American University of Central Asia

Applied Mathematics and Informatics Program

Multivariate Symbolic Aggregate Approximation for ECG Analysis

Moritz M. Konarski

A Thesis Submitted to the Applied Mathematics and Informatics Program of American University of Central Asia in Partial Fulfillment of the Requirements for the Degree of **Bachelor of Arts**

Author
Moritz M. Konarski

Certified by Thesis Supervisor
Professor Taalaibek Imanaliev

Accepted by
Sergey Sklyar
Head of Applied Mathematics and Informatics Program, AUCA

April 6, 2021
Bishkek, Kyrgyz Republic

ABSTRACT

This abstract will be written once the paper is more finished.

Keywords: acute cardiac ischemia, ECG, mathematical modeling, MSAX

ACKNOWLEDGEMENTS

This paper was supported by grant XX.

I also thank …

TABLE OF CONTENTS

1 INTRODUCTION

In the year 2016, over 9.4 million people worldwide died of ischemic heart disease (IHD). IHD is responsible for 16.6% of all deaths, making it the most common cause of death globally. All forms of cardiovascular disease make up 31.4% of all deaths (17.9 million). Death caused by IHD disproportionally affects people over 50 years of age, with 91% of deaths for men and 95% of deaths for women occurring in that age range. In Kyrgyzstan, 13% of all deaths in 2016 were caused by IHD [1].

Ischemic heart disease is characterized by restricted blood flow to an area of the heart, causing it to not receive enough blood and oxygen. Blood flow restriction is caused by a blockage (or narrowing) in a blood vessel supplying the heart muscle. An artery can be blocked by a blood clot, but the most common cause is plaque buildup, which is called atherosclerosis. If the circulation to the heart is completely blocked, the cells in the heart muscle begin to die. This is called myocardial infarction, more commonly known as a heart attack. The deprivation of oxygen the heart experiences leads to the characteristic chest pain commonly associated with heart attacks [2].

Arrhythmia is a common form of heart disease where the rhythm of heart beats is irregular. It either changes too quickly, is too high, too slow, etc. Simply, it is a variation of the normal heart rate that is not justified in any way. The present in ECGs as heart-beats that are too close together or too far apart. They can also be characterized as a section of too-fast or too-slow heart beats.

IHD can be diagnosed before it leads to a heart attack. The diagnosis can be performed based on a patient's medical history, pharmacologically induced stress, or stress induced by physical exercise. During an exercise stress test, an electrocardiograph (sometimes combined with other methods) records the patient's heart activity, resulting in an electrocardiogram (ECG) [2]. The ECG is a diagnostic tool used to evaluate patients

At the most basic level, an ECG is simply a time series of data values. It is real-valued and the process is represents is continuous. Any recorded ECG is turned into a discrete time series through the sampling that the electrocardiograph performs. The fact that it is still real-valued means that for adequate sampling rates, the accuracy of the representation is acceptable.

Today, most ECGs are recorded for at least 2 leads, meaning that for each point in time, there are more than 1 available data point. This makes ECGs multivariate time series. Multivariate time series are difficult to analyze because their data exhibits strong connections between the leads. analyzing 2 or more variables of such a time series while acknowledging the connections between the variables can lead to a better extraction of information.

How can this be done for arrhythmia?

with (suspected) heart problems. It is a non-invasive, real-time, and cost-effective method that may be used to diagnose IHD. It is the most common tool used for cardiac analysis and diagnosis [3, 4, 5]. The most common form of the ECG is the 12-lead variant. The 12-lead ECG consists of 6 leads connected to the limbs and 6 leads connected to the torso of the patient. The leads record the differences in electrical potential between the places on the body that they are attached to. This reflects the differences in voltage that the heart experiences with each heart beat because those

voltage differences are conducted by the body.

Each heart beat is an electrical pulse originating in the sinoatrial node that travels through the muscle cells in the heart, leading them to contract. This electrical pulse travels beyond the heart and thus can be recorded on the surface of the body. The recording of it results in an ECG.

The measurements are taken in millivolts (mV). The ECG represents the state of the heart; a recorded ECG has the shape of a wave (the ECG wave) [4, 5]. If the state of the heart beat changes as the result of a disease like IHD (changing the measurable potentials or their occurrence over time), the ECG is able to record these changes.

