Application of Symbolic Piecewise Aggregate Approximation (PAA) Analysis to ECG Signals

Article

3 authors, including:

Serhan Ozdemir
Izmir Institute of Technology
57 PUBLICATIONS **381** CITATIONS
SEE PROFILE

Bora İ Kumova
Duale Hochschule Baden-Württenberg Mosbach
27 PUBLICATIONS **86** CITATIONS
SEE PROFILE

APPLICATION OF SYMBOLIC PIECEWISE AGGREGATE APPROXIMATION (PAA) ANALYSIS TO ECG SIGNALS

Burcu Kulahcioglu*

Serhan Ozdemir†

Bora Kumova*

Artificial *Iyte*lligence & Design Lab
†Mechanical Engineering Department,
*Computer Engineering Department
Izmir Institute of Technology, 35430 Izmir, Turkey
burcukulahcioglu@iyte.edu.tr, serhanozdemir@iyte.edu.tr, borakumova@iyte.edu.tr

ABSTRACT
Symbolic Time Series Analysis (STA) is an emerging methodology that involves coarse graining of the signals. Repeating segments of the time series are associated with symbols, thereby reducing the complexity of the series. It facilitates data mining tasks to be performed easily such as indexing, clustering, classification, summarization, and anomaly detection. This study involves symbolization through Symbolic Aggregate Approximation (SAX) with Piecewise Aggregate Approximation (PAA). The same ECG series is symbolized first by PLA and then PAA. Coarsing the series by PLA proved to be more problematic than PAA. At coarser scales, details are lost in noise with PLA, whereas local features become clearer with PAA. However during the analyses of ECGs of various subjects, it is understood that PAA fails when the series is not perfectly periodic as in rotating machinery. This fact is contrasted with the synthetic ECG which is manipulated to be perfectly periodic to juxtapose the results of the two trials. It is deduced that PAA delivers better pattern detection when signals are truly periodic.

KEY WORDS
Symbolic Time Series Analysis (STA), Piecewise Aggregate Approximation (PAA), Symbolic Aggregate Approximation (SAX), ECG, Coarse Graining

1. Introduction

The studies on Symbolic Time Representations have emerged in the 1980s. The early effort mainly focused on chaotic signals. Having become popular, STA found other fields of application. To name a few, JN Crutchfield introduced computational mechanical approach into Symbolic Time Series Analysis. He studied pattern discovery methods with C.R. Shalzi and K.L.Shalzi. Asok Ray has performed some other important studies in the field. He analyzed symbolic dynamics of complex systems for anomaly detection. A symbolic representation method SAX (Symbolic Aggregate Approximation) is invented by Eamonn Keogh and Jessica Lin, which is also employed in this study. It is used with PAA (Piecewice Aggregate Approximation) to symbolize a time series into strings which provides an easier representation to work with.

This study examines the suitability of symbolic PAA analysis to aperiodic signals. We first symbolize an ECG data in order to detect anomalies. The ECG data set of a student is examined which has an anomaly where the student had risen an arm. This anomaly is tried to be detected by using the obtained symbol string. Then, a few more experiments are performed by using the series obtained from different students. In order to test the idea, the experiment is repeated on a synthetic periodic ECG data set that is generated to be perfectly periodic, which is in contrast to the real ECG data.

2. Symbolic Time Series Analysis

Symbolic Representation of time series aims to provide efficiently and accurately performing pattern discovery, data mining tasks, and anomaly detection. It also compresses data that needs less space to allocate by coarse-graining the time series. However; it introduces loss of information depending on the number of symbols and the size of the segments to symbolize.

In a simple example given in [1], the time series of “one year of hourly power consumption from an industrial process” is symbolized. The time series is given in Figure 1.1. The on-off type graph corresponds to the minute-long consumption. The series is more concise and understandable when it is converted into symbols as it is seen in Figure 1.2.

