Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1353–1356
DOI: 10.15439/2017F303
ISSN 2300-5963 ACSIS, Vol. 11

The North Sea Bicycle Race ECG Project: Time-Domain Analysis

Dominika Długosz, Aleksandra Królak
Łódź University of Technology,
Institute of Electronics
ul. Wólczańska 211/215, 90-924 Łódź, Poland,
Email: 195887@edu.p.lodz.pl,
aleksandra.krolak@p.lodz.pl

Trygve Christian Eftestøl, Stein Ørn, Tomasz Wiktorski
University of Stavanger,
Faculty of Science and Technology,
Department of Electrical and Computer Engineering,
4036 Stavanger, Norway,
Email: trygve.eftestol, stein.orn, tomasz.wiktorski@uis.no

***Abstract*—Analysis of electrocardiogram and heart rate provides useful information about health condition of a patient. The North Sea Bicycle Race is an annual competition in Norway. Examination of ECG recordings collected from participants of this race may allow defining and evaluating the relationship between physical endurance exercises and heart electrophysiology. Parameters reflecting potentially alarming deviations in the latter are to be identified in this study. This paper presents results of a time-domain analysis of ECG data collected in 2014, implementing K-Means clustering. A double stage analysis strategy, aimed at producing hierarchical clusters, is proposed. The first phase allows rough separation of data. Second stage reveals internal structure of the majority clusters. In both steps, discrepancies driving the separation could stem from three sources. The clusters were defined predominantly by combinations of features: heartbeat signals correlation, P-wave shape, and RR intervals; none of the features alone was discriminative for all the clusters.**

I. Introduction

THE North Sea Bicycle Race (Nordsjørittet) is an international competition organized annually in Rogaland, western Norway, between cities: Egersund and Sandness. It is open to a wide spectrum of competitors, from amateurs to professionals. In 2014, ECG data were collected from over a thousand participants on three days: the day of the race, the day before and after, as part of the North Sea Race Endurance Study (NEEDED). Continuation of this project with extended set of recorded data is planned for years 2017-2019.

Analysis of electrocardiogram (ECG) is a valuable tool in monitoring and diagnosis of patients for various cardiac conditions. The procedure of automatic ECG signal analysis can be performed in time or frequency domain and is usually divided into two steps: feature extraction and classifier designation [1]. There are various methods for feature extraction discussed in the literature. The aspects of Principal Component Analysis (PCA) related to ECG signal processing are discussed in [2], application of customized wavelet transform (WT) in ECG discriminant analysis is described in [3], while the use of Hilbert transform for feature extraction from ECG signal was examined in [4]. Comparison of support vector machine (SVM) algorithm and artificial neural network approach (ANN) for classification of arrhythmias in ECG signal is presented in [5]. Deep learning method for active classification of electrocardiogram signals was applied in the research described in [6], while the clustering method for QRS complexes classification was applied in [7].

Measurement of ECG and heart rate (HR) during daily activity is a potential tool for early diagnosis of cardiac diseases and may provide individualized guidance to exercise and physical training. This project aims at identifying ECG and HR parameters useful for differentiating normal and abnormal patterns during prolonged, high intensity endurance exercise.

II. The Dataset and Software

The database consisted of 3158 ten-second ECG recordings. After rejecting participants for whom some of the recordings were missing, 996 complete sets of 3 recordings were obtained. The collection was further reduced by cases of erroneous ECG segmentation. Further analysis was conducted for 989 participants (2967 ECG recordings). The data were processed and analyzed using Python programming language with packages: `BioSPPy`, `SciPy`, and `scikit-learn`.

III. Data Pre-processing and Feature Extraction

The dataset provided 8-channel ECG recordings, containing signals from leads I, II, and six precordial leads. In this project, only lead-I signal was analyzed. The channel of interest was extracted and subjected to pre-processing and measurements. The procedure aimed at visualization of changes in the ECG signal over the three days and extraction of features relevant for comparison of data obtained from different participants.

