A Fast Algorithm for Facilitating Heartbeat Annotation in Long-Term ECG Signals

Ana Santos Rodrigues[1], Mantas Lukoševičius[1,2], Vaidotas Marozas[1,3]

[1]Biomedical Engineering Institute, Kaunas University of Technology. Kaunas, Lithuania
[2]Faculty of Informatics, Kaunas University of Technology. Kaunas, Lithuania
[3]Electronics Engineering Department, Kaunas University of Technology. Kaunas, Lithuania

Abstract

Machine-learning models for automatic interpretation of ECG signals require properly labeled datasets. However, manual annotation is arduous and time-consuming, particularly in long-term recordings. We propose a fast semiautomatic algorithm for facilitating long-term ECG analysis and annotation, free from fatigue-caused errors or exorbitantly long computations. Heartbeats are compressed to short strings using Symbolic Aggregation approXimation (SAX). Every unique string becomes a pre-cluster of all the beats represented by this string, dramatically reducing the amount of memory and computations required. Next, hierarchical clustering of signal-averaged beats of every pre-cluster is performed. Clusters can then be conveniently analyzed and annotated by the researcher. Our algorithm shows a precision and Fowlkes-Mallows index of 0.91 for both metrics on the Physionet's MIT-BIH database, and runs, on average, in less than one hour on a modern PC for three-day-long ECGs. The proposed algorithm is accurate and efficient enough to facilitate practical long-term ECG signal beat analysis and annotation.

1. Introduction

With wearable health devices gaining traction nowadays, large amounts of data can be collected in clinical and consumer settings, creating valuable cardiovascular research opportunities. Dealing with big data is, however, challenging. Machine learning algorithms expedite analysis and research but require large annotated datasets and, in many cases, human supervision to develop and train such algorithms. The annotation process becomes arduous and time-consuming in long-term signals, and important details may be missed amidst monotonous signals. Accordingly, a system that balances the benefits of automated algorithms with the attention of a human expert, in which the data is structured to be efficiently analyzed and annotated, would accelerate the cardiovascular research pipeline.

Various algorithms have been proposed for heartbeat analysis and clustering [1, 2]. The authors of [2] selected representative heartbeats from each generated cluster for labeling, decreasing the time needed for annotation. While thriving in relatively short ECG signals (up to 30 min long in [1] and 24 h long in [2]), the existing heartbeat clustering algorithms analyze every individual beat in the signal, consuming many computational resources, thus making them impractical for long-term recordings.

In this work, we propose a fast algorithm for facilitating heartbeat analysis and annotation in long-term ECG signals, free from fatigue-caused errors or exorbitantly long computations. A discretization technique [3] is employed to compress heartbeats to short strings. Beats represented by equal strings are grouped into the same pre-cluster, lessening the computational demands. Instead of every individual heartbeat, the human expert is presented with the hierarchical clustering results of the generated pre-clusters for manual investigation and annotation.

2. Methods

Figure 1 illustrates our approach for annotating heartbeats in long-term ECG signals.

2.1. Preprocessing

The ECGs initially undergo preprocessing, comprised of filtering, detrending, and heartbeat delineation.

Filtering. High-frequency noise and baseline wandering are removed using low- and high-pass FIR filters, with respective cut-off frequencies of 40 Hz and 0.6 Hz.

Detrending. ECGs are detrended to remove any trace of baseline wandering. Periods contaminated with large amplitude deviations, such as motion artifacts, are discarded. ECGs are normalized in amplitude (z-score).

Heartbeat delineation. Beats are roughly delineated based on their respective RR-interval (Figure 2) since our

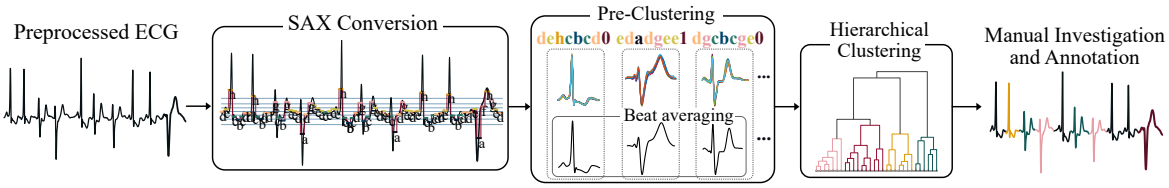

Figure 1. An overview of our approach for annotating heartbeats in long-term ECG signals.

clustering algorithm does not entail precise and computationally expensive PQRST detectors. Good-quality beats are aligned using R-peak as a reference point and zero-padded to equalize their length.