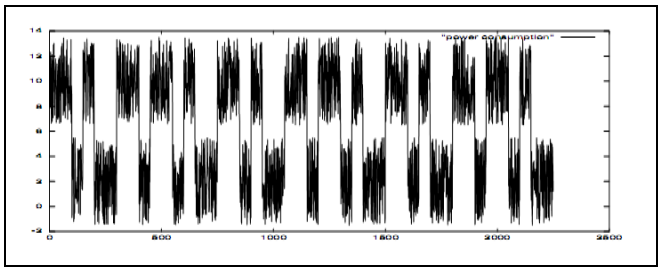

Figure 1.1 Simple Two-states Electric Consumption Example

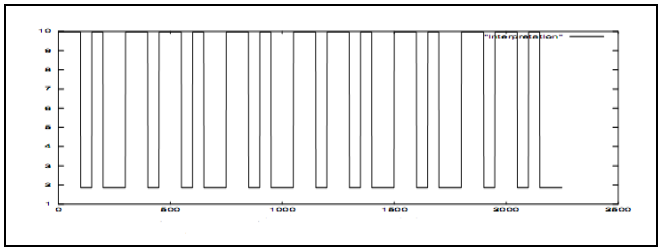

Figure 1.2 Reconstructed Data of Simple Two-states Electric Consumption Example

Symbolization of a time series can be executed by applying the following steps: Signal separation, phase space reconstruction, partitioning/segmentation of the time series, symbolization of the time series, construction of probabilistic finite state machine and construction of probability matrix [2].

SAX (Symbolic Aggregate Approximation): It is a symbolization method that reduces a time series of length n to a string with length w where w<<n [3]. By converting the time series into a symbol string S, it is possible to apply string operations and use distance measures easily. It also makes it easy to construct markov machines from time series.

3. Application of Symbolic Time Series on ECG Data

The first ECG used in this study belongs to a student at rest measured at 250 Hz for 70 seconds, Figure 3.1. It consists of 15080 data points. Between the points 13000 and 14000, the student had risen an arm to have a deliberate disturbance in the time series as seen. Here is the anomaly that is aimed to detect by comparing to the other segments.

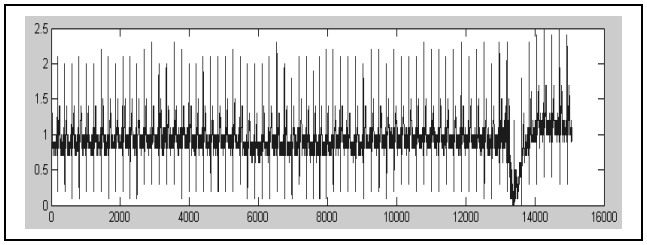

Figure 3.1 ECG Time Series to be Symbolized

In this study, it is assumed that there is no noise on the original data. Hence, no signal seperation method is applied.

At first; it is first proposed to use PLA method with Bottom-Up approach for segmentation stage. If too many data points are to be merged, then a rough time series with considerable amount of information loss, Figure 3.2. Few data points will have a time series nearly identical to the original one, as in Figure 3.3. Figure 3.4 seems to be the right combination.

