32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

Evaluation of an ECG Heartbeat Classifier Designed by Generalization-Driven Feature Selection

Mariano Llamedo and Juan Pablo Martínez

Abstract— **In this work we studied the classification performance of feature models selected with a floating algorithm, focusing in the generalization capability. The features were extracted from the RR interval series, from all ECG leads and different scales of the wavelet transform. The generalization was studied using Physionet databases. In all databases the AAMI recommendations for class labeling and results presentation were followed. A floating feature selection algorithm was used to obtain the best performing and generalizing models in the training and validation sets for different search configurations. The best model found includes 8 features, was trained in a partition of the MIT-BIH Arrhythmia database, and was evaluated in a completely disjoint partition of the same database. The results obtained were: global accuracy of 93%; for normal beats, sensitivity (S) 95%, positive predictive value (P^+) 98%; for supraventricular beats, S 77%, P^+ 39%; for ventricular beats S 81%, P^+ 87%. This classifier model has less features and performs better than other state of the art methods with results suggesting better generalization capability.**

I. INTRODUCTION

Cardiovascular diseases are currently the biggest single cause of death in developed countries according to their public health agencies. The analysis of the electrocardiographic signal (ECG) provides a noninvasive and inexpensive technique to analyze the heart function for different cardiac conditions. One important analysis performed in the ECG is the classification of heartbeats, which is important for the study of arrhythmias.

Many algorithms for ECG classification were developed in the last decade [1]–[4], but only few of them have completely comparable methodologies and therefore results [2], [3]. The Association for the Advancement of Medical Instrumentation (AAMI) recommendations [5] for class labeling and results presentation have eased this problem, and at the present time it is broadly accepted [2]–[4]. Regarding to the classes of interest, the AAMI recommendation suggests 5 classes: supraventricular (S) and ventricular (V) ectopic beats, fusion of normal and ventricular beats (F), a class including paced beats, fusion of paced and normal beats and beats that cannot be classified (Q), and finally a normal or bundle branch block beats (N) [5]. It is remarkable that all previous works were interested in discriminating between N and V classes, but only few of these works studied the multiclass classification problem [2], [3]. In terms of the data division, some works performed a beat-oriented division no matter which subject the heartbeats belong to, with the inconvenience that sometimes heartbeats from some subjects were included in both the training and testing datasets [4]. It was shown in [2] that this approach leads to an optimistic bias of the results, being more advisable a patient-oriented division, as it will also happen in the application scenario where the algorithm is to be used.

The objective of this work is to develop and evaluate a heartbeat classification model including the most discriminating features, in order to achieve the best performance in a multidatabase context. The algorithm developed will be completely automatic, compliant with AAMI recommendations, based on a simple classifier and robust features with physiological meaning. The developed classifier will be compared with [2], the best performing multiclass classifier to our best knowledge.

II. METHODOLOGY

A. ECG databases

In this work we used the well-known MIT-BIH Arrhythmia database (MIT-BIH-AR) for training and testing purposes. Additionally, the MIT-BIH Supraventricular Arrhythmia database (MIT-BIH-SUP) was used for validation purposes. All databases are described and freely available on Physionet [6]. We adopted the training ($DS1$) and test ($DS2$) set division scheme used in [2] to allow comparison. The AAMI Q class (unclassified and paced heartbeats) was discarded since it is poorly represented in all databases. A similar limitation occurs with the fusion (F) AAMI class, but instead of discarding the heartbeats of this class, a class-labeling modification to the AAMI recommendation was adopted. It consists in considering fusion (of normal and ventricular beats) and ventricular classes, as the same ventricular class. We will refer to this modification as AAMI2 labeling. The division scheme and class presence is summarized in Table I.

B. Signal processing

The sampling frequency of the MIT-BIH-SUP was first resampled to 360 Hz to become compatible with respect to the MIT-BIH-AR. This was performed with a tenth order lowpass filter without observing any notorious distortion. All recordings in all databases were first preprocessed to remove artifacts as described in [2]. No energy or amplitude normalization was done, as we were interested in some amplitude related features. Many of the considered features

Mariano Llamedo is with the Electronic Department, National Technological University, Buenos Aires, Argentina. (llamedom@electron.frba.utn.edu.ar).

J. P. Martínez is with the Communications Technology Group, Aragón Institute of Engineering Research, University of Zaragoza. Spain. (jpmart@unizar.es).

