Assessing the Robustness of Algorithms for Detecting Transient Ischemic ST Segment Changes

F Jager, GB Moody*, S Divjak, and RG Mark*

Faculty of Electrical and Computer Engineering, Ljubljana, SLO
(*) Harvard-M.I.T. Division of Health Sciences and Technology, Cambridge, MA, USA

Abstract

This paper proposes a protocol and techniques to assess the robustness of algorithms for detecting transient ischemic ST segment changes. An algorithm is robust if it is not critically dependent on the distribution or noise content of input signals, or the exact values of its analysis parameters. The protocol includes bootstrap evaluation of algorithm performance, a noise stress test and a sensitivity analysis of the most important algorithm parameters. We illustrate these methods by a case study in which we assessed the robustness of our Karhunen-Loève Transform based ischemic ST change detection algorithm.

1. Introduction

Assessing the properties of ST analyzers as well as predicting their behavior in the real world during ambulatory ECG monitoring or during in-patient monitoring is a difficult task. Although performance assessment using standard inputs [1] can provide much useful information about an analyzer's behavior, such tests often do not include methods for assessing robustness. While performance measurements typically characterize how the standard inputs are analyzed, it is important to understand to what extent performance depends critically on the choice of inputs. An analyzer whose performance varies little over a range of inputs may be said to be robust with respect to variation of input. Similarly, if parameters of the analysis do not critically affect performance as they are adjusted within some range, the analyzer is robust with respect to variation of these parameters. It is often the case that robustness is achieved at a cost in absolute performance. Nevertheless, robust methods are generally preferred, because they are less likely to fail catastrophically than are non-robust methods. There is no generally accepted methodology for assessing the robustness of ST analyzers; this paper proposes techniques and procedures for doing so.

2. Assessing robustness

The protocol proposed to assess the robustness of ST change detection algorithms includes:

1. *Bootstrap estimation of performance distributions*, to determine if performance is critically dependent on the choice of the database used for testing;
2. *Noise stress tests*, to determine the minimum signal-to-noise ratio (SNR) at which performance remains acceptable; and
3. *Sensitivity analysis*, to determine if analysis parameters are critically tuned to the test database.

An ST change detection algorithm is considered to be robust if the performance measurements obtained during these procedures remain above predefined *critical performance boundaries*.

The "bootstrap" procedure [2] can be used to estimate, for example, the 5% confidence limits of the performance (the lowest expected performance) of the ST detection algorithm. The procedure requires only the assumption that the reference database is a well-chosen representative subset of the population of examples for the problem domain. Thus the bootstrap predicts the algorithm's performance in the real world. The algorithm is robust if its lowest expected performance is still above the critical performance boundaries. The bootstrap is also useful for comparing the robustness of different algorithms. A narrower distribution of performance (a smaller interval between the 5% confidence limit and the raw statistic), as estimated by the bootstrap, indicates a more robust algorithm. Such an algorithm yields nearly the same performance given very different circumstances (input signals), and is therefore relatively insensitive to the choice of the database.

0276-6547/94 $4.00

Noise tolerance and ability to reject noise are also important aspects of the behavior of an ST analysis algorithm. A quantitative and reproducible technique of adding noise to ECG records allows us to determine the effects of noise on the algorithm's performance. Synthesized noise does not guarantee the same characteristics (e.g., non-stationarity) as are observed in a clinical environment. The noise stress test [3] is a method that consists of adding real noise (electrode motion artifacts, baseline wander, and muscle noise) to "clean" ECG signals. A useful concept when characterizing the results of a noise stress test is the lowest SNR at which the algorithm can still operate acceptably. This critical SNR is smallest for those algorithms that are most robust with respect to noise.

Sensitivity analysis addresses the question of how performance varies given small changes in analysis parameters. Performance of a robust analysis algorithm should not deteriorate below acceptable levels if such changes are made; if this does happen, it suggests that the design of the algorithm may be tuned to the test database.

3. Case study

We assessed the robustness of our Karhunen-Loève Transform (KLT) based two-channel ischemic ST change detection algorithm [4, 5], using the performance measures proposed in [1, 5] and the European Society of Cardiology ST-T database (ESC DB) [6] as the test database. Since no performance requirements have been previously published, we set critical performance boundaries (goals; see Table 1), based on performance measurements obtained in this way from a variety of current analysis algorithms, including our own. These levels of performance are possible but not trivial to achieve, and in our estimation represent the standard of performance to be expected of clinically useful ST analysis algorithms at present.