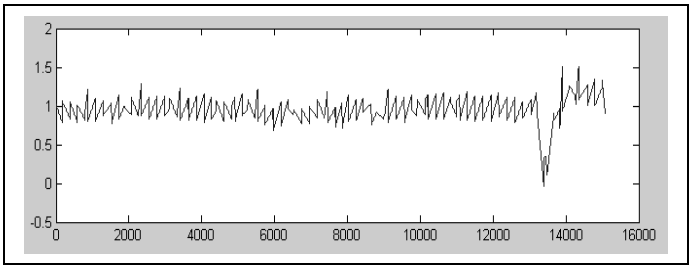

Figure 3.2 PLA Applied to ECG Time Series by Merging 15000 Data Points

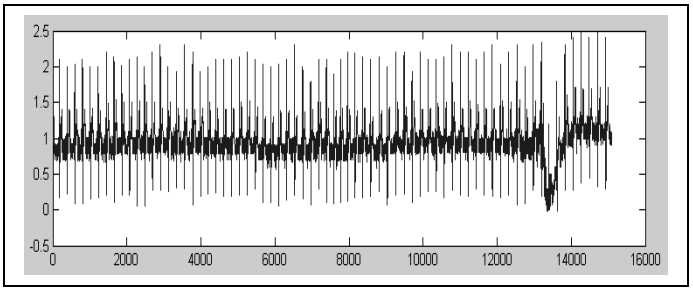

Figure 3.3 PLA Applied to ECG Time Series by Merging 10000 Data Points

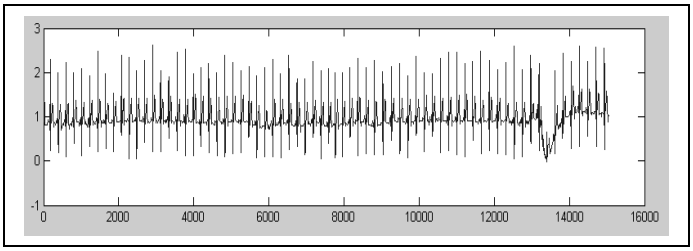

Figure 3.4 PLA Applied to ECG Time Series by Merging 14500 Data Points

The same series is now treated by PAA. This method divides the series into some number of segments and approximates each one to a constant value which is the average of the data values in this segment. Hence, selection of a suitable segment size is important to have a good approximation of the series. If a large segment size is implemented then a rough time series approximation will appear as in PLA, Figure 3.5. Small segment size will lead to retaining too much of the original data, Figure 3.6. Figure 3.7, the PAA approximation of the time series appears to be an optimum.

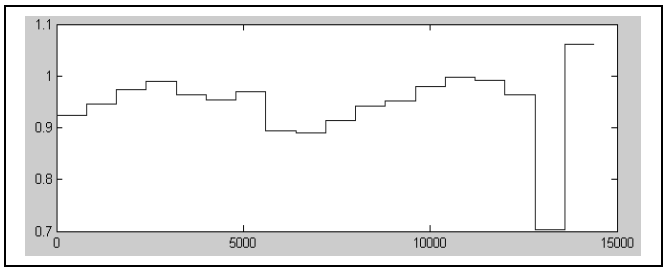

Figure 3.5 PAA Applied to ECG Time Series with Segment Size of 800

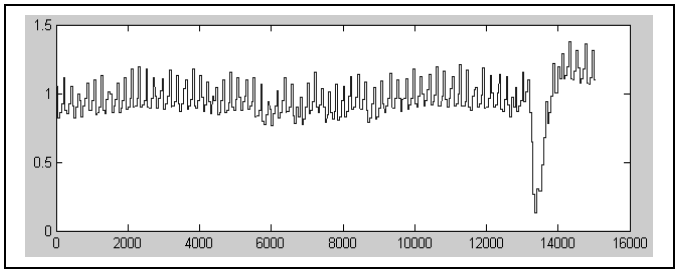

Figure 3.6 PAA Applied to ECG Time Series with Segment Size of 50

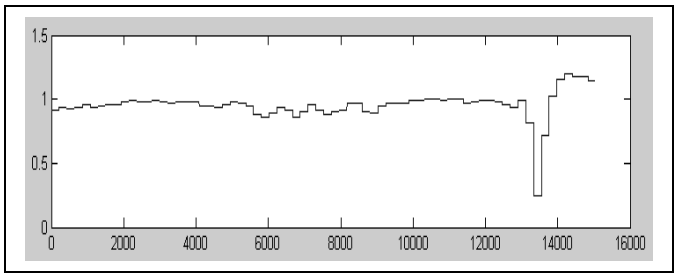

Figure 3.7 PAA Applied to ECG Time Series with Segment Size of 215

Having the time series is segmented by using PAA method, it can be converted into a symbol string by using SAX. SAX is a symbolization method that involves placing a symbol for each segment obtained by using PAA. In order to do that, it is essential to specify the number of symbols and the interval of the values for each symbol. The number of symbols to be used is generally determined by an expert having knowledge about the studied domain. Using too many symbols will cause to end up with a string keeping too much of the original data and will not simplify the series; on the other hand, too less symbols will cause considerably amount of information loss. After deciding on the number of symbols, some histograms of the data values can be helpful for specifying the intervals for each symbol.