Mariano Llamedo and J. P. Martínez are with the CIBER of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Spain.

TABLE I

SCHEME OF THE DIVISION OF THE MIT-BIH-AR DATABASE. RECORDINGS WITH PACED BEATS WERE EXCLUDED. HEART BEATS CLASSES ARE N: NORMAL, S: SUPRAVENTRICULAR, V: VENTRICULAR AND F: FUSION. BELOW THE DETAILS OF THE DATASET RECORDING DIVISION PERFORMED IN THE MIT-BIH-AR DATABASE FOR BUILDING TRAINING ($DS1$) AND TESTING ($DS2$) SETS.

	MIT-BIH-AR Arrhythmia					
Dataset	purpose	N	S	V	F	#Rec
$DS1$	train	45673	929	3755	412	22
$DS2$	test	44053	1833	3202	388	22
MIT-BIH-SUP	validation	161902	12083	9897	193	78

Dataset	MIT-BIH-AR recordings
$DS1$	101, 106, 108, 109, 112, 114, 115, 116, 118, 119, 122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223, 230
$DS2$	100, 103, 105, 111, 113, 117, 121, 123, 200, 202, 210, 212, 213, 214, 219, 221, 222, 228, 231, 232, 233, 234

are based on the discrete wavelet transform (DWT) of the ECG signal. We used the derivative of a smoothing function (quadratic spline) as the prototype wavelet, resulting the different scales of the DWT as a smoothed derivative of the ECG. As a result, the DWT retains at certain scales the useful information present in the ECG in form of absolute maxima and zero-crossings. See [7] for details in the DWT implementation for ECG delineation. Following the conclusions of [7], the resulting WT framework allows an analysis robust to the typical interferences present in routine ECG recordings, so the features derived from the DWT could inherit this desirable property.

C. Heartbeat classification: classifier and features

Under the assumption of independent and normally distributed data, using the maximum *a posteriori* criterion (MAP) we reach to the quadratic discriminants functions, broadly used for classification purposes. The general quadratic discriminant functions for feature vectors $\mathbf{x}$, and the i-th class can be written as

$$\begin{aligned} g_i(\mathbf{x}) &= -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{x} + \boldsymbol{\mu}_{\mathrm{i}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{i}}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i \\ &\quad -\frac{1}{2}\log(|\boldsymbol{\Sigma}_i|) + \log(P(\omega_i)) \end{aligned} \quad (1)$$

The classification rule assigns $\mathbf{x}$ to the class i which results in the maximum posterior probability $g_i(\mathbf{x})$. Being $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$ and $P(\omega_i)$ the mean vector, covariance matrix and prior probability of the i-th class. The values of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ were computed as the sample mean and covariance matrix. The values for the prior probabilities $P(\boldsymbol{\omega}_i)$ were considered the same for all classes. In the case that the covariance matrix $\boldsymbol{\Sigma}$ is considered to be unique for all classes ($\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_j, \forall i \neq j$), the quadratic discriminant classifier (QDC) becomes linear in $\mathbf{x}$ leading to the linear discriminant classifier (LDC) where $\boldsymbol{\Sigma}$ can be estimated as the weighted sample covariance

$$\boldsymbol{\Sigma} = \frac{\sum_{i=1}^{C} w_i \sum_{m=1}^{M_i}(\mathbf{x}_i(m) - \boldsymbol{\mu}_i).(\mathbf{x}_i(m) - \boldsymbol{\mu}_i)^{\mathrm{T}}}{\sum_{i=1}^{C} w_i.M_i}. \quad (2)$$

The class-weighting possibility is interesting due to the heavy class-size unbalance inherent to this application, where the normal class is in general one order of magnitude more represented that other classes. We refer as LDC to the linear classifier where $\mathbf{w}_i = \mathbf{w}_j, \forall i \neq j$, any other weight scheme will be referred as compensated linear classifier (LDC-C).

Following the conclusions of previous works [1], [2], we included in our model both interval and morphological features. As interval features we used features from the RR sequence $RR[i-1]$, $RR[i]$ and $RR[i+1]$ to describe the local time evolution of the heart rhythm. In order to assess the local variation of the heart rhythm, the feature $RR_V[i] = \sum_{j=-1}^{1} |dRR[i-j]|$ (being $dRR[i] = RR[i] - RR[i-1]$) comprehends the variation in the surrounding heartbeats. We also included estimates of the local and global rhythm by the mean RR interval in the last 1, 5, 10 and 20 minutes (RR_P being $P \in \{1, 5, 10, 20\}$, the interval in minutes of aggregation).

