Irregularity Detection in ECG signal using a semi-fiducial method

João Carvalho Student
joao.carvalho@ua.pt
Armando J. Pinho Prof
ap@ua.pt
Susana Brás Collaborator
susana.bras@ua.pt

IEETA
University of Aveiro
Aveiro, Portugal Portugal

Abstract

Irregularity detection is important for several ECG applications, both in the field of health or even other subjects like biometric identification. In this paper, we have proposed a method based in the *Hamming Distance* for finding irregularities on an ECG signal, that outperforms an alternative proposed in a paper by Keogh *et al*. in [6], at least for this specific dataset. We start by explaining the preprocessing steps required to work with ECG signals using non-fiducial methods and make a comparison of the results obtained by both methods.

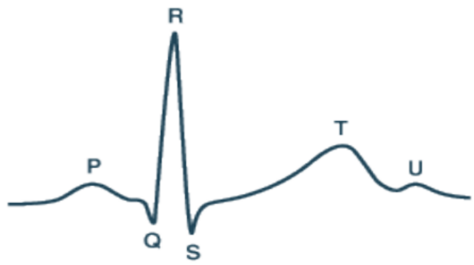

Figure 1: P, Q, R, S, T, U peaks of an ideal ECG heartbeat (from [2]).

1 Introduction

It is supposed that, at rest, the ECG signal from a complete cardiac cycle is similar to the previous and to the next cycle. However, due to external or internal interferences, this may not be true [1, 3, 8].

Developing an algorithm to identify where those interferences occur may be of great interest for biometry identification, as well as other applications, as it may allow to incorporate that algorithm into a decision support system, where parts of the signal that contain irregularities may be treated differently than regular signal.

2 Overview

In [6], the authors describe a method for finding noise using compression tools. They do this by computing the dissimilarity distance of a given segment of an ECG against the whole ECG. The *Compression-based Dissimilarity Measure* (CDM) between two string x and y [6] is defined as:

$$\mathrm{CDM}(x,y) = \frac{C(xy)}{C(x)+C(y)}, \tag{1}$$

where xy represents the strings x and y concatenated.

Two important facts about the CDM are: **(a)** it is close to one when x and y are not related; **(b)** if x and y are related, as strongly related they are, the lower the $\mathrm{CDM}(x,y)$ is, but it never reaches zero.

In short, what their method does is to “simply measure how well a small local section can match the global sequence" [6]. It is easy to see how this can be useful to find irregularities in the signal (assuming they are present only in some small portions of the signal).

Since the NRC has been shown to work very well on ECG signals [3] and it also respects both **(a)** and **(b)**, we did an implementation using it as a replacement for the CDM defined by Keogh *et al*. Finally, we compared that approach with our proposed method, based on the *Hamming Distance*.

2.1 Database

Despite the fact the proposed algorithm aims at the detection of ECG noise in real signals, we had to manage a control set of ECG signals to test and validate the algorithm. These were already acquired on a previous study.

The database is composed by 67 hours of ECG signal from which 19 hours (7 signals with 2h40 each, approximately) correspond to a clean ECG signal collected using a simulator with artificial/synthetic noise added. To gather the values from the simulator, we used VitalJacket [4], which provides an ECG signal with a sampling frequency of 500 Hz. The collecting protocol consisted in gradually increasing and decreasing the heart rate with 20 minutes interval. In total, **the signal is composed by 8 heart rate steps in the following order: 60, 80, 100, 120, 140, 120, 100, 80 beats per minute**. The noise values were randomly generated using a uniform discrete distribution between the minimum and maximum values (117 and 159, respectively) of the collected signal. This allowed us to have a controlled signal with noise in specific zones, to test the algorithm's behavior in different situations.

To each signal was then added some noise in specific zones, as we will see in some examples.

3 ECG Signal Preprocessing

The ECG signal is not periodic, but it is highly repetitive [1]. There are three major components of a complete heartbeat captured by an ECG signal: the *P* wave, the *QRS* complex and the *T* wave. The points *P*, *Q*, *R*, *S*, *T* and *U* (which is less used) are called *points of interest*, also known as *fiducial points*, of the ECG. An example can be seen in Fig. 1.

