Adell Kiani Majd

ANALYSIS OF SIMILARITY AMONG ARTERIAL BLOOD PRESSURE WAVEFORMS

UNIVERSIDADE DO ALGARVE

FACULDADE DE CIÊNCIAS E TECNOLOGIA

2016

Adell Kiani Majd

ANALYSIS OF SIMILARITY AMONG ARTERIAL BLOOD PRESSURE WAVEFORMS

Mestrado Integrado em Engenharia Eletrónica e Telecomunicações

Trabalho efetuado sob a orientação de:

Professora Doutora Maria da Graça Cristo dos Santos Lopes Ruano

UNIVERSIDADE DO ALGARVE

FACULDADE DE CIÊNCIAS E TECNOLOGIA

2016

ANALYSIS OF SIMILARITY AMONG ARTERIAL BLOOD PRESSURE WAVEFORMS

Declaração de autoria de trabalho

Declaro ser o(a) autor(a) deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

Assinatura do candidato:_____

Acknowledgment

First of all, I gratefully acknowledge the persistent support and encouragement from my advisor, Professor Maria da Graça Ruano. During my two years of Master study, she not only provided constant academic guidance to my research, she also gave me valuable suggestions on how to overcome the difficulties that I met in my life and really these few words are not enough to express my deepest gratitude.

I wish to express my deep gratitude to Professor Hamid Shahbazkia, who gave me valuable suggestions on my research. His comments on my thesis are precious and valuable discussion and ideas.

I want to thank Célia Oliveira and all mobility office team members for all the educational support during these two years' master period.

I also want to thank my parents for their efforts and supports to providing me the best possible education. Finally, I would like to thank my wife, Niloufar, for her love and patience. She has been always encouraging me during my study.

Resumo

As séries temporais são uma classe importante de objetos de dados que surgem de várias fontes e a sua análise geralmente envolve enormes quantidades de informações que exigem o uso de técnicas de mineração de dados. A medição da similaridade em séries de longo prazo desempenha um papel importante na busca por padrões semelhantes, classificação, agrupamento, previsão e descoberta de conhecimento. No contexto clínico qualquer estimativa de valores futuros baseada em seus valores passados pode ser útil no prognóstico de doenças.

Nesta tese são descritos diferentes métodos para medir a similaridade entre séries temporais de sinais de pressão arterial (ABP) e são fornecidos resultados experimentais. Para classificar um registro ABP dentro de uma classe de doenças particulares (um cluster), o procedimento típico é a determinação prévia da similaridade do registro ABP com um sinal de referência caracterizando uma doença cardiovascular (CVD) e depois, identificando a força dessa similaridade, possibilita-se uma classificação verdadeira positiva da doença (ou não). Vários métodos de mensuração da similaridade entre séries temporais são referidos na literatura, sendo os mais comumente empregados objeto desta pesquisa. Uma vez que o objetivo foi a aplicação dos resultados de similaridade para realizar agrupamento dos sinais ABP (clustering), vários métodos de similaridade foram investigados particularmente no que diz respeito ao seu desempenho ao prosseguir para a etapa seguinte de agrupamento de acordo com a patologia.

Assim, esta tese relata o uso de sete métodos de similaridade diferentes, cinco trabalhando no domínio do tempo e dois no domínio baseado em transformação, e explora o seu uso quando o clustering pelo método de Partitioning Around Medoids é implementado. Como os registros de dados são ruidosos e os sinais sofrem de variações devido a outras fontes além das do coração, seis tipos de variações foram impostas ao sinal de referência e foram testados 20 graus de possíveis variações. As séries temporais consideradas neste estudo foram de 10 segundos de duração, referindo-se a eletrocardiogramas (ECG) saudáveis, a sinais de ECG com segmentos ST de longo prazo, a ECG's relativos a fibrilação atrial e ainda a uma coleção de ECGs de diagnóstico. Foram considerados três agrupamentos, cada um envolvendo registros saudáveis e patológicos, em diferentes proporções.

Os resultados demonstram que a Transformação de Wavelet Discreta usando uma decomposição de wavelet de Haar com as transformações de Karhunen-Loève, além de reduzir a carga de processamento computacional, possibilita o agrupamento com uma precisão entre 76% e 84% entre as três classes diagnósticas consideradas.