The characteristic shape of an ECG for two heart beats is shown in Figure 1.1; the figure is taken from [6]. The figure has been annotated to show the significant features of an ECG. The peaks (or waves) P, Q, R, S, T, and U, as well as the segments between them, are the focus of ECG analysis. Multiple points together form what is called a complex; the QRS complex is a good example of this. Using these waves, the heart activity can be described and analyzed. In an ECG, the P-wave is the result of the atria depolarizing, which is the process of blood entering the heart as the first step in a heart beat. The QRS complex represents ventricular depolarization, the contraction of the heart causing it to pump blood. The T-wave is the return of the ventricle to its polarized state. The U-wave is only present in roughly 25% of the population and may be caused by mechanical-electric feedback. The RR interval can be used to calculate the heart rate because it represents one complete heart beat [6]. The shape of the P, Q, R, S, T, and U waves as well as the duration of various intervals between them are used as indicators of cardiac diseases.

Fix the repetition.

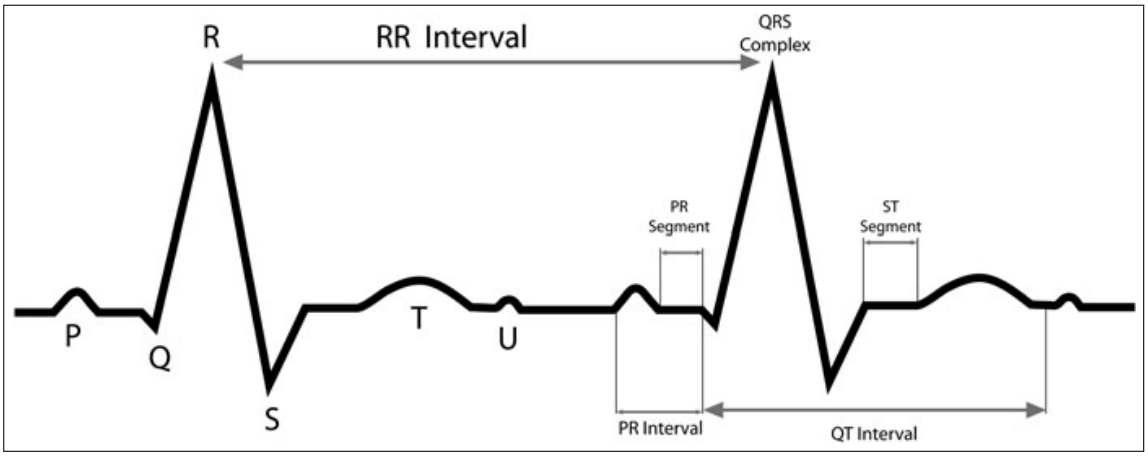

Figure 1.1: A schematic of an ECG waveform, annotated; from [6]

Using an ECG to diagnose a cardiac condition is difficult in practice. Small changes in the components of the ECG can be indicators of diseases and those changes can be overlooked, even by trained and specialized physicians. The chance to make a mistake is even higher for non-specialized physicians and trainees [3, 5].

furthermore, with ECGs being commonplace and ambulatory ECGs increasing in popularity, there is more and more ECG data to analyze. Many of these ECGs are 24h long or longer. Analyzing these ECGs manually is very time consuming in addition to the technical difficulties that ECG analysis includes.

This can be seen in the fact that for most online ECG databases, each ECG is analyzed by at least 2 independent cardiologists. If their diagnoses do not agree, they confer to find a correct diagnosis. This implies that the diagnosis is difficult and not necessarily a 1-person job.

For the diagnosis of IHD, changes in the ST-segment and T-wave are of particular interest. An elevation of the ST-segment compared to a normal heart beat is one of the main indications of IHD and myocardial infarction. A downward depression of the ST-segment, especially in combination with chest pain, is another indication of IHD. The changes in the ST-segment are thought to be caused by current flow between healthy heart muscle and ischemic heart muscle [6, 7].

How can arrhythmia be diagnosed? How is it normally done?

The diagnosis of IHD on the basis of an ECG is time sensitive. If a patient has IHD or suffers from a heart attack, treatment has to be started as soon as possible. Some forms of treatment are most effective in the first 3 hours after symptom onset and lose most of their effectiveness after 9 to 12 hours. The diagnosis required for treatment to begin should thus be as quick as possible. The ECG delivering information in real-time is an advantage here, even though there are more time consuming methods that can deliver more accurate results than an ECG [8].

