Multi-Class Classification of Pathologies Found on Short ECG Signals

Georgi Nalbantov[1], Svetoslav Ivanov[1], Jeffrey van Prehn[1]

[1]Data Science Consulting Ltd, Pernik, Bulgaria

Abstract

The ability to detect several key cardiac pathologies simultaneously, based on ECG signals, is key towards establishing a real-world application of AI models in cardiology. Such a multi-label classification task requires not only well-performing binary classification models, but also a way to combine such models into an overall classification modeling structure. We have approached this task using materials from Classification of 12-lead ECGs for the PhysioNet/Computing in Cardiology Challenge 2020. Duplicate ECG strips have been removed. An annotation tool for labeling ECG wave points and intervals/templates has been created in MATLAB®, and used for labeling pathological intervals, as well as noisy intervals and inconsistencies between the ECG data and the pre-assigned labels. Several one-vs-rest binary classifiers were built, where morphological features specific to each pathology had been generated from the signals. The binary classifiers were augmented by a multi-class classifier using an Error Correcting Output Codes (ECOC) methodology. Our approach achieved a challenge validation score of 0.616, and full test score of 0.194, placing us 23 (team DSC) out of 41 in the official ranking.

1. Introduction

The ability to detect various cardiac pathologies individually and simultaneously in ECG signals is key towards establishing a real-world application of AI models in cardiology. The PhysioNet/Computing in Cardiology Challenge 2020 ('the Challenge') focused on automated, open-source approaches for classifying cardiac abnormalities from 12-lead ECGs [1,2]. Our best entry in the Challenge applied a multi-stage binary pathology detection scheme, using an Error Correcting Output Codes (ECOC) [3] method for multi-class classification sub-problems. We have subsequently used a bootstrap-averaging (as in Bagging [4]) approach, applied over our overall learning scheme, for robustness of the predictions.

2. Methods

The learning steps to approach the task in the Challenge ran as follows. At the first step, data pre-processing, we scanned a vast number of the training data for noisy and (where apparent) mislabeled signals within each of the 12 ECG leads. Subsequently, we either relabeled or excluded some of those signals. All signals have been re-sampled to 250 Hz. At the second step, feature generation, we applied an R-peak detector, 'gqrs' [5], to leads I and II of the ECG strips. Thus we obtained (predicted) R-peak locations along each strip, and were able to define ECG inter-beat intervals (IBI). A total of 867 features have been generated, some of which on complete strips, some within each IBI. Specifically, we have used a feature set proposed in [6] which was used in the PhysioNet/Computing in Cardiology Challenge 2017 on Atrial Fibrillation (AF) classification from a short single lead ECG recording [7], the Global Electrical Heterogeneity features which were included in the MATLAB® baseline model from [2] as well as morphology-specific features, which we have developed to be appropriate for the pathologies provided in the Challenge. However, for 'pacing rhythm' and 'nonspecific intraventricular conduction disorder' no specific features have been created. A lot of the features were based on ECG wave detections - P-wave, Q-wave, and T-wave - for which we used the so-called wavedet detector developed by [8]. The wavedet detector was applied separately on leads I, II, aVL, V1, V4 and V6 of each signal. The AF features were generated on leads I, II and V1. Specific features, based on all ECG leads if not stated otherwise, next to the above general features, were assigned for the following pathologies:

1. '1st degree av block' and 'prolonged pr interval' : distance between beginning/mid/end of P-wave and R;
2. 'atrial fibrillation' : variance of the distances between beginning/mid/end of P-wave and R;
3. 'bradycardia' / 'tachycardia': length of IBI's;
4. '(complete) right bundle branch block' / 'left bundle branch block' : T-wave (lack of) inversion in leads V1, V2, V5, and V6; length of QRS complex;
5. 'left anterior fascicular block', 'left axis deviation', and 'right axis deviation' : height of R-peak in leads I,II,III,

