A Single Channel ECG Quality Metric

J Behar[1], J Oster[1], Q Li[1], G D Clifford[1]

[1]Dept of Engineering Science, University of Oxford, Oxford, UK

Abstract

We describe a framework for automated electrocardiogram (ECG) quality assessment which works in both normal and arrhythmic situations, on an arbitrary number of ECG leads and for time periods of as short as five seconds. Originally developed for the Physionet/Computing in Cardiology (CinC) Challenge 2011, we present here an extension to our original works with improved quality metrics. We manually annotated the 18000 single lead from the Challenge dataset as well as 9452, 10s segments (of both leads) from every subject in the MIT-BIH arrhythmia database as clinically acceptable or not. To balance the classes, noisy segments from the Noise Stress Test Database were added to clean data segments. A support vector machine was then trained to classify the data as clinically acceptable or not. A 97.1% accuracy was achieved on the test set of the extended database of 10s recordings, dropping almost linearly to 92.4% for 5s recordings. Retraining the classifier on all the challenge data, the classifier gave 93% accuracy on the MIT-BIH arrhythmia database. The results are promising and indicate that our method may be applied to Holter and intensive care unit monitoring.

1. Introduction

Electrocardiogram (ECG) recordings are often severely corrupted by noise and artefact. The corruption often has similar frequency content with the signal (thus limiting filtering in the frequency domain) and similar morphology to the physiological signal (thus limiting filtering in the time domain). See [1] for a detailed review of ECG noise and artefact types and how they impact the signal.

In common ECG recording scenarios, poor ECG quality increases the number of false alarms (i.e alarms with no clinical significance) leading to an increase in the workload of intensive care staff [2] and eventually their desensitization [3]. False arrhythmia alarms are often due to single channel ECG artefacts and low voltage signals [4]. ICU false alarm rates as high as 86% have been reported [5].

Holter monitors are used for ambulatory ECG monitoring in order to evaluate patient cardiac problems during normal daily activities, including sleeping. The monitor is often worn for more than 24 hours and now achieve up to three days continuous monitoring allowing delocalizing ECG monitoring from hospitals to home environments [6]. Holter recordings contain a significant amount of noise, particularly because of motion. Pollution of the ECG due to noise is therefore likely to skew statistics when automatically processing the record or lead the technician to spend time analysing artefactual ECG segments.

With an ageing population and the prevalence of chronic diseases numbers of telehealth based technologies (the delivery of health related services and information via telecommunications technologies [7]) have begun to proliferate. The aim is to deliver home-based and high quality care in a cost-effective way. In particular it is now becoming common for patients to play an active role in their healthcare management by taking their vital signs on a regular basis. The data are recorded (and transmitted) for further analysis by a trained physician.

However, with non-experts recording data, or long term recordings where significant artifacts can occur, it is difficult to ensure that the recorded physiological data are diagnostically useful. This was the subject of the Physionet/Computing in Cardiology (CinC) Challenge 2011 [8], which was aimed at producing an automated algorithm running in near real-time on a mobile phone in order to provide useful feedback to the user with respect to the quality of the acquisition. In the case of a bad acquisition the user will be asked to retake the recording. The CinC Challenge 2011 was oriented toward the usage of the system in developing countries where a community health worker would be taking the ECG but the problem is rather general and can apply to any noisy recoding scenario, such as Holter monitoring or a patient taking their own vital signs.

The aim of the CinC Challenge 2011 was to classify 12-lead 10 second ECG recordings. However, since many recordings contained transient artifact on just one or a few leads, classification of all 12 leads can be problematic. We choose therefore to relabel each lead and develop a single lead approach which is adaptable to any number of leads. The CinC Challenge 2011 led to the development of various indices and methodologies for assessing the quality of an ECG with excellent accuracy levels (around 94%) [8].

The Challenge did not however address two key issues: i) What happens if the window length is reduced? and ii) What would be the performance of the method on pathological (e.g. arrhythmic) records? This article describes an extension to our previously published work on ECG signal quality [9, 10] which addresses these issues.