A organização desta tese é a seguinte. Uma breve representação de séries temporais está incluída no capítulo 1. Uma breve descrição de vários métodos de similaridade e métodos de agrupamento são apresentados nos capítulos 2 e 3. As experiências realizadas e os resultados obtidos são descritos no capítulo 4. Finalmente, a conclusão deste trabalho é apresentada no capítulo 5, onde a lista de publicações resultantes desta tese está incluído.

Keywords: Séries temporais; Correspondência de dados; Medidas de similaridade; Distância Euclideana; Transformada de Wavelet; Transformada de Fourier; Coeficiente de Correlação; Distância de Mahalanobis; PAM Clustering.

Abstract

Time series are an important class of data objects that arise from various sources and their analysis typically involves huge amounts of information requiring usage of data mining techniques. Measuring similarity in long time series plays an important role in searching for similar patterns, classification, clustering, prediction and knowledge discovery. In clinical context any estimation of future values based on its past values can be useful in disease prognosis.

In this thesis different methods of measuring similarity between time series of arterial blood pressure (ABP) signals are described and experimental results are provided. To classify an ABP record within a particular diseases' class (a cluster), the typical procedure is the prior determination of the similarity of the ABP record with a reference signal characterizing a cardiovascular disease (CVD) and then identifying the strength of that similarity to enable a true positive classification of the illness (or not). Several methods of measuring similarity among time-series are referred in literature, the most commonly employed one were object of this research. Since the goal was the application of the similarity results to perform clustering of the ABP signals, similarity methods were investigated particularly in what concerns their performance when proceeding for the clustering following step.

So, this thesis reports the usage of seven different similarity methods, five working in the time domain and two in the transform-based domain, and explores their usage when clustering by Partitioning Around Medoids is implemented. As data records are noisy and signals suffer from variations due to other sources than heart, six types of variations were imposed on the reference signal and 20 degrees of possible variations were tested. The time series considered on this study were 10 seconds length, referring to healthy, electrocardiogram (ECG) long term ST's, atrial fibrillation and a collection of diagnostic ECGs. Three clusters were considered, each involving healthy and pathological records, in different proportions.

Results demonstrate that the Discrete Wavelet Transform using a Haar wavelet decomposition with the Karhunen-Loève transforms, besides reducing the computational processing load enables clustering with an accuracy between 76% and 84% among the three diagnostic classes considered.

The organization of this thesis is as follows. A short representation of Time-series is in chapter.1. A brief description of various similarity methods and clustering methods are given in chapters 2 and 3. Experiments performed and results obtained are described in chapter 4. Finally, the conclusion of this work is presented in chapter 5 where the list of publications resultant from this thesis is included.

Keywords: Time series; Data matching; Similarity measure; Euclidean distance; Wavelet transform; Fourier transform; Correlation coefficient; Mahalanobis distance; PAM Clustering.

Index

Index of Figures

Index of Tables

Abbreviations List

ABP	Arterial Blood Pressure
AF	Atrial Fibrillation
CC	Correlation Coefficient transform
Cov	Covariance
DCT	Discrete cosine Transform
DFT	Discrete Fourier Transform
DTW	Dynamic Time Warping
DWT	Discrete Wavelet Transform
ECG	Electrocardiogram
ED	Euclidean Distance
FT	Fourier Transform
GWN	Gaussian White Noise
ICU	Intensive Care Units
KLT	Karhunen-Loève Transform
MSD	Medical streaming data
MTS	Multivariate time series
PAM	Partitioning Around Medoids
PCA	Principal Component Analysis
PTB	Physikalisch-Technische Bundesanstalt
SVD	Singular Value Decomposition
HRV	Heart rate variability

Chapter 1

INTRODUCTION

1.1 Structure of the thesis

This thesis is organized into five chapters.

The present chapter, presents a general introduction and a brief representation of what time series correspond to. Second chapter exposes the methods typically employed for similarity measurements and presents an overview on each method; time domain methods and transformed- based methods are considered.

Chapter.3 is concerned with clustering and the description of Partitioning Medoids clustering method, the clustering method employed in this thesis and reported as the second experiment on chaper4.

