Reducing False Arrhythmia Alarms Using Robust Interval Estimation and Machine Learning

Christoph Hoog Antink, Steffen Leonhardt

Philips Chair for Medical Information Technology, RWTH Aachen University, Germany

Abstract

Reducing false arrhythmia alarms in the intensive care unit is the objective of the PhysioNet/Computing in Cardiology Challenge 2015*. In this paper, an approach is presented that analyzes multimodal cardiac signals in terms of their beat-to-beat intervals as well as their average rhythmicity. Based on this analysis, several features in time and frequency domain are extracted and used for subsequent machine learning.*

Results show that alarm-specific strategies proved optimal for different types of arrhythmia and that obtained scores varied: While the score for reducing false ventricular tachycardia alarms was 68.91*, false extreme tachycardia alarms could be suppressed with perfect accuracy. Overall, a top score of 75.55 / 75.18 could be achieved for real-time / retrospective false alarm reduction.*

1. Introduction

False alarms form an enormous problem in the intensive care unit (ICU) of today. Thus, the *PhysioNet/Computing in Cardiology Challenge 2015* aims at reducing them; please see [1] for details about the background and the mechanics of the competition as well as its data.

Our approach is based on robust interval estimation exploiting signal self-similarity. This estimator was originally developed for the analysis of ballistocardiographic signals [2], but variations have since been successfully applied to unobtrusive vital sign estimation from multimodal sources [3] as well as robust detection of heart beats in multimodal ICU data [4]. Using a moving window, the self-similarity of the cardiac signals is analyzed. While it was necessary to locate individual heart beats for the *PhysioNet/Computing in Cardiology Challenge 2014* [5], features are extracted *directly* from the estimated intervals and used to train several machine learning approaches here. As it is essential to this year's challenge, the general concept of interval estimation is briefly reviewed in the next section, while a detailed explanation can be found in [3].

2. Interval estimation

One of the most straightforward approaches to assess self-similarity is the short-time autocorrelation (STA) function. Let $x(n)$ be a time-discrete signal and

$$\omega_i(\nu) = x(n_i + \nu) \tag{1}$$

be an analysis window with index i centered around n_i. For better readability, the index is omitted in the following derivation. A common definition of the STA for each lag η for a window of constant length L is given by

$$S_{\mathrm{STA}}(\eta) = \frac{1}{L} \sum_{\nu=-L/2}^{L/2-\eta} \omega(\nu)\omega(\nu+\eta). \tag{2}$$

If the interval η_{opt} between exactly two heart beats is to be estimated, the length of the analysis window L has to be set in a way that the window contains only two beats, i.e. $L \approx 2\eta_{\mathrm{opt}}$. If $L \ll 2\eta_{\mathrm{opt}}$, no estimation is possible; if $L \gg 2\eta_{\mathrm{opt}}$, averaging over multiple beats occurs. This can be overcome by introducing the lag-adaptive short-time autocorrelation (LASTA)

$$S_{\mathrm{LASTA}}(\eta) = \frac{1}{\eta} \sum_{\nu=0}^{\eta} \omega(\nu)\omega(\nu-\eta), \tag{3}$$

which ensures that the exact number of samples necessary for each candidate lag η is considered, see also [3]. Another metric to asses self-similarity is the average magnitude difference function (AMDF). The modified AMDF

$$S_{\mathrm{AMDF}}(\eta) = \left(\frac{1}{\eta} \sum_{\nu=0}^{\eta} |\omega(\nu) - \omega(\nu-\eta)| \right)^{-1} \tag{4}$$

also uses the lag-adaptive window and is inverted so that it assumes larger values for lags that indicate more self-similarity [3]. As a third metric, the maximum amplitude pairs (MAP) function considers the amplitude of the signal and can thus be considered as indirect peak-detection,

$$S_{\mathrm{MAP}}(\eta) = \max_{\nu \in \{0,\ldots,\eta\}} \left(\omega(\nu) + \omega(\nu-\eta)\right). \tag{5}$$

For each lag η, the maximum of all sums of sample-pairs that are spaced exactly η time steps apart is calculated. It was shown that the presented similarity estimators exhibit a complimentary noise characteristic and results can be improved by fusing the estimators based on a Bayesian approach [2], which reduces to

$$\tilde{S}_{\mathrm{fused}}(\eta) = S_{\mathrm{LASTA}}(\eta) \cdot S_{\mathrm{AMDF}}(\eta) \cdot S_{\mathrm{MAP}}(\eta). \quad (6)$$

