Signal Quality Indices and Data Fusion for Determining Acceptability of Electrocardiograms Collected in Noisy Ambulatory Environments

GD Clifford[1], D Lopez[1,2], Q Li[1], I Rezek[1,2]

[1]Dept. of Engineering Science, University of Oxford, Oxford, UK
[2]Imperial College, London, UK

Abstract

An algorithm to detect poor quality ECGs collected in low-resource environments is described (and was entered in the PhysioNet/Computing in Cardiology Challenge 2011 'Improving the quality of ECGs collected using mobile phones'). The algorithm is based on previously published signal quality metrics, with some novel additions, designed for intensive care monitoring. The algorithms have been adapted for use on short (10s) 12-lead ECGs. The metrics quantify spectral energy distribution, higher order moments and inter-channel and inter-algorithm agreement. Six metrics are produced for each channel (72 features in all) and presented to machine learning algorithms for training on the provided labeled data (Set-a) for the challenge. (Binary labels were available, indicating whether the data were acceptable or unacceptable for clinical interpretation.) We re-annotated all the data in Set-a as well as those in Set-b (the test data) using two independent annotators, and a third for adjudication of differences. Events were balanced and the 1000 subjects in Set-a were used to train the classifiers. We compared three classifiers: Naïve Bayes, a Support Vector Machine (SVM) and a Multi-Layer Perceptron artificial neural network classifiers. The SVM and MLP provided the best (and almost equivalent) classification accuracies of 99% on the training data (Set-a) and 95% on the test data (Set-b). The binary classification results (acceptable or unacceptable) were then submitted as an entry into the PhysioNet Computing in Cardiology Competition 2011. Before the competition deadline, we scored 92.6% on the unseen test data (0.6% less than the winning entry). After improving labelling inconsistencies and errors we scored 94.0%, beating the top score.

1. Introduction

The explosion of mHealth in both abundant and resource-constrained countries is both a cause for concern and for celebration [1]. While mHealth clearly has the potential to deliver information and diagnostic decision support to the poorly trained, it is not appropriate to simply translate the technologies which the trained clinician uses into the hands of non-experts. In particular, it is important that the explosion of access does not lead to a flooding of the medical system with low quality data and false negatives. Clearly for mHealth to expand, a paradigm shift in how data is analysed must occur. Data must be vetted at the front end, using automated algorithms, to provide robust filtering of low quality data.

This article addresses the specific problem of vetting the quality of electrocardiograms (ECGs) collected by an untrained user in ambulatory scenarios. The system described here is intended to provide real-time feedback on the diagnostic quality of the ECG and prompt the user to make adjustments in the recording of the data until the quality is sufficient that an automated algorithm or medical expert may be able to make a clinical diagnosis. This is the subject of the PhysioNet/Computing in Cardiology Challenge 2011.

2. Methods

2.1. Data selection and labelling

Data to support development and evaluation of challenge entries were collected by the Sana project and provided freely via PhysioNet. The data set includes 1500 ten-second recordings of standard twelve-lead ECGs; age, sex, weight, and possibly other relevant information about the patients; and (for some patients) a photo of the electrode placement taken using the mobile phone. Some of the recordings were identified initially as acceptable or unacceptable, but subsequently challenge participants annotated their own annotations to establish a *gold* (or perhaps *silver*) standard reference database of the quality of the recordings in the challenge data set.

The challenge data are standard 12-lead ECG recordings (leads I, II, II, aVR, aVL,aVF, V1, V2, V3, V4, V5, and V6) with full diagnostic bandwidth (0.05 through 100 Hz). Each lead was sampled at 500 Hz with 16-bit resolution.

The leads were recorded simultaneously for a minimum of 10 seconds by nurses, technicians, and volunteers with varying amounts of training recorded the ECGs, to simulate the intended target user.

The data were divided into two sets in a ratio of 2:1, with the larger for training (Set-a) for which binary annotations (acceptable or unacceptable) were available, and one for testing (Set-b) for which annotations were not available. The competition required users to submit a list of the files in Set-b together with an estimated classification and an automated scorer posted results immediately.

ECGs collected for the challenge were reviewed by a group of annotators with varying amounts of expertise in ECG analysis, in blinded fashion for grading and interpretation. Between 3 and 18 annotators, working independently, examined each ECG, assigning it a letter and a numerical rating (A (0.95): excellent, B (0.85): good, C (0.75): adequate, D (0.60): poor, or F (0): unacceptable) for signal quality. The average numerical rating, $\hat{s}$, was calculated in each case, and each record was assigned to one of 3 groups:

- Group 1 (acceptable): If $\hat{s} \geq 0.70$, and $N(F) \leq 1$.
- Group 2 (indeterminate): If $\hat{s} \geq 0.70$, and $N(F) \geq 2$.
- Group 3 (unacceptable): If $\hat{s} < 0.70$.