In this study; ECG data is symbolized with extended SAX [4], which employs the minimum and maximum points of each segment in addition to the average values. Hence, for each segment, the minimum and maximum values are also symbolized as well as the average values in the order of their occurrences.

Since the PAA representation of our time series is already obtained, it should be determined the segment size, number of symbols to use and the breakpoints for each symbol. There are 15084 data points and 70 peaks in this example. In this study, segment size is calculated as:

Segment_size=number_of_datapoints / number_of_peaks

Hence, here the segment size is: 15084 / 70 ~= 215.

This study uses three symbols {a, b, c} for this time series. In order to use three symbols, it should be determined two breakpoints β_1 and β_2 such that:

$$\text{symbol} = \begin{cases} a & \text{if value} < \beta_1 \\ b & \text{if } \beta_1 < \text{value} < \beta_2 \\ c & \text{if value} > \beta_2 \end{cases}$$

Given this information, some histograms are examined to decide on the breakpoints. Figure 3.8 shows a typical histogram for the average values of data points for all segments.

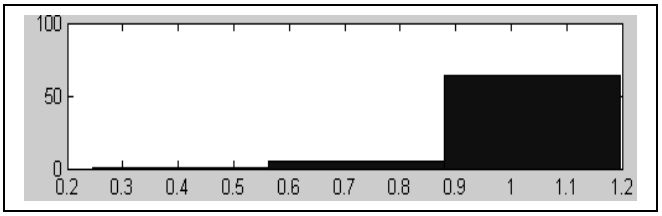

Figure 3.8 A Histogram for the Average Points of Segments

To give an idea of anomaly detection, an in depth analysis is conducted in between breakpoints. One is presented below as an example: Breakpoints 0.9 and 1.2 yield the following symbol string where each word with three-symbols corresponds to a segment.

aca aca aca bca aca bca bca bca bca bca bca bca cca cca caa cca aca bca bca caa caa caa caa cab cab caa cab caa caa cab cab caa cab cab cab cab abc aac bca bca aca aca cab cab cab cab cab cab cac cab cab cac cac cab cab cac cac cac cab cab cab cba **aca** caa cab cac cac cac cab cab

Euclidean Metrics is used as similarity measure between the segments, by using the formula:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + \ldots\ldots + (x_n - y_n)^2}$$

The error level of an anomalous segment is calculated by comparing its each symbol to an ordinary one's symbols. Higher euclidean distance yield less similarity, which we aim to detect anomalies.

If the number of symbols in the symbol alphabet is increased; then the resolution is increased, meaning the anomalous segment would have more hamming distance; hence more error from a normal segment. Eventually, the symbol sequences obtained are not as regular as expected. If the sequences are examined more carefully, a shift can be observed in the symbols.

Here is another experiment performed by using the ECG series of a different student (see in Figure 3.9). The series is irregular and has some anomalies that can also be seen in the PAA approximation shown in Figure 3.10. In this example, there are 15084 data points and 96 peaks; the segment size for this example is:
15084 / 96 ~= 157.