3.1 R-peak detection

The development of a robust automatic *R-peak* detector is essential, but it is still a challenging task, due to irregular heart rates, various amplitude levels and shapes of *QRS* morphologies, as well as all kinds of noise and artifacts. [5]

We have decided to use a *partially fiducial* method for segmenting the ECG signal and, since this was not the major focus of the work, we used a preexisting implementation to detect *R-peaks*, based on [5]. This method detects the *R-peak* by calculating the average point between the *Q* and *S* peaks (the *QRS complex*) – this may not give the real local maximum of the *R-peak*, but it produces a very close point. Some validations were done against *R-peaks* detection performed by humans in order to validate this step.

The process used for detecting the *QRS* complexes is somewhat similar to the one described in [5]. It uses some bandpass filtering and differentiation operations used to enhance *QRS* complexes and to reduce out-of-band noise. A nonlinear transformation based on energy thresholding, Shannon energy computation, and smoothing processes is used to obtain a positive-valued feature signal which includes large candidate peaks corresponding to the *QRS* complex regions.

3.2 Quantization

In order to convert the real-valued ECG signal into a symbolic time series, the first step we had to perform was to reduce its dimensionality. This was achieved by using a modification of the Piecewise Aggregate Approximation, PAA, method [7].

Since, in our case, the *R-peaks* of the ECG signal were already detected, we used this to our advantage and, instead of splitting the complete

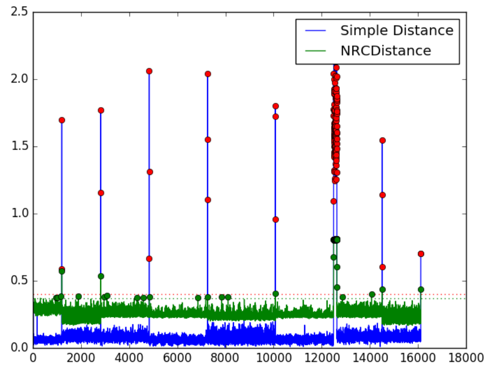

Figure 2: The transition between 120 and 100 beats per minute was replaced by 90 seconds of noise (signal number 5).

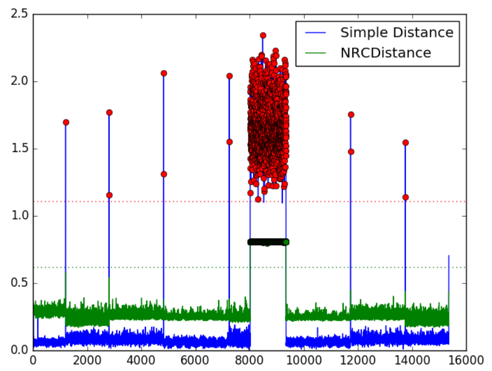

Figure 3: In the 140 heart rate step, 15 minutes of noise were introduced (signal number 3).

signal into w dimensions, as the original method suggests, we applied it only inside of each *R-R-interval* (intervals between consecutive *R-peaks*) individually, with the values inside each of those intervals previously normalized. This means that, for example, if one heartbeat takes more time than another, then the real value obtained by PAA may correspond to more real values on the original signal. On our case, the number of samples, w, chosen for each heartbeat was 200. This means that each heartbeat on the ECG, on this phase, was represented by 200 real values (instead of the original 1000 real values per second).

After completing this process for all the heartbeats of a signal, since the purpose is to have a symbolic representation of the series – not a real valued one, the *Symbolic Aggregate approXimation* (SAX) was applied to each heartbeat PAA series individually.

From this explanation, it is already implicit that one parameter used by this method as input is the *alphabet size* (the number of different symbols allowed as output), that we want to use. From experiments using a different database, we realized a choice of an alphabet size of 6 is appropriate for most of our applications and, therefore, this was the alphabet size we used for this tests as well. Using the process described, each complete heartbeat is outputted as a 200 length string. We refer to a string like that as a *word* or *SAX-word*.

3.3 Proposed Method

The proposed method is very straight forward, which is why we called it the *Simple Distance method* (in fact, it basically consists on the *Hamming Distance* applied to consecutive *SAX-words*).