As morphological features we considered the QRS width, from the vectocardiogram (VCG) loop constructed with the two available leads we calculated the maximum modulus of the QRS loop and the angle of the loop at this position. Other morphological features were calculated from peak amplitudes and positions from the fourth scale of the DWT, since this scale has good projection of the ECG information (between 12.25–22.5 Hz). From the same scale of the DWT, the autocorrelation sequence for both leads ($r_x(k)$ and $r_y(k)$) and the inter-lead cross-correlation signal ($r_{xy}(k)$) were calculated within a time window which starts 130 ms before the fiducial point and ends 200 ms after. Then we calculated for the 3 correlation sequences the location and value of the absolute maximum, and for r_x and r_y the location of the first zero-crossing, as shown in Figure 1. One remarkable aspect is that features calculated from the correlation signals will essentially be synchronized in time, even if the fiducial point is not accurately determined.

The complete set of features consists of 39 features, related to the heart rhythm and QRS complex morphology. It is well known that low dimensional models generalize better to examples not presented during the training phase, resulting in a more robust and realistic classifier. In order to obtain the smallest and best performing model, a sequential floating feature selection algorithm (SFFS) was used [8].

D. Experiment Setup

In this work we are interested in finding a small, well performing and generalizing model in a multidatabase context. The experiment can be divided in three steps:

1) The first step is to find the best performing model from the pool of available features in the training ($DS1$ of MIT-BIH-AR) and validation (MIT-BIH-SUP) sets, as is shown in Figure 2a. For the model (or features) selection we used all data except $DS2$ of the MIT-BIH-AR, which was reserved for the final performance evaluation (test set) as in [2]. Each iteration of the SFFS algorithm, the current model

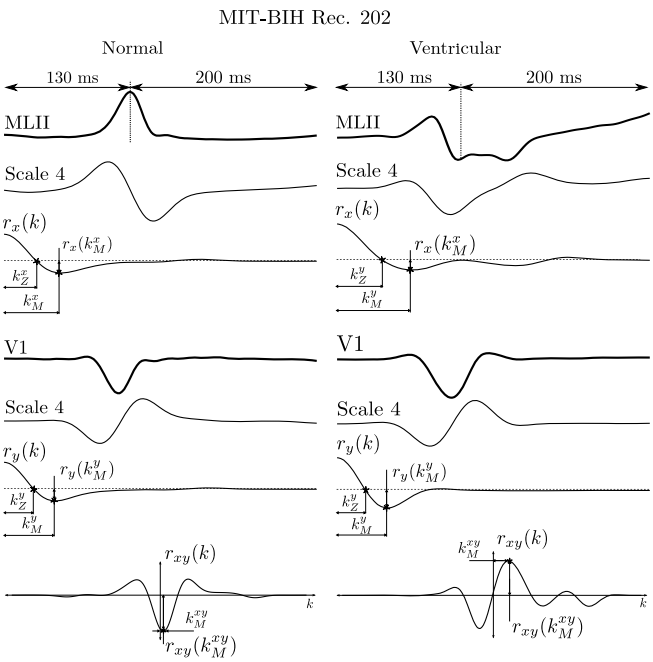

Fig. 1. Illustration of the features calculated from the wavelet correlation signals for the same normal and ventricular beats. The autocorrelation signal of the QRS complex at scale 4 is shown for both leads (r_x and r_y) as well as the cross-correlation signal (r_{xy}) at the bottom. The zero-crossings and peaks of interest are indicated with an asterisk.

was trained in $DS1$ of MIT-BIH-AR and its performance was evaluated in the MIT-BIH-SUP database, ensuring in this manner the generalization of the selected models. As the data divisions in both databases does not share any recording, the features selected should retain the generalization properties. Several parameter configurations were studied for the SFFS algorithm, like the effect of the classifier (LDC, LDC-C and QDC) and the optimization criterion for the search. The optimization criteria used were the mean class positive predictive value (J_{P^+}) and the mean class sensitivity (J_S). The weight compensation used in the experiments for the LDC classifier is $w_N = 1$, $w_S = 10$ and $w_V = 10$.

