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

1. Introduction and Background
2. Methods
3. Results
4. Conclusions

Motivation and Relevance

- heart diseases kill more people a year than any other disease
- ischemic heart disease (IHD) makes up 16% of global deaths
- IHD can be diagnosed using a stress test and electrocardiogram (ECG or EKG)
- manual ECG analysis is slow and error-prone

→ computerized ECG analysis can help

What is an ECG?

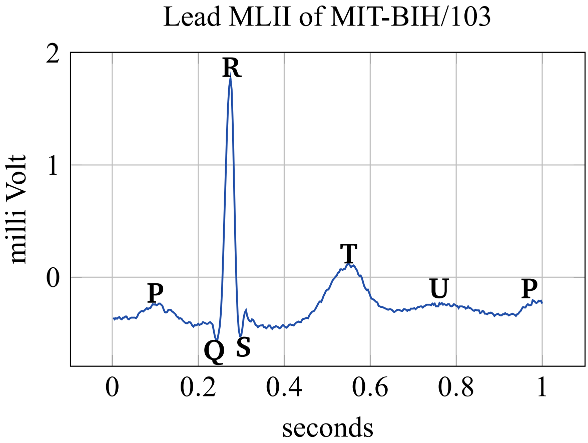


Figure 1: Annotated ECG of one heartbeat

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

Steps of ECG Analysis I

General steps:

(1) signal acquisition and filtering

(2) data transformation or preparation for processing

(3) waveform recognition

(4) feature extraction

(5) classification or diagnosis

Steps of ECG Analysis II

This work's focus:

(1) signal acquisition and filtering

→ **data transformation** or preparation for processing

(3) waveform recognition

(4) feature extraction

→ **classification** or diagnosis

Research Questions & Hypothesis

- using the MIT-BIH ECG database, determine the parameters maximizing HOT SAX and HOT MSAX recall
- compare recall value for best parameters

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

Previous Research

- 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

This Work’s Novel Contributions

- application of MSAX to ECG discord discovery
- the HOT MSAX algorithm
- the expansion of HOT SAX to multivariate time series through HOT MSAX

SAX and MSAX – Overview

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

SAX and MSAX – Step (2)

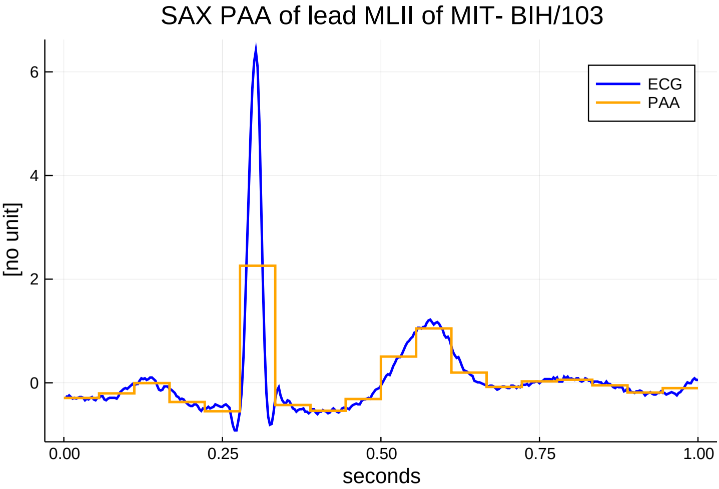


Figure 2: ECG with PAA (MIT-BIH/103, $w = 18$, $T = 360$)

SAX and MSAX – Step (3)