The idea of our approach is to store all the n *words* (or *SAX-words*) of an ECG on a size n array and compute the $n-1$ distances between those consecutive n words. Since we want to have a sample of size n, an interpolation from size $n-1$ to size n should be performed. After that, some decisions can be made using the values obtained for that metric.

3.4 Results and Future Work

From our experiments, we found that a threshold which produced more consistent results for both metrics was $\bar{x}-2\sigma$ (average value minus two times the standard deviation). We were able to do it by experimentation because we knew the zones where noise was supposed to be found beforehand.

In Fig. 2, it is possible to see that the *Simple Distance* measure detects the areas where random noise was inserted very precisely, while the NRC detects a lot of false positives. In Fig. 3 the *NRC* is not able to detect any noise at all, however, the *Simple Distance* can detect both the zones where the heart beat rate was changed, which, even tough it is not noise, may be considered a point of interest, depending on the application.

Even though the threshold choice worked properly for this dataset, it should not be static and, therefore, some future work should be done in order to adjust it in a dynamic way. We also plan on making further tests using different datasets, as well as possible changes to the method itself, which is still very basic.

4 Acknowledgments

This work was partially supported by national funds through the FCT - Foundation for Science and Technology, and by european funds through FEDER, under the COMPETE 2020 and Portugal 2020 programs, in the context of the projects UID/CEC/00127/2013 and PTDC/EEI-SII/6608/2014.

References

[1] Foteini Agrafioti and Dimitrios Hatzinakos. ECG biometric analysis in cardiac irregularity conditions. *Signal, Image Video Process.*, 3(4):329–343, sep 2008. ISSN 1863-1703. doi: 10.1007/s11760-008-0073-4. URL `http://link.springer.com/10.1007/s11760-008-0073-4`.

[2] M Bassiouni and W Khalefa. A study on the Intelligent Techniques of the ECG-based Biometric Systems. *Recent Adv. Electr. Eng.*, 2015. URL `http://www.inase.org/library/2015/crete/COCI.pdf{\#}page=26`.

[3] Susana Brás and Armando J Pinho. ECG biometric identification: A compression based approach. In *2015 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, pages 5838–5841, aug 2015. doi: 10.1109/EMBC.2015.7319719. URL `http://www.ncbi.nlm.nih.gov/pubmed/26737619`.

[4] João P. Silva Cunha, Bernardo Cunha, António Sousa Pereira, William Xavier, Nuno Ferreira, and Luis Meireles. Vital-Jacket: A wearable wireless vital signs monitor for patients' mobility in cardiology and sports. In *2010 4th Int. Conf. Pervasive Comput. Technol. Healthc.*, pages 1–2. IEEE, 2010. doi: 10.4108/ICST.PERVASIVEHEALTH2010.8991.

[5] P. Kathirvel, M. Sabarimalai Manikandan, S. R. M. Prasanna, and K. P. Soman. An Efficient R-peak Detection Based on New Nonlinear Transformation and First-Order Gaussian Differentiator. *Cardiovasc. Eng. Technol.*, 2(4):408–425, oct 2011. ISSN 1869-408X. doi: 10.1007/s13239-011-0065-3. URL `http://link.springer.com/article/10.1007/s13239-011-0065-3/fulltext.html`.

[6] E Keogh, S Lonardi, and CA Ratanamahatana. Towards parameter-free data mining. In *KDD '04 Proc. tenth ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pages 206–215, 2004. URL `http://dl.acm.org/citation.cfm?id=1014077`.

[7] J Lin, E Keogh, S Lonardi, and B Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *DMKD '03 Proc. 8th ACM SIGMOD Work. Res. issues data Min. Knowl. Discov.*, pages 2–11, 2003. URL `http://dl.acm.org/citation.cfm?id=882086`.

[8] Ikenna Odinaka, Po-Hsiang Lai, Alan D. Kaplan, Joseph A. O'Sullivan, Erik J. Sirevaag, and John W. Rohrbaugh. ECG Biometric Recognition: A Comparative Analysis. *IEEE Trans. Inf. Forensics Secur.*, 7(6):1812–1824, dec 2012. ISSN 1556-6013. doi: 10.1109/TIFS.2012.2215324. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6287579`.