Detection of Inferior Myocardial Infarction: A Comparison of Various Decision Systems and Learning Algorithms

J Spilka[1,2], V Chudáček[1], J Kužílek[1], L Lhotská[1], M Hanuliak[2]

[1]Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic
[2]Medical Technologies CZ a.s., Czech Republic

Abstract

In this work we have focused on classification of inferior myocardial infarction (MI). We compared the best known scoring/coding/decision systems (the Selvester QRS score, the Novacode, and the Siemens 440/740) and several learning algorithms (Ripper, C4.5, and SVM). The decision systems were developed with different purposes (the Selvester for estimation of MI size, the Novacode for clinical and epidemiologic studies, and the Siemens for ECG device Siemens 440/740). In this work we combined these systems with additional simple rules and compared performance to: (i) decision systems alone, (ii) base classifiers (Ripper, C4.5, and SVM). Our database consisted of 2596 ECG records annotated by experienced cardiologist.

Among decision systems the Selvester and the Siemens had F-measure 54% and 51%, respectively. Meaning that about 50% of MI's were correctly classified. Even lower F-measure of 39% was obtained by Novacode. Better results were achieved using rule miner Ripper with F-measure of 68%, however, due to a number of rules created, the resulting model was hard to interpret. Last, combination of decision systems with additional simple rules created by AdaBoost yielded the best performance with F-measure 71%, sensitivity (Se) 78%, and specificity (Sp) 95%.

1. Introduction

Coronary artery disease, the number one killer in the developed world, is a heart disease when coronary arteries are either partially occluded, resulting to myocardial ischemia, or totally occluded resulting to myocardial infarction (MI). MI could be a minor event, perhaps not even recognized, or it may be a major attack with results varying from acute pains, hemodynamic deterioration to sudden death. MI can be revealed by a number of different signs, including biochemical markers, imaging or pathological characteristics, but the most important initial clinical test for MI diagnosis still remains ECG.

Decision rules are used to assess morphological changes at ECG caused by myocardial ischemia and infarction (ST-T changes for acute state, Q and T wave changes for an infarcted myocardium). Rules were suggested by cardiologists and originated from theoretical foundations and experience. It were adopted later into computerized scoring/coding/decision systems (hereinafter commonly referred as decision systems); the best known are: the Selvester score, the Novacode, and Siemens 440/740. To the best of our knowledge, there have been two papers that compared performance of decision system. First Pahlm et.al. [1] compared performance of the Selvester, the Novacode, and the Cardiac Injury Score regarding to accuracy of MI size estimation. Second, more recent paper [2], compared the Novacode and the Minnesota code in large epidemiologic study. In this work we focus on comparison of three decision systems, which were developed for different purposes, with respect to cardiologists annotation and, furthermore, we attempted to combine the decision systems together with simple rules, created by AdaBoost, to improve the accuracy of MI detection.

2. Methods

2.1. Experimental data

The database of 12-lead resting ECG was provided by Medical Technologies CZ a. s. It contained 6332 records collected during 2004 – 2007 using ECG device 12BTL-08 LC EKG, BTL, Czech Republic, sampling frequency 500 Hz, resolution 3.9 μV for the least significant bit. The interpretation of ECG was performed by experienced cardiologist; 2333 records were assessed as normal; the rest was abnormal. Inferior MI was present in 510 cases.

2.1.1. Data preprocessing

The initial database of 6332 ECG's, we removed 502 records that were unsuitable for diagnostics e. g. ECG was corrupted by noise, ECG was out of physiological limits, or leads were swapped. Furthermore, since we focused on classification accuracy of decision systems, we

ISSN 0276–6574

removed confounding diagnoses such as various types of blocks, ventricles hypertrophy, WPW syndrome, or MI located elsewhere with exception of multiple MI. The reason was to eliminate possible sources of error and not to bewilder decision systems by other pathologies. The final dataset consisted of $\{n_{\text{normal}} = 2333; n_{\text{MI}} = 263\}$.

2.1.2. Morphological features

The morphological features of ECG were extracted from averaged beats computed from 10 seconds of 12-lead rest ECG; premature ventricular beats were not included into averaged beat. Features described important waves and intervals used for diagnostics of MI; the list of 12 features for each lead was as follows (amplitudes ($_{\text{amp}}$) in [μV], durations ($_{\text{dur}}$) in [ms]): Q_{amp}, Q_{dur}, Q_{pos}, R_{amp}, R'_{amp}, R''_{amp}, R_{dur}, S_{amp}, S_{dur}, QRS_{amp}, $\text{R}_{\text{amp}}/\text{Q}_{\text{amp}}$ratio, $\text{R}_{\text{amp}}/\text{S}_{\text{amp}}$ratio. Note that when Q wave was not present the ratio was substituted by R_{amp}.