Chapter.4 describes the experiments developed to compare the performance of similarity measuring methods and their performance while clustering data through Partitioning Medoids approach. Two experiments are reported: the first is related to the comparative assessment of similarity methods by applying different time series variations and comparing the sensitivity results among the methods; the second experiment is devoted to the evaluation of those similarity methods while integrated in the clustering strategy. This chapter starts by explaining how to use similarity measurement definitions reported in chapter.2 and presents a brief explanation about preprocessing techniques to be previously applied in the datasets. This chapter follows with a description of the PhysioNet databases that were used in the experiments. Finally, the results obtained are explained and some conclusions are drawn at the end of each experiment.

The last chapter, chapter.5, the concluding remarks related to this thesis are presented and also possible future research lines are suggested. A list of publications derived from this thesis are included at the end of the chapter.

1.2 Thesis Background

Time series are defined as ordered sequence of values of a variable at equally spaced time intervals. They are used to obtain an understanding of the underlying structure that produced the observed data and quite often to fit a model and proceed to forecasting, monitoring or even feedback and feedforward control [1].

Time series similarity measurement methods are methods of measuring the degree of similarity between two-time series. If we can work with a highly efficient method of measuring similarity and find the relationship among the time series, it will greatly increase precision of the analysis in time series databases and helps improving the accuracy and efficiency in classification, prediction and cluster analysis [2] [3].

Many researchers have devoted their time studying similarity measuring methods. Application of similarity matching algorithms is included in the main area of Univariate and Multivariate time series (MTS) analysis according to the number of variables considered to generate the data collection. These research topics are commonly used in various multimedia, medical and financial applications [4], it is one of the main subject in earthquake prediction research [5], changes' detection of vegetation indices in the land ecosystem research [6], stock prices data and money exchange rate analysis [7] [8] [9], bioinformatics [10] and medical streaming data (MSD) [11], arrhythmia detection [12] [13] and lots of other applications in sciences considering different methods of similarity measuring.

Each of these publications are based on different approaches for similarity search, in terms of working in time domain or in a reduced space of variables by means of transformed spaces like frequency representation [14]. There are many similarity and distance measuring methods, namely Dynamic Time Warping (DTW) distance [10] , Mahalanobis distance [13], transforming and Dimension reduction techniques like Discrete Fourier Transform [15] or Karhunen-Loève Transform [16], Singular Value Decomposition Transform [17], Principal Component Analysis [4] [18], Discrete Wavelet Transform (DWT) [19].

The main objective of this work is to compare the performance of different methods of measuring similarity between long time series representing heart rate variability aiming at precise and efficient cardiovascular disease clustering.

1.3 Representation of Time-series

Long Time-series data is the simplest representation of temporal data and refers to those changes of real values in time or space that resulted from being sampled at a fixed time interval. Mathematically, time series are represented as an ordered set of *m* real-valued variables $Y_t = x_{t1}, x_{t2}, \dots, x_{tm}$ each representing a value at a time point *tm* [20].