Moreover, self-similarity is principally modality-independent and this concept can be extended towards multiple channels and modalities:

$$S_{\mathrm{fused}}(\eta) = \tilde{S}_{\mathrm{fused,ECG}}(\eta) \cdot \tilde{S}_{\mathrm{fused,PPG}}(\eta) \cdot \ldots \quad (7)$$

Thus, for every window position i, the optimal interval can be obtained via

$$\eta_{i,\mathrm{opt}} = \arg\max_{\eta} \left[S_{i,\mathrm{fused}}(\eta) \right]. \quad (8)$$

Additionally, a quality metric can be estimated,

$$Q_i = \frac{S_{i,\mathrm{fused}}(\eta_{i,\mathrm{opt}})}{\sum_{\eta=1}^{L} S_{i,\mathrm{fused}}(\eta)}, \quad (9)$$

which is the ratio of the peak height to the area under the curve. It indicates, how much self-similarity this window actually exhibits, i.e., how trustworthy the estimated interval is.

3. Feature extraction

For the subsequent machine learning, all features were calculated in three different variations, namely
ECG: fusing only the available ECG signals (channel one and two),
BP: fusing only the available pressure-based cardiac signals, i.e., all channels named `PLETH` or `ABP`, and
ALL: fusing all cardiac-related signals, i.e., all channels *not* named `RESP`.
This resulted in a set of 27 interval estimation based features, namely

(1 - 3) $\min(\eta_{i,\mathrm{opt}})$,
(4 - 6) $\max(\eta_{i,\mathrm{opt}})$,
(7 - 9) $\mathrm{mean}(\eta_{i,\mathrm{opt}})$,
(10 - 12) $\sum_i \eta_{i,\mathrm{opt}}$,
(13 - 15) $\mathrm{mad}(\eta_{i,\mathrm{opt}})$,
(16 - 18) $\mathrm{std}(\eta_{i,\mathrm{opt}})$,
(19 - 21) $\mathrm{std} \,/\, \mathrm{mean}(\eta_{i,\mathrm{opt}})$,
(22 - 24) $\mathrm{mean}(Q_i)$,
(25 - 27) $\mathrm{median}(Q_i)$.

Here, std indicates standard deviation, while mad stands for the median absolute deviation function. In addition to these 27 features derived from beat-to-beat interval estimations, six features were calculated using regular autocorrelation in a fixed window to estimate the signals average rhythmicity in different interval ranges.
(28) High-frequency ECG: Relative maximum of the autocorrelation function of all ECG signals. Evaluated for a lag of 0 - 2000 ms in a 16 second window prior to the alarm. Set to zero if the corresponding lag is smaller than 200 ms to exclude artifacts.
(29) High-frequency BP: Like 28 but using all available pressure-based signals.
(30) Low-frequency ECG: Like 28 but set to zero if the corresponding lag is smaller than 900 ms to focus on slow rhythms.
(31) Low-frequency BP: Like 30 but using all available pressure-based signals.
(32) Average rhythmicity: Absolute maximum of the average of autocorrelations of all cardiac-related signals. Evaluated for a lag of 0 - 1500 ms in a 5 second window prior to the alarm. Set to zero if the corresponding lag is smaller than 80 ms to exclude artifacts.
(33) Peak rhythmicity: Like 32 but calculating the absolute maximum of maximums of all available autocorrelations.

Finally, to improve the recognition of variations in the rhythmicity, $S_{i,\mathrm{fused}}(\eta)$ was evaluated graphically. In Figure 1, two time courses for a false and a true ventricular tachycardia alarm are shown. These "two-dimensional

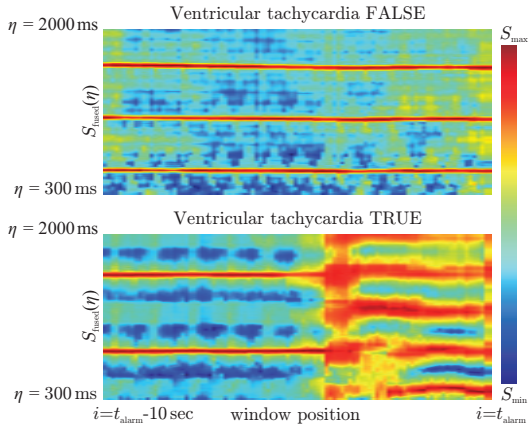

Figure 1. Time courses for a false and a true ventricular tachycardia alarm as two-dimensional correlogram. While the y-axis constitutes η, the color represents $S_{i,\mathrm{fused}}(\eta)$.

correlograms" or "correlogram images" can be interpreted similar to a spectrogram. Note the true alarm exhibits a relatively slow rhythm compared to the false alarm at first but changes towards a faster, oscillating rhythm approximately four seconds prior to the alarm. Further note the harmonic structure. In digital image processing, the classification of images is a common problem and several approaches to extract features exist. Here, an approach based on spatial

frequencies was chosen. For this, the two-dimensional correlograms are transformed using the spatial Fourier Transform (2D-FFT), where the phase information is discarded. Next, principal component analysis (PCA) is separately performed for the five alarm categories. Finally, all correlograms represented in the spatial Fourier domain are projected onto the first N_{PCA} eigenvectors to generate N_{PCA} additional features.