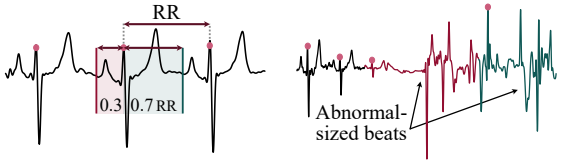

Figure 2. RR-based beat delineation (left): a beat is outlied at 30% and 70% of its RR interval. Abnormal-length beats are categorized as noise (right).

2.2. SAX Conversion

Heartbeats are individually compressed to short strings using Symbolic Aggregation approXimation (SAX) [3]. SAX converts a time series into a set of equiprobable symbols (strings) that approximate the original time series. SAX embodies two parts: discretization via Piecewise Aggregate Approximation (PAA) [3], followed by symbol assignment (Figure 3). PAA reduces the dimensionality of a time series of length n by splitting it into w segments of equal size z. Each i-th segment is represented by the mean value of the data contained within the segment.

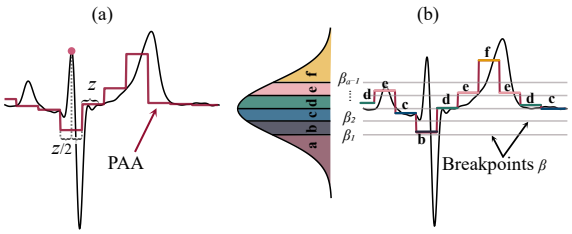

Figure 3. SAX Conversion: (a) PAA transformation of a heartbeat. PAA coefficients are calculated starting at $\frac{z}{2}$ samples centered at the R-peak; (b) Symbol assignment. PAA coefficients are mapped out according to the breakpoints β as follows. The symbol **a** is ascribed to coefficients $< \beta_1$; **b** to $\geq \beta_1$ but $< \beta_2$, and so forth.

The PAA-transformed time series is divided into a equiprobable symbols, where a is an arbitrary alphabet size of $2 \leq a \leq 20$ [3]. Since normalized time series have a Gaussian distribution [3], the Gaussian curve can be split into equal-size regions, $\boldsymbol{\beta} = \beta_1, \ldots, \beta_{a-1}$, such that the area under the curve between β_i and β_{i+1} is $\frac{1}{a}$. $\boldsymbol{\beta}$ are the breakpoints for assigning the symbols and are provided by a statistical table.

2.3. Pre-clustering

Every unique string becomes a pre-cluster of all the beats represented by this string. The number of pre-clusters depends on the chosen a and z parameters. In this work, we chose $a = 4$ and $z = 0.1\,\text{s}$. Incrementing either a or z increases the number of pre-clusters and, thus, computational demands. Beats within the same pre-cluster are signal-averaged. The morphology of certain abnormal beats, such as atrial premature beats (APBs), resembles normal beats, except in low-amplitude components (e.g., P-waves), which SAX can envelop. Enveloping such components causes abnormal beats to be represented by matching strings as beats of a different class (Figure 4), and thereby wrongly pre-clustered. To minimize pre-clustering errors, for every i-th beat, we add an extra symbol representing the RR information of two neighboring beats: $RR_i = RR_{(i,i-1)}/RR_{(i+1,i)}$.

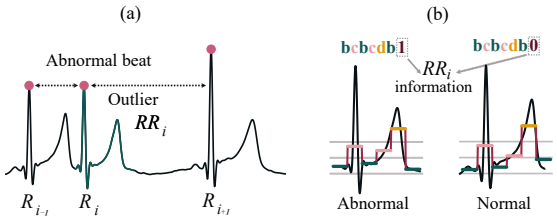

Figure 4. (a) RR_i of an abnormal beat (APB). (b) Adding of RR_i information to distinguish strings of APB from normal beats. The symbol '**1**' is added to beats whose RR_i is an outlier and '**0**' otherwise.