Arrhythmia is a condition that can lead to heart failure, thus it is necessary to recognize and correctly diagnose it. then it can be treated by pacemakers and other methods.

The widespread use of ECGs and the time-sensitive nature of their application as diagnostic tools makes errors, delays, or inconsistencies in their interpretation unacceptable. A recent approach to minimizing this problem is the application of computer technology in ECG recording, storage, and analysis. The main steps of computerized ECG analysis are [4] (1) signal acquisition and filtering, (2) data transformation or preparation for processing, (3) waveform recognition, (4) feature extraction, and (5) classification or diagnosis.

This research will investigate steps (3) and (4) through the use of different feature extraction algorithms.

A method that combines dimensionality reduction with data transformation is SAX. SAX is a method that relies on PAA and the discretization. SAX is only applied to a univariate time series and does not take into account the connections between different variables of the same time series.

MSAX is an extension of SAX for multivariate time series. It has an in-built mechanism that enables it to be more responsive to events that happen in multiple leads at once.

The idea of this research is to apply MSAX to 2 leads of an ECG and to examine whether or not this increases the accuracy of detection of anomalies in the ECG. These detected anomalies could be examined with priority by a cardiologist to speed up the process of analysis and diagnosis by them. Thus this process could help improve the speed of diagnosis of heart conditions that show up on ECGs.

HOT-SAX? or kNN? Explain this here

update this with respect to MSAX and what I will actually do

The ECG data will be retrieved from the European ST-T Database.

Include MIT-BIH database here

This database provides ECG recordings that can be used as trial data to test feature extraction

algorithms. The European ST-T Database contains annotations made by cardiologists indicating the ST-segment, T-wave, and their changes. They also include information about the suspected disease [9, 10]. This information can be used to determine the effectiveness of the feature extraction algorithms.

What benefit has the dimensionality reduction and discretization that SAX exhibits?

2 STATE OF COMPUTERIZED ECG ANALYSIS

Recent advances in computer technology have enabled the use of computers in every aspect of ECG acquisition, processing, analysis, and storage. In light of these developments, the American Heart Association published recommendations for the interpretation and standardization of the ECG. They recommend that the low-frequency cutoff for low-frequency filtering of an ECG should be 0.05 Hz or 0.67 Hz for filters that do not exhibit phase distortion. For high-frequency filtering they recommend a cutoff of at least 150 Hz. For the storage of digital ECG samples (at 500 samples per second), it is recommended use use compression with an error of less than 10 microvolt [4]. kligfield2007

xie2020 Xie *et al.* [5] provide an overview of the current approaches to computerized ECG analysis. The standard approach to using computerized methods in ECG analysis is comprised of four steps (1) denoising of the raw ECG signal(s), (2) feature engineering, (3) dimensionality reduction, and (4) classification. To denoise an ECG, digital filters are often used. Their drawbacks are that they only filter out very specific frequencies. Because noisy ECGs contain different types of contaminations, digital filters can be inaccurate. Using wavelet transforms for denoising has the advantage that noise can be more precisely targeted and the clean signal reconstructed afterwards. Choosing appropriate wavelet parameters can be challenging and methods to optimize this process have been proposed. Empirical mode decomposition is the third option generally employed to denoise an ECG. It does not require the user to set parameters but it can lead to a mixing of oscillations of different time scales.

After the signal has been appropriately denoised, feature engineering is performed. Feature engineering is the process of extracting features that are relevant for diagnosis from the many points the ECG signal contains. The main features targeted for extraction are the PQRST features mentioned in the introduction. The fast Fourier Transform provides a way of analysing the frequency domain of the ECG signal, enabling the detection of the QRS complex and other features. The missing time information in the fast Fourier Transform can lead to difficulties in detecting time-dependent features. The short-time Fourier Transform adds time information to the fast Fourier Transforms data. This can increase the accuracy of the feature extraction. This transform has the drawback that there is a tradeoff between the time and frequency resolutions. Wavelet transforms can also be used for feature extraction. They have the advantage that they are suitable for all frequency ranges. Choosing the right wavelet base for the desired application can be a challenge. The discrete wavelet transform is the most widely used wavelet transform, thanks to its computational efficiency. Statistical methods are also used to extract features from ECGs; those methods are generally less affected by noise in the signal.