(N_F is the number of grades that were marked as F.) Approximately 70% of the collected records were assigned to group 1, 30% to group 3, and fewer than 1% to group 2, reflecting a high degree of agreement among the annotators. Challenge participants were also given the opportunity to grade the ECGs in the challenge data sets.

Our team also annotated all data in both Set-a and Set-b using two independent annotators with no previous experience in annotating ECGs, and adjudicated by one engineer with over a decade experience examining and processing ECGs. We submitted our two independently annotated classifications for both Set-a and Set-b, but not the adjudicated data (because it was not available by the deadline of the 20th July 2011).

Our team also annotated individual leads although due to time constraints, no adjudication of discrepancies was made for individual leads. Furthermore, an extended classification scheme, detailed in table 1 was employed, which does not render all recordings with a disconnected lead to be unacceptable. This was deemed necessary since a single missing lead should not necessarily be cause for rejection.

To map our annotations back to the competition annotations, B-, C- and D- became D, and B+ and C+ became B and C respectively. Note also that each class of acceptability was mapped to a numerical quality rating between -1 (worst quality) to +1 (best quality) in order to provide a less quantized set of targets for the MLP and to allow our continuous classifiers the option to predict individual classes. The ECGs were not preprocessed prior to annotation.

Quality	Class	Description give to annotators
1.00	A	An outstanding recording with no visible noise or artifact; such an ECG may be difficult to interpret for intrinsic reasons, but not technical ones
0.75	B+	A good recording with transient artifact or low-level noise that does not interfere with interpretation; all leads recorded well
0.5	B-	Same as above with missing lead(s)
0.25	C+	An adequate recording that can be interpreted with confidence despite visible and obvious flaws, but no missing signals
-0.25	C-	Same as above with missing lead(s)
-0.5	D+	a poor recording that may be interpretable with difficulty, or an otherwise good recording with one or more missing signals
-0.75	D-	A poor recording that may be interpretable with difficulty
-1.00	F	an unacceptably poor recording that cannot be interpreted with confidence because of significant technical flaws

Table 1. Augmented labelling system used in this study

2.2. Pre-processing of ECGs

Each channel of ECG was filtered to remove baseline wander and low frequency noise using a high pass filter with a cut-off at 1 Hz. QRS detection was performed on each channel individually using two open source QRS detectors (*eplimited* and *wqrs*) since *eplimited* is less sensitive to noise (see Li *et al.* [2]).

2.3. Signal Quality Indices

Six signal quality indices (SQIs) were chosen based on earlier work [2] and run on each of the $m = 12$ leads separately, producing 72 features per recording:

1. iSQI: The percentage of beats detected on each lead which were detected on all leads.
2. bSQI: The percentage of beats detected by *wqrs* that were also detected by *eplimited*.
3. fSQI: The ratio of power P(5-20Hz)/P(0-f_nHz), where f_n=62.5 Hz is the Nyquist frequency.
4. sSQI: The third moment (skewness) of the distribution.
5. kSQI: The fourth moment (kurtosis) of the distribution.
6. pSQI: The percentage of the signal x_m which appeared to be a flat line ($dx_m/dt < \epsilon$ where $\epsilon = 25\mu$V).

2.4. Classification of ECG

The resultant 72 features were then used to train various machine learning algorithms to classify the data as acceptable (1) or unacceptable (-1). To compare possible inconsistencies in labelling between the sets we compared results for training on Set-a and testing on Set-b against training on Set-b and testing on Set-a. We compared three different classifiers; Naïve Bayes (NB), a support vector machine (SVM), and a multi-layer perceptron (MLP) artificial neural network.

We tested two classification approaches: a single classifier trained on all 12 leads combined and 12 separate classifiers trained on the individual leads. In the 12-lead classifier the input data consisted of 72 features (6 per lead, see section 2.2) whilst the single lead classifiers were trained on the 6 features extracted for each lead individually. All classifiers were provided with the same class-labels *1:Acceptable or -1: Unacceptable*, as described in section 2.1.