2.4. Hierarchical Clustering

Hierarchical clustering is employed to cluster the average heartbeats of every pre-cluster. The dissimilarity matrix D_m is calculated by combining dissimilarity matrices

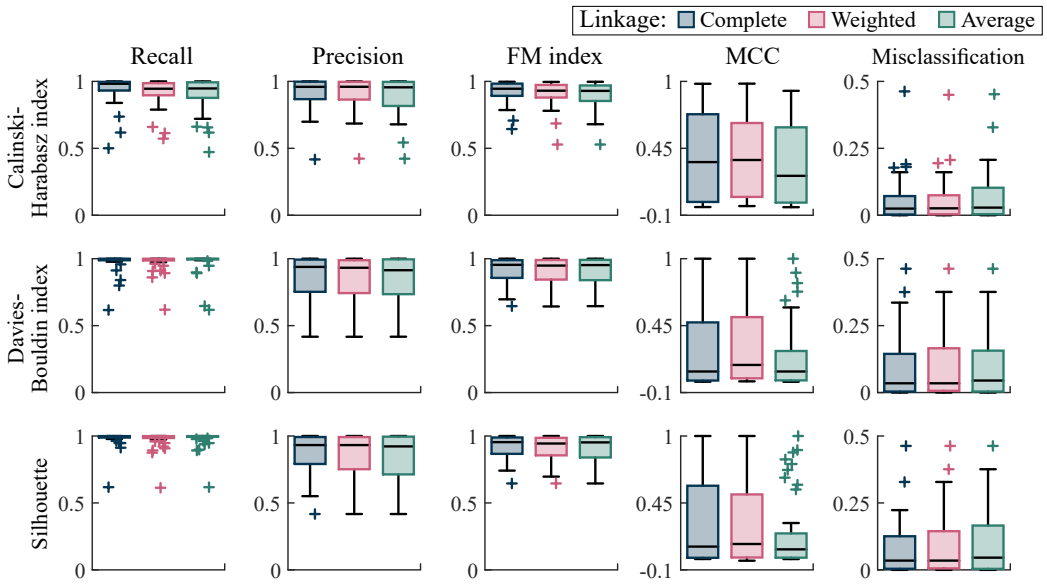

Figure 5. Boxplots of the performance metrics obtained by our algorithm in the MIT-BIH database using various linkage and internal clustering validation indexes.

of various heartbeat components:

$$D_m = D'_{PQ} + D'_{QRS} + D_b^{0.5}. \quad (1)$$

D'_{PQ} and D'_{QRS} measure the *Spearman* distance between the gradient of PQ-interval and QRS-complex, and D_b is the *complexity-invariant* distance [4] of the whole heartbeat. To choose the optimal number of clusters, we assess the performance of three internal validation indexes, Calinski-Harabasz, Davies-Bouldin, and silhouette [5], together with three types of linkage: complete, weighted, and average.

2.5. Data and Performance Evaluation

We test the accuracy and efficiency of our algorithm on two separate datasets. Accuracy is tested on the Physionet MIT-BIH database with 13 manually labeled classes. To test efficiency, we apply our algorithm to annotate our long-term ECG recordings (two- to three-days-long).

We evaluate accuracy through recall, precision, Fowlkes-Mallows (FM) index, Matthews Correlation Coefficient (MCC), and misclassification (beats that ended up in clusters where the dominant class is another). True Positives (TP) are the # of pairs of beats with the same class and cluster. True Negatives (TN) are the # of pairs of beats with different classes and clusters. False Positives (FP) are the # pairs of beats with different classes but the same cluster. False Negatives (FN) are the # of pairs of beats with the same class but different clusters. Except for MCC, all metrics range in [0,1]. MCC ranges in [-1,1], where MCC = 1 indicates a perfect prediction, MCC = 0 equals to a random prediction, and MCC = -1 corresponds to a total disagreement between the prediction and the labels.

3. Results

The compression rate for the MIT-BIH database was 26.2%, yielding on average 556.4±318.9 unique strings per recording. Recordings have an average of 2 258.6±440 heartbeats. Beat misclassification was 0.97±1.51% after SAX, with junctional and escape beats being the most susceptible to being improperly pre-clustered.

Figure 5 shows boxplots of the performance evaluation metrics on the MIT-BIH database. The results reveal a large variability of MCC for all combinations of linkage and internal validation indexes, despite high precision and FM index, with a few MCC values reaching near or below 0, likely caused by an increase in FNs (Figure 6). The combination of complete or weighted linkage with the Calinski-Harabasz index produces the best precision, FM index, MCC, and misclassification results, with a respective mean of 0.91, 0.91, 0.37, and 5.6%.

As for efficiency, in one 3-days-long recording, more than 480 000 beats were reduced to less than 30 000 unique strings and further to 30 clusters, which were then annotated as three classes plus noise. The computations took less than an hour on an AMD Ryzen™5 3.6 GHz CPU with six cores (12-threads) and 16 GB of RAM.

