A Topology Informed Random Forest Classifier for ECG Classification

Paul Samuel Ignacio, Jay-Anne Bulauan, John Rick Manzanares

University of the Philippines Baguio, Baguio City, Philippines

Abstract

This paper accompanies Team Cordi-Ak’s entry to the classification of 12-lead ECGs for the PhysioNet/Computing in Cardiology Challenge 2020. Our approach leverages mathematically computable topological signatures of 12-lead ECGs as proxy for features informed by medical expertise to train a two-level random forest model in a multi-class classification task. We view ECGs as multivariate time series data and convert different segments and groupings of leads to point cloud embeddings. This stores both local and global structures of ECGs, and encodes periodic information as attractor cycles in high-dimensional space. We then employ topological data analysis on these embeddings to extract topological features based on different summaries available in the literature. We supplement these features with demographic data and statistical moments of RR intervals based on the Pan-Tompkins algorithm for each lead to train the classifier. Our two-level random forest classifier received a score of $0.219 \pm 0.002$ using 3-fold cross-validation on the full training data and a score of $0.304$ on the subset of the full test data, and we ranked 53rd out of 100 teams that successfully participated in the official phase of this year’s Challenge.

1. Introduction

Cardiovascular diseases lead the causes of death worldwide [1]. Early and accurate diagnosis of cardiac conditions are prerequisite to the development of an appropriate and personalized treatment program [2]. This in turn increases the chances of survivability or successful management of the specific cardiac condition. The diagnosis of cardiac conditions relies on the rigorous and manual analysis of a patient’s 12-lead electrocardiogram reading as different cardiovascular diseases have different causes and require different interventions [3]. Evidently, this is time-consuming and requires interpretation provided by highly skilled personnel with similarly high degree of training.

The PhysioNet/Computing in Cardiology Challenge 2020 focused on automated, open-source approaches for classifying cardiac abnormalities from 12-lead ECGs [4]. Our entry to this challenge leverages mathematically computable topological signatures of 12-lead ECGs as proxy for features informed by medical expertise to train a random forest model in a multi-class classification task. As has been shown for detecting Atrial Fibrillation using single-lead ECGs [5], this approach verifies the existence and viability of signal in the topology of ECGs for improving diagnosing cardiac conditions. Upscaling this to the use of all 12 leads of a standard ECG to diagnose multiple heart conditions improves accessibility to automated diagnostics by reducing expert-dependent input in feature extraction.

2. Methods

We use a standard pipeline for examining time series data using topological data analysis. We first generate point cloud embeddings from the ECG data, then extract topology-based features using tools from algebraic topology, and finally employ these features to train a two-level random forest classifier. Figure 1 illustrates this pipeline.

Figure 1. Pipeline. Time series data is converted to point cloud embeddings from which topological signatures can be extracted via TDA and used for machine learning algorithms.

2.1. Point Cloud Embedding

To generate point cloud embeddings, we consider each ECG reading as a sequence of multi-dimensional vectors.
where leads correspond to coordinates. Each time-slice of a 12-lead ECG represents a vector in high-dimensional Euclidean space, and each periodic signal in an ECG is embedded as an attractor cycle.

We unwrap characteristics of ECG readings by considering point cloud embeddings generated from different groupings of lead segments. The idea is to record characteristics that become more pronounced after filtering out excessive or redundant information from multiple leads. We consider three segments consisting of 1800 time points respectively from the beginning, middle, and ending portions of the ECG from which different lead groupings will be extracted to generate point clouds. See Figure 2. It is worth noting that due to the variable lengths in the ECG readings, the middle and end segments across ECG readings may refer to different time points. However, we argue that because of the overall periodic behavior of the cardiac cycle, and since the length of the segments being considered spans multiple periods, the point cloud embeddings generated from the captured middle and end segments provide good representations of the topology of the corresponding ECG segment. All things considered, the inclusion of the middle and end segments provides additional topological information about the ECGs.

Within each segment, we then consider different groupings of leads to generate several point cloud embeddings. To reduce computational costs, we represent each lead groupings by 300 equally spaced time slices within the segments. The first grouping uses all 12 leads in the ECG and represents the overall topology of the ECG segment. The other groupings are constructed based on two criteria: i. groups must collectively span all 12 leads; ii. some groups may represent collections of leads described in the literature as references for diagnosing specific cardiac conditions belonging to the original 8 classes identified in the unofficial phase of the challenge. Table 1 provides the different lead grouping considered in every segment.

2.2. Feature Extraction

We examine each point cloud embedding through the lens of topological data analysis, particularly via persistent homology [7]. For a quick introduction to this approach with accompanying application to similar data, please see [6]. Succinctly, we induce a distance-parameterized monotonic sequence of abstract simplicial complexes, each revealing a parameter-specific set of topological signatures about the underlying point cloud. In this application, these signatures pertain to connectivity and periodicity information about the corresponding ECG segment. We track topological information that persist across different parameter values and record it as a collection of bars, called persistence barcodes. In this work, we are only concerned with persistent homological features from dimensions 0 and 1, and use the python package RIPSER [8] to compute the barcodes. We derive statistical features from persistence barcodes and other persistence-based summaries such as landscapes [9], and entropy [10]. We summarize these features below.

1. Dimension 0 and 1 Barcode Statistics.
   (a) Number of dimension 0 and dimension 1 bars.
   (b) Mean, Standard Deviation, Skewness, and Kurtosis for persistence of dimension 0 bars.
   (c) Mean, Standard Deviation, Skewness, and Kurtosis for birth time, death time, and persistence of dimension 1 bars.
   (d) Sums of persistence in dimensions 0 and 1.