3.1. Bootstrap distributions of the aggregate performance statistics

In order to illustrate comparisons of bootstrap performance distributions for different ST detection algorithms, we also derived the bootstrap distributions for our earlier time-domain ST change detection algorithm [5, 7]. Table 1 shows the raw aggregate performance statistics and the 5% confidence limits for the minimum expected performance statistics (based on 10,000 bootstrap trials) for both algorithms in comparison to the critical performance boundaries. Of these, the KLT-based algorithm yields better performance. Its lowest expected performance (the 5% confidence limits)

[%]	*Goal*	*KLT algorithm*	*Time algorithm*
Statistic		*Raw* (5%)	*Raw* (5%)
IE Se [g]	> 80	85.2 (80.6)	83.2 (77.8)
IE +*P*[g]	> 80	86.2 (81.1)	83.9 (78.3)
ID Se [g]	> 70	75.8 (69.6)	73.1 (66.6)
ID +*P*[g]	> 70	78.0 (72.5)	77.7 (71.6)
$p_{(100\mu V)}$	< 15	9.8 –	9.8 –
IE Se [a]	> 80	87.1 (82.2)	85.6 (80.6)
IE +*P*[a]	> 80	87.7 (82.9)	86.9 (82.0)
ID Se [a]	> 70	78.2 (73.2)	78.0 (73.3)
ID +*P*[a]	> 70	74.1 (69.3)	74.4 (69.4)

Table 1. Performance of the ST change detection algorithms obtained on the ESC Database versus critical performance boundaries. The bracketed figures are 5% confidence limits (based on 10,000 bootstrap trials); those that do not meet the goals (critical performance boundaries) are boxed. (*IE* - ischemic ST episode, *ID* - ischemic ST duration, [g] - gross, [a] - average, $p_{(100\mu V)}$ - discrepant ST measurement percentage)

[%]	*Gross*	*IE Se*	*IE* +*P*	*ID Se*	*ID* +*P*
KLT alg.	ΔP	4.6	5.1	6.2	5.5
Time alg.	ΔP	5.4	5.6	6.5	6.1
	Average	*IE Se*	*IE* +*P*	*ID Se*	*ID* +*P*
KLT alg.	ΔP	4.9	4.8	5.0	4.8
Time alg.	ΔP	5.0	4.9	4.7	5.0

Table 2. Widths of the bootstrap performance distributions (ΔP) between the 5% confidence limits and raw statistics for the ST detection algorithms.

is close to (gross *ID Se*) or exceeds the critical performance requirements, while this is not the case for the time-domain algorithm. Table 2 summarizes the widths of the bootstrap distributions (ΔP) between the 5% confidence limits and raw statistics for both algorithms. Bootstrap distributions for gross as well as average aggregate performance statistics are narrower for the KLT-based algorithm, so it appears to be more robust with respect to the choice of inputs than the time-domain algorithm.

3.2. Noise stress test

We created a noise stress test database containing all the records of the ESC Database, to which noise from MIT Noise Stress Test Database was added. For this test, the first five minutes of the ECGs were left "clean" for each record, to permit the detector an opportunity

to measure the baseline ST segment deviation level accurately. Following this period, the test stresses the detector by adding noise throughout the entire record. We wanted to simulate real circumstances so all three kinds of noise were represented equally. Records to which a given type of noise was added were chosen at random. SNRs were chosen from 6 dB to 36 dB with a step of 6 dB. The protocol includes two variants. In the first variant, noise was added to signals immediately prior to preprocessing phase which represents arrhythmia detector analysis, while in the second variant, noise was added to signals immediately after the preprocessing phase. In our case, the ARISTOTLE arrhythmia detector [8], written by the second author, was used. Using both types of test permits us to determine to what extent the performance of our ST analysis algorithm may be limited by that of the arrhythmia detector in the presence of noise.

The noise stress test results (Figure 1) shows stable gross and average *IE Se* and *ID Se*, even for SNR = 6 dB. Gross and average *IE* $+P$ and *ID* $+P$ stay above critical performance requirements until SNR = 24 dB or even 18 dB. Only slightly worse performance was obtained when noise is added prior to ARISTOTLE's analysis. In this case, the significantly discrepant ST measurement percentages, $p_{(100\mu V)}$, were 12.5% (for SNR = 30dB), 16.0% (24dB), and 23.1% (18dB).

