

MSAX for ECG Analysis

- └ Introduction
- └ ECG Basics
 - └ What is an ECG?

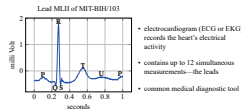


Figure 1: Annotated ECG of one heartbeat

- P wave: blood entering the heart
- QRS complex: heart contraction pumping blood
- T: return of ventricle to polarized state
- U: present in 25%
- muscle contractions caused by electric pulses
- electric pulse can be measured on the skin
- electrodes form leads (need 2 to measure anything)
- most types of heart disease can be detected
- diagnosis and analysis is performed by trained cardiologists
- **datasets available online; contain 2 or more leads (the most significant ones)**
- **I will be using online datasets for my analysis**
- heart diseases are some of the most deadly ones, thus ECG are really important

MSAX for ECG Analysis

└ Introduction

└ ECG Basics

└ ECGs as Time Series

Definition

A discrete time series is an ordered sequence which, at discrete points in time, has n values each. If $n = 1$, the series is univariate and if $n > 1$, it is multivariate.

- digital ECGs are discrete multivariate time series:
 - have > 1 value at each point, often $n = 12$
 - recorded at discrete, evenly spaced time points
- time series analysis methods can be applied to ECGs

- modern ECGs have at least 2, most have 12
- digital ones have set sampling frequencies, even the machines have set frequencies
- multivariate: measure more than 1 lead per time point
- discrete: set sample frequency in the machines
- discrete: because measured at discrete moments in time
- time series: they are data measured at equal time intervals
- n measurements per point in time (i.e. leads)
- $n = 1$ is univariate, $n > 1$ is multivariate

MSAX for ECG Analysis

└ Introduction

└ ECG Analysis

└ ECG Analysis

- standard method: manual analysis by cardiologist
- automated or computer-assisted ECG analysis seeks to replace that
- multiple stages: (1) signal acquisition; (2) data transformation, processing, filtering; (3) waveform recognition, feature extraction; (4) classification
- current research focus: artificial neural networks

- is relatively slow; time is of the essence
- lots of training required
- error prone
- maybe not feasible for long ECGs
- can speed up process
- can pick up details humans miss
- digitizing paper ECGs or recording digital ones
- filtering to remove various types of noise
- reduce complexity of the data
- select important features and neglect irrelevant ones to ease analysis
- often added, figure out if there is some disease present or not
- balance between accuracy and complexity needed
- ann: hand all the steps discussed to a NN; use as good classifier too

MSAX for ECG Analysis

└ Introduction

└ ECG Analysis

└ SAX, MSAX, and HOTSAX

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

- ecg as letters that mean same thing as original
- guaranteed to behave like the original data
- works on univariate time series
- has been used on ECGs
- takes the correlation between ecg leads into account
- cov mat: covariance between each lead and variance on diag
- uses sax representation to make the finding of discords easier
- can use MSAX just as well

MSAX for ECG Analysis

└ Introduction

└ ECG Analysis

└ Time Series Discords

- these can be diseases, noise, etc
- the discord does not discern

Definition

A time series discord is the subsequence of a time series that is most different from all other subsequences.

k time series discords are the k most different subsequences.

- discords represent anomalies in an ECG
- HOTSAX enables fast discord discovery

MSAX for ECG Analysis

- └ Introduction
- └ Hypothesis
- └ Hypothesis

HOTSAX with MSAX will increase the number of relevant discords detected compared to HOTSAX with SAX.

- **mention that MSAX to ECGs in particular is new**
- **mention that HOTSAX with MSAX is new**
- THIS METHOD WILL NOT BE SUPER ACCURATE; MANY ECG changes are relatively small and would get lost in the SAX process

MSAX for ECG Analysis

└ Methods

└ SAX and MSAX

└ Step 1: Z-Normalization

- say that the process is the same as MSAX based on SAX
- this is assumed and this worked for other people who applied SAX to ECGs
- to compare time series, normalization is the accepted step
- what is this
- takes into account the correlation between leads

Assumption

The time series values are normally distributed.

SAX

- normalize univariate time series
- uses scalar mean and variance

MSAX

- normalize multivariate time series
- uses vector mean and covariance matrix

MSAX for ECG Analysis

└ Methods

└ SAX and MSAX

└ Step 2: Dimensionality Reduction

- this reduces complexity
- PAA form of time series is shorter and simpler
- it still somewhat corresponds to the original

PAA

Piecewise Aggregate Approximation (PAA) takes T time series points, splits them into w ($w < T$) segments, and averages each of them.

