Rule-Based Methods for ECG Quality Control

Benjamin E Moody

Harvard-MIT Division of Health Sciences and Technology
Massachusetts Institute of Technology, Cambridge, MA, USA

Abstract

It is possible, using a smart phone or similar device, to collect ECGs from patients in remote locations, storing the results to be analyzed later. In this situation, however, the person collecting the ECG may not have the time or the necessary training to evaluate the quality of the recording at the time it is collected. It is useful for the device itself to analyze the recorded signals and provide feedback to the user about their quality.

This paper explores a number of heuristic rules that can be used to detect the most common problems in ECG recordings. These rules are designed to be simple enough that they can easily be tested in real time on a mobile phone. A combination of several of these rules is able to correctly detect a majority of poor-quality ECGs, as demonstrated using the PhysioNet/CinC 2011 Challenge database.

1. Introduction

With smart phones and similar mobile computing devices becoming widely available, even in the world's poorest countries, it is possible to use one of these devices as a mobile platform for collecting ECGs. One of these devices can be used to record ECGs from patients in rural areas, collecting the results for later analysis. However, doing this requires the device to provide some feedback to the user about the quality of the recording; if there is a problem, such as an improperly attached electrode, it is much easier to fix if the user can be made aware of the issue immediately.

The 2011 PhysioNet/Computing in Cardiology Challenge [1] is to develop an algorithm that can provide this sort of feedback. Given a short (10-second) sample of a 12-lead ECG, the algorithm should be able to determine whether the recording is "good enough" to be used for later analysis. The objective is for the algorithm to attempt to match, as closely as possible, the opinions of a group of human reviewers, who have classified each record as either "acceptable" or "unacceptable."

In the Challenge database, there are many records that stand out as being obviously faulty, in ways that should be easy for a computer to detect, and it seems clear that some simple heuristics could be used to reduce the size of the problem substantially. Even though they cannot hope to fully solve the problem of assessing "signal quality," such methods can be used to quickly identify both very-high-quality and very-low-quality recordings. In the less common, questionable cases, a more sophisticated analysis may be needed; in the end, the user can be allowed to decide whether to keep a possibly-faulty ECG or to try again.

2. Possible algorithms

A set of algorithms designed to detect the most common problems in the Challenge records are discussed below. These algorithms were tested based on the published classifications for the Challenge training set. Note that the "score" of an algorithm is measured as the fraction of records that are classified correctly. This training set included 773 "acceptable" records, 225 "unacceptable" records, and 2 unclassified (which were not counted for scoring.)

Based on the published classifications, it appears that most reviewers felt a record was still "acceptable" if all but one signal was of decent quality. Therefore, the following algorithms only consider a record unacceptable if at least two signals are found to be faulty.

2.1. Constant sections of a signal

By far the most common flaw in the Challenge records is for a signal to be completely flat for part or all of a record, generally indicating that one or more electrodes are not attached. In contrast, a real ECG is never constant; there is always some low-level noise, if nothing else.

To check for constant sections, all we need to do is track, for each signal, the most recent value, and the number of consecutive samples with that value. If the number of consecutive samples exceeds some threshold, the signal is marked as unusable.

In the Challenge training set, 123 of the 225 "unaccept-

able" records (and none of the "acceptable" ones) had two or more signals that each contained a constant section of at least 200 milliseconds, giving this method an overall score of 0.897.

2.2. Overall range

Another very simple test is to look at the maximum and minimum values for a signal. If the range is too small, the signal is likely to be completely useless (perhaps because an electrode is not properly attached.) If the range is too large, this could result from one or more large noise spikes, or from a severely drifting baseline; although these are not necessarily fatal problems, reviewers for the Challenge did generally mark such cases as unacceptable.

In the Challenge training set, "acceptable" signals did occasionally have an extremely small or extremely large range, but if the range was smaller than 0.2 millivolts, or larger than 15 millivolts, the signal was more likely to be "unacceptable." Using these values as cutoffs, 98 of the "unacceptable" records, and none of the "acceptable" ones, had at least two signals with too small a range; 46 of the "unacceptable" records and 12 of the "acceptable" ones had at least two signals with too large a range. Overall, this method had a score of 0.908.

2.3. Frequency of large changes

In a normal ECG, most parts of the signal are fairly "quiet," in that there are no sudden changes. If we look at a small section of the signal (on the order of a few tens of milliseconds), in most cases the range of samples is very small. At the same time, there will be some sections — the actual physiological events that we're interested in observing — that do contain large, sudden changes.