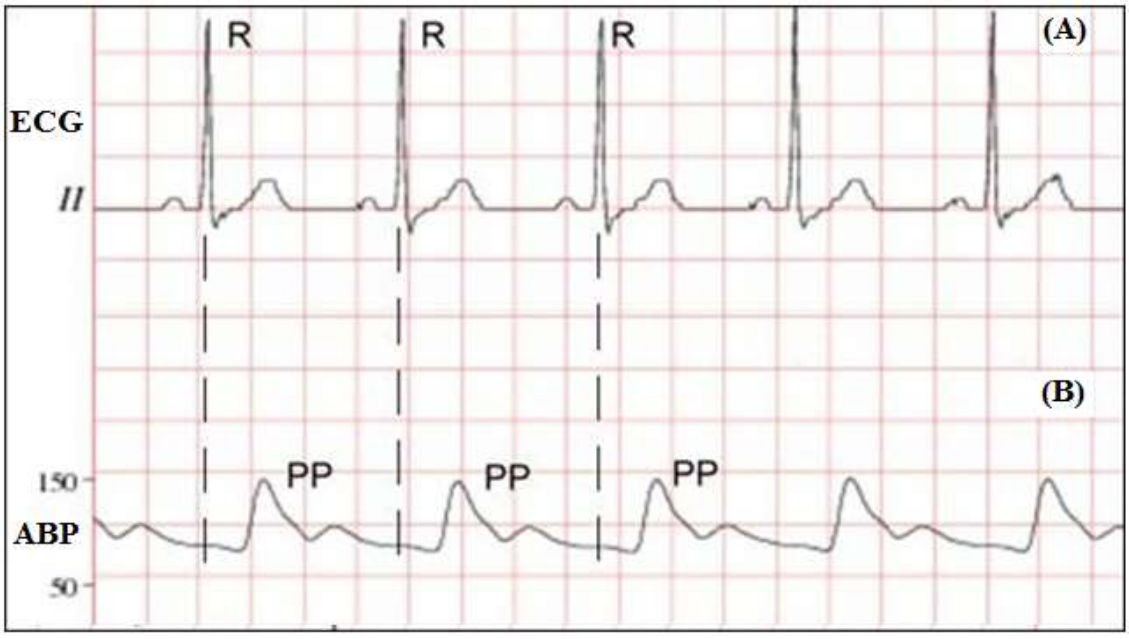


Figure 1: Representation of Time-series

(A. Electrocardiogram (ECG) signal, B. Arterial Blood Pressure (ABP) signals)

Long time series usually are so extensive and growing so fast that it becomes impossible for a single person to utilize it all effectively. Also we are typically not interested in the exact values of each time series data point so the time series analysis comprises methods capable of extracting some useful and meaningful statistics and other characteristics of the data. Time

series can be described using a variety of qualitative terms and features such as seasonal, trending, noisy, non-linear, chaotic and patterns which are contained within the data [21, 22].

Here is short explanation about the most common features of time series data or types of time series patterns that may be used for characterizing the time series [21, 23]:

1- Seasonal effect (seasonal variation or seasonal fluctuations)

Many of the time series data show a seasonal variation which is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week, or, in case of biologic signals, the heart rate) such as sales and temperature reading. This type of variation is easy to understand and can be easily measured or removed from the data. It could be defined as a pattern that repeats itself over fixed intervals of time and can be found by identifying a large autocorrelation coefficient at the seasonal partial.

2- Trend (secular trend or long term variation):

It is a longer term change in the mean level, this is, when there is a long-term increase or decrease in the data. The trend may be linear or non- linear (curvilinear).

3- Skewness:

It is a measure of symmetry, or more precisely, the lack of symmetry. It is used to characterize the degree of asymmetry of values around the mean value.

4- Kurtosis:

It is a measure of whether the data are peaked or flat relative to a normal distribution. A data set with high kurtosis tends to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case of low kurtosis.

How to effectively manage and use vast amounts of data contained in time series, the effective discovery and understanding of the data sequence and knowledge behind the law, in order to extract meaningful statistics and other data characteristics, has been more and more challenging to data mining researcher's [5], particularly thinking about the huge amount of data nowadays available.

Analysis of time series usually comes across some underlying problems, such as large volume of data, non-finite or even discrete numerical range, non-constant sampling rate, various noise interference forms [2]. So before applying any analysis techniques, some pre-processing is necessary namely normalization and noise removal.
We follow with a brief description of the background theory behind the similarity methods that will be addressed in chapter.2.

1.4 Definition of similarity measurement

In almost all research on time series concerning clustering, classification, feature extraction, trend forecasting, and decision support, the efficacy of measured of similarity between two time series plays a fundamental role.

The similarity measure $d = D(X, Y)$ between time series X and Y measures the distance d between X and Y. $D(X, Y)$ is a function of both time series (inputs) presenting as result (output) the distance d between these series. This distance has to be nonnegative, that is, $d \geq 0$. Zero distance indicates a complete match between X and Y while high value of d indicates that there is no association between the two time series.

The distance is said to be a metric, if $D(X, Y)$ satisfies the additional symmetry property $D(X, Y) = D(Y, X)$ and also the triangle inequality $D(X, Y) \leq D(X, Z) + D(Z, Y)$ [24, 25].

Different methods of calculating d will be described in chapter.2.

Chapter 2

TIME SERIES SIMILARITY MEASURING METHODS

2.1 Introduction

This chapter presents an explanation about the time series similarity approaches that will be used on future chapters. These methods were selected among many of the available ones because the main goal of the thesis is the evaluation of the similarity methods that can present better accuracy when performing clustering procedures. The main idea of this thesis is based on [26], to conclude assessing similarity methods performance; so in this thesis the similarity methods were selected to allow results comparison with [26].

Determining similarity between time series can be processed in time or in transformed domains (transform base methods). The time domain methods work with raw time series (with preprocessing step) and have less computational complexity. The transformed methods are based on transformation of the time series and have the ability to reduce the size of the signals. They also reveal more details of the signal but at the expense of higher computational burden.

