Weighted Conditional Random Fields for Supervised Interpatient Heartbeat Classification

Gaël de Lannoy*, Damien François, Jean Delbeke, and Michel Verleysen

***Abstract*—This paper proposes a method for the automatic classification of heartbeats in an ECG signal. Since this task has specific characteristics such as time dependences between observations and a strong class unbalance, a specific classifier is proposed and evaluated on real ECG signals from the MIT arrhythmia database. This classifier is a weighted variant of the conditional random fields classifier. Experiments show that the proposed method outperforms previously reported heartbeat classification methods, especially for the pathological heartbeats.**

***Index Terms*—Classification, conditional random fields (CRFs), electrocardiogram (ECG), physiobank, unbalance.**

I. Introduction

THE analysis of ECG signals provides critical information on the cardiac function of patients. Cardiac disease conditions can be diagnosed by identifying abnormal heartbeats in the ECG signal. In such applications as clinical monitoring or pharmaceutical phase-one studies, long-term recordings of the ECG signal are required to this end. These long-term recordings are typically obtained using the popular Holter recorders. Holter ambulatory systems record at least 24 h of heart activity, resulting in data that contain thousands of heartbeats. The analysis is usually performed offline by cardiologists, whose diagnosis may rely on just a few transient patterns. Because of the high number of beats to evaluate, this task is very time consuming and reliable visual inspection is difficult. Computer-aided classification of pathological beats is, therefore, of great importance to help physicians perform correct diagnosis.

Nevertheless, the task is not trivial because heartbeat data share two specific characteristics: 1) a strong class unbalance (the vast majority of the heartbeats are normal healthy beats while just a small number of beats are pathological) and 2) time dependences between observations (the beats are extracted from ECG time series). The most challenging characteristic is the class unbalance. In such situations, standard automatic classifiers generally perform poorly because they are designed to generalize from training data and to output the simplest hypothesis that best fits the data, based on Occam's razor. As a result, the classifier tends to treat the pathological beats as noise and the learning process often leads to a dummy classifier always predicting the healthy class. Cost-sensitive classifiers such as the weighted support vector machine (wSVM) classifier [1] or the weighted linear discriminant analysis (wLDA) classifier [2]–[4] have, therefore, been proposed in the field of heartbeat classification to overcome the class unbalance.

Nevertheless, the second characteristic is left untreated in previous works. Heartbeat data contain sequential observations since there is a time dependence between subsequent heartbeats. Clearly, if a given beat is a healthy beat, there are more chances that the subsequent beat will also be a healthy one. To the opposite, if a pathological beat has occurred, there are more chances that another pathological beat will also occur in the future. In this study, a classifier which is both robust to the class unbalance and able to integrate the time dependences between observations is proposed. This classifier is a weighted variant of the conditional random fields (CRFs) classifier. The performances of the proposed model are validated on real ECG signals from the MIT arrhythmia database.

The rest of this paper is structured as follows. Section II tries to stress best practice rules for constructing reliable heartbeat classification systems. Section III provides a theoretical background over the methods used and proposed in this paper. Section IV details the construction of the experimental dataset. Section V holds the methodology followed by the experiments and Section VI shows the results. Eventually, Section VII draws some conclusions.

II. Heartbeat Classification

In this section, the guidelines defined by the American Association for Medical Instrumentation (AAMI) and the interpatient classification paradigm for constructing reliable ECG classification algorithms are presented. Next, the state of the art in heartbeat classification following these two recommendations is detailed.

A. AAMI Standards

Several features characterizing the heartbeats and several classification models have been investigated previously in the literature for computer-aided heartbeat classification. However, as first detailed by [2], very few reported works follow the standards defined by the AAMI, which makes it very difficult to

Manuscript received July 1, 2011; revised September 5, 2011; accepted September 26, 2011. Date of publication October 10, 2011; date of current version December 21, 2011. The work of G. de Lannoy was supported by a Belgian FRIA Grant. *Asterisk indicates corresponding author.*

*G. de Lannoy is with the Machine Learning Group, Université Catholique de Louvain, B-1348 Louvain-La-Neuve, Belgium, and also with the Neuroscience Institute, Université Catholique de Louvain, B-1200 Bruxelles, Belgium (e-mail: gael.delannoy@uclouvain.be).