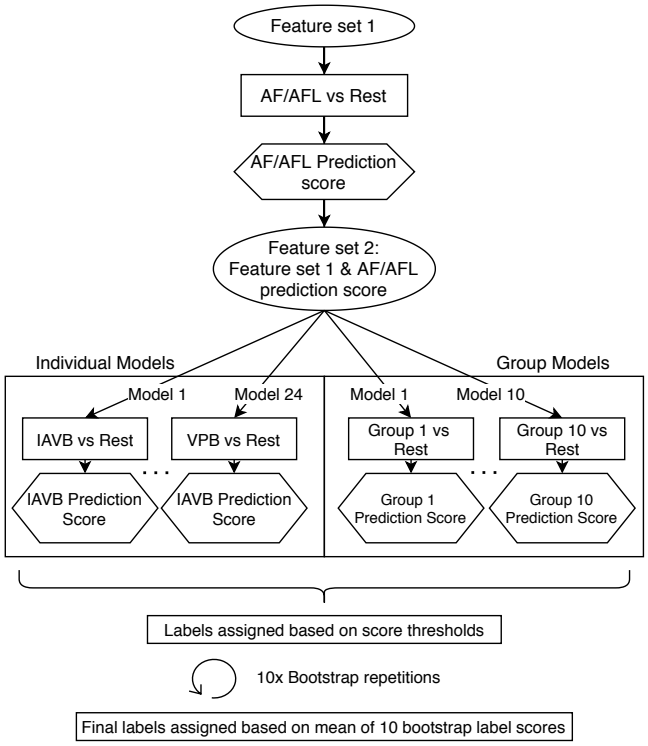

Figure 1. *Our methodology comprises a learning scheme of 6 steps: (1) binary classification of 'regular vs. irregular rhythm'; (2) addition of the resulting raw prediction score to the feature set; (3) run one-vs-rest binary models; (4) run group-vs-group (of labels) binary models; (5) repeat the procedure 10 times on bootstrap versions of the training set; (6) assign average score per label/pathology.*

aVR, aVL, aVF, where the R-peak locations are computed only from leads I and II and super-imposed on the rest of these leads;

6. 'low qrs voltages' : voltage difference between the lowest and highest points in the QRS complex;

7. 'premature atrial contraction' and 'premature ventricular contractions' : morphology of the signal both within one IBI and two consecutive IBI intervals.

8. 't wave abnormal' : we have used features only for detecting two types of abnormalities: inverted T-waves and flattened T-waves.

Once the set of IBI features had been generated, we employed a multi-label learning scheme. See Figure 1 for a graphical representation of the scheme. In each step of the scheme we employed a binary classifier, which is AdaBoost [9], implemented in MATLAB® [10]. We note that the classes in each binary classifier consisted either of one label or of several labels (pathologies) combined into one label. Accordingly, the labels in this scheme we first divided into regular vs. irregular rhythms, where the irregular rhythms were 'atrial fibrillation' and 'atrial flutter'. We have taken a decision not to include 'sinus arrhythmia' among the irregular rhythms. This binary classifier we referred to as "regular rhythms vs. irregular rhythms". Once a prediction for the presence of 'irregular rhythms' had been generated, we added the prediction score (a value between 0 and 1) to the feature set for the rest of the (further) classifiers. This approach was inspired by Classifier Chains [11], with the modification that we add the prediction scores instead of the predicted labels. We obtained these scores in the training phase using 5-fold cross validation. The second step of the learning scheme consists of running a one-vs-rest binary classifier for each of the so-called 'scored' pathologies in the Challenge. In addition, we generated several multi-class classification problems for frequently occurring (individually covering at least 1.8% of the instances) pathology or normal class combinations in the training dataset. This resulted in 10 combinations (groups) being selected, which covered 60% of the training set. For these multi-class problems we applied an Error Correcting Output Codes (ECOC) [12] methodology, which resulted in one set of frequently occurring combinations of labels to be predicted for a particular ECG strip.

The ECOC methodology in essence works as follows: each class label is described in terms of several (boolean) aspects, with the aim to provide redundancy. In fact, each label is coded as a vector of zeros and ones, representing the true label. To predict the label of an instance, one predicts each of the aspects. The prediction is represented by a vector of zeros and ones. The prediction vector is decoded by looking up the closest vector (according to some distance measure) for which a true label is known. To apply ECOC in the context of the challenge, each combination of pathologies that occurs in the training set, is mapped one-to-one to a true label. Note that this results in more than a thousand labels. For a particular label, the occurrence or non-occurrence of each pathology in that label defines part of the label's coding: one bit for each pathology. This scheme is known in the literature as Binary Relevance [13, p. 60]. Each of the remaining bits represents a frequently-occuring combination of pathologies. We use Hamming Distance for decoding. The MATLAB® implementation of ECOC, ECOClib [14] and our own customizations of it were used in several submissions.

Once scores for each separate pathology (or lack of such) were obtained, the following rule (known as Proportional Cut Method in the literature [13, p. 50]) was used to assign thresholds on these scores, so that a label is being assigned only if the respective threshold is surpassed (if no threshold is surpassed, the label with the highest score is assigned): Choose a threshold (one threshold per pathology) such that the distribution of predicted labels in a test set is the same as the observed one in the training set. Note that here, the test set is not the hidden test set of the Chal-

lenge, but a stratified holdout sample of the data available for training in which care was taken not to include ECG strips from the same patient in both the training and test set. After the thresholds are determined in this way, they remain fixed for use on the hidden test set. The reasons for fixing the thresholds are that (1) the class distribution of the hidden test data might differ from the class distribution of the training data and (2) hidden test instances are classified one after the other, so the distribution of the test data is not known to the classification code. To apply this rule we used code published by [15].