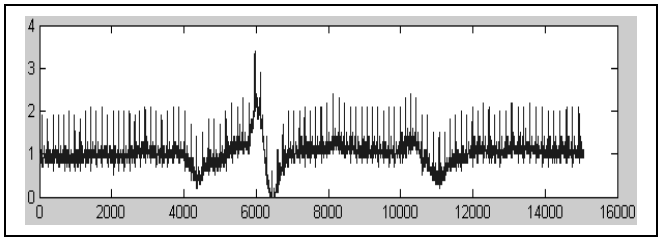

Figure 3.9 ECG Time Series to be Symbolized

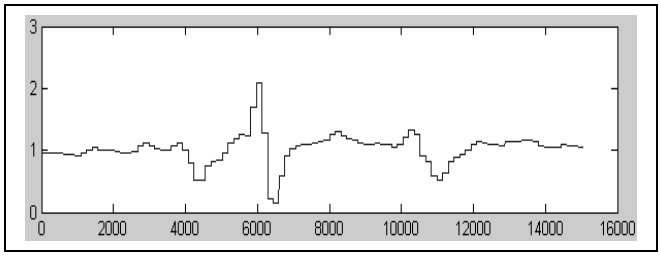

Figure 3.10 PAA Applied to ECG Time Series with Segment Size of 157

Applying SAX to this PAA approximation by using three symbols and the breakpoints, 0.75 and 1.5 as the following string is obtained:

cca caa cab cab cab cab cab cab cab cab cab cab cab cab bca bca bca bca bcb bcb bca bca bca bbc bcb bca bca **aba aac** aca aca abc bca bbc bbc bbc bbc **bcc ccc cba baa aab aac** abc bbc bcb bca bcb bca bcb bbc bbc bcb bcb bcb bcb bcb bcb bcb bca bcb bca bca bcb bbc bcc cbb caa caa **caa bca acb aca** caa cab bcb cab bcb bcb cbb cbb cab bcb bcb cbb cbb ccb bcb bca bca bbc bbc bbc bbc cab cbb

As is seen, as the irregularity of the ECG data increases, the accuracy of the obtained symbol strings lessens, because it causes more frequent symbol shifts to occur.
The experiment is also repeated on a third ECG data set taken from a different student given in Figure 3.11. This student has not risen his arm and there is just a small anomaly at the beginning of the series. The PAA approximation of this example is shown in Figure 3.12 which has 15074 data points and 67 peaks, hence it has
15074 / 67 ~= 225 as segment size.

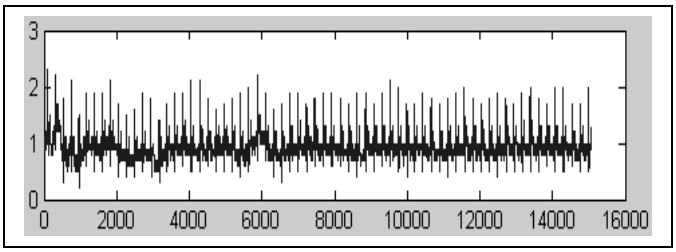

Figure 3.11 ECG Time Series to be Symbolized

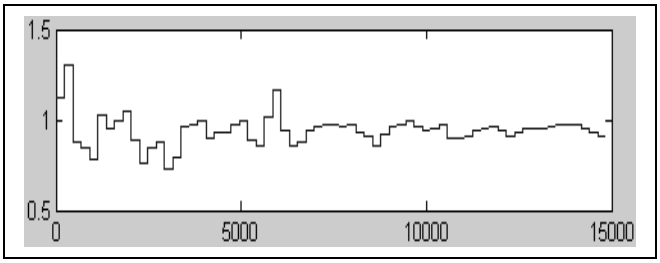

Figure 3.12 PAA Applied to ECG Time Series with Segment Size of 225

For the sake of simplicity, instead of repeating all steps with similar strings, we are just giving our comments on the results. In these three analyses, it can be seen that even though there is considerably large hamming distance between an anomalous segment and an ordinary one, similar error levels can be obtained between the two ordinary segments because of the symbol shifts.
This is caused by the aperiodicity of the ECG, which is corroborated with a synthetic ECG that is manipulated to be perfectly periodic, Figure 3.13. The same ECG with a synthetic anomaly may be seen in Figure 3.14. In this case, anomaly is better detected since it became conspicuous among the truly periodic and repeating symbols.