4. Discussion

This work proposes a fast algorithm for facilitating heartbeat analysis and annotation in long-term ECG sig-

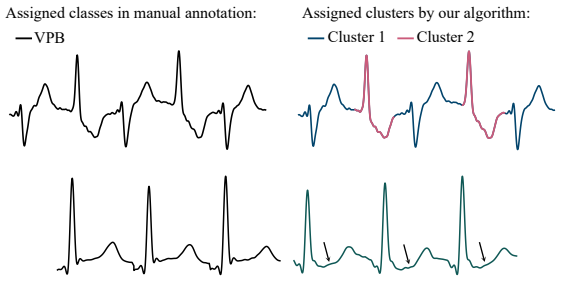

Figure 6. Examples of clustering results. Five beats labeled with the same class of ventricular premature beats (VPB) from the MIT-BIH database are assigned to two different clusters by our algorithm (top). This is a case of FNs that is not fundamentally undesirable for manual annotation. Normal beats from two distinct time periods with subtle ST-changes from our long-term ECG recordings (bottom).

nals. Heartbeats are compressed to short strings using SAX and pre-clustered before performing hierarchical clustering of averaged beats of every pre-cluster. A human expert then annotates the generated clusters.

A similar annotation strategy was adopted in [2], in which only selective heartbeats representative of each cluster are labeled. Albeit precise, the clustering algorithm is computationally demanding as it examines all beats within the recording. In contrast, our algorithm pre-clusters equal-string beats, performing hierarchical clustering only to signal-averaged beats of every pre-cluster. Moreover, signal-averaging may also boost clustering results. It enables the inclusion of various low-amplitude components, as P-waves and ST-segments, into the dissimilarity matrix that would otherwise be occluded by noise.

Our algorithm shows promising results. While low MCC values tend to suggest poor clustering results, we verified that this is not always the case, particularly if paired with high precision values. MCC decreases when the number of FNs or FPs increases. An increase in FNs, i.e., same-class beats categorized into different clusters, is not fundamentally undesirable for manual annotation. Our algorithm categorized beats labeled with the same class into various clusters if differences in beat morphology exist (see Figure 6). In ambulatory cardiovascular research, even modest changes in the ST-T complex can hold meaningful value. For instance, ST-T morphology variations are pivotal for investigating noninvasive markers of electrolyte fluctuations in hemodialysis patients in out-of-hospital settings. Another scenario of low MCC is ECGs with minimal abnormal beats (e.g., five APBs occur amidst 2 000 normal beats). Misclassifying one or two beats in such recordings increases FPs but is unlikely to compromise cardiovascular research.

Albeit beneficial for annotating long-term ECG recordings, pre-clustering using SAX can increase the incidence of FPs by enveloping beats of different classes into the same pre-cluster. Misclassification is higher in beats whose distinctive morphological feature lies in subtle low-amplitude components, such as APBs, escape, or junctional beats. Higher a and z values could ameliorate beat misclassification by allowing such components to be discernible by SAX but at the expense of higher computational cost without necessarily producing better clustering results. Finding the best compromise between a and z parameters, computational demands, and accuracy is a subject of our future research.

5. Conclusions

The proposed algorithm facilitates practical long-term ECG signal beat analysis and annotation. Furthermore, it allows researchers to explore how the heartbeats fall into various classes naturally and even study the existence of unexpected sub-classes.

Acknowledgments

This work has received funding from the European Regional Development Fund (No. 01.2.2-LMT-K-718-01-0030) under grant agreement with the Research Council of Lithuania (LMTLT).

References

[1] Lagerholm M, Peterson C, Braccini G, Edenbrandt L, Sornmo L. Clustering ECG complexes using hermite functions and self-organizing maps. IEEE Transactions on Biomedical Engineering 2000;47(7):838–848.

[2] Kiranyaz S, Ince T, Pulkkinen J, Gabbouj M. Personalized long-term ECG classification: A systematic approach. Expert Systems with Applications 2011;38(4):3220–3226.

[3] Lin J, Keogh E, Wei L, Lonardi S. Experiencing SAX: a novel symbolic representation of time series. Data Mining and Knowledge Discovery apr 2007;15(2):107–144.

[4] Batista GEAPA, Keogh EJ, Tataw OM, de Souza VMA. CID: an efficient complexity-invariant distance for time series. Data Mining and Knowledge Discovery 2013;28(3):634–669.

[5] Liu Y, Li Z, Xiong H, Gao X, Wu J. Understanding of internal clustering validation measures. In 2010 IEEE International Conference on Data Mining. IEEE, 2010; .

Address for correspondence:

Ana Santos Rodrigues
Biomedical Engineering Institute, KTU
K. Baršausko g. 59-A453, LT-51423 Kaunas, Lithuania
ana.rodrigues@ktu.lt