2.2. Decision systems

The decision systems are described with respect to inferior MI. The adjustment of (codes selection for Siemens; score estimation for Selvester) was made in order to obtain best results on the training/testing set where performance was estimated by 10-fold cross-validation (CV).

Selvester QRS score The Selvester QRS scoring system [3, 4] was derived from computer simulation. Overall score is computed from decision table including 50 rules; summation of individual points is multiplied by three giving percentage of injured left ventricle. We used only scores for inferior leads; score exceeding 1 point was considered as indication of MI.

Novacode The Novacode [5] is successor of Minnesota code [6]. The score system uses different thresholds to quantify severity of an event, though we used it in a dichotomous manner. Codes 5.1. – 5.4. asses Q wave MI and were used for prediction of myocardial injury.

Siemens 440/740 The Siemens 440/740 [7] was used for ECG interpretation in Siemens 400/700 series. Codes used: 1(a), 7(a), and 8(a), section A2.1.8.1. [7].

2.3. Feature selection and classification

Each of inferior lead (II, III, and aVF) was described by 12 features as listed above. In total we had 36 features plus one reference class. The feature set distribution was skewed towards normal class and some classifiers tends to favor this class because of high prior probability. In order to avoid this behavior, we balanced training set using Synthetic Majority Over Sampling Technique (SMOTE) [8], number of nearest neighbors used: $k = 5$. We performed feature selection using filter method Correlation Feature Selection (CFS) [9]. This method selects features that are in strong relationship with a class while having low inter-correlation. Selected features were used for training classifiers Ripper [10], C4.5 [11], and Support Vector Machine [12] (polynomial kernel, C = 1).

2.3.1. AdaBoost learning

The reason of using AdaBoost [13] was to learn a simple classifiers that were different from the original decision systems. The diversity of weak classifiers is corner stone of AdaBoost. As the weak classifier we used a simple thresholding (e.g. we searched a best threshold for $\text{Q}_{\text{dur}}$from $\min(\text{Q}_{\text{dur}})$ to $\max(\text{Q}_{\text{dur}})$). Instead of minimization of classification error we minimized F_{β}-measure. The parameter β weights importance between precision and recall – variation of β leads to different rules to be chosen; $\beta \in \langle 0, 1)$ recall is preferred; $\beta \in (1, 2\rangle$ precision is preferred. Three classifiers were learned – specific (cAdaSpec with $\beta < 1$), balanced (cAdaB, $\beta = 1$), and sensitive (cAdaSens, $\beta > 1$). Then, rules were extrapolated in the way that a rule should be fulfilled at least for two inferior leads. Resulting classification $H \in \{1, -1\}$ for a record x was estimated as $H(x) = sign(\sum_{t=1}^{T} \alpha_t I[single\_rule(x)])$, where T is number of rules, α is weighting factor, and $I[]$ is statement that equals 1 when a $single\_rule$, e.g. $Q_{\text{dur}} < 40$, is satisfied otherwise results in -1.

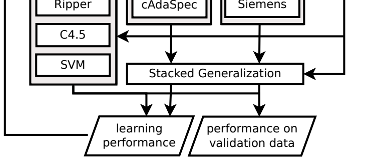


Figure 1. Learning and validation – cAdaSens, cAdaB, and cAdaSpec are (sensitive, balanced, and specific) rules created by AdaBoost algorithm.

Stopping criteria for AdaBoost learning were empirically estimated on progression curves of recall, precision, and F-measure (computed on training/testing dataset by 10-fold CV). The new, simple, classifiers were then stacked together with existing decision systems using decision tree C4.5 and rules miner Ripper. Resulting model was compared to the base classifiers (Ripper, C4.5, SVM).

The 5-fold cross-validation approach was used for data division into training, testing, and validation sets. The results on validation set were averaged thus giving overall performance, see Figure 1 for details of the overall setup.

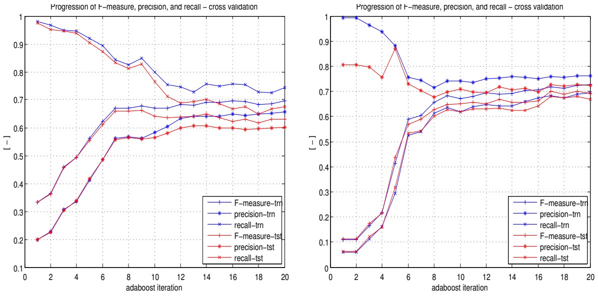
3. Results

First, we tested performance of decision systems (the Selvester score, the Novacode, the Siemens 440/740); results are shown in Table 1. With respect to F-measure the Siemens and the Selvester achieved above 50%. The Siemens had also good sensitivity of 63% but lower precision than the other two.