The final stage of the learning scheme, after scores for each pathology (or lack of such) had been generated and detection thresholds on the scores applied, involved repeating the procedure 10 times using bootstrap samples taken from 80% (at random) of the training set. The main idea was to achieve more robust scoring. Thus, 10 scores were obtained, one for each training bootstrap sample, and the thresholds for each pathology were applied to the average score (per pathology) obtained from the 10 bootstrapped training datasets.

3. Results

The so-called Challenge metric score for our best-ranked model was 0.194 on the undisclosed full test set of the Challenge (0.824 on test database 1 of size 1463, 0.301 on test database 2 of size 5167 and 0.062 on test database 3 of size 10000). This result has put our team, DSC, a team-position rank number 23 among the official scores with ranking. The best team achieved a score of 0.533. Our cross-validation score on the training set on the same metric was 0.4. We note that our submissions strategy for the Challenge test-set scoring involved 10 submissions (maximum allowed), each of which cover partially the aspects of our overall learning scheme presented in the Methodology section. Our aim was to find out which aspects of our methodology contributed best to improvements of our naive baseline score of 0.346 of a multi-class, single-label ECOC model taking into account the cost-matrix provided by the challenge. Introducing a multi-label scheme, improved the score to 0.578. Bagging improved this result further to 0.611. Adding specific morphological features increased the score to 0.616 on the partial validation hidden test set of the official phase. Apart from the official submissions, we also ran several experiments on the training data using cross-validation. As a result, we have observed that most importantly, it is possible to greatly reduce the number of features to 1-5 features per pathology. This greatly improves the interpretability of the models. The presence or absence of specific morphological features tuned to detect morphologies in the ECG signal play the most important role in reducing the number of features. Just adding these specific features to the feature set does not make a huge difference. Second of all, the thresholding strategy plays a very significant role towards obtaining a high score, though it has an intrinsic flow, as discussed in the next section. Last but not least, the addition of (raw) prediction scores for 'irregular rhythm' to all one-vs-all binary classification problems proved beneficial.

4. Discussion and Conclusions

We have implemented a multi-label multi-step strategy to classify pathologies (or lack of such) based on a huge training set of 12-lead ECG strips. A total of 23 pathologies as well as a 'sinus rhythm' have been included in the Challenge. We believe that including the prediction for a certain group of pathologies, in our case pathologies representing irregular rhythms, as an input feature in (subsequent) one-vs-rest binary classification models proved to be beneficial. One alternative was, for example, to use a binary classifier for regular vs irregular rhythm, and then at a second stage to apply a one-vs-others classification scheme, where 'others' is not all the rest of the ECG labels but rather those with regular rhythm. This strategy proved to be inferior in our cross-validation testing. The strategy for assigning thresholds on the label prediction scores was chosen so as to make sure that the expected distribution of predicted labels in a test set to be the same as the observed one in the training set. This is valid only if the assumption holds true. In real life applications where the populations differ such a strategy is expected to be inferior. Furthermore, we do not take into account the cost matrix provided by the Challenge in choosing the thresholds. This could be incorporated by choosing the thresholds using cross-validation. We have applied just one classifier for all of our (binary) classification problems, namely AdaBoost. This is bound to be suboptimal. In addition, we have not performed feature selection for each classification problem, which is another drawback of our learning scheme that needs to be improved, since various feature-selection algorithms exist. Our ECOC approach to assigning multiple predicted labels to a test patient has an implicit downside that these labels should have occurred in the training dataset. Fundamentally, we believe that the purpose of making model-driven pathology detections on ECG strips is to enable their implementation in real-world conditions. This means that noisy intervals in ECG signal recordings have to be excluded before pathology prediction to avoid high false alarm rates. Since "noise" was not provided as a separate label for strips, we believe that the usefulness of our method is limited for real-world applications. In addition, other pathologies do exist, including major ones, which have not been included in the list of 24 pathologies (although some of the them have been provided in a so-called unscored list of pathologies). Since our model did not include such possibility, it would by construction pro-

