A Support Vector Machine Approach for Reliable Detection of Atrial Fibrillation Events

Roberta Colloca[1,2], Alistair EW Johnson[1], Luca Mainardi[2], Gari D Clifford[1]

[1] Dept. of Engineering Science, University of Oxford, UK
[2] Dip. Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy

Abstract

There is a need for accurate and reliable detectors of asymptomatic atrial fibrillation (AF). Several ECG-based algorithms have been described in the literature, but no open comparison of features on out of-sample data has been published. Therefore, ten R-peak related features were selected from detectors available in literature and their classification performances were assessed both univariately and when combined using a Support Vector Machine (SVM). The MIT-BIH AFDB was used as the training set, and the MIT-BIH Normal Sinus Rhythm Database (NSRDB) and the MIT-BIH Arrhythmia Database were used for out-of-sample test performance assessment.

During the training phase, the optimal number of beats for accurate detection was determined using cross validation. The SVMs hyper-parameters were optimized with a grid search. On the training set the SVM had a Sensitivity (Se) of 99.07% and a Positive Predictive Value (PPV) of 98.27%.

During independent testing on the MIT-BIH NSRDB the SVM had a Sp=99.72% which was superior to any single feature or previous detector. The SVM also provided a Sp=99.70% on series 100 of the MIT-BIH Arrhythmia Database and a Sensitivity of 100% on series 200 of the same datase. A good Specificity (82.00%) and Accuracy (85.45%) were also obtained. Results are superior to any previously reported, for both training and testing and robust across multiple databases.

1. Introduction

Atrial fibrillation is the most common heart rhythm disorder: lifetime risk for AF is approximately 25% in men and women over 40 years of age and its prevalence is destined to rise with an ageing population [1]. Current diagnostic methods to detect AF are mostly symptom-based, thus leading to a large under-representation of the AF burden, which is associated with an increased risk of hospitalization, stroke and death [2]. If we consider that the silent nature of this arrhythmia is as likely to generate complications as symptomatic AF, screening of the general population should be taken into account, as it would detect additional AF cases compared to routine practice. Earlier detection would result in improved outcomes, given the effectiveness of existing treatments in managing symptoms and reducing further complications. To accomplish wider screening, a screening system that requires minimal training and is robust to noise is required. This work therefore only considers detectors based on peak-detection, since morphology-based detectors are prone to low levels of noise.

A variety of AF detectors employing different features extracted from the RR interval series have been proposed in literature, but no open comparison of those features on out of-sample data is still available. Therefore, in this work, we selected ten RR features from the top performing algorithms and their performances in detecting AF episodes were tested on variety of public (MIT-BIH) databases. The classification performances of the selected features were evaluated both univariately and multivariately using a Support Vector Machine (SVM).

2. Methods

We focused our attention on published AF detectors which were based on Ventricular Response analysis (R-peak detection), aimed to reveal the irregular, rapid-varying nature of RR intervals during AF. Only those with high performance and possessing robustness to artefacts and noise were chosen.

2.1. RR features

Among the available methods, we selected RR features from the three most accurate methods reported: Lake and Moorman [3], Linker [4] and Sarkar *et al.* [5]. The methods were selected for their superior performances measured on MIT-BIH databases and the relative independence of their employed features. From these algorithms we considered ten mostly independent features for further evalua-

ISSN 2325-8861

tion and comparison. They were: CosEn and SampEn, described by Lake and Moorman [3] and based on the magnitude of the sample entropy within a segment of the RR series; AFEvidence, OriginCount, PACEvidence, IrregularityEvidence, as described by by Sarkar *et al.* [5] which are based on the irregularity of the Δ RR interval series; MAD described by Linker [4], which is a RR variance-based AF feature. We also considered the features of median heart rate, minimum RR interval and mean RR interval within the analysis segment.

2.2. Databases

For training and testing the algorithms, standard datasets available on Physionet were used. The AF detectors were trained on the MIT-BIH Atrial Fibrillation Database (AFDB), which includes 25 long–term (10 hour) ECG recordings of adult humans with (mostly paroxysmal) AF.