2.2 Time Domain Methods

2.2.1 Introduction

The simplest algorithms for measuring similarity between time series are the time domain approaches. Within this class of methods the Minkowski and Euclidean distance (ED), Dynamic Time Warping (DTW), Correlation Coefficient transform (CC) (based on Pearson's correlation) and Mahalanobis distance, will be implemented and are below described in detail.

2.2.2 Minkowski distance

The Minkowski distance between two time series X(t) = {x(1), x(2), ..., x(N)} and Y(t) = {y(1), y(2),..., y(N)} is the length of the path connecting each pair of the points. This distance understood as a measure of similarity, should be interpreted as representing less similarity for greater distance and vice versa [27]. The most commonly used and simplest time domain distance measurements in classification approaches are derived from the Minkowski distance. Eq.1 is generally employed for both the Euclidean distance (D_{Ed}) and the Manhattan distance (D_{Man}) [6].

$$D_{Minkowski}\big(X(t), Y(t)\big) = \Big(\sum_{t=1}^{N} |X(t) - Y(t)|^{p} \Big)^{\frac{1}{p}} \tag{1}$$

In the case of p=1, Eq.1 represents the Manhattan distance and for p=2 it produces the Euclidean distance (Eq.2) characterized as being of easy usage to calculate similarity between time series of the same length [6] [28] [29].

2.2.3 Euclidean distance

As mentioned this measurement is simple to understand and easy to compute (see Eq.2). However, its major disadvantage is the fact of being heavily affected by size of the signals and sensitive to small dispersion differences, so it is important to do some preprocessing to standardize signals before proceeding with this tool. Normalization and Standardizing scores are especially important if variables have been measured on different scales [30] .

$$D_{\text{Euclidean}}\big(X(t), Y(t)\big) = \sqrt{\sum_{t=1}^{N} |X(t) - Y(t)|^{2}} \tag{2}$$

Figure.2 shows two-time series X and Y presenting different ranges of amplitude scale besides resembling similar in shape. The Euclidean distance between these two-time series will be large. To avoid this kind of problem one should apply an offset translation and

amplitude scaling, which requires normalizing the signals before applying the distance operator [22]**.**

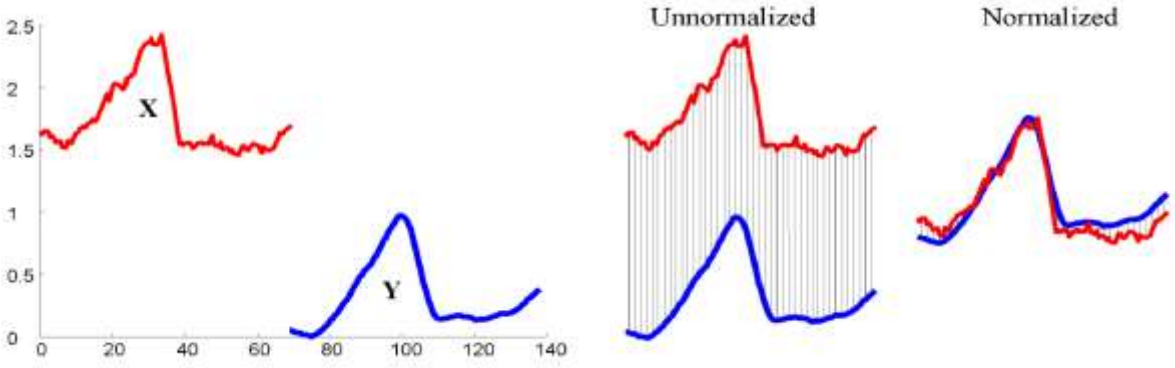


Figure 2: Necessity to normalize time series before measuring the distance between them. Two-time series X and Y have approximately the same shape, but have different offsets [22].

Even with this preprocessing step, measuring similarity with the Euclidean distance may still be unsuitable for some time series domains since it does not show similarity of two time series that are stretched or compressed. To cope with this problem in time domain, researchers suggested [31], [32] the usage of Dynamic Time Warping distance measurement (DTW).