A. Data pre-processing

The lead-I ECG signal was subjected to filtering to suppress high-frequency noise and remove baseline drift using bandpass Finite Impulse Response (FIR) filter with cutoff frequencies of 3 and 45 Hz. Next, locations of R-peaks were detected applying Engelse-Zeelenberg approach modified by Lourenco et al. [8]. For singular cases in which this method failed to reliably identify the peaks (less than 3 peaks found in a 10-s recording), the detection was repeated with Christov method [9]. The identified R-peaks were used as reference during extraction of heartbeat templates, defined in a time

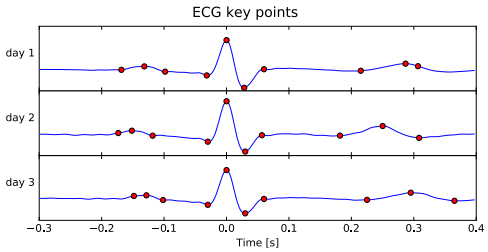

Fig. 1. ECG key points detection - exemplary results.

window of 0.3 s. before and 0.4 s. after the spike. Algorithms implemented in the `BioSPPy` package were used.

Finally, heartbeat templates extracted from a single recording were averaged to improve signal-to-noise ratio [10]. Additionally, parameters referring to the heart rate (mean duration and standard deviation of RR intervals) were derived.

B. Heartbeat templates measurements

In order to measure the ECG waveforms, methods for searching key points (peaks of P, Q, R, S, and T waves, as well as onsets and endpoints of some of them) in the heartbeat templates were developed. The location of R-peak in the signal was fixed at time 0.0 s (see Fig. 1). P wave top was defined as a maximum before the occurrence of R, excluding 0.05 s directly preceding the latter. Mirror-reflected procedure was applied for determining the top of the T wave. The Q and S points were found as local minima within a fixed, short time window before and after the R. The S wave endpoint was defined as a point where the positive slope after S is <90% of its value at S. Onsets and endpoints of P and T waves were found following the idea described by Laguna et al. [11]. Exemplary results of the ECG key point search are presented in Fig. 1. Each subplot presents an averaged heartbeat template for the respective day of measurements for the same participant. The points were used to measure intervals and amplitudes of ECG signals. For estimation of amplitudes, the level of Q was regarded as the baseline. ST elevation was defined as difference in amplitude between the endpoint of the S wave and the onset of the T wave.

C. Morphological comparison of heartbeat templates

A set of parameters was derived from comparison of morphology of the extracted heartbeats, either a full set of beats from one signal or a set of 3 averaged beats from the 3 days for a given participant. To exclude correlation changes stemming from heart rate variation between the days, processing was done on QRS complexes, whose shape did not exhibit any heart-rate dependency. A basic measure to compare the heartbeats is Pearson r coefficient. Its value was computed for every pair of heartbeats within the analyzed set. To ensure that exclusively the shape of the beats is compared, with no influence of residual baseline drift, the coefficient was calculated using first differences of the signals.

Another aspect in beat contour analysis is the idea of morphological classification [12], [13]. QRS complexes from the 1st day were iteratively compared using Pearson r coefficient. Similar peaks were grouped into a class; if similarity threshold was exceeded, a new class was created. Beats within each class were averaged to serve as templates for comparison with signals from the 2nd and 3rd day. Beats from days 2 and 3 were assigned to this of the 1st day classes to which they were the most similar. In case Pearson coefficient for a beat and each of the classes' templates was below the threshold, the beat was considered an outlier.

D. Features definition

Ten ECG features were derived from the measurements using the above described approaches:

- Shape coefficient of P wave - ratio of height of the wave to its width; the used features expressed change in this value from day 1 to day 2 or 3 (`P_shape_12` and `P_shape_13` respectively).
- Difference in duration of QT interval on day 2 or 3 with respect to day 1 (`QT_12` and `QT_13` respectively).
- Difference in duration of RR interval on day two or three with respect to day 1 (`RR_12` and `RR_13` respectively).
- Change (difference) in mean correlation of heartbeat templates from the 2nd or 3rd recording with respect to correlation in the 1st day (`correlation_12` and `correlation_13` respectively).
- Maximal ST elevation (`max_ST_elev`) - maximum from values measured on the three days. ST elevation itself, not necessarily its change from day to day, should be regarded as an alarming ECG feature. [14]
- Percentage of morphological outliers - percentage of beats from days 2 and 3 not matching any beat class defined in day 1 for the given participant (expressed with relation to total number of beats from the three days), as defined in the previous section (`morph_outliers`).

Features based on differences between days are defined by subtracting value on day 2 or 3 from value on day 1. Positive values of these features indicate a decrease with respect to day 1 (shorter intervals or decline in correlation).