During the testing phase the MIT-BIH Normal Sinus Rhythm Database (NSRDB) and the MIT-BIH Arrhythmia Database (ArrDB) were employed. The NSRDB is composed by 18 long-term ECG recordings: 18 subjects, including 5 men, aged 26 to 45 and 13 women, aged 20 to 50. None exhibits significant arrhythmias, except sporadic ectopy. The ArrDB contains 48 short-term (30 minutes) ECG recordings. This dataset can be divided in two series: series 100 of the ArrDB includes 23 recordings containing different types of arrhythmias (but not AF). Series 200 of the ArrDB includes 8 AF subjects, but also events of ventricular bigeminy and trigeminy, atrial flutter and other arrhythmias, which are likely to confound an AF detector.

2.3. Evaluation protocol

2.3.1. Performance metrics

The AF detectors were compared using the following performance metrics: *Sensitivity:* $Se = TP/(TP + FN)$, *Specificity:* $Sp = TN/(TN + FP)$, *Accuracy:* $Acc = (TP + TN)/N$, *Positive Predictive Value:* $PPV = TP/(TP + FP)$, where *TP* is the number of True Positives, *TN* is the number of True Negatives, *FP* is the number of False Positives and *FN* is the number of False Negatives. The total number of observations is denoted by *N*.

2.3.2. Window length

The analysis window length (WL) is an important parameter to take into consideration when AF methods are compared. A short WL allows for faster calculations and is more suitable to address the arduous challenge of paroxysmal AF events, which usually have unpredictable onset and short duration. On the other hand, a longer WL provides a more robust estimation of the RR segment content, since more data are available to and the signal to noise ratio is therefore higher, although, the computational cost is higher. During training, values of WL from 12 to 300 were considered for each feature. However, AF events whose duration is less than 30s are usually not considered to have clinical relevance.

2.3.3. Minimum number of AF beats

It is important to specify which is the minimum number of AF beats (minNbeat) within a N-beat segment necessary to classify that segment as AF event. This choice has impact on the shortest AF episode which can be detected. Therefore, both for the training and the testing phase, results were evaluated for minNbeat = 10, 30 or 50 beats. Values of WL smaller than the minimum number of AF beats required were not considered during the training phase. For example, if minNbeat=30, the minimum value of WL was also set to be equal to 30.

2.3.4. Univariate analysis

Detection algorithms described by Lake and Moorman [3], Linker [4] and Sarkar *et al.* [5] were trained on the MIT-BIH AFDB. A 5-fold cross validation was performed to assess the optimal WL, using the minNbeat values reported in section 2.3.3. The area under the receiver operating characteristic curve (AUROC) was the target metric for maximization. Successively, each model was retrained on the entire AFDB to determine the optimal threshold.

2.3.5. AF classification using an SVM

The features CosEn, SampEn, AFEvidence, OriginCount, PACEvidence, IrregularityEvidence, median heart rate, minimum RR interval, mean RR interval were combined by using an SVM with a radial basis function kernel. The SVM training was preceded by a grid search (tuning phase) to optimize the SVM hyper-parameters: γ, which controls the kernel width, and C (cost of misclassification) for each window length. The grid search was performed in combination with a 5 fold cross validation to reduce overfitting, while efficiently maximizing the amount of data used for model development.

3. Results

Table 1 presents the performance of the top three algorithms in the literature (on the training set - the AFDB) and the corresponding optimal WL for each method. Table 2 reports results obtained by the SVM for the MIT-BIH AFDB for the optimal WL for each value of minNbeat being considered. Note that the Se, Acc and PPV are rel-

minNbeat=10	CosEn	MAD	AFEvidence
Se	95.57	94.64	95.04
Acc	96.56	92.45	96.72
PPV	96.39	89.49	97.40
WL	41	282	82
minNbeat=30	CosEn	MAD	AFEvidence
Se	96.45	95.74	97.20
Acc	97.13	92.85	97.29
PPV	96.80	89.20	96.52
WL	60	282	76
minNbeat=50	CosEn	MAD	AFEvidence
Se	97.11	96.42	97.25
Acc	97.49	92.98	97.80
PPV	96.96	88.82	97.58
WL	91	282	99

Table 1. Univariate feature performance (%) on the MIT-BIH AF database. WL is in beats.