SAX

- apply PAA to time series

MSAX

- apply PAA to each of the time series individually

MSAX for ECG Analysis

└ Methods

└ SAX and MSAX

└ Step 3: Discretization

- result is called word
- N is the alphabet size
- big thing here is that this gives defined probability to each letter; makes no sense for real numbers (like PAA values)
- simplifies time series even more
- creates discrete categories, can be more useful

SAX Discretization

Find breakpoints splitting $N(0, 1)$ into B equiprobable segments.
Assign a letter to each area, starting with a to the left-most segment.
PAA segments get letters based on which area they are in.

SAX

- discretize the time series
- results in one word

MSAX

- discretize each time series
- results in one word with one letter for each time series

MSAX for ECG Analysis

└ Methods

└ SAX and MSAX

└ Step 4: Distance Measure

- distance is based on letter pairs
- SAX: sqrt of sum of squared distance
- MSAX: sqrt of sum of squared distance; also sum all leads
- this lower-bounds the euclidean distance, meaning that results in SAX should hold true for the real data too

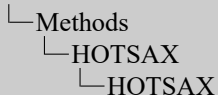
MINDIST

A distance measure is defined to compare two SAX words. It is defined for a pair of letters, distances between words are sums of distances between letters.

Table 1: Difference matrix for $B = 4$

	a	b	c	d
a	0	0	0.67449	1.34898
b	0	0	0	0.67449
c	0.67449	0	0	0
d	1.34898	0.67449	0	0

MSAX for ECG Analysis



- "brute-force" discord discovery is slow, needs T^2 operations
- HOTSAX speeds up discord discovery by considering:
 - discords are rare, start with rarest segment
 - similar segments have similar distances, consider together
- HOTSAX detects anomalies, it is not a classifier
- it uses SAX and MSAX for dimensionality reduction

- this is the basic idea that can speed up the process
- it is not guaranteed to do so, but it does not decrease efficiency
- this speeds up the process even more as we have fewer elements
- because of lower bounding, it still gives accurate results

MSAX for ECG Analysis

- └ Preliminary Results
- └ Implementation
- └ Implementation

- SAX, MSAX, HOTSAX was implemented in Julia (scientific programming language)
- used annotated digital ECGs from the MIT-BIH arrhythmia database
- HOTSAX was performed for different w , B , subsequence lengths
- results were exported to CSV file and analyzed using the R programming language

- fast, type support, great libraries, JIT compilation
- ecgs have all heart beats annotated
- know which are normal, diseases, noise, etc
- 48 recordings of 30 minutes
- w - paa segments; B - alphabet size; subsequence length for HOTSAX

MSAX for ECG Analysis

- └ Preliminary Results
 - └ Preliminary Results
 - └ Preliminary Results

- compared SAX and MSAX using the top $k = 80$ discords (816 sets of discords total)
- analyzed the relevance of results with recall (sensitivity)
- recall for MSAX is higher compared to SAX
- if SAX is applied to two leads and the results combined, it slightly outperforms MSAX

- how many relevant items are selected
- recall = true positive / (true positive + false negative)
- this is done because for medical things it is more useful to look at a couple too many segments than not enough

MSAX for ECG Analysis

- └ Preliminary Results
- └ Preliminary Results

- msax: average = 17.5%
- sax: average = 6.4%

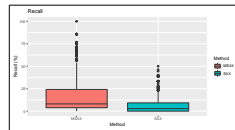


Figure 4: Boxplot comparing Recall for MSAX and single-lead SAX

MSAX for ECG Analysis

- └ Preliminary Results
- └ Preliminary Results

- msax: average = 17.5%
- sax: average = 18.5%

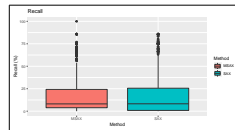


Figure 5: Boxplot comparing Recall for MSAX and dual-lead SAX

MSAX for ECG Analysis

- └ Preliminary Results
 - └ Remaining Tasks
 - └ Remaining Tasks

- perform tests for statistical significance of the result
- analyze the outliers visible in the boxplots
- compute more sets of discords with different parameters
- explore the influence of the parameters on the result

- for example t-test, biserial correlation