IV. Feature Set Analysis

Analysis of the derived set of features was performed by unsupervised clustering. Clustering on the entire dataset tends to yield one or more larger clusters and a few 'far outliers' groups, containing points significantly separated from the majority. Therefore a two-stage procedure was developed: after first-attempt analysis and clustering, the outliers' clusters (containing <10% of the total number of observations) were removed and the analysis was repeated to reveal structure of the majority clusters. Each of the two stages consisted of two main elements: principal component analysis (PCA) and K-means clustering combined with silhouette analysis.

A. Principal Component Analysis

PCA is a statistical operation aimed at reduction of dimensionality of the clustering data [15], frequently applied prior to K-means clustering. It allows reducing computational effort by

decreasing number of dimensions to be analyzed and suppressing possible correlation between the original features [16]. PCA was applied after data normalization on both stages of the analysis. Six principal components, explaining 80% of the data variance, were retained. The data mapped on the PC space was passed to clustering and silhouette analysis.

B. Clustering with Silhouette Analysis

Since no prior assumptions on the structure of the data were made, and the K-means clustering requires specified number of clusters as an input, silhouette analysis was launched on the dataset. It allows validating consistency of computed clusters by comparing cohesion of each sample and its separation from other clusters. The resulting silhouette score is a fraction between -1 and 1, where 1 represents good sample classification and -1 indicates that the sample might have been assigned to an improper cluster. Average silhouette score of all samples allows assessing general consistency and validity of the clustering [17]. Silhouette analysis on the PC-transformed data was performed for number of clusters ranging from 2 to 7 to choose the one with highest average score. K-means clustering with the chosen number of clusters was applied to the dataset mapped to the reduced PC space. The result was presented and analyzed graphically in the PC and the original feature space for both processing stages.

V. Results and Discussion

The results of clustering in the original feature space and PC space are presented in scatter plot of observations in two dimensions of the feature space (Fig. 2-5). The results of PCA are shown as bar plots of components' eigenvectors (Fig. 6).

In the first stage of the analysis, the majority (over 90%) of observations were assigned to cluster 1, while the other two clusters are smaller (Fig. 2 and 3). Cluster 2 is separated from the other two with respect to the PCs 4 and 5 (Fig. 2a), defined mainly by percentage of morphological outliers and ST elevation (Fig. 6a). As confirmed by Fig. 3a, this cluster is composed of the observations with high values for morphological outliers percentage, while in most cases the values are close to 0. Cluster 0 in PC 4&5 projection is overlapped partially with clusters 1 and 2. However, it is clearly separated when observed from PCs 1 and 3, exhibiting high dependence on beat correlation (Fig. 3b). Analysis of respective projection of the data reveals majority of the points being concentrated around the (0,0) point, indicating little change in intra-recording beat correlation between the days. For some participants - assigned to cluster 0 - correlation on both the 2nd and 3rd day was considerably high when compared to day 1. This is typically not accompanied by increased percentage of morphological outliers since the latter uses day 1 as a reference. Since the clusters 0 and 2 encompassed minor portion of the observations (1.2% and 4.8% respectively), they were excluded from further analysis. The second stage of the procedure was conducted on the points originally assigned to cluster 1. In second phase results, the three major clusters (0, 1, and 4) can be discriminated by looking i.a. at principal

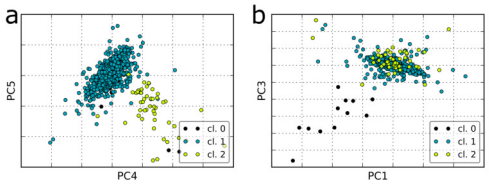

Fig. 2. Result of clustering on the full dataset, in the PC space - projection on: (a) PCs 4 and 5; (b) PCs 1 and 3. (normalized).