SVM	minNbeat 10	minNbeat 30	minNbeat 50
Se	98.12	98.45	99.07
Acc	98.42	98.62	98.84
PPV	98.26	98.36	98.27
WL	65	95	140

Table 2. SVM performance (%) on the MIT-BIH AFDB, by varying minNbeat. WL and minNbeat are in beats.

atively unnafected by values of minNbeat if the window length parameter (WL) is adjusted in a relative manner. Note also that all three algorithms provide similar performances, with the exception of MAD, which has a 7% lower PPV and requires a much longer analysis window.

Results on the test sets are reported in Tables 3, 4 and 5. On the NSRDB, all the methods maintained a good performance in rejecting normal sinus rhythm, with the exception of CosEn, which exhibited a Sp=5% lower than all other techniques being tested. When minNbeat= 10 beats, AFEvidence exhibited the best specificity (99.15%), followed by the SVM approach (98.91%) and MAD (98.37%). Note that the latter method uses a smaller window length (65 beat-segment) compared to MAD and AFEvidence. By increasing minNbeat to more clinically relevant scenarios, the SVM approach exceeded the performance of the best single feature (AFEvidence), providing a Sp=99.71% and Sp=99.72%, for minNbeat equal to 30 and 50 beats, respectively.

On series 200 of the MIT-BIH ArrDB, Lake and Moorman [3] displayed the highest Se (98.97%) and the smallest window size (41 beat-segment), followed by the SVM (Se=96.35%), which maintains a better Acc=85.53%, when minNbeat=10 beats. However, as noted earlier, AF episodes of less than 30s duration are not clinically significant. When considering minNbeat=50 beats, the SVM had a Sp=99.70% on series 100 of the MIT-BIH ArrDB and a Se=100% on series 200. The successful detection of the all true AF segments when using the SVM is obtained without degrading the accuracy (85.45%), with better performance compared to the most accurate single feature of AFEvidence (Acc=84.97%).

minNbeat	CosEn	MAD	AFEvidence	SVM
10	93.07	98.37	99.15	98.91
30	94.10	98.37	98.63	99.71
50	94.58	98.37	99.39	99.72

Table 3. Specificity (%) obtained on the MIT-BIH NSRDB at different values of minNbeat (in beats).

minNbeat	CosEn	MAD	AFEvidence	SVM
10	91.99	94.30	98.76	99.17
30	92.98	94.30	97.39	99.59
50	92.56	94.30	99.15	99.70

Table 4. Specificity (%) obtained on the series 100 of the MIT-BIH Arrhythmia Database with different values of minNbeat (in beats).

minNbeat 10	CosEn	MAD	AFEvidence	SVM
Se	98.97	92.00	96.05	96.35
Sp	75.06	68.15	81.76	82.78
Acc	79.67	73.91	84.68	85.53
PPV	48.65	47.92	57.48	58.73
minNbeat 30	CosEn	MAD	AFEvidence	SVM
Se	98.94	93.48	99.33	99.20
Sp	74.82	67.08	77.93	83.56
Acc	79.26	72.95	81.93	86.60
PPV	46.98	44.79	50.86	59.33
minNbeat 50	CosEn	MAD	AFEvidence	SVM
Se	99.19	97.62	99.12	100.00
Sp	74.64	66.67	81.73	82.00
Acc	79.17	72.95	84.97	85.45
PPV	46.95	42.71	55.39	56.85

Table 5. AF detector performance (%) on the series 200 of the MIT-BIH ArrDB when minNbeat is equal to 10, 30 and 50 beats.

All the presented methods show low values of PPV on the series 200 of the MIT-BIH ArrDB (see Table 5). This is due to the elevated number of FP in this dataset, which affects all methods, with significantly better performance for AFEvidence (PPV=55.39%) and the SVM (PPV=56.85%). We observed common difficulties in rejecting TN segments for record 106, containing a high number of premature ventricular contractions ($\sim$ 26% of all beats) and records 200, 201, 207, 208, 214, 222, 228. These records are all characterized by the presence of ventricular bigeminy and/or ventricular trigeminy and/or atrial bigeminy and/or supraventricular tachycardia.