D. François and M. Verleysen are with the Machine Learning Group, Université Catholique de Louvain, B-1348 Louvain-La-Neuve, Belgium (e-mail: damien.francois@uclouvain.be; michel.verleysen@uclouvain.be).

J. Delbeke is with the Neuroscience Institute, Université Catholique de Louvain, B-1200 Bruxelles, Belgium (e-mail: jean.delbeke@uclouvain.be).

Digital Object Identifier 10.1109/TBME.2011.2171037

assess the relative merits of the methods and of the proposed extracted features [5]. The AAMI defines the four clinically relevant heartbeat classes.

1) N-class includes beats originating in the sinus node: normal beats, bundle branch block beat types, atrial, and nodal escape beats.
2) S-class includes supraventricular ectopic beats: (aberrant) atrial, nodal, and supraventricular premature beats.
3) V-class includes ventricular ectopic beats: premature ventricular contraction and ventricular ectopic beats.
4) F-class includes beats that result from fusing normal and ventricular ectopic beats.

For a given classification algorithm, the AAMI outlines the necessity to use a performance metric which reveals the classification performances for each of these four classes.

B. Interpatient Paradigm

Supervised classifiers learn their parameters from labeled data, called the training set. Most of the previously reported heartbeat classification methods require beats from a new tested patient in the training set which is used to learn the parameters of the classifier. This is referred to as "intrapatient" classification [1]. This means that each time a new patient arrives, an expert has to manually label a portion of the beats from the patient's ECG signal, train the classifier, and then obtain a prediction on the rest of the beats. By contrast, "interpatient" classification consists in classifying the beats of a new tested patient according to a training set previously built and labeled from other patients. This is a much harder task of generalization.

The results that can be achieved with intrapatient methods are naturally better than when interpatient classification is performed, because the classifier is trained using data from the patient itself. Nevertheless, the patient labeled beats are usually not timely available in real clinical situations. Furthermore, because pathological beats can be very rare, there is no guarantee that the few training beats that would be labeled for this patient would contain representatives for each class; and the classifier could possibly fail in predicting something it has not learned.

Despite these major drawbacks, the large majority of previously reported work focuses on intrapatient classification. A comprehensive review of intrapatient classification methods can be found in [6] and in [7] for recent results.

C. State of the Art

For the reasons detailed in the previous section, this paper focuses on interpatient classification of heartbeats following the AAMI guidelines. The first study to establish a reliable interpatient classification methodology is [2], where a wLDA model is trained to classify the beats in the four classes defined by the standards of the AAMI [5]. This algorithm was later improved using the same classifier and other features first in [4] and later by the same authors in [3]. The common point between these algorithms is the use of the wLDA classifier, which has three strong limitations. First, it is a linear classifier which will fail to detect nonlinear decision functions. Second, the linear discriminant analysis (LDA) classifier is based on a Gaussian assumption over class distributions which is not always validated. Finally, the estimation of its parameters becomes difficult in the case of strongly correlated features because of the singularity of the covariance matrix.

For this reason, more powerful classifiers such as support vector machines (SVMs) have also been considered. In [8], hierarchical SVMs are used but the reported algorithm does not improve the results of [2]. Later, de Lannoy *et al*. [1] proposed an algorithm based on an SVM classifier optimizing a weighted cost function. This algorithm increased the performances of [2] for the pathological classes. Recently, Doquire *et al*. [9] investigated the use of feature selection techniques with the mutual information (MI) criterion and further improved the results of [1].

In summary, as far as interpatient classification is concerned, two kinds of classifiers have previously been considered: 1) the wSVM classifier and 2) the wLDA classifier. Both algorithms assume independent observations, and lack the use of time dependences between heartbeats. In this paper, a new classifier that is able to use the time dependences is proposed. This classifier is presented in the next section.

III. Theoretical Background

In this section, the CRFs classifier for sequential data is first presented. Next, the advantages offered by the L_1 regularization of its objective function are detailed. Finally, a weighted variant of this classifier that is robust to the class unbalance is proposed.