2.2.4 Dynamic Time Warping

As mentioned, in practice, Euclidean distance has some drawbacks, such as, it does not allow different sequence length and sampling rates, shifting in time axis, even though these time series are similar to each other. Thus the Euclidean distance is difficult be directly used to solve the problem. To cope with these problem, modifications have been introduced based on the principle of Dynamic Time-warping (DTW) to allow more precise distance calculations, however it is computationally expensive [5] [27] [31] [32].

As shows in Figure.3, with this method it is possible to measure similarity of signals that are "stretched" or "compressed", so, they can be compared. The only point that should be

considered is that the time series being compared are of exactly the same dimensionality (length) [21, 33].

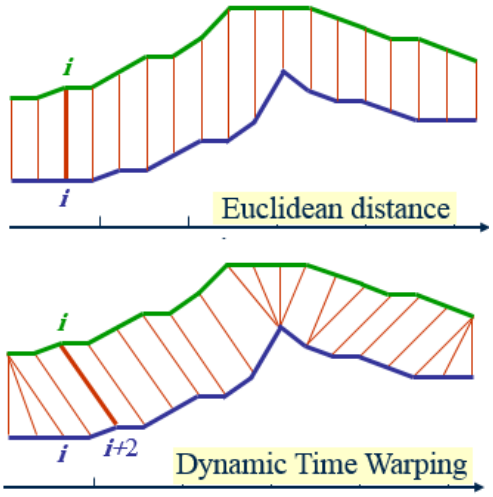


Figure 3: Dynamic time-warping Vs Euclidean distance

Dynamic time Warping distance between two time-series X and Y is defined as Eq.3.

$$D^2_{DTW}(X,Y) = D^2_{i=1:N}(X_i, Y_i) + min\begin{cases} D^2_{DTW}\left(X_i, Rest(Y)\right) \\ D^2_{DTW}(Rest(X), Y_i) \\ D^2_{DTW}\left(Rest(X), Rest(Y)\right) \end{cases} \quad (3)$$

Eq.3 is used to minimize measured distance in two similar time series with a little difference in terms of the stretching or squeezing in time axis.

Figure.4 shows two similar time series X and Y, but out of phase. To align the sequences and computing the DWT distance, we construct a $n \times m$ warping matrix. The cell (i, j) is correspond to the alignment of element x_i with y_i. First of all the distance $D(i, j)$ between each two point is calculated then the optimal warping path from cell $(0, 0)$ to $(n-1, m-1)$ (the *Rest* of the points) is calculated to find the minimum (*min*) distance, shown with solid squares in the figure. Note that the "corners" of the matrix shown are dark gray, are excluded from the search path, because the optimum distance is searched. The result of alignment is shown in Figure.4 (b) as red color path [22]. This dynamic programming

technique is impressive in its ability to discover the optimal of an exponential number alignment. Further explanation can be found in [34].

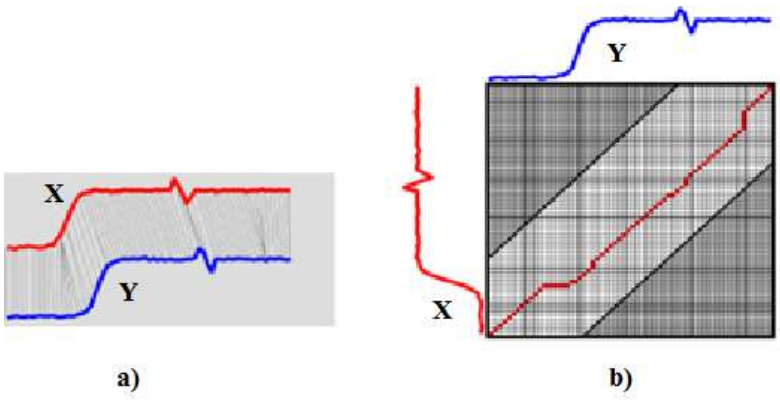


Figure 4: a) Dynamic time warping layout. b) An example of a warping path with local constraint

DTW produces a more intuitive similarity measure, allowing similar shapes to match even if they are out of phase and are not perfectly synchronized in the time axis. The main drawback of this similarity method is the time consuming effort dedicated to the calculation of the path of minimal cost but it is a good method to cope with varying lengths in Euclidean space and signals with out-of-phase similarities [35].