Second, we constructed classifiers using learning algorithms Ripper, C4.5, and SVM for selected features by CFS method: {R$_{\mathrm{dur}}$(II), R”$_{\mathrm{amp}}$(II), Q$_{\mathrm{amp}}$(III), R$_{\mathrm{dur}}$(III), R’$_{\mathrm{amp}}$(III), Q$_{\mathrm{amp}}$(aVF), Q$_{\mathrm{dur}}$(aVF), R’$_{\mathrm{amp}}$(aVF)}; results are present in Table 2. Best performance was achieved by Ripper with Se/Sp 82/93% and F-measure of 68%.

Third, we constructed new classifiers with AdaBoost taking single rules as weak learner. Stopping criteria were experimentally estimated; when either precision (cAdaSens) or recall (cAdaSpec) dropped markedly the learning was stopped. The progression of precision, recall, and F-measure for cAdaSens and cAdaSpec are shown in Figure 2. Stopping criteria: cAdaSens 4th, cAdaSpec 5th, and cAdaB 1th iteration. Individual classifiers consisted of following rules:

- cAdaSens: $H = sign(0.22 \cdot I[R_{\mathrm{amp}}/Q_{\mathrm{amp}} < 46] + 0.3 \cdot I[R_{\mathrm{amp}}/Q_{\mathrm{amp}} < 228] + 0.18 \cdot I[Q_{\mathrm{amp}} < -100] + 0.37 \cdot I[R_{\mathrm{amp}}/Q_{\mathrm{amp}} < 80])$.
- cAdaB: $H = sign(1.25 \cdot I[R_{\mathrm{amp}}/Q_{\mathrm{amp}} < 4.9])$.
- cAdaSpec: $H = sign(1.1 \cdot I[Q_{\mathrm{amp}} < -269] + 0.58 \cdot I[Q_{\mathrm{dur}} \geq 24] + 0.07 \cdot I[R_{\mathrm{amp}}/Q_{\mathrm{amp}} < 6.7] + 0.1 \cdot I[R_{\mathrm{amp}}/Q_{\mathrm{amp}} < 8.7]) + 0.57 \cdot I[R_{\mathrm{amp}}/Q_{\mathrm{amp}} < 4.9])$.



(a) (b)

Figure 2. Estimation of stopping criteria on *trn* (training) set in blue and *tst* (testing) set in red. (a) cAdaSens: stopping at 4th iter., (b) cAdaSpec: stopping at 5th iter.

Table 3 shows results of stacked generalization model (SG) that combines decision systems with AdaBoost classifiers. SG model created by Ripper yielded better performance in both cases (i) than using decision systems alone and (ii) in comparison with Ripper, C4.5, and SVM. The created models were decision tree (C4.5) and rules (Ripper) shown in Figure 3.

Table 1. Classification performance of decision systems.

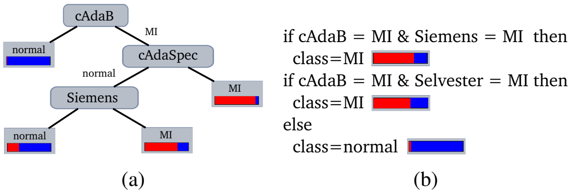
All in [%]	Selvester	Novacode	Siemens
sensitivity	51	28	63
specificity	95	98	91
precision	58	63	43
F-measure	54	39	51

Table 2. Performance of Ripper, C4.5, and SVM on the validation set; estimated by 5-fold cross-validation.

All in [%]	Ripper	C4.5	SVM
sensitivity	82	78	86
specificity	93	93	86
precision	58	57	41
F-measure	68	66	56

Table 3. Performance of stacked generalization (SG) models of decision system and AdaBoost rules; estimated by 5-fold cross-validation.

All in [%]	SG model C4.5	SG model Ripper
sensitivity	63	78
specificity	97	95
precision	68	66
F-measure	65	71



if cAdaB = MI & Siemens = MI then
class=MI
if cAdaB = MI & Selvester = MI then
class=MI
else
class=normal

(a) (b)

Figure 3. Stacked generalization model. (a) decision tree (C4.5), (b) rules (Ripper). Bars in the tree’s leaves and rules show estimate of accuracy of MI prediction (blue – normal; red – MI)

4. Discussion and conclusions

First, we tested performance of decision systems; regarding overall assessment using F-measure the Selvester QRS score and the Siemens had F-measure of 54% and 51%, respectively. It means that the number of correctly classified MI's was almost the same as the number of false positive and false negative examples. The Novacode had even worse F-measure of 31%, though better results of the Novacode's ancestor (Minnesota code) were published by [14]. They reported Se/Sp of 62/88% which contradicts our results of 28/98%. The Selvester score was designed to have specificity of 95% [4] and although we lowered points indicating MI the high specificity was maintained with even better sensitivity of 51% than the published 41% by Haisty et. al. [15]. Haisty et. al. also suggested that the Selvester code needs further improvements. The improvements were tackled by [16] and resulted in addition of new rules to existing ones, however, in our case, by modification according to [16] we did not obtain better results. Therefore we created new rules using AdaBoost and combined them with existing decision systems. This combination yielded to the best performance with Se/Sp 78/95% and F-measure 71%. The created model was cAdaB in combination with Siemens and Selvester.