2. Methods

2.1. Data selection and labelling

Data from the CinC Challenge 2011 [8] were used. The data set includes 1500 ten-second (10s) recordings of standard twelve-lead ECGs (leads I, II, II, aVR, aVL, aVF, V1, V2, V3, V4, V5, and V6) with full diagnostic bandwidth (0.05-100 Hz). The lead were sampled at 500 Hz with 16-bit resolution. The recordings were performed for a minimum of 10s by nurses, technicians, and volunteers with varying amounts of training recording ECGs. The data were balanced by generating additional bad quality data from the good quality records by adding noise from the Noise Stress Test Database [11] (as described in [10]). This resulted in 20000, 10s single ECG leads for the training set and 10000 for the test set with half of the leads of good quality and half of bad quality. This constituted the first dataset, and is referred to as the extended CinC database. We denote Set-a‡ and Set-b‡ the balanced training and test sets respectively. To analyse how the accuracy changed as the recording length is changed, we considered only the first n seconds ($n \in [5 - 10]s$) for each lead. The classifier was trained with the SQIs computed on the first n seconds of Set-a‡ and tested on the corresponding first n seconds of Set-b‡.

The second dataset was built from the MIT-BIH arrhythmia database [12, 13]. The database includes 48 complete two leads ECG records with reference annotations. The records have a diagnostic bandwidth of 0.1-100 Hz with 12-bit resolution and were digitized at 360 Hz. We identified locations of atrial premature beats (A) and premature ventricular contraction beats (V) and segmented 10s records centred on each beat, ensuring no overlap between segments. We also located the beginning of six arrhythmias; atrial fibrillation (AFIB), supraventricular tachyarrhythmia (SVTA), atrial flutter (AFL), sinus bradycardia (SBR), ventricular tachycardia (VT), ventricular flutter (VFL). Data were segmented for 10s after the onset of each arrhythmia. In the event of an arrythmia being less than 10s the segment was discarded. Both of the two leads available were used. In most records, the first channel is the modified limb lead II and the second channel is usually a precordial lead V1 (occasionally V2 or V5, and in one instance V4). This resulted in 9452 individual 10s leads which were manually annotated for quality. Within the 9452 single leads 269 were annotated as being of poor diagnostic quality and the remaining 9183 as good diagnostic quality.

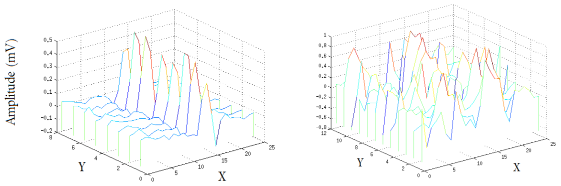

Figure 1. Example of pcaSQI with a good quality record (left plot, pcaSQI=0.95) and a bad quality record (right plot, pcaSQI=0.63). x axis: sample number, y axis: cycle (beat) number.

2.2. Pre-processing and signal quality indices

Each channel of ECG was downsampled to 125 Hz. QRS detection was performed on each channel individually using two open source QRS detectors (*eplimited* [14] and *wqrs* [15]). The *eplimited* algorithm is less sensitive to noise [16]. Six SQIs were calculated for each lead. Four of the SQIs (1 to 4 below) provided good results in earlier work [10, 16]. Two additional SQIs (5 and 6 below) were also chosen for evaluation in this new work. The chosen SQIs were:

1. pSQI: The relative power in the QRS complex: $\int_{5Hz}^{15Hz} P(f)\,\mathrm{df}$ / $\int_{5Hz}^{40Hz} P(f)\,\mathrm{df}$.

2. kSQI: The fourth moment (kurtosis) of the signal.

3. basSQI: The relative power in the baseline: $\int_{1Hz}^{40Hz} P(f)\,\mathrm{df}$ / $\int_{0Hz}^{40Hz} P(f)\,\mathrm{df}$.

4. bSQI: The percentage of beats detected by *wqrs* that matched with one from *eplimited*.

5. rSQI: The ratio of the number of beats detected by *eplimited* and *wqrs*.

6. pcaSQI: The ratio of the sum of the eigenvalues associated with the five principal components over the sum of all eigenvalues obtained by principal component analysis (PCA) applied to the time-aligned ECG cycles detected in the window by the *eplimited* algorithm, segmented 100ms either side of the R-peak (see Figure 1).