4. Machine learning

Three different types of machine learning approaches were evaluated using MATLAB's *Statistics and Machine Learning Toolbox*, namely
(A) binary classification trees (BCTs) [6],
(B) discriminant analysis classifiers (DACs) [7], and
(C) support vector machines (SVMs) [8].
Trees were trained using features 1 - 33 and pruning was used to optimize the cross validation error (CVE) by reducing features. DACs as well as SVMs were trained using combinations of all available features. Linear and pseudo quadratic discriminant analysis classifiers were evaluated. In the case of the former, regularization is used to optimize the CVE and reduce features. SVMs were trained using linear and radial basis functions.

An overview of the algorithm can be found in Figure 2. For clarity, preprocessing by resampling to 100 Hz, bandpass filtering with 1-30 Hz and normalization to zero mean and unit standard deviation is not shown. Moreover, for feature 1-32, intervals below a certain Q_{th} were excluded and correlogram images were normalized.

5. Results

Before submitting to the evaluation system, all approaches were evaluated and optimized in terms of their CVE. While this provided a good initial estimate of the algorithms performance, vast differences between the CVE and the actual score obtained on the hidden dataset could be observed. For example, while the CVE using SVMs was lower, the scores obtained were among the worst with a numerical average of 50.54 across all alarm categories. The highest average score (74.55) could be achieved with DACs trained with a combination of 2D-FFT / PCA features and features 32 and 33. This was followed by DACs trained with the 2D-FFT / PCA features only (73.05).

In general it could be observed that for the five alarms, different strategies lead to optimal individual scores, see Table 1 for results of the overall real-time optimal strategy.

5.1. Extreme tachycardia

For this alarm, perfect results could be obtained with

$$\min\left(\eta_{i,\mathrm{optALL}}\right) < 475\,\mathrm{ms} \Rightarrow A_{\mathrm{ETC}} = true. \qquad (10)$$

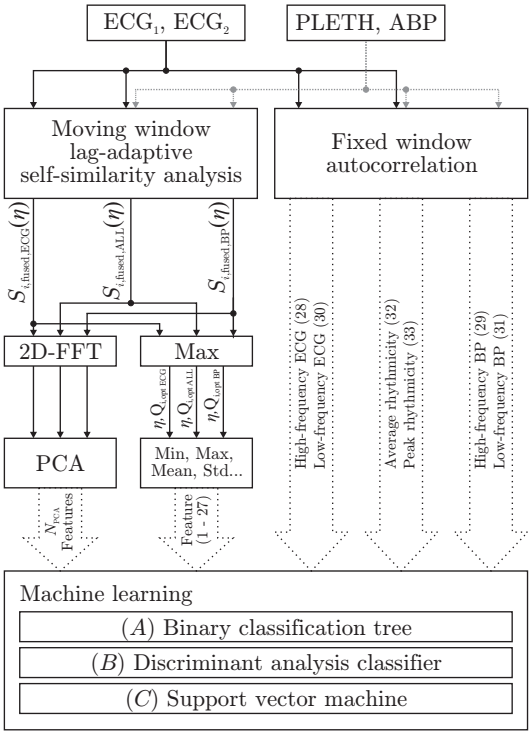

Figure 2. Overview of the algorithm. Preprocessing of the signals is not shown.

	TPR	TNR	Score
Asystole	56%	94%	74.33
Bradycardia	100%	57%	74.23
Tachycardia	100%	100%	100.00
Ventricular Flutter / Fib.	67%	92%	72.86
Ventricular Tachycardia	90%	71%	68.91
Real-time	93%	77%	75.55
Retrospective	92%	79%	74.82

Table 1. Real-time optimal results on the hidden test set.

If the smallest detected interval fusing all available cardiac signals is smaller than 475 ms (which corresponds to 126 BPM), the alarm is true. Else, it is a false alarm.

5.2. Asystole

Here, a score of 74.33 (TPR 56%, TNR 94%) could be achieved using a small (three feature) BCT, see Figure 3.