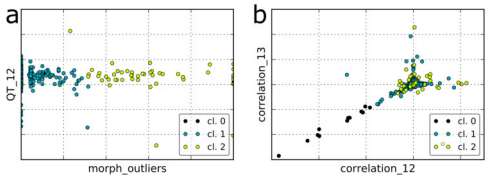

Fig. 3. Result of clustering on the full dataset, in the original feature space - projection on: (a) QT interval difference (days 1 and 2) and percentage of morphological outliers; (b) correlation difference-related features.

components 1 and 4 (Fig. 4a), dependent on maximal ST elevation and features related to QT and RR interval (Fig. 5a). Clusters 2, 3, and 6, can be distinguished by projection onto PCs 2 and 3, defined predominantly by features associated with shape of the P wave, correlation, and morphological outliers percentage (Fig. 6 and Fig. 4b). Statistical significance of the latter was slightly lower than in the first stage of the analysis (considering its contribution to the first two PCs); however, it is still one of main components differentiating cluster 2 from others (as shown in Fig. 5b). This is particularly interesting when compared to correlation representation of the clustering result (Fig. 5c). Cluster 2 is constituted by points for which decreased correlation was indeed observed, but predominantly either on day 2 or 3; rarely on both days. On the other hand, closer look at the P shape allows discriminating cluster 3 (Fig. 5e). For participants belonging to this cluster, P wave was flattened (lower height-to-width ratio) in days 2 and 3 with respect to day 1. The change in shape was more prominent than observed in the other groups. Finally, cluster 5 is distinctly separated with respect to PCs 5 and 6 (Fig. 4c). It was not reflected in any of the first components due to relatively small size of this cluster (ca. 0.5% of all observations), which diminishes its impact on the total variance of the dataset. Original features that contribute the most to this component include those related

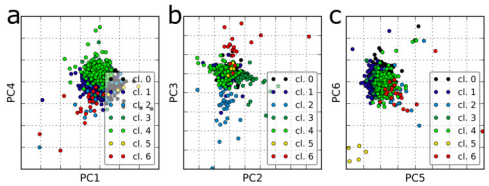

Fig. 4. Result of clustering on the restricted dataset, in the principal component space - projection on: (a) PCs 1 and 4; (b) PCs 2 and 3; (c) PCs 5 and 6.

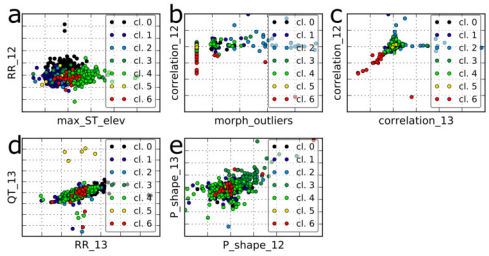

Fig. 5. Result of clustering on the restricted dataset, in the original feature space - projection on: (a) RR interval difference (days 1 and 2) and maximal ST elevation; (b) correlation difference between days 1 and 2 and percentage of morphological outliers; (c) the correlation difference-related features; (d) differences in QT and RR intervals between days 1 and 3; (e) the P-shape-related features.

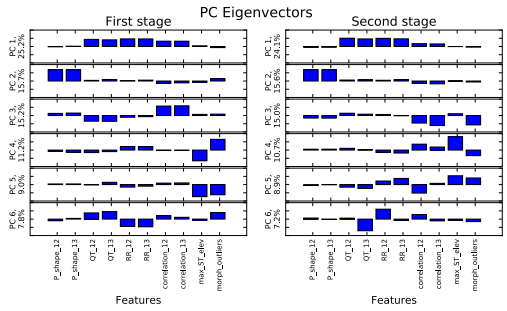

Fig. 6. PCA results: eigenvectors and explained variance portions of the six components; results of the first (left) and second (right) stage of the analysis.

with QT and RR intervals. As presented in Fig. 5d, decrease in duration of QT interval is in general correlated with increase in RR interval. For cluster 5, however, this trend does not apply. Further detailed investigation is needed to determine whether the phenomenon is a question of improper key point localization or a sign of potential cardiac issue.

VI. Conclusion

The NEEDED study focuses on characterization patterns associated with a prolonged endurance exercise. One of its major goals is identification of parameters related to ECG and heart rate which could be used to distinguish between regular and deviated heart performance. Several potentially discriminative features were recognized. Further investigation and validation with additional data is needed to verify their relevance in detection of electrocardiophysiological abnormalities.

The study was conducted using 2 stage clustering and principal component analysis. It was found that no single feature or principal component would provide separation between all the clusters globally. Each cluster could be described by a combination of two to four features making it distinguishable. Determination of features defining the partition was facilitated by analysis of eigenvectors of the principal components. However, PCA is only based on variance of the dataset, which is not equivalent to separation between clusters. Discriminative features are always reflected in high values in PCs' eigenvectors, but the reverse is not always true.

