Multivariate Symbolic Aggregate Approximation for ECG Analysis

Moritz M. Konarski
Supervised by Prof. Taalaibek M. Imanaliev

Applied Mathematics and Informatics Program,
American University of Central Asia

Partially supported by AUCA FRG
"Mathematical Model in Acute Cardiac Ischemia Evaluation"

May 31, 2021
Bishkek, Kyrgyz Republic

American University of Central Asia

Outline

Introduction

- ischemic heart disease (IHD) makes up 16% of global deaths but can be diagnosed using an electrocardiogram (ECG or EKG)
- manual ECG analysis is slow and error-prone, computers can help
- long ECGs are problematic even for computers → simplification through representation
- such representations should be simpler, but still correspond to the original data

What is an ECG?

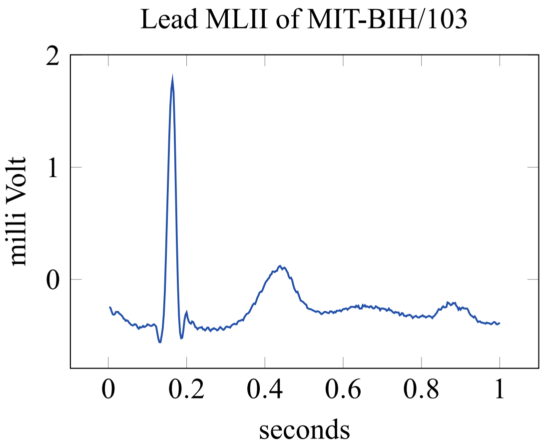


Figure 1: ECG of one heartbeat.

- records the heart's electrical activity
- contains up to 12 leads (simultaneous measurements)

Representation and Classification

- for simplification, time series can be represented, e.g. as SAX or MSAX
- then, the representation can be analyzed instead of the raw data
- HOT SAX can be used to classify ECG segments into discord and non-discord
- this work uses HOT MSAX to combine MSAX and HOT SAX
- effectiveness of the methods will be judged by recall and precision

Research Questions & Hypothesis

- Using the MIT-BIH ECG database, what parameters maximize HOT SAX and HOT MSAX recall?
- Which is better: optimal HOT SAX or optimal HOT MSAX?

HOT MSAX should have higher recall than HOT SAX if both use their best parameters

Novel Contributions

- application of MSAX to ECG discord discovery and medical data in general
- the HOT MSAX algorithm, a modification of HOT SAX that uses MSAX
- the expansion of HOT SAX to multivariate time series through HOT MSAX

Method Background

- Lin *et al.* (2003):
  Symbolic Aggregate Approximation (SAX): simplified, symbolic representation
- Keogh *et al.* (2005):
  Heuristically Ordered Time series using SAX (HOT SAX): discord discovery algorithm using SAX
- Anacleto *et al.* (2020):
  Multivariate SAX (MSAX): expands SAX to multivariate time series

SAX and MSAX – Overview

SAX	MSAX
Application	
univariate time series, e.g. a single ECG lead	multivariate time series, e.g. multiple ECG leads (whole ECG)
Steps	
(1) univariate z-normalization	(1) multivariate z-normalization
(2) PAA dimension reduction (w)	(2) PAA dimension reduction (w)
(3) SAX discretization (a)	(3) SAX discretization (a)

SAX and MSAX – Step (2)

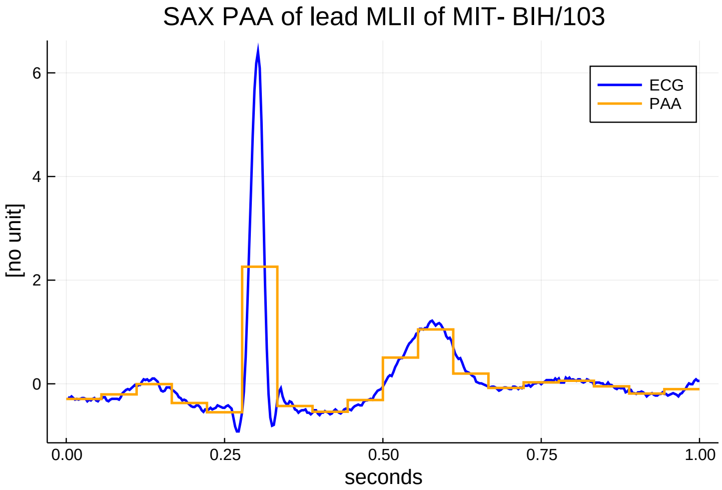


Figure 2: ECG with PAA, $w = 18$

SAX – Step (3)