A. CRFs

Let us define a P-dimensional observation sequence $x'_t = [x_t^p]_{p=1}^P \in \mathbb{R}^P$ and the associated labels $y_t \in \{1, 2, \ldots, K\}$ where K is the number of classes and $1 \leq t \leq T$ is the time index with T being the total number of sampled observations in the sequence. CRFs are a form of discriminative model first proposed by [10] that relies on the first-order Markov assumption over labels. The probability distribution defined by CRFs is

$$p(y|x) = \frac{\prod_{t=1}^{T} \psi(y_{t-1}, y_t, x)}{\sum_{\mathbf{y}} \prod_{t=1}^{T} \psi(y_{t-1}, y_t, x)} \tag{1}$$

where $\sum_{\mathbf{y}}$ is the sum over all possible y sequences and x is the whole observation sequence. In the original CRF model [10], $\psi(y_{t-1}, y_t, x)$ is chosen as a parametric logistic function

$$\psi(y_{t-1}, y_t, x) = \exp\left(\sum_{kj} \chi_{kj} f_{kj}(y_{t-1}, y_t, x) + \sum_{kp} \omega_{kp} g_{kp}(y_t, x)\right) \tag{2}$$

where $1 \leq k \leq K$ and $1 \leq j \leq K$ are indexes ranging over the number of labels, $1 \leq p \leq P$ is an index over the number of features, $\chi = \{\chi_{11}, \chi_{12}, \ldots, \chi_{kj}, \ldots, \chi_{KK}\}$ are transition weights and $\omega = \{\omega_{11}, \omega_{12}, \ldots, \omega_{kp}, \ldots, \omega_{KP}\}$ are emission weights. The f_{kj} are called transition feature functions and the g_{kp} are called emission feature functions.

CRFs are typically trained by maximizing the conditional log likelihood $\mathcal{L}(\chi, \omega)$

$$\max_{\chi,\omega} \mathcal{L}(\chi, \omega) \tag{3}$$

$$= \max_{\chi,\omega} \log(p(y|x)) \tag{4}$$

$$= \max_{\chi,\omega} \sum_{t=1}^{T} \sum_{kj} \chi_{kj} f_{kj}(y_{t-1}, y_t, x) + \sum_{t=1}^{T} \sum_{kp} \omega_{kp} g_{kp}(y_t, x) - \log(Z(x)). \tag{5}$$

The likelihood function in (3) cannot be maximized in closed form, so numerical optimization is used. Since it is a convex function, quasi-Newton methods or conjugate gradient optimization methods using only first-order derivatives are directly applicable [11].

We now detail the computation of the normalizer $\log(Z(x))$ since it will later be implied in the embedding of the class unbalance within the CRF model. The computation of $\log(Z(x))$ requires a summation over all possible label sequences and the number of possible sequences grows exponentially with the sequence length. Nevertheless, the forward–backward algorithm originally used in hidden Markov models (HMMs) can be used to reduce the computational cost from $O(K^T)$ to $O(TK^2)$ [11]. Let us define the forward variable $\alpha_t(k) = p(x_1, x_2, \ldots, x_t, y_t = k)$, the probability of the partial observation sequence $\{x_1, \ldots, x_t\}$ until time t and state k at time t, which is solved recursively (see [12] for details). In CRFs, the regularizer term is computed in a similar way to $p(x)$ in HMMs as

$$Z(x) = \sum_{k=1}^{K} \alpha_T(k). \tag{6}$$

B. L_1 Regularization

In recent years, there has been a growing interest in the L_1-norm regularization, which is equivalent to a Laplacian prior on parameters [13]. This type of regularization enforces sparsity in the parameters and yields models that are more easily interpreted [14]. In particular, the L_1-regularized logistic regression model has proven to be very efficient [15]. CRFs can actually be cast as a multiclass logistic regression model with extra parameters for the first-order Markov dependences between labels. For this reason, the L_1 regularization of the CRF model yields the same benefits and has been investigated with success [13]. Sparsity is especially useful in sequence models having two sets of parameters: transition parameters and emission parameters. The L_1 penalty indeed achieves feature selection by encouraging sparsity in the emission parameters and in addition leads to a sparse transition matrix. Hence, if the L_1-norm regularization is not used, nonexisting transitions between states may not be strictly set to zero by the learning process and then induce errors during inference.