vide a wrong label to a pathology outside the 24 scored pathologies. Moreover, the fact that many participants in the Challenge expressed doubts in the validity of the provided labels for parts of the training set brings the notion that those labels might have been provided in a subjective way, e.g. by doctors looking at ECG strips and pointing out detected by eye morphologies, rather than an objective way, where the presence of a pathology is demonstrated unequivocally by other sorts of formal tests (sometimes not based on ECG signals). This makes the learning task extremely hard. What if a doctor fails to detect a given pathology (and thus mis-labels ECG strips)? What if a noisy signal is provided, say for a patient with (permanent) 'atrial fibrillation' in a noisy signal environment and a model predicts that 'atrial fibrillation' should be present while a doctor may not spot it due to the noise? We believe that in order to move closer to a real-world application of our models, providing longer strips with proven pathologies, where doctors can detect a pathology on part of the ECG strip and cannot detect the pathology on other parts of the strips, is a must. Last, we were not sure why the so-called 'CPSC' training set of the Challenge contained a substantial number duplicate strips. It could have been a test for the Challengers to find this out, and therefore we would like to report this in the current paper. Ignoring this fact may introduce a huge bias in cross-validation performance. We also note that we have purposefully avoided any use of deep learning / neural network techniques so as to make sure our results are replicable, meaning that the learning process cannot be ''stuck' in a local minimum and thus provide different predictions if the learning process is re-performed.

To conclude, this paper provided a summary of some of our approaches for modeling the presence or lack of presence of 23 cardiac pathologies based on provided 12-lead ECG strips by the Physionet 2020 Challenge. We have introduced two novelties in the analysis of such 12-lead strips, which have been borrowed from other literature/areas, namely (1) ECOC approach for handling multi-class problems, and (2) a stepwise multi-label approach where the predictions from one stage are used as inputs for another stage.

References

[1] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Componentsof a New Research Resource for Complex Physiologic Signals. Circulation 2000;101(23):e215–e220.

[2] Perez Alday EA, Gu A, Shah A, Robichaux C, Wong AKI, Liu C, Liu F, Rad BA, Elola A, Seyedi S, Li Q, Sharma A, Clifford GD, Reyna MA. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Physiol Meas 2020;In Press.

[3] Dietterich TG, Bakiri G. Solving Multiclass Learning Problems via Error-Correcting Output Codes. Journal of Artificial Intelligence Research 1994;2:263–286.

[4] Breiman L. Bagging predictors. Machine Learning 1996; 24(2):123–140.

[5] Silva I, Moody G. An Open-source Toolbox for Analysing and Processing PhysioNet Databases in MATLAB and Octave. Journal of Open Research Software 09 2014;2:e27.

[6] Datta S, Puri C, Mukherjee A, Banerjee R, Choudhury AD, Singh R, Ukil A, Bandyopadhyay S, Pal A, Khandelwal S. Identifying Normal, AF and Other Abnormal ECG Rhythms Using a Cascaded Binary Classifier. In Computing in Cardiology, CinC 2007, Rennes, France, September 24-27, 2017. 2017; 601–604.

[7] Clifford GD, Liu C, Moody B, Lehman LH, Silva I, Li Q, Johnson AE, Mark RG. AF Classification from a Short Single Lead ECG Recording: the PhysioNet/Computing in Cardiology Challenge 2017. In Computing in Cardiology, CinC 2007, Rennes, France, September 24-27, 2017. 2017; 217–220.

[8] Martinez JP, Almeida R, Olmos S, Rocha AP, Laguna P. A Wavelet-Based ECG Delineator: Evaluation on Standard Databases. IEEE Transactions on Biomedical Engineering 2004;51(4):570–581.

[9] Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences 1997;55(1):119–139.

[10] MATLAB. version 9.8.0 (R2020a). Natick, Massachusetts: The MathWorks Inc., 2020.

[11] Read J, Pfahringer B, Holmes G, Frank E. Classifier Chains for Multi-label Classification. In Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II, volume 5782 of Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. ISBN 978-3-642-04174-7, 2009; 254–269.

[12] Kajdanowicz T, Kazienko P. Multi-label Classification Using Error Correcting Output Codes. International Journal of Applied Mathematics and Computer Science 01 2012; 22:829–840.

[13] Read J. Scalable Multi-label Classification. Ph.D. thesis, University of Waikato, 2010.

[14] Escalera S, Pujol O, Radeva P. Error-Correcting Output Codes Library. Journal of Machine Learning Research 2010;11(20):661–664.

[15] Kimura K, Sun L, Kudo M. MLC Toolbox: A MATLAB/Octave Library for Multi-Label Classification. ArXiv 2017;1704.02592.

Address for correspondence:

Jeffrey van Prehn
Lippestraat 24, Voerendaal, The Netherlands
jvanprehn@gmail.com