The presented method produces hierarchical structure of clusters. This allows 2-level investigation of the data structure, with separate insight in huge discrepancies and more subtle trends. Furthermore, the hierarchy is also followed in analysis of statistical significance of the features. Combined with additional data, it could be used in differentiation between natural groups among the population and allow early detection of certain cardiac abnormalities.

Future works include a fusion of time-domain and frequency-domain analysis of the ECG data. The new dataset will be supplemented with information of i.a. patients' age, gender, the race completion time, and possibly indication of medical condition. This will allow to verify the significance of the ECG features derived and investigated in this paper.

References

[1] X. Dong, C. Wang, and W. Si, "ECG beat classification via deterministic learning," *Neurocomputing*, vol. 240, pp. 1–12, May 2017.

[2] F. Castells, P. Laguna, L. Sornmo, A. Bollmann, and J. Roig, "Principal component analysis in ECG signal processing," *EURASIP JOURNAL ON ADVANCES IN SIGNAL PROCESSING*, 2007.

[3] A. Daamouche, L. Hamami, N. Alajlan, and F. Melgani, "A wavelet optimization approach for ECG signal classification," *Biomedical Signal Processing and Control*, vol. 7, pp. 342–349, July 2012.

[4] D. Benitez, P. Gaydecki, A. Zaidi, and A. Fitzpatrick, "The use of the Hilbert transform in ECG signal analysis," *Computers in Biology and Medicine*, vol. 31, no. 5, pp. 399–406, 2001. 399.

[5] M. Moavenian and H. Khorrami, "A qualitative comparison of Artificial Neural Networks and Support Vector Machines in ECG arrhythmias classification," *EXPERT SYSTEMS WITH APPLICATIONS*, vol. 37, pp. 3088–3093, Apr. 2010.

[6] M. A. Rahhal, Y. Bazi, H. AlHichri, N. Alajlan, F. Melgani, and R. Yager, "Deep learning approach for active classification of electrocardiogram signals," *Information Sciences*, vol. 345, pp. 340–354, June 2016.

[7] M. Lagerholm, C. Peterson, G. Braccini, L. Edenbrandt, and L. Sornmo, "Clustering ECG complexes using Hermite functions and self-organizing maps," *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING*, vol. 47, pp. 838–848, July 2000.

[8] A. Lourenco, H. Silva, P. Leite, R. Lourenco, and A. Fred, "Real Time Electrocardiogram Segmentation for Finger based ECG Biometrics (PDF) - Semantic Scholar."

[9] I. I. Christov, "Real time electrocardiogram QRS detection using combined adaptive threshold," *BioMedical Engineering OnLine*, vol. 3, p. 28, 2004.

[10] A. Gautam, Y. D. Lee, and W. Y. Chung, "ECG Signal De-noising with Signal Averaging and Filtering Algorithm," in *2008 Third International Conference on Convergence and Hybrid Information Technology*, vol. 1, pp. 409–415, Nov. 2008.

[11] P. Laguna, R. Jane, and P. Caminal, "Automatic detection of wave boundaries in multilead ECG signals: Validation with the CSE database," *Computers and biomedical research*, vol. 27, no. 1, pp. 45–60, 1994.

[12] P. W. Macfarlane, B. Devine, and E. Clark, "The university of Glasgow (Uni-G) ECG analysis program," in *Computers in Cardiology, 2005*, (Lyon), pp. 451–454, Sept. 2005.

[13] "Glasgow 12-lead Analysis Program - Physician's Guide."

[14] K. Wang, R. W. Asinger, and H. J. Marriott, "ST-segment elevation in conditions other than acute myocardial infarction," *New England Journal of Medicine*, vol. 349, no. 22, pp. 2128–2135, 2003.

[15] U. Demsar, P. Harris, C. Brunsdon, A. S. Fotheringham, and S. McLoone, "Principal Component Analysis on Spatial Data: An Overview," *Annals of the Association of American Geographers*, vol. 103, pp. 106–128, Jan. 2013.

[16] B. Hariharan, J. Malik, and D. Ramanan, "Discriminative Decorrelation for Clustering and Classification," in *Computer Vision - ECCV 2012*, pp. 459–472, Springer, Berlin, Heidelberg, Oct. 2012.

[17] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.