2) The second step is the selection of the best performing model, among the best models obtained with the SFFS for the different parameter configurations in the previous step. For that purpose, we compare the results obtained in the union set of $DS1$ of MIT-BIH-AR dataset and the MIT-BIH-SUP database, using a recording-based k-fold cross validation with $k = 10$ recordings, as it is shown in Figure 2b.

3) Finally the performance of the selected model is evaluated in $DS2$ for comparison with [2], as shown in Figure 2c.

All experiments described in this work will focus to achieve automatic classification between the three AAMI2 classes (N, S and V'), since the fusion class is poorly represented in the databases used. The restrictions imposed by the recording-oriented division of the data, and the fact that only a few recordings concentrates the majority of the examples of the fusion heartbeats, makes unfeasible

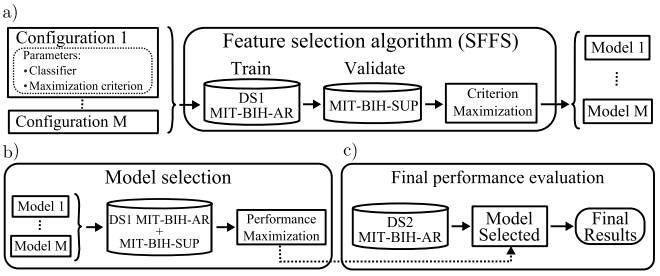

Fig. 2. In the picture a block diagram describing the experiments performed in this work is shown. In panel A the feature selection algorithm is summarized, indicating the train and validation dataset division, as well as the different parameters of the algorithm. In panel B is shown the methodology to obtain the best performing model among the different searches performed. Finally in panel C, the best performing model is selected for the final performance evaluation in the test dataset.

to perform the feature selection using the original AAMI labeling.

III. Results

The results of the experiments described in the previous section are presented in tables II and III. In table III, balanced performance calculation means that all rows (or classes) in the confusion matrix were scaled to sum the same. This is equivalent to repeat examples of the less represented classes, in order to balance the class presence. The best model found was an LDC-C with 8 features, trained in the $DS1$ of the MIT-BIH-AR. The features were $\ln(RR[i])$, $\ln(RR[i+1])$, $\ln(RR_1)$, $\ln(RR_{20})$, k_Z^x, k_Z^y, k_M^x and k_M^y.

IV. Discussion and conclusions

The best model found consists of 8 features, and as can be noted, the selected features are computed without exception from time interval measurements. This could be explained given that the used databases, do not always include the same pair of ECG leads in each recording, and naturally the classification performance of features which are calculated from amplitudes are heavily degraded. The first four features in the model are clearly connected to the evolution of heart rhythm, while the other four can be understood as surrogate measurements of the QRS width, and therefore the QRS morphology. As a result, the model found has the evident advantage of a lower size, which results in a computational saving and lower error in the parameter estimation during the training phase. In addition, it only relies on the QRS fiducial point detection, making the classifier model robust to signals where the delineation of the ECG waves is not reliable.

In this work we have developed and evaluated a heartbeat classification system focusing in the generalization capability during the feature selection; as a result we obtained improvements in size and performance. In order to do this, we included in the development the MIT-BIH-SUP database [9], freely available in Physionet [6]. The limitation of a not as well represented fusion class, is overcome by adopting the alternative labeling AAMI2 proposed in this work. The AAMI2 labeling could have a physiological interpretation since the AAMI fusion class comprehends those heartbeats

TABLE II

SUMMARY OF THE BEST PERFORMING MODELS FOUND WITH THE SFFS ALGORITHM SEPARATING ALL AAMI2 CLASSES ACCORDING TO FIGURE 2B. THE BEST PERFORMING MODEL (IN BOLD) IS SELECTED FOR THE FINAL PERFORMANCE EVALUATION. THE RESULTS ARE EXPRESSED IN PERCENTAGES.