A simple way to test this is to look at overlapping intervals: first, the interval from $t = 0$ to $t = 2k$; next, the interval from $t = k$ to $t = 3k$; then, $t = 2k$ to $t = 4k$; and so forth. A value of $k = 32\text{ms}$ seems to work well. In each interval, compute the minimum and maximum values; this is still a very fast operation because each sample is only part of at most two of these intervals. An interval is then considered "quiet" if the range is less than some threshold. A high-quality signal should contain mostly, but not entirely, quiet intervals.

In the Challenge training set, this method proved to be useful for identifying *good* signals (rather than poor signals as with the methods above.) Using a threshold of 0.1 millivolt, 100 of the 773 "acceptable" records, and none of the "unacceptable" ones, contained between 64% and 96% quiet intervals. (This gives this method the unimpressive overall score of 0.326.)

3. Combined algorithm

All of the above methods can be considered useful to some extent, and a real-world application would need to incorporate all of these tests (probably with some refinements) as well as others. In the case of the Challenge, the objective was simply to give a "yes" or "no" answer for each record, and the "score" is simply the fraction of correct answers (according to a somewhat arbitrary notion of correctness.) There are many ways that the above methods could be combined into a single algorithm; in the end, the best scoring method is highly dependent on the frequency of the various types of problems that happen to be present in the Challenge database.

After testing a variety of possible combinations of the above methods and others, the best score for the Challenge training set was obtained by the following:

- If a signal is constant for an extended period (at least 200 milliseconds), mark it as bad. If two or more signals are marked as bad, mark the record as unacceptable.
- If all signals have an appropriate fraction of quiet intervals (between 64% and 96%), mark the record as acceptable.
- If a signal has an overall range of less than 0.2 millivolts, or more than 15 millivolts, mark it as bad. If two are more signals are marked as bad, mark the record as unacceptable.
- Otherwise, the record's status is uncertain.

For the Challenge training set, most of the "uncertain" records (664 of 744) were in fact considered acceptable by human reviewers, so the best-scoring method was to call these acceptable.

This combined algorithm attained a score of 0.913 on the Challenge training set (misclassifying 9 acceptable and 78 unacceptable records.) It also attained a score of 0.896 on the Challenge test set (set "B"), for which the reference classifications have not been published.

4. Conclusions

The algorithms described above are very simple to implement, requiring minimal computation, no floating-point math, and very little memory. These methods are effective enough to detect a large fraction of poor-quality ECGs, and a significant fraction of high-quality ones. These methods could be used, perhaps with some refinements, as the first part of a quality control system; ECGs that clearly pass or fail could be accepted or rejected immediately, while the more questionable cases could be subjected to more detailed (slower) analysis.

Trying to judge these algorithms according to the Challenge rules is somewhat unrealistic. In a real-world application, the program's performance would not correspond to the number of correct answers. The real-world

penalty for incorrectly marking a "good" record as "bad" is fairly small (at most a minute or two wasted), whereas the penalty for incorrectly marking a "bad" record as "good" is potentially much larger.

Furthermore, in a real-world application, the goal would not simply be to look at an ECG and pronounce it "acceptable" or "unacceptable." Instead, the program should provide useful suggestions, telling the operator what the problem appears to be. Ideally, the program should start out by aiming high, reporting potential problems even if they wouldn't necessarily render the record "unacceptable;" for instance, a single disconnected electrode is easy to detect and easy to fix. On the other hand, if the operator has tried several times and found the same problem every time, then the ECGs should probably be accepted despite their flaws.

(Of course, for this to work well requires some very careful user-interface design. The system needs to be designed to ensure that the operator actually reads its warning messages, and pays attention to them, rather than succumbing to the universal human urge to click "OK" until the problem goes away.)

A great deal of further work would be useful. These algorithms have been tuned to work well for the Challenge data sets, but the parameters are probably not ideal for a system, like that described above, that is intended to interact with the user and provide detailed feedback. The algorithms could perhaps be designed to adapt themselves to the data they are given, rather than having all their parameters hard-coded. And many additional tests could be devised to potentially detect other, less common problems that these algorithms do not look for.

References

[1] PhysioNet/Computing in Cardiology Challenge 2011. http://physionet.org/challenge/2011/.

Address for correspondence:

Benjamin E Moody
MIT Room E25–505
45 Carleton St
Cambridge, MA 02139 USA
benjaminmoody@gmail.com