2.3. Machine learning for classifying quality of ECG

The data for the extended CinC Challenge dataset were divided into 2/3 for the training set (20000 records, Set-a‡) and 1/3 for the test set (10000 records, Set-b‡). The parameters of the different SQIs were tuned on the training set: Set-a‡. For the MIT-BIH arrhythmia database we used the 30000 leads from the extended CinC Challenge database as the training set and all the MIT-BIH arrhythmia leads as the test set. For each 10s record, the six SQIs were computed and used as the input features of a support vector machine (SVM) classifier. We used the libSVM library [17] with a Gaussian (non linear) kernel defined by: $k(\mathbf{x}_n, \mathbf{x}_m) = \exp(\gamma\|\mathbf{x}_n - \mathbf{x}_m\|^2)$, where γ controls the width of the Gaussian and plays a similar role as the degree of the polynomial kernel in controlling the flexibility of the resulting classifier [18]. $\mathbf{x}_n$ and $\mathbf{x}_m$ are two vectors expressed in the initial feature space. The SVM with a Gaussian kernel has two parameters: C and γ, where C is a constant that controls the trade-off between minimizing training errors and controlling the model complexity. Based on [10] we used $C = 25$ and $\gamma = 1$.

3. Results

In the following sensitivity (Se) measures the proportion of poor quality leads that have been correctly identified as poor, specificity (Sp) measures the proportion of good quality records that have correctly been identified as acceptable, and accuracy (Ac) corresponds to the proportion of signals that have correctly been classified.

3.1. CinC extended database

We first evaluated the two additional SQIs introduced. Adding rSQI and pcaSQI resulted in an increase in accuracy of 0.8% on the training set and 0.7% on the test set as shown in Table 1.

Table 1. Accuracy when adding rSQI and pcaSQI.

SQIs	training Set-a‡	test Set-b‡
(1,2,3,4)	0.971	0.964
(1,2,3,4,5)	0.974	0.969
(1,2,3,4,6)	0.976	0.97
(1,2,3,4,5,6)	0.979	0.971

The best results achieved on the extended database were $Ac = 0.979$, $Se = 0.976$, $Sp = 0.981$ on the training set and $Ac = 0.971$, $Se = 0.977$, $Sp = 0.965$ for the test set.

Table 2 shows the effect of reducing the window size from 10s to 5s in 1s decrements. The lower limit is 5s, because pcaSQI needs at least a few ECG cycles to be computed. Note that the performance is not largely diminished by reducing the window size.

3.2. MIT-BIH arrhythmia database

On the arrythmia data we achieved $Ac = 0.978$, $Se = 0.977$ and $Sp = 0.978$ on the training set (i.e the combination of Set-a‡ and Set-b‡) and $Ac = 0.931$ on the arrhythmia dataset. However a significant difference between the outcomes on the first lead (limb II) and the second lead (usually a modified lead V1) was noted; the classification of the first channel (lead II) only, resulted in $Ac = 0.97$ in contrast to the second channel (precordial leads) with $Ac = 0.893$ (see Table 3).

Table 2. Single lead SVM classifier results when reducing window length. Results are given on test Set-b‡.

	5s	6s	7s	8s	9s	10s
Ac	0.924	0.942	0.956	0.962	0.967	0.971
Se	0.935	0.953	0.966	0.972	0.976	0.977
Sp	0.914	0.932	0.946	0.954	0.959	0.965

Table 3. Classification results on the arrhythmias dataset.

	Ac	Se	Sp
All leads	0.931	0.796	0.935
lead 1	0.97	0.933	0.97
lead 2	0.893	0.787	0.899

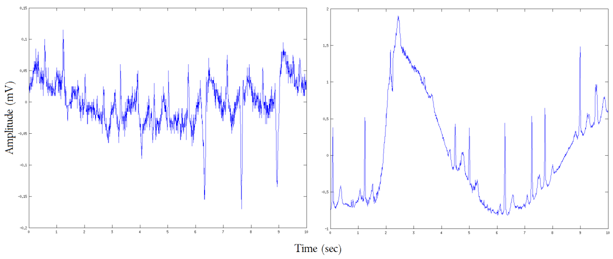

Figure 2. Examples of records from the arrhythmia dataset that were labelled as good quality but classified as bad.

4. Discussion and conclusions

An overall accuracy of 97% on the augmented Challenge data was achieved. The accuracy did not diminish significantly as the window length was reduced and results on the arrhythmia data were acceptable. Moreover, the reduction in accuracy was likely to be due to the transient nature of the noise which results in an incorrect class label being used in training. (When reducing the window length transient events on a given bad quality lead might not be within the first n seconds, but would still have been annotated as bad).