Configuration parameters			Model Evaluation								
	Opt.	Resultant	Normal		Suprav.		Ventr.		Total		
Classifier	Crit.	# Features	S	P^+	S	P^+	S	P^+	A	S	P^+
LDC-C	J_{P^+}	**8**	**93**	**98**	**78**	**40**	**68**	**70**	**91**	**80**	**70**
QDC	J_{P^+}	7	80	98	7	12	89	22	77	59	44
LDC	J_S	10	92	98	74	37	70	67	89	78	67
QDC	J_S	9	87	98	43	32	80	33	84	70	55
de Chazal et al. [2]		48	87	98	57	30	63	36	84	69	55

TABLE III

PERFORMANCE COMPARISON BETWEEN THE MODEL SUGGESTED IN THIS WORK AND THE REFERENCE CLASSIFIER [2] SEPARATING AAMI2 CLASSES IN $DS2$ OF MIT-BIH-AR. BOTH MODELS WERE TRAINED IN $DS1$ OF THE SAME DATABASE. FIRST THE CONFUSION MATRICES FOR BOTH MODELS ARE SHOWN, AND BELOW THE CLASS AND TOTAL PERFORMANCES ARE SUMMARIZED. THE PERFORMANCES ARE EXPRESSED IN PERCENTAGES FOR BOTH, BALANCED AND UNBALANCED CLASS PRESENCE IN THE DATASET.

de Chazal et al. [2]

		Algorithm			
		n	s	v'	Total
Truth	N	40718	1863	1677	44258
	S	307	1361	169	1837
	V'	235	845	2529	3609
	Total	41260	4069	4375	49704

This work

		Algorithm			
		n	s	v'	Total
Truth	N	41950	2002	236	44188
	S	216	1422	197	1835
	V'	473	222	2911	3606
	Total	42639	3646	3344	49629

Performance			Normal		Suprav.		Ventr.		Total		
calculation mode	Classifier	# Features	S	P^+	S	P^+	S	P^+	A	S	P^+
Unbalanced	This work	8	95	98	77	39	81	87	93	84	75
	de Chazal et al. [2]	48	92	99	74	33	70	58	90	79	63
Balanced	This work	8	95	79	77	88	81	88	84	84	85
	de Chazal et al. [2]	48	92	80	74	73	70	84	79	79	79

which results from the simultaneous occurrence of normal and ventricular heartbeats.

From the results obtained for the model selection presented in Table II, several models that outperform the reference classifier in the train and validation datasets were obtained. The selected model corroborates the generalization capability when is evaluated in heartbeats not considered during the development phase, as shown in table III. It is worth to note than the performance achieved by both compared classifiers in Table II is lower for all classes than the obtained in the final performance reported in Table III. This phenomenon was also reported in the original work [2] suggesting further corroboration of the performance achieved. The validity of the generalization capability of the proposed model, is somehow restricted to the available data, and should be corroborated in future works by including new databases in the analysis. Despite this limitation, the degree of generalization should be better than the works reviewed, which only were developed considering the MIT-BIH-AR database. The results presented in this work are an improvement regarding to the size of the classification model and the performance achieved.

ACKNOWLEDGMENTS

This work was supported by projects TEC-2007-68076-C02-02 from CICYT and GTC T-30 from DGA (Spain). The CIBER of Bioengineering, Biomaterials and Nanomedicine is an initiative of ISCIII.

REFERENCES

[1] Y. H. Hu, S. Palreddy, and W. Tompkins, "A patient-adaptable ecg beat classifier using mixture of experts approach," *IEEE Transactions on Biomedical Engineering*, vol. 44, pp. 891–899, 1997.

[2] P. de Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ecg morphology and heartbeat interval features," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 1196–1206, 2004.

[3] K. Park, B. Cho, D. Lee, S. Song, and J. Lee, "Hierarchical support vector machine," in *Computers in Cardiology 2008*, vol. 35. IEEE Computer Society Press, 2008, pp. 229–232.

[4] T. Ince, S. Kiranyaz, and M. Gabbouj, "A generic and robust system for automated patient-specific classification of ecg signals," *IEEE Transactions on Biomedical Engineering*, vol. 56, pp. 1415–1426, 2009.

[5] *Testing and reporting performance results of cardiac rhythm and ST-segment measurement algorithms*, American National Standard, ANSI/AAMI/ISO EC57, American National Standard Std., 1998–(R)2008.

[6] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000, circulation Electronic Pages: http://circ.ahajournals.org/cgi/content/full/101/23/e215.

[7] J. Martínez, R. Almeida, S. Olmos, A. Rocha, and P. Laguna, "A wavelet-based ecg delineator: Evaluation on standard databases," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 570–581, 2004.

[8] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15(11), pp. 1119–1125, 1994.

[9] R. Mark, G. Moody, and S. Greenwald, "Mit-bih supraventricular arrhythmia database," http://www.physionet.org/physiobank/database/svdb/, 1990.