Building classifiers using imbalanced classes, i.e. when one class greatly outnumbers the other classes, causes bias and results in a poor generalisation ability of the classification model. When prior probabilities (and a Bayesian training paradigm) are not used to overcome this problem, the alternative is to balance the training classes. In a balanced data set an equal numbers of examples is selected from each of the classes and this allows us to find a more accurate model.

We compared the performance of four machine learning algorithms: Linear Discriminant Analysis (LDA), Naïve Bayes (NB), Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) neural network.

We chose the radial basis function as our kernel for the SVM and trained the classifier (determining the values for the Lagrange multipliers) based on a sequential minimal optimisation (SMO) algorithm [3]. The slack variables' trade-off parameter C was optimised by grid search within the range of 1 to 10^3 and the scale of the RBF kernel was optimised by grid search within the range of 0.1 to 8.

The MLP was trained using a scaled conjugate gradient algorithm [3] while varying the number of nodes (N_{ij}) in the hidden layer from 3 to 30 and choosing the topology that gave the highest accuracy on a validation set. For this purpose, the training set was divided into 70% training, 15% validation and 15% testing.

Note that the input comprised of 72 nodes and the output layer was a single node, representing the probability of good (+1) or bad (-1) quality data. Since we have at most 1000 training examples in Set-a and we want the number of free parameters (weights in the MLP) to be approximately one tenth of this or less [3], then the number of hidden nodes must be restricted to about 13 ($< 1000/74$ if we include the bias weights). With bootstrapping of the less frequent class, higher values of N_{ij} are permissible.

For single lead classification, since none of our metrics except skewness were lead-specific, we were able to use over 12,000 training examples, and the restriction on the number of possible hidden nodes rose to 162.

Both Nïve Bayes classifier and LDA parameters were adjusted in the usual maximum likelihood framework [3]. The prior class probability was set to be uniform, which is justified because we balanced classes.

2.4.1. Classifier Fusion

In this work we employed two approaches to fusing algorithms. The first approach involved using the three multi-lead methods with highest results on Set-b to perform a majority vote (denoted 'VOTE'in the results).

The second fusion approach was needed for the single-lead classifiers (to produce on class per 12-lead recording). The chosen approach involved dividing the sum of the classifier outputs of each individual channel by 12. The Receiver Operator Characteristic Curve was then calculated on the training data and an optimal threshold was calculated. An additional step was also added, to override the result when a flat line was detected.

3. Results

The results of applying each classifier to Set-a and Set-b data are given in tables 2, 3 and 4. The SVM and MLP provided the best (and almost equivalent) classification accuracies of 0.99 on the training data (Set-a) and 0.95 on the test data (Set-b) using our own labels. This produced a PhysioNet (PNet) score of 0.926 (0.6% less than the winning entry) on the unseen test data (Set-b) labels. Note that we swapped the training and testing sets around to compare annotation consistencies between the two sets. Note the drop in performance when Set-b is used for training, indicating that there are inconsistencies between the two data sets, or that only 500 training patterns is insufficient to train the classifiers. For the MLP, and entry 4 of the competition, the number of hidden nodes was 25 (achieving an accuracy of 0.988 on training Set-b and a challenge score of 0.922). For entry 3 the number of hidden nodes was 12, achieving an accuracy of 0.972 on training Set-a, 0.952 testing on Set-b and a challenge score of 0.902.

The best test results on our own balanced annotations (training accuracy on Set-a of 0.996, testing on Set-b of 0.954) were achieved using an MLP with 16 hidden nodes (see table 3).

After the competition deadline passed, we attempted to cohere our labels with the unseen competition labels (by relaxing our criteria for rejecting leads). This allowed us to submit an entry which provided an entry score of 94.0%, beating the competition winning score. However, we note that this increased score was achieved with a MLP which achieved 99% accuracy in training, but only 88% accuracy on independent testing. Note also that the single channel approach yields a slightly lower accuracy, see table 4.

Table 2. Competition entries with accuracy of classifiers on different data and annotations. † indicates algorithm trained on Set-b, * indicates competition annotations used. Note entry 2 method was essentially the same as entry 1.

Entry ↓	Train Set-a	Test Set-b	Train Set-b	Test Set-a*	Test Set-a	PNet Score
2 SVM	0.880	0.834	0.974	0.894	0.898	0.830
3 MLP	0.972	0.952	0.982	0.918	0.926	0.902
4 MLP†	N/A	0.988	0.988	0.916	0.934	0.922
5 SVM†	N/A	1.000	1.000	0.837	0.844	0.926

Table 3. Classifier accuracy. Note that for voting, the results are simply from the majority vote of the SVM, MLP and LDA classifiers. ‡ indicates balanced data. ♯ indicates new annotations.