5.3. Extreme bradycardia

While a TPR of 100% could be achieved, the optimal TNR was 57%, leading to a score of 74.23. Here, a combination of 2D-FFT / PCA (ECG data, $N_{\mathrm{PCA}} = 17$), average rhythmicity (32) and peak rhythmicity (33) were used to train a regularized linear discriminant classifier. Without the rhythmicity features, a score of 71.13 was achieved.

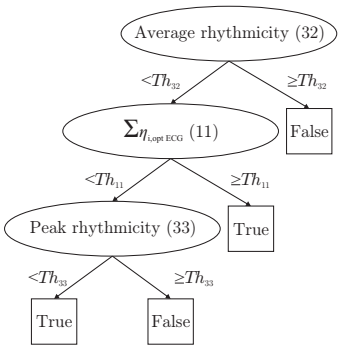

Figure 3. Binary classification tree for optimal asystole false alarm reduction.

5.4. Ventricular fibrilation or flutter

A score of 72.86 (TPR 67%, TNR 92%) was obtained using the same strategy as for bradycardia with the difference that only the first three principal components of the blood pressure based data were used to train a regularized linear discriminant classifier. Again, the score was notably lower (67.00) without the rhythmicity feature.

5.5. Ventricular tachycardia

The optimal yet lowest score, 68.91, with a TPR of 90% and a TNR 71% was achieved here using 2D-FFT / PCA features only. In particular, 14 principal components of the fused data were used to train a regularized linear discriminant analysis classifier.

6. Discussion

First, it is interesting to note that very different strategies proved optimal for the five alarms of interest.

Second, it is notable that the most straightforward approach (tachycardia) leads to the best overall score. Consistently, the strategy for asystole alarms allows physiological interpretation and leads to the second best score. Here, the most distinguishing feature for a false alarm is a high average rhythmicity. Next, a large sum of detected intervals (which corresponds to a low heart-rate) is an indicator for a true alarm. Finally, a low peak rhythmicity is an indicator for a true asystole alarm, whereas a high value indicates a false alarm.

On the other side, ventricular tachycardia could be classified best with the most abstract feature. While the result is still fair, it is comparably weak.

7. Conclusion

In this paper, a method for the reduction of false arrhythmia alarms using multi sensor data fusion was presented. In particular, beat-to-beat intervals were estimated, several features were extracted and different machine learning techniques were evaluated. Interestingly, the best results were obtained when physiologically motivated features and straightforward machine learning approaches could be used.

In the future, with the use of a sophisticated method of feature selection, the results using SVMs could be improved. Moreover, the description of the two-dimensional correlogram via 2D-FFT and PCA is likely suboptimal. Here, the development of physiologically motivated features seems appropriate.

Finally, no dedicated signal quality analysis and artifact exclusion strategy was used. This was done intentionally to show the robustness of the interval estimation. However, it might be performed indirectly at the machine learning stage, where a lack in rhythmicity can be present due to a loss in cardiac activity or a missing signal. This should be explicitly included in future algorithms. Moreover, artifacts could be annotated manually in the training data and / or robust methods of machine learning could be used.

References

[1] Clifford GD, Silva I, Moody B, Li Q, Kella D, Shahin A, Kooistra T, Perry D, Mark RG. The physionet/computing in cardiology challenge 2015: Reducing false arrhythmia alarms in the icu. Computing in Cardiology Sep 2015;.

[2] Brüser C, Winter S, Leonhardt S. Robust inter-beat interval estimation in cardiac vibration signals. Physiological Measurement 2013;34(2):123.

[3] Hoog Antink C, Gao H, Brüser C, Leonhardt S. Beat-to-beat heart rate estimation fusing multimodal video and sensor data. Biomed Opt Express Aug 2015;6(8):2895–2907.

[4] Hoog Antink C, Brüser C, Leonhardt S. Detection of heart beats in multimodal data: a robust beat-to-beat interval estimation approach. Physiological Measurement 2015; 36(8):1679.

[5] Silva I, Moody B, Behar J, Johnson A, Oster J, Clifford GD, Moody GB. Robust detection of heart beats in multimodal data. Physiological Measurement 2015;36(8):1629.

[6] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press, 1984.

[7] Fisher RA. The use of multiple measurements in taxonomic problems. Annals of eugenics 1936;7(2):179–188.

[8] Schölkopf B, Smola AJ. Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT press, 2002.

Address for correspondence:

Christoph Hoog Antink
Chair for Medical Information Technology
Helmholtz-Institut, RWTH Aachen
Pauwelsstr. 20 / D-52074 Aachen / Germany
hoog.antink@hia.rwth-aachen.de