2. Dimension 0 and 1 Truncated Barcode Statistics.
   (a) Mean, Standard Deviation, Skewness, and Kurtosis for persistence of dimension 0 bars.

Table 1. Lead Groupings. Leads are grouped either to span all 12 leads of a standard ECG, or to represent collections of leads used in practice as references for diagnosing specific cardiac conditions.

<table>
<thead>
<tr>
<th>Grouping</th>
<th>Leads</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-12</td>
</tr>
<tr>
<td>2</td>
<td>1-8</td>
</tr>
<tr>
<td>3</td>
<td>5-12</td>
</tr>
<tr>
<td>4</td>
<td>1-6</td>
</tr>
<tr>
<td>5</td>
<td>7-12</td>
</tr>
<tr>
<td>6</td>
<td>2-8</td>
</tr>
<tr>
<td>7</td>
<td>1-11</td>
</tr>
</tbody>
</table>

Table 1. Lead Groupings. Leads are grouped either to span all 12 leads of a standard ECG, or to represent collections of leads used in practice as references for diagnosing specific cardiac conditions.
2.3. Classifier Training

We use a two-level random forest classifier. A first level random forest is trained to classify between scored and unscored classes based on the Challenge-provided lists. The second level uses two random forests, one each for scored and unscored classes.

For each ECG recording in the training set, we scan the accompanying labels and determine if at least one label belongs to the scored class. In such case, the first label of an ECG that belongs to the scored class is taken as its assigned label, and the ECG is included in the training subset for the scored classes. Otherwise, the ECG is placed in the training subset for the unscored classes. Using a binary list, we keep track of which ECGs eventually belong to the scored and unscored classes.

Each second-level forest is trained using the appropriate training subset, and features for each forest are again ranked by importance using Scikit-learn’s built-in functions and a shortlist is used to re-train the forest.

Table 2 shows the non-optimized parameter values for each random forest. We note that parameters for the random forest that classifies the unscored classes have smaller values to control computational costs due to significantly larger number of unscored classes.

<table>
<thead>
<tr>
<th>RF Level</th>
<th># Features</th>
<th># Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>400</td>
<td>1000</td>
</tr>
<tr>
<td>2-1</td>
<td>400</td>
<td>600</td>
</tr>
<tr>
<td>2-2</td>
<td>200</td>
<td>200</td>
</tr>
</tbody>
</table>

Table 2. Parameter Values for Random Forest Training. All forests have a maximum depth of 20 and use the square root function for the maximum number of features.

3. Results

The performance of our two-level random forest classifier on a stratified 3-fold cross validation over the full training set is reported in Table 3.

<table>
<thead>
<tr>
<th>Fold</th>
<th>AUROC</th>
<th>AUPR</th>
<th>AAcc</th>
<th>F1</th>
<th>Challenge Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.845</td>
<td>0.319</td>
<td>0.208</td>
<td>0.257</td>
<td>0.220</td>
</tr>
<tr>
<td>2</td>
<td>0.848</td>
<td>0.327</td>
<td>0.214</td>
<td>0.256</td>
<td>0.219</td>
</tr>
<tr>
<td>3</td>
<td>0.845</td>
<td>0.321</td>
<td>0.205</td>
<td>0.253</td>
<td>0.217</td>
</tr>
<tr>
<td>Average</td>
<td>0.846</td>
<td>0.322</td>
<td>0.209</td>
<td>0.255</td>
<td>0.219</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>0.002</td>
<td>0.004</td>
<td>0.005</td>
<td>0.002</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Table 3. Performance of the proposed topology informed random forest classifier on a stratified 3-fold cross validation over the training set.

Testing performed by the Challenge organizers on the trained classifier on a subset of the full test data during the official phase yields a Challenge score[^3] of 0.304, which puts our team at rank 53 out of 100 teams that successfully participated in the official phase of this year’s Challenge. To put this score in perspective of the objective of our approach, the Challenge-released baseline python classifier, which only uses features from statistical moments of RR intervals from a single lead, received a testing score of 0.076.

4. Discussions

We elected to employ a random forest classifier as we believe it mimics what is practiced at large in the community of experts in coming up with a collective and community-accepted diagnosis of cardiac conditions. However, as the significant portion of our work is about extracting mathematically computable topological features from ECGs, which is fairly recent in ECG diagnosis research, other machine or deep learning algorithms may be trained using these features to develop new models.

As our main objective was to demonstrate the existence and viability of signal in the topology of ECGs for diagnosing cardiac conditions, we made an informed choice to not perform fine-tuning nor parameter optimization for the classifier to avoid introducing additional factors in the classifier performance.

Because of the size of the full training set, extracting the features using our approach takes a significant amount of time. In fact, it is unfortunate that majority of our submit-
5. Conclusions and Future Plans

In this work, we set out to test whether topological information embedded within 12-lead ECG readings contain signal that can be tapped to improve cardiac diagnosis. These topological signatures, which are mathematically computable and require minimal input from highly skilled medical personnel, improve accessibility to automated diagnostics by reducing expert-dependent input in feature extraction. Despite the computational obstructions that we encountered, including the failure of our “proper” model entries to get official scores due to time limits, we found that these topological signatures do improve ECG diagnosis as evidenced by marked positive differences in the evaluation scores of our successful initial entries relative to the baseline model that uses similar non-topological features as our model. We plan to optimize our model and resubmit for evaluation after the conference to provide a better analysis of our approach.

Acknowledgments

We would like to thank Prof. David Uminsky for the useful discussions and suggestions to improve our approach. We also acknowledge support from the University of the Philippines Baguio.

References