The optimization of the L_1-regularized CRF log likelihood is

$$\max_{w} \mathcal{L}(w) - \lambda \|w\|_1 . \tag{7}$$

where λ is a regularization constant, $\mathcal{L}()$ is the log-likelihood function as in (3) and $w = \{\chi, \omega\}$ is the set of all model parameters. It is more convenient in practice to minimize the negative regularized log likelihood defined as

$$\min_{w} f(w) = \min_{w} -\mathcal{L}(w) + \lambda \|w\|_1 . \tag{8}$$

The drawback is that the objective function $f(w)$ in (8) is no longer continuously differentiable for $w_i = 0$. Nevertheless, subgradients can be used to extricate the task of dealing with the nondifferentiable gradients [16]. The subgradient at a point of nondifferentiability is defined as the interval by the derivatives at the limit of each side of that point [17]. At a local minimizer $\tilde{w}$ of (8), we have the following optimality conditions:

$$\begin{cases} \nabla_i \mathcal{L}(\tilde{w}) + \lambda \text{sign}(\tilde{w}_i) = 0, & |\tilde{w}_i| > 0 \\ -\lambda \leq \nabla_i \mathcal{L}(\tilde{w}) \leq \lambda, & \tilde{w}_i = 0 \end{cases} \tag{9}$$

with $\nabla_i \mathcal{L}(w) = \frac{\partial \mathcal{L}(w)}{\partial w_i}$. The second optimality condition comes from the nondifferentiability of the absolute value function when its argument is zero. In this case, the subgradient is used. From these conditions, the gradient for each w_i computed during the optimization process is

$$\nabla_i f(w) = \begin{cases} \nabla_i \mathcal{L}(w) + \lambda \text{sign}(w_i), & |w_i| > 0 \\ \nabla_i \mathcal{L}(w) + \lambda, & w_i = 0, \nabla_i \mathcal{L}(w) < -\lambda \\ \nabla_i \mathcal{L}(w) - \lambda, & w_i = 0, \nabla_i \mathcal{L}(w) > \lambda \\ 0, & w_i = 0, -\lambda \leq \nabla_i \mathcal{L}(w) \leq \lambda. \end{cases} \tag{10}$$

C. Weighted CRFs (wCRFs)

In this section, the general framework for building unbalance embedded classifiers is presented. Next, we show how the specific case of the CRF classifier can be cast under this framework. Consider the general framework of classification models whose optimization takes the following form:

$$\min_{w} \sum_{t=1}^{T} \text{L}(y_t, f(x_t, w)) + \lambda \|w\| . \tag{11}$$

where w are the parameters of the model, $\text{L}(y_t, f(x_t))$ is a loss function measuring the discrepancy between the true label vector y_t and the model output $f(x_t)$ for each training instance t, and the right-hand side of the sum is a regularization term. Traditional classification algorithms choose an approximation of the accuracy (the overall classification rate) as loss function. For example, in SVMs, the loss function is the hinge loss and the regularizer is the L_2 norm of w. However, in unbalanced applications, the accuracy is not a suitable metric since the small class has less effect on accuracy than the majority class [18]. For example, with an unbalance of 99 to 1, a classifier that classifies everything in the majority class will be 99% accurate, but it will be completely useless as classifier. Even worse, in

many applications, the minority class is of prime importance (e.g., in medical diagnosis, false negatives can have dramatic consequences while false positives are of course undesired but still not life threatening) but the class will be completely ignored by such a classifier.

The idea in unbalanced-embedded approaches is to design cost-enabled classifiers that include distinct class misclassification costs in their objective function. More cost can then be given to errors in small classes to unbias the classifier. Any model that can be cast as an optimization in the form of (11) can be modified to integrate distinct class error costs as follows. First, the sum over all observations in (11) is reorganized into two nested sums over the classes and over the observations in each class

$$\min_{w} \sum_{k=1}^{K} \sum_{\{t|y_t=k\}} \mathrm{L}(y_t, f(x_t, w)) + \lambda \left\| w \right\|. \tag{12}$$

Next, K cost parameters c_k are added to weight the terms associated with each class by a factor c_k

$$\min_{w} \sum_{k=1}^{K} c_k \sum_{\{t|y_t=k\}} \mathrm{L}(y_t, f(x_t, w)) + \lambda \left\| w \right\|. \tag{13}$$

The wSVM classifier previously proposed for heartbeat classification in [1] can simply be cast in this form by choosing the hinge loss as loss function and the L_2 norm as regularizer.