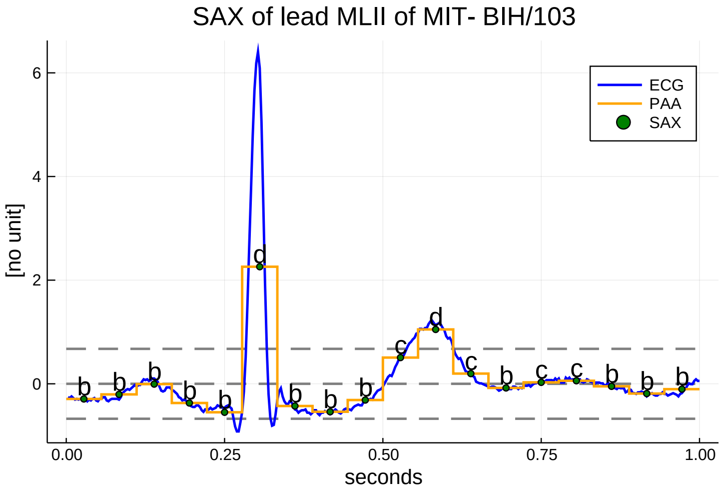


Figure 3: ECG with SAX, $w = 18$, $a = 4$

MSAX – Step (3)

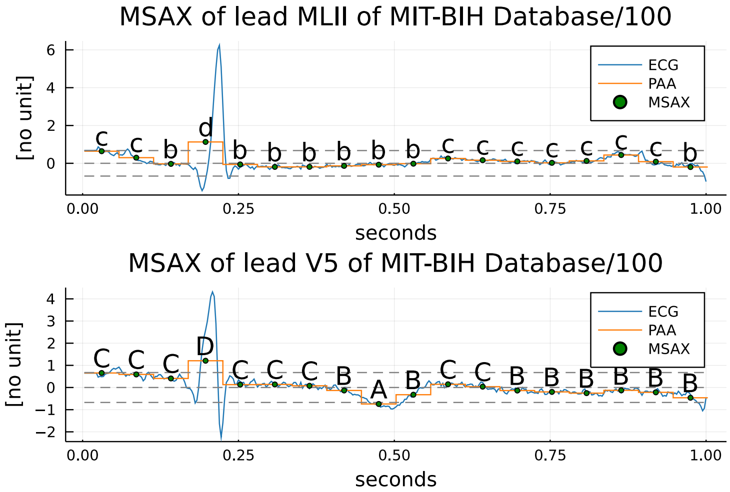


Figure 4: ECG with MSAX, $w = 18$, $a = 4$

SAX and MSAX – Distance Measure

- needed to compare two SAX/MSAX segments
- sum of distances between symbols
- symbol distance is based on difference of breakpoints
- lower-bounds Euclidean Distance, i.e. corresponds to "real" distance

HOT SAX and HOT MSAX – Overview

- HOT SAX: find discords in SAX-represented time series
- classifies time series segments into “discord” and “non-discord”
- HOT MSAX: uses MSAX instead of SAX
- HOT MSAX can work with multivariate time series

HOT SAX and HOT MSAX – Heuristic

- two parameters: m and k
- two assumptions:
  - time series discords are rare
  - segments similar to discords may also be discords
- speed up discord discovery:
  - consider rarest segments first
  - consider similar segments together

Analysis Process

(1) perform HOT SAX and HOT MSAX for many parameter combinations for all ECGs in the MIT-BIH database

(2) find recall, precision for each combination

(3) set recall threshold of 95%, then sort by precision

(4) choose top 10 of those parameters for each method

(5) choose best parameters for each method using box plot, interquartile range, outliers