The proposed algorithm also appears to work well on arrhythmic data, although it seems to be lead location dependent. We obtained an $Ac = 0.93$ for the arrhythmia dataset but noted an important difference in accuracy between the lead locations; the second lead was noisier than the first in many cases and although QRS detection was judged feasible by the annotator their classification were challenging. This is an important issue, since any algorithm that classifies arrhythmias as poor quality would lead to a sequential rejection of abnormal data until clean sinus data is observed. This would reject all abnormalities and result in all subjects appearing healthy! It is expected that in an ICU context the combination of these ECG indices together with information provided by other pulsative waveforms such as arterial blood pressure will provide a robust system for reducing the number of false alarms as in [4].

We note however, that the results are not likely to be the same with respect to each arrhythmia type. Moreover, the arrhythmias were not represented in the same proportion. For example we had 12 VT 10s segments and 5974 V beat segments. Therefore our analysis was skewed towards the more frequent arrhythmias. However, we postulate that retraining our algorithm on each arrhythmia type will improve the results.

We should also note that many of the records that were labelled good quality but classified as bad were actually borderline (see Figure 2 for an example) or had a strong baseline wander which resulted in a very low value of basSQI. I.e. it is difficulty to distinguish between slow electrode motion and fast baseline wander. The results could perhaps be enhanced by training the classifier with clean ECGs with a broader range of baseline examples, since this type of artifact were under-represent in our training set.

Acknowledgement

JB is supported by the UK Engineering and Physical Sciences Research Council and the Balliol French Anderson scholarship fund. JO is supported by the Royal Society under a Newton Fellowship. QL is supported by Mindray North America. GC is supported by the University of Oxford.

References

[1] Friesen G, et al. A comparison of the noise sensitivity of nine QRS detection algorithms. IEEE Trans Biomed Eng 1990;37:8598.

[2] Allen J, Murray A. Assessing ECG signal quality on a coronary care unit. Physiol Meas 1996;17:24958.

[3] Chambrin M. Alarms in the intensive care unit: how can the number of false alarms be reduced? Crit Care 2001; 5:1848.

[4] Aboukhalil A, et al. Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform. Jour of Biomed Inform 2008;41:442451.

[5] Lawless ST. Crying wolf: false alarms in a pediatric intensive care unit. Crit Care Med 1994;22:981–5.

[6] Romero I, Berset T, et al. Motion artifact reduction in ambulatory ECG monitoring: An integrated system approach. Wireless Health conf 2011;.

[7] European health telematics association (EHTEL). Sustainable telemedicine paradigms for future-proof healthcare. European Commission DG Enterprise Industry Feb 2008; Version 1.0.

[8] Silva I, Moody G B, Celi L. Improving the Quality of ECGs Collected Using Mobile Phones: The PhysioNet/Computing in Cardiology Challenge 2011. Comput in Cardiol 2011;38:273–276.

[9] Clifford G D, Lopez D, Li Q, Rezek I. Signal quality indices and data fusion for determining acceptability of electrocardiograms collected in noisy ambulatory environments. Comp in Card 2011;38:285–288.

[10] Clifford G, Behar J, Li Q, Rezek I. Signal quality indices and data fusion for determining acceptability of electrocardiograms collected in noisy ambulatory environments. Phys Meas 2011;In review.

[11] Moody GB, Muldrow WE, Mark RG. A noise stress test for arrhythmia detectors. Computers in Cardiology 1984; 11:381–384.

[12] Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. IEEE Eng in Med and Biol 2001; 20:45–50.

[13] Goldberger AL, et al. Components of a new research resource for complex physiologic signals. Circ 2000; 101:215–220.

[14] Hamilton PS. Open Source ECG Analysis Software Documentation. http://www.eplimited.com/osea13.pdf, 2002.

[15] Zong W. Single-lead QRS detector based on length transform. http://www.physionet.org/physiotools/wfdb/app/wqrs.c, 2010.

[16] Li Q, Mark RG, Clifford GD. Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter. Physiol Meas 2008; 29(1):15–32.

[17] Chang C C, Lin C J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2011;2:27:1–27:27.

[18] Ben-Hur A, Weston J. A User's Guide to Support Vector Machines. Technical report http://pyml.sourceforge.net/doc/howto.pdf, 2012.

Address for correspondence:

Joachim Behar: joachim.behar@eng.ox.ac.uk