We now show how to cast the CRF classifier in a form similar to (11). Remember the objective function of the CRF model

$$\max_{\chi,\omega} \sum_{t=1}^{T} \sum_{kj} \chi_{kj} f_{kj}(y_{t-1}, y_t, x) + \sum_{t=1}^{T} \sum_{kp} \omega_{kp} g_{kp}(y_t, x) - \log(Z(x)) \tag{14}$$

$$= \min_{\chi,\omega} - \sum_{t=1}^{T} \left(\sum_{kj} \chi_{kj} f_{kj}(y_{t-1}, y_t, x) + \sum_{kp} \omega_{kp} g_{kp}(y_t, x) \right) + \log(Z(x)). \tag{15}$$

The difficulty to cast (15) in a form similar to (11) lies in writing the regularizer term $\log(Z(x))$ as a sum over observations. It can actually be achieved by using the scaling trick used for the computation of the forward variable. In practical implementations, the values of the forward $\alpha_k(t)$ variable head exponentially to zero. For sufficiently large T (i.e., ten or more), the dynamic range of α will exceed the precision range of any machine. Hence, the only reasonable way of performing the computation is either to work in the log domain or to incorporate a scaling procedure [12]. In the scaling procedure, at each time step, the forward variables are normalized to sum to one as follows:

$$z_t = \sum_{k=1}^{K} \alpha_t(k) \tag{16}$$

$$\widehat{\alpha}_t(k) = \frac{\alpha_t(k)}{z_t}. \tag{17}$$

The only drawback is that we cannot merely sum up the $\widehat{\alpha}_T(k)$ terms for computing $Z(x)$ since these are already scaled. Nevertheless, in the context of HMMs, [12] has shown that the computation of $\log(Z(x))$ can be rewritten as a product over observations thanks to the scaling factors z_t

$$\log(Z(x)) = -\sum_{t=1}^{T} \log(z_t). \tag{18}$$

see [12] for mathematical details.

It is thus only feasible to compute the logarithm of $Z(x)$ but not $Z(x)$ since it would be out of the dynamic range of the machine anyway. Substituting (18) into (15) yields the desired formulation

$$\min_{\chi,\omega} \sum_{t=1}^{T} \left(\sum_{kj} \chi_{kj} f_{kj}(y_{t-1}, y_t, x) + \sum_{kp} \omega_{kp} g_{kp}(y_t, x) - \log(z_t) \right). \tag{19}$$

The wCRF objective function then becomes

$$\min_{\chi,\omega} \sum_{k=1}^{K} c_k \sum_{\{t|y=k\}} \left(\sum_{kj} \chi_{kj} f_{kj}(y_{t-1}, y_t, x) + \sum_{kp} \omega_{kp} g_{kp}(y_t, x) - \log(z_t) \right). \tag{20}$$

IV. Experimental Dataset

In this section, several experiments concerning the supervised classification of heartbeats are conducted. The database used in the experiments is first presented together with the features that are extracted from the heartbeat time series. Next, the methodology followed by the experiments is described. Finally, the results are presented.

A. ECG Database and Preprocessing

Data from the MIT–BIH arrhythmia database [19] are used in the experiments. The database contains 48 half-hour long ambulatory recordings obtained from 48 patients, for a total of approximately 110 000 heartbeats manually labeled into 15 distinct types. Following the AAMI recommendations, the four recordings with paced beats are rejected, and the paced beats in other recordings are also rejected.

The interpatient dataset configuration defined in [2], which has since been used in each interpatient classification systems [4], [8], is used in this paper. The 44 available recordings are divided in two independent datasets of 22 recordings each with approximately the same ratio of heartbeat classes. The first dataset is the training set, and is used to build the model. The second dataset is the test set, and is used to obtain an independent measure of the performances of the classifier.

The sampled ECG signals are first filtered to remove unwanted artifacts using the filtering procedure proposed in [2].

TABLE I
DISTRIBUTION OF HEARTBEAT CLASSES IN THE TWO INDEPENDENT DATASETS

	N	S	V	F	Total
Training	45809	942	3784	413	50948
	89.91%	1.85%	7.43%	0.81%	100%
Test	44099	1836	3219	388	49542
	89.01%	3.71%	6.50%	0.78%	100%

Two median filters are designed for this purpose. The first median filter is of 200-ms width and removes the QRS complexes and the P waves. The resulting signal is then processed with a second median filter of 600-ms width to remove the T waves. The signal resulting from the second filter operation contains the baseline wanderings and can be subtracted from the original signal. Powerline artifacts are then removed from the baseline corrected signal with a 60-Hz band-stop filter.