For comparison, among base classifiers the rule miner Ripper algorithm had Se/Sp 82/93%, and 68% F-measure. However, the Ripper created a very complex rules even though we used only control and MI records for learning. It is likely that rather than modeling distribution of population the Ripper modeled available data.

One import question should be consider in the future work – will the stacked generalization model (rules incorporating cAdaB, Siemens, and Selvester) perform well even when confounding diagnoses will be included?

Acknowledgements

This work has been supported by the ČVUT Grant SGS10/279/OHK3/3T/13 and by the research program MSM 6840770012 "Transdisciplinary Research in the Area of Biomedical Engineering II" of the CTU in Prague, sponsored by the Ministry of Education, Youth and Sports of Czech Republic.

References

[1] Pahlm US, Chaitman BR, Rautaharju PM, Selvester RH, Wagner GS. Comparison of the various electrocardiographic scoring codes for estimating anatomically documented sizes of single and multiple infarcts of the left ventricle. Am J Cardiol Apr 1998;81(7):809–815.

[2] Zhang Z, Prineas RJ, Eaton CB. Evaluation and comparison of the Minnesota Code and Novacode for electrocardiographic Q-ST wave abnormalities for the independent prediction of incident coronary heart disease and total mortality (from the Women's Health Initiative). Am J Cardiol Jul 2010;106(1):18–25.e2.

[3] Selvester RH, Sanmarco RE, Solomon JC, Wagner GS. "The ECG: QRS change," in Myocardial Infarction: Measurement and Intervention. Martinus Nijhoff, The Hague, The Netherlands, 1982; 23–50.

[4] Hindman NB, Schocken DD, Widmann M, Anderson WD, White RD, Leggett S, Ideker RE, Hinohara T, Selvester RH, Wagner GS. Evaluation of a QRS scoring system for estimating myocardial infarct size. V. Specificity and method of application of the complete system. Am J Cardiol Jun 1985;55(13 Pt 1):1485–1490.

[5] Rautaharju PM, Park LP, Chaitman BR, Rautaharju F, Zhang ZM. The Novacode criteria for classification of ECG abnormalities and their clinically significant progression and regression. J Electrocardiol Jul 1998;31(3):157–187.

[6] Blackburn H, Keys A, Simonson E, Rautaharju P, Punsar S. The electrocardiogram in population studies. a classification system. Circulation Jun 1960;21:1160–1175.

[7] Macfarlane PW, Lawrie VTD. Comprehensive Electrocardiology: Vols 1-3: Theory and Practice in Health and Disease. McGraw Hill Higher Education, 1988.

[8] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 2002;16:321–357.

[9] Hall MA. Correlation-based feature selection for machine learning. Technical report, The University of Waikato, 1998.

[10] Cohen W. Fast effective rule induction. In Proc. of the 12th Intl. Conf. on Machine Learning. Tahoe City, CA., 1995; 115–123.

[11] Quinlan J. C4.5: Programs for machine learning. Morgan Kaufmann, 1992.

[12] Vapnik VN. Statistical Learning Theory. Wiley-New York, 1998.

[13] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. In European Conference on Computational Learning Theory. 1995; 23–37.

[14] Uusitupa M, K KP, Raunio H, Rissanen V, Lampainen E. Sensitivity and specificity of Minnesota Code Q-QS abnormalities in the diagnosis of myocardial infarction verified at autopsy. Am Heart J 1983;106(4 Pt 1):753–7.

[15] Haisty WK, Pahlm O, Wagner NB, Pope JE, Wagner GS. Performance of the automated complete Selvester QRS scoring system in normal subjects and patients with single and multiple myocardial infarctions. J Am Coll Cardiol Feb 1992;19(2):341–346.

[16] Horacek BM, Warren JW, Albano A, Palmeri MA, Rembert JC, Greenfield JC, Wagner GS. Development of an automated selvester scoring system for estimating the size of myocardial infarction from the electrocardiogram. J Electrocardiol Apr 2006;39(2):162–168.

Address for correspondence:

Jiří Spilka

Department of Cybernetics, Czech Technical University in Prague, Technicka 6, 166 37 Praha 6, Czech Republic
spilka.jiri@fel.cvut.cz