizing maps. *IEEE Transactions on Biomedical Engineering*, 47(7):838–848, 2000.

Liberczuk S. and Bergamini L.A.P. Heart beat parametric modeling based on monte carlo fitting techniques. *XXI Congreso Argentino de Bioingeniería, X Jornadas de Ingeniería Clínica*, 2017.

McSharry P.E., Clifford G.D., Tarassenko L., and Smith L.A. A dynamical model for generating synthetic electrocardiogram signals. *IEEE Transactions on Biomedical Engineering*, 50(3):289–294, 2003.

Ministerio de Salud. http://www.msal.gob.ar/ent/index.php/vigilancia/areas-de-vigilancia/mortalidad. 11-07-2018.

Pan J. and Tompkins W.J. A real-time qrs detection algorithm. *IEEE transactions on biomedical engineering*, (3):230–236, 1985.

Philips W. and De Jonghe G. Data compression of ecg's by high-degree polynomial approximation. *IEEE Transactions on Biomedical Engineering*, 39(4):330–337, 1992.

PTB Database. https://www.physionet.org/physiobank/database. 22-03-2018.

Sornmo L., Borjesson P.O., Nygards M.E., and Pahlm O. A method for evaluation of qrs shape features using a mathematical model for the ecg. *IEEE Transactions on Biomedical Engineering*, (10):713–717, 1981.

Suppappola S., Sun Y., and Chiaramida S.A. Gaussian pulse decomposition: an intuitive model of electrocardiogram waveforms. *Annals of biomedical engineering*, 25(2):252–260, 1997.