After the features of the ECG have been extracted, it is often necessary to reduce the number of features. The reason for this is that a large number of features, despite their high accuracy, require a high amount of computation to classify. This lengthy computation can negate the advantages gained by high accuracy. This process sacrifices a certain amount of information and sometimes precision, but significantly speeds up the classification. Feature selection is a process that attempts to select a subset of the original data that adequately describes the whole data. Feature selection can

be performed by a filter that filters out unnecessary attributes based on some metric. This methods is relatively simple, but the filtering process removes data and thus negatively impacts the precision of further steps. Feature extraction on the other hand uses dimensionality reduction methods to keep as much of the original information as possible. Principal component analysis preserves as much of the variance in the original data as it can. Other algorithms focus on separating classes of data, pattern recognition, or retaining the structure of the original data.

The final stage of the ECG processing is the classification stage. In this stage judgements are made based on the prepared input data and the result should be a disease diagnosis. In the early stages of computerized ECG analysis classification was performed by algorithms based on human actions when reading an ECG. Those algorithms were basic and not particularly accurate. Currently, the classification at the end of the preparation process is performed by a machine learning algorithm. Such models include the k-nearest-neighbors model which classifies points into groups but which is very expensive to calculate for high-dimensional data. Support vector machines are used for pattern recognition and are able to work with small samples. Artificial neural networks are robust and can work with complex problems, they are generally more accurate than support vector machines. The newest approach is to forego the stages discussed here and use a single neural network to perform all the required tasks ”end-to-end”. These networks are fed raw data and the denoising, feature extraction, selection, and classification is performed internally by the model [5]. xie2020

The end-to-end approach to ECG analysis is a relatively new development and is being actively researched. The more traditional method using denoising, features engineering, and classification as separate steps is also still relevant. The combination of denoising and feature extraction with a machine learning classifier can lead to very good results. prasad2018 Prasad and Parthasarathy [11] use the fast Fourier Transform to extract features from an ECG and then employ a multi-objective genetic algorithm to detect abnormal ECG signals with high accuracy. vaneghi2012 Vaneghi *et al.* [12] compare 6 common feature extraction techniques with respect to their detection of ventricular late potentials. The compared methods are the autoregressive method, wavelet transform, eigenvector, fast Fourier Transform, linear prediction, and independent component analysis. valupadasu2012 Valupadasu and Chunduri [13] use the fast Fourier Transform to analyze the energy level in different frequencies in the ECG of patients with IHD. They find that the energy is distributed differently, allowing the distinction of ECGs with IHD from those without IHD. kaur2016 Kaur, Rajni, and Marwaha [14] analyzed ECG signals with both the wavelet transform and principal component analysis. They found that the wavelet transform outperformed principal component analysis for the detection of heart beats in an ECG. Their model achieved an error rate of 0.221% of incorrectly classified heart beats.

What are the basics of time series analysis? How does this relate to multivariate time series analysis?

History of SAX, ESAX, TSAX, SAX_TD, MSAX. How are they each used, what are their advantages, disadvantages? What were the motivations for their creation?

Review literature with respect to kNN or HOT-SAX

3 METHODS

this will lay out the methods used in this research and explain the previous information in more mathematical detail

3.1 What are time series

time series are an ordered list of measurements generally taken at regular time intervals. Insert mathematical definition here.

picture of a single lead of a time series

3.2 Multivariate time series

multivariate time series are time series that include more than one variable for each point in time.

picture of ecg with graph of 2 concurrent leads.

3.3 Where the ECGs come from

explain the data bases; how the annotation get there; how can they be used? how can a usable annotated version be extracted?

3.4 Preprocessing of the ECGs

Explain why normalization takes place, what the rationale is, and why it matters. How is it done?

Distinguish between normalization for univariate and multivariate time series.

3.5 Applying SAX and MSAX

how does PAA work? what is the idea behind it? what are the advantages and drawbacks?

How is the discretization done?

Investigate the probability distribution for the SAX segments, is it close enough to the expected normal gaussian distribution?

How are the alphabets created and how does the distance work?

example of a discretized time series, both uni- and multivariate

3.6 Choice of parameters

What are good parameters to choose for w, a, and the others? are there algorithms, or is it a guessing game?

3.7 Anomaly Detection or classification

which method is being used for detection or classification? explain it mathematically, how does it work for SAX/MSAX specifically?

should I include implementation details here? or are those irrelevant?