The location of R spikes and the associated beat types are provided with the database. These R locations serve as beat identifiers and the heartbeats are recognized in the signals accordingly. The MIT–BIH heartbeat labels are then grouped in the four classes defined by the AAMI recommendations (see Section II-A). Table I shows the number of beats in each class and their frequencies in the two datasets. The class unbalance is obvious. Beats having a R–R interval smaller than 150 ms or higher than 2 s most probably involve segmentation errors and are discarded.

B. Feature Extraction

A large variety of popular feature groups previously proposed for heartbeat classification are extracted from the heartbeat time series. The feature groups involved in this study are R–R intervals (used in almost all previous works), segmentation intervals [2], [20], morphological features [2], [7], Hermite basis function (HBF) expansion coefficients [8], [21], [22], and higher order statistics [8], [23]. We also introduce two additional features groups corresponding to the normalized R–R intervals and the normalized segmentation intervals. Each group is populated with the following individual features.

1) Segmentation intervals (24 features): ECG characteristic points, corresponding to the onset and ending of P, QRS, and T waves, are annotated in each beat using the unsupervised algorithm in [24]. A large variety of 24 features are then computed from the annotated characteristic points.
    a) QRS wave: Boolean flag indicating whether both Q and S points have been annotated, area, maximum, minimum, positive area, negative area, standard deviation, skewness, kurtosis, length, QR length, and RS length.
    b) P wave: Boolean flag indicating whether its onset and ending have been annotated, area, maximum, minimum, and length.
    c) T wave: Boolean flag indicating whether its onset and ending have been annotated, area, maximum, minimum, length, QT length, and ST length.

    When the characteristic points needed to compute a feature failed to be detected in the heartbeat segmentation step, the feature value is set to the patient's mean feature value.
2) R–R intervals (eight features): this group consists of four features built from the original R spike segmentations provided with the MIT–BIH database: the previous R–R interval, the next R–R interval, the average R–R interval in a window of ten surrounding R spikes, and the patient's mean R–R interval. The same four features are also computed using the R spikes detected by the segmentation algorithm.
3) Morphological features (19 features): ten values are measured by uniformly sampling the ECG amplitude in a window defined by the onset and ending of the QRS complex, and nine other features in a window defined by the QRS ending and the T-wave ending. As the ECG signals are already sampled, linear interpolation is used to estimate the intermediate values of the ECG amplitude. Here again, when the onset or ending points needed to compute a feature were not detected, the feature value is set to patient's mean feature value.
4) HBF coefficients (20 features): the parameters for the HBF expansion coefficients are chosen as in [8]: the order of the Hermite polynomial is set to 20 and the width parameter σ is estimated so as to minimize the reconstruction error for each beat.
5) High-order statistics (30 features): The second-, third-, and fourth-order cumulant functions are computed. The parameters as defined in [22] are used: the lag parameters range from -250 to 250 ms centered on the R spike and ten equally spaced sample points of each cumulant function are used as features, for a total of 30 features.
6) Normalized R–R intervals (six features): these features correspond to the same features as in the R–R interval group except that they are normalized by their mean value for each patient. These features are, thus, independent from the mean normal behavior of the heart of patients, which can naturally be very different between individuals, possibly misleading the classifier. The normalization is obviously not applied to the R–R feature corresponding to patient's mean itself, for a total of six features.
7) Normalized segmentation intervals (21 features): this group contains the same features as in the segmentation group, except that they are normalized by their mean value for each patient. The normalization is obviously not applied to Boolean segmentation features. Here again, the objective is to make each feature independent from the mean behavior of the heart of a patient, because it can naturally be very different between individuals.

Several studies have shown that using the information from both leads can increase the classification performances [2], [4]; all features are, therefore, computed independently on both leads (except the four R–R intervals and the three normalized reference R–R intervals computed from the original segmentations since they are common to both leads), for a total of 249 individual features.

TABLE II
TOP TEN FEATURES AS RANKED BY THE MI CRITERION

Pos.	Description	Lead	wSVM	wCRF+L_1	wCRF
1	Previous R-R (normalized)	Ref.	•	•	•
2	T wave amplitude (normalized)	1	•	•	•
3	2nd-order statistic at -40msec	1	•	•	•
4	2nd-order statistic at +40msec	1	•	•	•
5	2nd-order statistic at -166msec	1	•	•	•
6	2nd-order statistic at 166msec	1		•	•
7	T wave interpolation at 50% of wave length	1			•
8	Previous R-R	Ref.			•
9	Next R-R (normalized)	Ref.			•
10	T wave interpolation at 60% of wave length	1			

The features from this list that are selected by the classifiers are also shown.