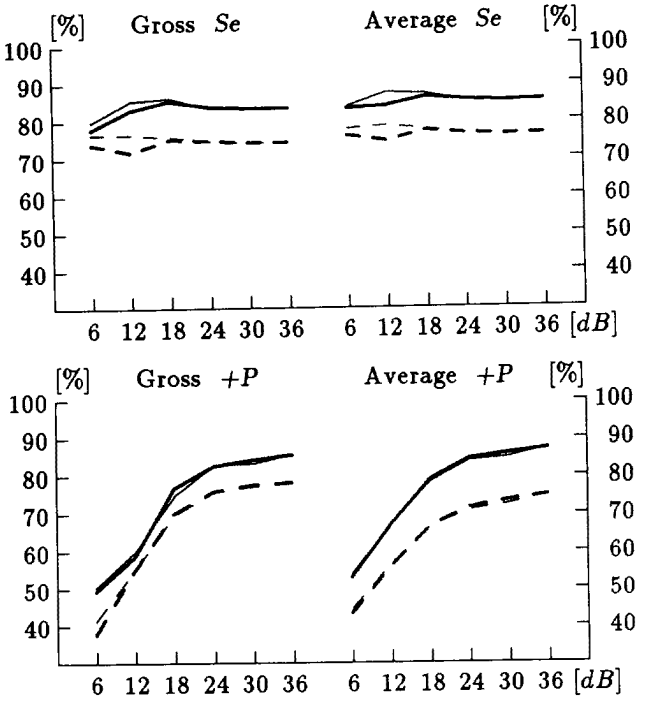

Figure 1. Performance of the KLT algorithm obtained during noise stress test. (*Solid* - ischemic ST episode, *IE*, statistics; *dashed* - ischemic ST duration, *ID*, statistics; *thin* - noise added after ARISTOTLE's analysis; *bold* - noise added prior to ARISTOTLE's analysis)

3.3. Sensitivity Analysis

Sensitivity analysis was performed by modifying the dimensionality of ST and QRS feature vectors, modifying the feature-space boundaries and decision thresholds of the algorithm, and randomly perturbing the exact position of the fiducial point.

When studying the influence of modifying the dimensionality of feature vectors (number of KL coefficients), we retained the original noise-detection procedure (5 KL coefficients). We wanted to study the influence on the quality of feature representation and subsequently on those parts of the algorithm performing pattern recognition. We varied the number of KL coefficients from 2 to 8. Other analysis parameters were in each case recalculated according to feature space dimensionality ratio. The results of the test demonstrate that performance is not critically dependent with respect to variation in the number of KL coefficients between 2 and 6. Performance varied significantly, but the variations were smooth. Performance statistics remained above critical performance requirements when the dimensionality of feature vectors, N, varied from 4 to 6 (Table 3).

Studying the influence of modifying the most important feature-space boundaries and decision thresholds of the KLT algorithm involved procedures for noise detection (feature-space distance threshold, residual error threshold), for correcting the reference ST segment level (updating feature-space boundary, time constant), for detecting significant axis shift (step change threshold, flatness threshold), and for detecting ST change episodes (initial center and initial width of the guard zone, lower decision threshold and the time constant) [4]. Changing these parameters one-by-one up to $\pm 10\%$ did not influence performance of the algorithm significantly (Table 3). The most sensitive parameter is the initial center of the guard zone, denoted Λ_{c_0}. Change in this decision threshold by -10% resulted, in the worst case, in a change of gross *IE* $+P$ by -7%; and of gross *ID* $+P$ by -7% from raw statistics. Changing the decision threshold Λ_{c_0} for $+10\%$ resulted, in the worst case, in a change of gross *IE Se* by -8%; and of gross *ID Se* by -8%. Otherwise, performance remained strictly above critical performance boundaries.

To study the influence of fiducial point jitter, simulated uniformly distributed jitter in the interval $[-m, m]$, $m = 0, ..., 8$, original signal samples around the ARISTOTLE's fiducial point was introduced. The jitter was introduced immediately before applying the

[%]	Goal	N	N	Λ_{c_0}	Λ_{c_0}	[$\pm m$]
Statistic		4	6	-10%	$+10\%$	$m=2$
IE Se [g]	> 80	84.0	83.2	86.8	77.2	84.8
IE $+P$[g]	> 80	81.7	84.6	79.4	88.3	78.1
ID Se [g]	> 70	74.8	73.7	76.0	67.3	72.2
ID $+P$[g]	> 70	73.4	76.5	71.4	80.6	63.9
$p_{(100\mu V)}$	< 15	9.8	9.8	9.8	9.8	10.9
IE Se [a]	> 80	86.2	86.1	87.7	81.0	87.5
IE $+P$[a]	> 80	85.2	86.1	83.5	89.3	82.7
ID Se [a]	> 70	77.8	76.7	77.8	72.1	77.8
ID $+P$[a]	> 70	72.7	73.3	70.7	77.0	70.8

Table 3. Performance of the KLT algorithm obtained during sensitivity analysis (see text).