Three Datasets

(1) S-SAX: data for HOT SAX algorithm, considering each lead separately

(2) D-SAX: data for HOT SAX, considering both leads combined

(3) MSAX: data for the HOT MSAX algorithm

Overview of Results

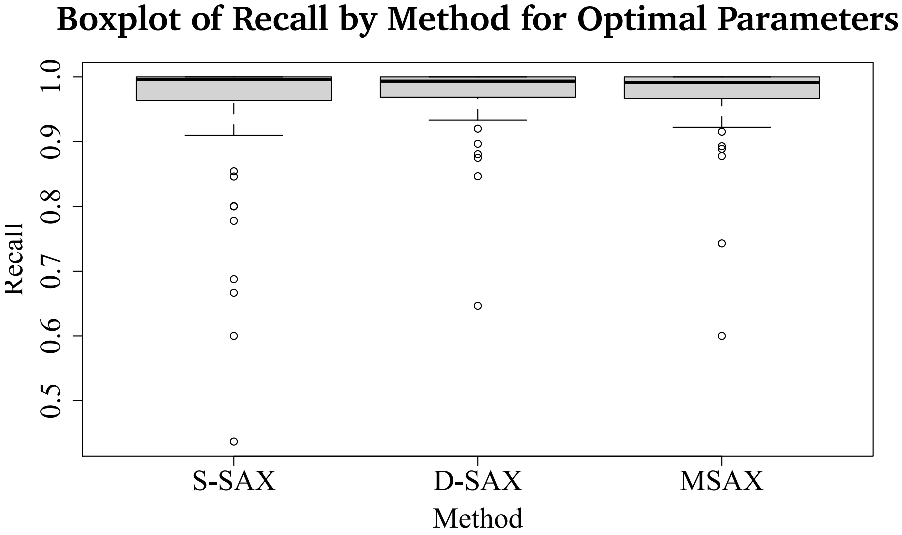
Table 1: Coarse Overview of Results. Shown are sets of parameters for each method that satisfy the recall threshold of 95%.

Method	Total Sets	Sets Satisfying recall $\geq$95%
S-SAX	4,968	99 (1.2%)
D-SAX		192 (3.9%)
MSAX		255 (5.1%)

Best Parameter Sets by Method

Table 2: Best Parameter sets for each of the methods. Best overall parameters highlighted in bold.

Method \ Parameter	k	w, m	a
S-SAX	-1	36	21
D-SAX	-1	**12**	24
MSAX	-1	**12**	**17**


Boxplot of Recall by Method for Optimal Parameters
Recall
1.0
0.9
0.8
0.7
0.6
0.5
S-SAX
D-SAX
MSAX
Method

Comparing Best Parameter Sets – Recall

Table 3: Statistical measures for recall of optimal parameter sets. Best overall values highlighted in bold.

Method \ Measure	IQR	Median	Outliers
S-SAX	0.035	**99.60%**	11
D-SAX	**0.030**	99.35%	**6**
MSAX	0.033	99.13%	**6**

Discussion

- no statistically significant difference in recall for the methods

→ hypothesis cannot be supported

- MSAX has both the smallest alphabet size and highest dimension reduction

→ this points to MSAX being more efficient in achieving the same results

- Anacleto, Vinga, and Carvalho [5] showed similar performance for ECG classification, supports this work's conclusion

Conclusion

- could not demonstrate superiority of HOT MSAX for best parameters
- contributed the HOT MSAX method to literature
- showed viability of a discord classifier for ECG analysis
- in future research: different ECG data, different detection criteria

References I

[1] *The top 10 causes of death*, 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (visited on 05/25/2021).

[2] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery - DMKD '03*, San Diego, California: ACM Press, 2003, pp. 2–11. DOI: 10.1145/882082.882086.

[3] C. Zhang *et al.*, "Anomaly detection in ECG based on trend symbolic aggregate approximation," *Mathematical Biosciences and Engineering*, vol. 16, no. 4, pp. 2154–2167, 2019. DOI: 10.3934/mbe.2019105.

[4] E. Keogh, J. Lin, and A. Fu, "HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence," in *Fifth IEEE International Conference on Data Mining (ICDM'05)*, Houston, TX, USA: IEEE, 2005, pp. 226–233. DOI: 10.1109/ICDM.2005.79.

References II

[5] M. Anacleto, S. Vinga, and A. M. Carvalho, "MSAX: Multivariate Symbolic Aggregate Approximation for Time Series Classification," in *Computational Intelligence Methods for Bioinformatics and Biostatistics*, P. Cazzaniga, D. Besozzi, I. Merelli, and L. Manzoni, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 90–97. DOI: 10.1007/978-3-030-63061-4_9.

Thank You!