TABLE III
PERFORMANCES OF THE WLDA, WSVM, WCRF, AND WCRF + L_1 MODELS ON THE TEST SET

Model	Features (#)	BCR	N	S	V	F
wLDA	Ranking (3)	58.32%	80.90%	54.90%	76.61%	20.88%
wSVM	Ranking (5)	82.45%	77.65%	84.48%	85.43%	82.47%
wCRF	Ranking (9)	81.29%	76.55%	80.72%	86.24%	81.96%
wCRF+L_1	Ranking (6)	85.39%	79.78%	92.59%	85.12%	84.54%

N, S, V and F are the accuracies of the normal, supraventricular, ventricular and fusion class respectively.

V. METHODOLOGY

The performances of the proposed wCRF classifier are compared to the previously reported models: the wLDA and the wSVM classifiers. The advantages offered by the L_1-norm regularization of the wCRF classifier are also evaluated. As mentioned in Section III-B, it is expected that leaving the CRF model unregularized will lead to unsatisfactory results. In this experiment, the hyperparameters of the wSVM and of the wCRF+L_1 classifiers are estimated by a "leave-one-patient-out" cross-validation procedure on the training set. The balanced classification rate (BCR), estimated by the geometrical mean of the class accuracies, is used as performance measure. The cost parameters are set to the inverse of the class priors in all the experiments.

The filter approach with a MI ranking criterion proposed in [9] is used to select the discriminative features. According to recent results [3], [9], the number of discriminative features is known to be typically much smaller than the number of features in our dataset. Therefore, only the top ten ranked features are considered in this experiment. The optimal number of features between one and ten is chosen by the "leave-one-patient-out" cross validation on the training set, as an additional hyperparameter.

The final models are obtained by training the four models on the complete training set with their selected hyperparameters, including the selected feature subset. The final performances are then evaluated on the test set—which has never been involved in any computation before—as a fair measure of their real generalization capabilities on unseen data.

VI. RESULTS

The classification results are presented in Table III. Table II holds the top ten features, as ranked by the MI criterion, and reveals which of these features were selected by each classifier. In the heartbeat classification task, errors in the pathological classes (i.e., missing a cardiac disease) can have dramatic consequences while errors in the normal class (i.e., incorrectly diagnosing a cardiac disease) are of course undesired but still not life threatening. The pathological classes are, therefore, of uttermost importance.

The wLDA model, with the ranking feature selection, achieves unsatisfactory results in this aspect with an accuracy below 55% for two pathological classes. The loss in performance with the LDA model can be explained by its strong assumptions such as the Gaussianity and the homoscedascity of classes which barely hold in this case. Also, when too many features start to be included in the model, the estimation of its parameters becomes unstable because of colinearity.

On the other hand, the nonlinear polynomial wSVM model achieves a BCR of 82.45% with only five features. In particular, the wSVM model yields an accuracy over 80% for all the pathological classes. The wCRF model obtains results slighly below the wSVM model. However, when the L_1 regularization is added to the wCRF model, the best overall results are obtained with a BCR of 85.39% and an accuracy of at least 85% for each pathological class. These results confirm that the wCRF model can benefit from the time dependences as long as a L_1 regularization term is included.

VII. CONCLUSION

The AAMI guidelines and the interpatient classification paradigm are two important aspects to consider for the design of reliable automatic heartbeat classifiers and for the evaluation of their relative merits. Previously reported classifiers for interpatient heartbeat classification are the wLDA and the wSVM classifier. Nevertheless, these two classifiers are unable to use the time dependences in the heartbeat dataset. In this paper, a weighted variant of the CRF classifier called wCRF which is able to integrate such time dependences between observations is proposed.

Classification experiments are conducted on real Holter recordings to compare the proposed wCRF classifier to previously reported interpatient classification algorithms. Results show that the wCRF classifier with a L_1 regularization term achieves better results with a BCR of 85.39% and an accuracy of at least 85% for each pathological classes. These results show that the information contained in the time dependence in class labels significantly increases the performances and that the L_1 regularization is useful to improve the estimation of the parameters of the CRF model.

ACKNOWLEDGMENT

The authors have no conflict of interest to disclose.