Method ↓	Train Set-a	Test Set-b	PNet Score	Train Set-b	Test Set-a	PNet Score
SVM	1.000	0.950	0.904	1.000	0.934	0.926
SVM‡	0.986	0.932	0.862	1.000	0.885	0.926
MLP	0.990	0.954	0.892	0.992	0.935	0.922
MLP‡	0.996	0.954	0.888	1.000	0.930	0.926
MLP♯	0.978	0.918	0.890	0.992	0.880	0.940
MLP‡♯	0.993	0.928	0.900	1.000	0.876	0.936
NB	0.911	0.936	0.890	0.942	0.907	0.880
NB‡	0.911	0.936	0.890	0.940	0.909	0.894
LDA	0.949	0.942	0.900	0.960	0.921	0.890
LDA‡	0.928	0.910	0.880	0.902	0.897	0.876
VOTE	0.994	0.948	0.902	0.996	0.934	0.922
VOTE‡	0.994	0.942	0.876	1.000	0.933	0.926

Table 4. Single lead classifier accuracy (balanced data).

Method ↓	Train Set-a	Test Set-b	PNet Score	Train Set-b	Test Set-a	PNet Score
SVM	0.973	0.96	0.89	0.984	0.934	0.914
NB	0.902	0.898	0.864	0.920	0.908	0.882
MLP	0.912	0.934	0.896	0.956	0.920	0.898

4. Discussion and Conclusions

The method for classifying the quality of ECGs presented in this paper is a novel and completely general approach to the problem of automatically identifying trustworthy signals or events. Essentially the covariant structure of how the noise manifests in the multi-lead ECG is learned to provide a context for when a signal is trustable or not. Reduced performance on the single lead approach illustrates how our technique learns the inter-lead relationships of the noise, and exploits these to provide more accurate classifications.

Training accuracies of 98% to 100%, with test set accuracies of above 95% indicate that extremely accurate classification of noisy ECGs is possible. With more accurate and consistent labelling, it is likely that the test accuracy will approach the training accuracy. In fact, this is a particular strength of the approach described here, in that as more data is made available, and more data is annotated, our algorithm can be updated (by incremental training on new examples) without the need for full re-training, and hence be used adaptively in a prospective manner, profiting from user feedback. We have therefore begun to implement this system on an Android phone application for adaptive use in the field.

It should be noted that initially it is perhaps more important to reject noisy ECGs that it is to retain clean ECGs, since the down side of a falsely-accepted noisy ECG is usually worse than rejecting a clean ECG (as long as this is not too frequent) since the user is available to re-take the reading. As time progresses and the user learns from the algorithm what is acceptable or not, they will be able to provide oversight for the algorithm and flag erroneous classifications. This learning feedback between human and computer may provide a general approach for all intelligent mobile applications.

It is also important to note that our approach does not suffer unduly from ignoring the prior distribution of acceptable and unacceptable data (ie. balancing the classes). Interpreting statistics on unbalanced data is wrong, since it skews the performance towards accepting ECGs, and does not give a balanced prediction of the utility of the algorithm on any given single recording.

A final important note is that the six quality metrics we chose, based on earlier work [2], are unlikely to be the optimal set of quality indicators, and may require changing for different contexts, equipment, diagnostic outcomes, or patient populations. However, the flexibility of this framework is in fact its strength. The general framework we have described in this paper is sufficiently flexible to allow us to use an arbitrary number of quality metrics, selecting those that are most appropriate for a given situation. In associated work [4] we describe a method which employs feature selection to determine the most relevant group of features, including both physiological and noise metrics. In this way, a multidimensional threshold can be found which uses the temporal association of noise, signal and their covariances, mimicking the human approach.

Acknowledgements

The authors would like to thank the Nuffield Foundation (grant number URB/39767), Ben Jackson (for additional annotations), the GSMA, Sana Mobile and PhysioNet.

References

[1] Waegemann CP. mhealth: the next generation of telemedicine? Telemed J E Health 2010;16(1):23–25.

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

[3] Bishop CM. Pattern Recognition and Machine Learning. Springer Verlag, 2006.

[4] Monasterio V, Clifford GD. Robust apnoea detection in the neonatal intensive care unit. Annals of Biomedical Engineering 2011;In Submission.

Address for correspondence:

Gari Clifford: gari@robots.ox.ac.uk