4. Discussion

In this study, we focused on AF detection algorithms appropriate for a screening application which would allow for mass screening and address the under-estimation problem of AF. We chose RR analysis-based features, since they are least affected by artifacts and noise normally present in ambulatory conditions. Nine AF predictors were used in an SVM, allowing the predictive power of each published algorithm to be combined. We also conisdered the effect of varying the mininum number of beats required to trigger a detection (minNbeat) and the optimal WL.

The excellent results obtained by using the SVM on the AFDB during the training phase (Se=99.07% and Acc=98.84%) were validated when using the NSRDB and the MIT-BIH Arrhythmia Database as test sets. Results obtained on these test sets as compared to the best single features were generally improved with the SVM, in particular in terms of Se. The SVM detected the most true AF events (Se=96.35%) while still correctly rejecting normal sinus rhythm (Sp=98.91%) when minNbeat=10 beats and with a relatively short window length of WL=65 beats.

The MAD feature provided consistently and significantly lower performances than all other algorithms on all databases, and CosEn exhibited a significantly lower performance than AFEvidence and the SVM on the NSRDB and ArrDB. The SVM performed consistently higher than all other tested algorithms, and in particular for low numbers of atrial beats in a given window. Given that we wish the window to be as short as possible for easier data collection (and often patient compliance as noise reduces the length of the useful window to less than the desired protocol), the SVM applied to a 95 beat window with minNbeat=30 is recommended. However, if the window is reduced below this length (in particular to a 65 beat segment) then a model switch to the SVM trained with WL=65, minNbeat=10 is ideal. Signal quality algorithms could be used to automatically adjust the window length and remove low quality segments.

All algorithms exhibited showed a high number of FP on the series 200 of the MIT-BIH ArrDB. The presence of irregular rhythms, such as ventricular bigeminy and trigeminy caused false positive detections, degrading the overall performance of the AF detectors. It should be noted that the relatively lower performances are at odds with the other test databases, and perhaps should be re-interpreted in light of the fact that we would expect R-peak based AF detectors to trigger on any high entropy time series. The arrhythmias observed in the ArrDB constitute such high entropy time series, and for the proposed application (a mass screening tool), it is probably not important to penalise algorithms for triggering on bigeminal and trigeminal rhythms. It may in fact be important to trigger on such rhythms and draw the patient's or clinician's attention to such a problem, since treatment may also be needed for these non-AF arrhythmias.

5. Conclusion

Fusion of AF-related features using machine learning provided a best-in-class model for AF detection, characterized by high accuracy, and in particular sensitivity to AF events and high specificity for normal sinus rhythm. In particular, the SVM provided a robust algorithm across multiple databases, especially on the out-of-sample databases. The effect of window length and minimum number of AF beats was also quantified. The presented algorithm could be considered for real-world AF screening applications. We are releasing a Java version of our SVM-based approach under an open source license [6].

Acknowledgements

AJ acknowledges the RCUK Digital Economy Programme grant number EP/G036861/1 (Oxford Centre for Doctoral Training in Healthcare Innovation).

References

[1] Cerutti S, Mainardi L, Sörnmo L. Understanding atrial fibrillation: the signal processing contribution. 2009.

[2] Wang TJ, Larson MG, Levy D, Vasan RS, Leip EP, Wolf PA, D Agostino RB, Murabito JM, Kannel WB, Benjamin EJ. Temporal relations of atrial fibrillation and congestive heart failure and their joint influence on mortality. Circulation ; 107(23):2920–2925.

[3] Lake DE, Moorman JR. Accurate estimation of entropy in very short physiological time series: the problem of atrial fibrillation detection in implanted ventricular devices. American Journal of Physiology Heart and Circulatory Physiology 2011;300(1):H319–H325.

[4] Linker DT. Long-term monitoring for detection of atrial fibrillation, December 8 2009. US Patent 7,630,756.

[5] Sarkar S, Ritscher D, Mehra R. A detector for a chronic implantable atrial tachyarrhythmia monitor. Biomedical Engineering IEEE Transactions on 2008;55(3):1219–1224.

[6] Oster J, Behar J, Colloca R, Li Q, Li Q, Clifford GD. Open source Java-based ECG analysis software and Android app for atrial fibrillation screening. Computing in Cardiology 2013;40:–.

Address for correspondence:

Dr Gari. D. Clifford, gari@robots.ox.ac.uk