REFERENCES

[1] G. de Lannoy, D. Francois, J. Delbeke, and M. Verleysen, "Weighted SVMs and feature relevance assessment in supervised heart beat classification," *Commun. Comput. Inf. Sci.*, vol. 127, pp. 212–225, 2011.

[2] P. D. Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ECG morphology and heartbeat interval features," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 7, pp. 1196–1206, Jul. 2004.

[3] M. Llamedo-Soria and J. P. Martinez, "Heartbeat classification using feature selection driven by database generalization criteria," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3, pp. 616–625, Mar. 2011.
[4] M. Llamedo-Soria and J. P. Martinez, "An ECG classification model based on multilead wavelet transform features," in *Proc. Comput. Cardiol.*, Sep. 2007, vol. 35, pp. 105–108.
[5] Association for the Advancement of Medical Instrumentation, Arlington, VA, *Testing and Reporting Performance Results of Cardiac Rhythm and ST Segment Measurement Algorithms*, ANSI/AAMI EC38:1998, 1998.
[6] G. D. Clifford, F. Azuaje, and P. McSharry, *Advanced Methods And Tools for ECG Data Analysis*. Norwood, MA: Artech House, 2006.
[7] F. Melgani and Y. Bazi, "Classification of electrocardiogram signals with support vector machines and particle swarm optimization," *IEEE Trans. Inf. Technol. Biomed.*, vol. 12, no. 5, pp. 667–677, Sep. 2008.
[8] K. S. Park, B. H. Cho, D. H. Lee, S. H. Song, J. S. Lee, Y. J. Chee, I. Y. Kim, and S. I. Kim, "Hierarchical support vector machine based heartbeat classification using higher order statistics and Hermite basis function," in *Proc. Comput. Cardiol.*, Sep. 2008, pp. 229–232.
[9] G. Doquire, G. de Lannoy, D. Francois, and M. Verleysen, "Feature selection for supervised inter-patient heart beat classification," in *Proc. 4th Int. Conf. Bio-Inspired Syst. Signal Process.*, Roma, Italy, Jan. 26–29, 2011, pp. 67–73.
[10] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 282–289..
[11] C. Sutton and A. McCallum, *An Introduction to Conditional Random Fields for Relational Learning*. Cambridge, MA: MIT Press, 2007.
[12] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[13] A. Smith and M. Osborne, "Regularisation techniques for conditional random fields: Parameterised versus parameter-free," in *Proc. Int. Joint Conf. Natural Lang. Process.*, 2005, pp. 896–907.
[14] A. Y. Ng, "Feature selection, l1 vs. l2 regularization, and rotational invariance," in *Proc. 21st Int Conf. Machine Learn.*, 2004, pp. 78–86.
[15] S. I. Lee, H. Lee, P. Abbeel, and A. Y. Ng, "Efficient L1 regularized logistic regression," in *Proc. 21st Nat. Conf. Artif. Intell.*, 2006, pp. 401–407.
[16] M. Schmidt, G. Fung, and R. Rosales, "Fast optimization methods for l1 regularization: A comparative study and two new approaches," in *Proc. 18th Eur. Conf. Mach. Learn.*, 2007, pp. 286–297.
[17] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge Univ. Press, 2004.
[18] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Proc. 15th Eur. Conf. Mach. Learn.*, 2004, pp. 39–50.
[19] 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.
[20] I. Christov, G. Gómez-Herrero, V. Krasteva, I. Jekova, A. Gotchev, and K. Egiazarian, "Comparative study of morphological and time-frequency ECG descriptors for heartbeat classification," *Med. Eng. Phys.*, vol. 28, no. 9, pp. 876–887, 2006.
[21] M. Lagerholm, C. Peterson, G. Braccini, L. Edenbrandt, and L. Sornmo, "Clustering ECG complexes using Hermite functions and self-organizing maps," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 838–848, Jul. 2000.
[22] S. Osowski, L. Hoai, and T. Markiewicz, "Support vector machine-based expert system for reliable heartbeat recognition," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 4, pp. 582–589, Apr. 2004.
[23] S. Osowski and L. Hoai, "ECG beat recognition using fuzzy hybrid neural network," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 11, pp. 1265–1271, Nov. 2001.
[24] J. P. Martinez, R. Almeida, S. Olmos, A. P. Rocha, and P. Laguna, "A wavelet-based ECG delineator: Evaluation on standard databases," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 4, pp. 570–581, Apr. 2004.

Authors' photographs and biographies not available at the time of publication.