3.8 how to test the effectiveness of these methods

statistical conventions for testing

how can this be done in connection to the annotations in the files?

4 RESULTS

4.1 Anomaly detection or classification accuracy

make some graphs and tables that compare SAX and MSAX.

How accurate are they for different parameters?

Is my hypothesis that MSAX, by incorporating more than one variable, more accurate that SAX, even if SAX is applied to both leads of the time series and then naively connected?

How can this be proved or rather disproven?

REFERENCES

who2018 [1] World Health Organization, *Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016*. Geneva, 2018. [Online]. Available: https://www.who.int/healthinfo/global_burden_disease/estimates/en/.

iom2010 [2] Institute of Medicine (US) Committee on Social Security Cardiovascular Disability Criteria, *Cardiovascular Disability: Updating the Social Security Listings*. Washington (DC): National Academies Press (US), 2010, pp. 101–131. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK209964/.

alghatrif2012 [3] M. AlGhatrif and J. Lindsay, "A brief review: History to understand fundamentals of electrocardiography," *J Community Hosp Intern Med Perspect*, vol. 2, no. 1, 2012. DOI: 10.3402/jchimp.v2i1.14383.

kligfield2007 [4] P. Kligfield *et al.*, "Recommendations for the standardization and interpretation of the electrocardiogram: Part I: The electrocardiogram and its technology: A scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society: Endorsed by the International Society for Computerized Electrocardiology," *Circulation*, vol. 115, no. 10, pp. 1306–1324, 2007. DOI: 10.1161/CIRCULATIONAHA.106.180200.

xie2020 [5] L. Xie *et al.*, "Computational Diagnostic Techniques for Electrocardiogram Signal Analysis," *Sensors (Basel)*, vol. 20, no. 21, 2020. DOI: 10.3390/s20216318.

asilewski2012 [6] J. Wasilewski and L. Poloński, "An introduction to ECG interpretation," in *ECG Signal Processing, Classification and Interpretation*, A. Gacek and W. Pedrycz, Eds. London: Springer, 2012, ch. 1, pp. 1–20. DOI: 10.1007/978-0-85729-868-3_1.

autaharju2009 [7] P. M. Rautaharju *et al.*, "AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: Part IV: The ST segment, T and U waves, and the QT interval: Ascientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society: Endorsed by the International Society for Computerized Electrocardiology," *Circulation*, vol. 119, no. 10, e241–250, 2009. DOI: 10.1161/CIRCULATIONAHA.108.191096.

herring2006 [8] N. Herring and D. Paterson, "ECG diagnosis of acute ischaemia and infarction: Past, present and future," *QJM: An International Journal of Medicine*, vol. 99, no. 4, pp. 219–230, 2006. DOI: 10.1093/qjmed/hcl025.

physionet [9] A. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. 215–220, 2000. DOI: 10.13026/C2NK5R.

taddei1992 [10] A. Taddei *et al.*, "The European ST-T database: Standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography," *Eur Heart J*, vol. 13, no. 9, pp. 1164–1172, 1992. DOI: 10.1093/oxfordjournals.eurheartj.a060332.

prasad2018 [11] B. V. P. Prasad and V. Parthasarathy, "Detection and classification of cardiovascular abnormalities using FFT based multi-objective genetic algorithm," *Biotechnology & Biotechnological Equipment*, vol. 32, no. 1, pp. 183–193, 2018. DOI: 10.1080/13102818.2017.1389303.

vaneghi2012 [12] F. M. Vaneghi *et al.*, "A comparative approach to ECG feature extraction methods," in *2012 Third International Conference on Intelligent Systems Modelling and Simulation*, 2012, pp. 252–256. DOI: 10.1109/ISMS.2012.35.

alupadasu2012 [13] R. Valupadasu and B. R. R. Chunduri, "Identification of cardiac ischemia using spectral domain analysis of electrocardiogram," in *2012 UKSim 14th International Conference on Computer Modelling and Simulation*, 2012, pp. 92–96. DOI: 10.1109/UKSim.2012.22.

kaur2016 [14] I. Kaur, R. Rajni, and A. Marwaha, "ECG signal analysis and arrhythmia detection using wavelet transform," *Journal of The Institution of Engineers (India): Series B*, vol. 97, no. 4, pp. 499–507, 2016. DOI: 10.1007/s40031-016-0247-3.