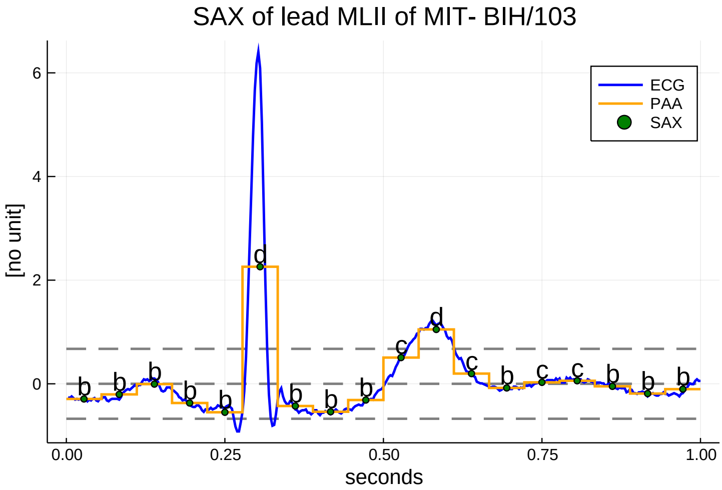


Figure 3: ECG with SAX (MIT-BIH/103, $w = 18$, $T = 360$)

SAX and MSAX – Distance Measure

- basically the sum of distances between symbols
- symbol distance is based on difference of breakpoints
- lower-bounds Euclidean Distance, corresponds to “real” distance

Table 1: Distance matrix for 5 symbols.

	a	b	c	d	e
a	0	0	0.59	1.09	1.68
b	0	0	0	0.51	1.09
c	0.59	0	0	0	0.59
d	1.09	0.51	0	0	0
e	1.68	1.09	0.59	0	0

HOT SAX and HOT MSAX – Overview

- HOT SAX: find discords in SAX-represented time series
- speeds up the “brute force” approach
- 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 assumptions:
  - time series discords are rare
  - similar segments are similarly rare
- speed up discord discovery:
  - consider rarest segments first
  - consider segments similar to the rarest first

Parameter Optimization

- perform HOT SAX and HOT MSAX for many parameter combinations
- find recall for each combination
- recall threshold of 95%, then sort by precision
- choose top 10 of those parameters for each method
- choose best parameters using box plot, interquartile range, outliers

Overview of Results

Table 2: Coarse Overview of Results. Shown are sets of parameters for each method that satisfy the recall threshold of 95%.

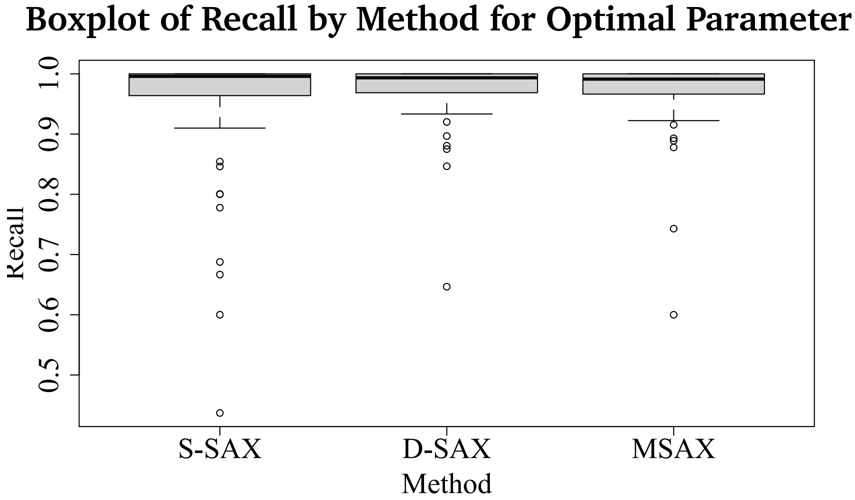
Method	Total Sets	Sets Satisfying recall $\geq$95%
S-SAX	4,968	99 (1.2%)
D-SAX		192 (3.9%)
MSAX		255 (5.1%)

Best Parameter Sets by Method

Table 3: Best Parameter sets for each of the methods. Best overall parameters highlighted in bold.

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

Comparing Best Parameter Sets – Recall



Comparing Best Parameter Sets – Recall

Table 4: 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
- HOT MSAX was found to be more efficient than HOT SAX
- showed viability of a discord classifier for ECG analysis

References I

[1] G. B. Moody and R. G. Mark, *MIT-BIH Arrhythmia Database*, physionet.org, 1992. DOI: 10.13026/C2F305.

[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, ISSN: 1547-1063. 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!