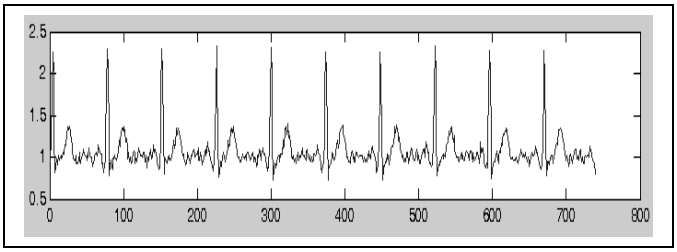

Figure 3.13 Synthetic ECG Series

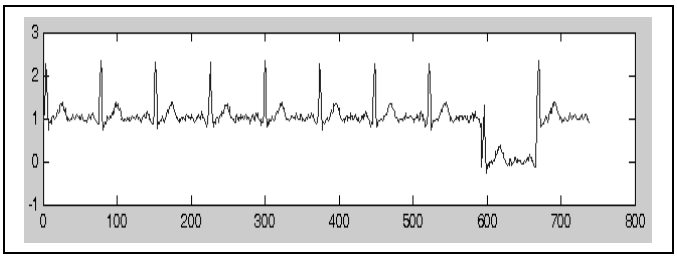

Figure 3.14 Synthetic ECG Series with anomaly

4. Discussion

In this study, we have analyzed ECGs through Symbolic Time Series Representation. It is attempted to detect anomalies using symbol strings. It is seen that the symbol strings did not yield the expected results. The error caused by the anomaly is similar to the error caused by some other less important parts of the symbol string. This dicrepancy is a result of the symbol shifts that tend to appear after processing a number of segments. These shifts occur because of the nature of the ECG data that has non-equal segment sizes. Additionally, as the irregularity the ECG data increases, the accuracy of the obtained symbol strings gets worse. This claim is supported with an artificially created completely periodic ECG data, where no such shifts occurred. This lack of shift exposes the true anomalies against a background of similar symbol strings. It is concluded that ECG data is a challenging series to deal with. It is also concluded that certain methods must be carefully chosen to realize the symbolic representation to achieve anomaly or fault detection. Since, it is shown that PAA along with extended SAX is not capable of handling aperiodic data.

References

[1] G. Hebrail, B. Hugueney, Symbolic representation of long time series. *Conference on Applied Statistical Models and Data Analysis (ASMDA2001)*, Compiègne, 2001

[2] A. Ray, Symbolic dynamic analysis of complex systems for anomaly detection, *Signal Processing*, *84*(7), 2004, 1115–1130.

[3] E. Keogh, S. Lonardi, B. Chiu, A Symbolic Representation of Time Series, with Implications for Streaming Algorithms, *Proc. 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, San Diego, CA, 2003

[4] B. Lkhagva, Y. Suzuki, K. Kawagoe, Extended SAX: Extension of symbolic aggregate approximation for financial time series data representation, *DEWS,* 2006, 4A-i8.

[5] E. Keogh, A tutorial on indexing and mining time series data, *ICDM '01 The 2001 IEEE International Conference on Data Mining*, San Jose

[6] A. Ray, Anomaly detection and failure migitation in complex dynamical systems, *Seminar at National Institute of Standards and Technology*, 2004

[7] C. R. Shalizi, K. L. Shalizi, J. P. Crutchfield, Pattern Discovery in Time Series, Part I: Theory, Algorithm, Analysis, and Convergence, *Journal of Machine Learning Research*, 2003

[8] L. Karamitopoulos, G. Evangelidis, Current trends in time series representation, A-68, *Panhellenic Conference on Informatics (P.C.I.) 2007*, Patras

[9] B. Chiu, E. Keogh, S. Lonardi, Probabilistic discovery of time series motifs, *Proc. 9th ACM SIGKDD international conference on Knowledge discovery and data mining,* Washington D.C., 2003, 493-498

[10] M. Falk, F. Marohn, R. Michel, D. Hofmann, M. Macke, A First Course on Time Series Analysis, (Chair of Statistics, University of Würzburg, 2006)

[11] T. K.Moon, W. C. Stirling, Mathematical methods and algorithms for signal processing (Prentice Hall, 2000)