KL basis functions to pattern vectors. Such a situation may be expected as a result of inaccurate time alignment of the pattern vector or of a suboptimal procedure for determining the position of the fiducial point. Performance of the KLT-based algorithm (Table 3) remained close to (gross IE $+P$, ID $+P$) and above the critical performance boundaries when introducing jitter in the window of ± 2 original signal samples (± 8 ms) around the fiducial point.

4. Discussion and conclusions

We have described methods for characterizing the robustness of ischemic ST change detection algorithms. We demonstrated the robustness of the KLT-based ischemic ST change detection algorithm, and compared it with our earlier time-domain algorithm. Bootstrap analysis showed that the performance of these algorithms is not critically dependent on the choice of database, and that the KLT-based analysis not only performs better than the time-domain analysis in absolute terms, but is also marginally more robust. The noise stress test demonstrated that the critical SNR at which the KLT-based algorithm can still operate acceptably equals 24 dB. Sensitivity analysis proved that the algorithm is not critically tuned to the test database (ESC DB).

Apart from comparing the robustness of the algorithms, the bootstrap is also useful for comparing the robustness of different performance statistics. Table 2 shows that distributions of average performance statistics are narrower than those of gross statistics for both algorithms. These average statistics appear to be more robust estimates of performance than the corresponding gross statistics, particularly for ischemic ST duration (ID) statistics. Aggregate gross ID statistics are extraordinarily sensitive to single errors. Significant errors in a small number of long episodes have a disproportionately negative influence on gross statistics, but not on the corresponding average statistic.

Performance results of the KLT-based algorithm when changing the dimensionality of feature vectors confirmed that the 5 KL coefficients for ST and QRS feature vectors were a good choice for distinguishing between noisy and non-noisy events [4]. Furthermore, this study suggests that a dimensionality of 5 is also the optimal choice for feature representation and pattern recognition part of the KLT-based analysis algorithm.

References

[1] Jager F, Moody GB, Taddei A, Mark RG. Performance measures for algorithms to detect transient ischemic ST segment changes. In: Computers in Cardiology 1991. Los Alamitos: IEEE Computer Society Press, 1991; 369–372.

[2] Efron B. Bootstrap methods: another look at the jackknife. Annals of Statistics 1979; 7:1–26.

[3] Moody GB, Muldrow WK, Mark RG. A Noise Stress Test for Arrhythmia Detectors. In: Computers in Cardiology 1984. Los Angeles: IEEE Computer Society Press, 1984; 381–384.

[4] Jager F, Mark RG, Moody GB, Divjak S. Analysis of Transient ST Segment Changes During Ambulatory ECG Monitoring Using the Karhunen-Loève Transform. In: Computers in Cardiology 1992. Los Alamitos: IEEE Computer Society Press, 1992; 691–694.

[5] Jager F. Automated Detection of Transient Ischemic ST-Segment Changes During Ambulatory ECG-Monitoring. Doctoral Thesis. Faculty of Electrical and Computer Engineering, University of Ljubljana. Ljubljana, 1994.

[6] Taddei A, Distante G, Emdin M, Pisani P, Moody GB, Zeelenberg C, Marchesi C. The European ST-T database: standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography. European Heart Journal 1992, 13; 1164–1172.

[7] Jager F, Mark RG, Moody GB. Analysis of transient ST segment changes during ambulatory ECG monitoring. In: Computers in Cardiology 1991. Los Alamitos: IEEE Computer Society Press, 1991; 453–456.

[8] Moody GB, Mark RG. Development and Evaluation of a 2-Lead ECG Analysis Program. In: Computers in Cardiology 1982. Los Angeles: IEEE Computer Society Press, 1982; 39–44.

Address for correspondence:
Franc Jager
Faculty of Electrical and Computer Engineering,
Tržaška 25, 61000 Ljubljana, SLOVENIA
Internet: `franc@manca.fer.uni-lj.si`