2.2.5 Correlation coefficient

The Pearson Correlation Coefficient (CC) is a well-known similarity measure that is invariant to shifting and scaling. Eq.4 shows the definition of Pearson Correlation Coefficient between two time-series X(t) and Y(t) [34].

$$r_{CC}(X(t),Y(t)) = \frac{\sum_{t=1}^{N}(X(t) - \mu_X)(Y(t) - \mu_Y)}{\sqrt{\sum_{t=1}^{N}(X(t) - \mu_X)^2}\sqrt{\sum_{t=1}^{N}(Y(t) - \mu_Y)^2}} \quad (4)$$

$$where \qquad \mu_X = \frac{1}{N}\sum_{i=1}^{N}(x_i) \quad and \quad \mu_Y = \frac{1}{N}\sum_{i=1}^{N}(y_i) \quad (5)$$

Where N is the length of the time series and μ is the mean value of each time series (see Eq.5) [12]. Eq.6 shows another equation commonly encountered in literature for calculating the correlation coefficient. This equation uses the Covariance (Cov) which is a measure of how much two random variables change together (see Eq.7) and also the standard deviation (σ) of X(t) and Y(t) respectively which by itself is a quantification of the amount of variation in a set of data values and can be computed by the square roots of variance (see Eq.8).

$$r_{CC}(X(t), Y(t)) = \frac{\mathrm{Cov}\,(\mathrm{X(t)}\,, \mathrm{Y(t)})}{\sigma_X \,.\, \sigma_Y} \tag{6}$$

$$\mathrm{Cov}\,\big(\mathrm{X(t)}, \mathrm{Y(t)}\big) = \frac{1}{n}\sum_{i=1}^{n}(\,x_i - \mu_X)(\,y_i - \mu_Y) \tag{7}$$

$$\sigma_X \;=\; \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\;x_i - \mu_X\;)^2} \qquad and \qquad \sigma_Y \;=\; \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\;y_i - \mu_Y\;)^2} \tag{8}$$

The correlation coefficient tells us whether the pattern of responses between time series are similar, it doesn't tell us anything about the distance between two-time series. The Pearson Correlation Coefficient range is $-1 \leq r_{CC} \leq +1$ where +1 indicates a perfectly match between two time-series and 0 indicates that there is no association between the two variables. A value less than 0 indicates a negative association this is, while one variable increases the other is decreasing.

In the case of comparing two time series, usually when the trends and evolution are intended to be evaluated, the similarity measures based on Pearson's correlation are used [35]. This method is symmetric in the sense that the correlation of X(t) with Y(t) is the same as the correlation of Y(t) with X(t).

This method presents the advantage of being unaffected by dispersion differences across variables (linear transformations) [30]. It means that multiplying a time series by a constant and/or adding a constant does not change the correlation coefficients of that time series variable with other variables [36].

2.2.6 Mahalanobis distance

The Mahalanobis distance is defined as a dissimilarity measure between two time-series with the same distribution and covariance matrix S introduced by P. C. Mahalanobis in 1936 [37]. It is defined as Eq.9:

$$D_{\text{Mahalanobis}}\big(X(t), Y(t)\big) = \sqrt{(X(t) - Y(t))^T\, S^{-1} (X(t) - Y(t))} \qquad (9)$$

The advantage of using Mahalanobis distance is that it takes into consideration the correlations, S, between the time series under study and computes the distance with respect to a base or reference point [38].
According to [35] Mahalanobis distance usually performs successfully with large data sets with reduced features, otherwise undesirable redundancies tend to distort the results.

2.3 Transformed-based Methods

2.3.1 Introduction

The most important problem of long time series is about high dimensionality, the similarity measuring of high dimensional time series is not possible based on human perception. To cope with this problem, we need dimension reduction techniques [24].

Usually reduction techniques can be used to reduce the size of the data in the time series lossless or without substantial loss of information (there might occur loss within a very small margin). Therefore, a concise and precise representation of the data is provided by these techniques. These transformations in data allow more efficient storage, transmission, visualization, and computation during the process of measuring similarity between long time series and diminish computation burden and processing complexity [22].

The Discrete Fourier Transform (DFT) is a classic data reduction technique and based on that the Discrete Wavelet Transform (DWT) is developed. Also, Singular Value Decomposition (SVD) based on traditional Principal Components Analysis (PCA) is an attractive data reduction technique [34] but which will not be addressed in this thesis.

A summary of the above referred transformed-based methods is presented in next sections.

2.3.2 Discrete Fourier Transform

In the signal processing techniques, Fourier Transform expresses a mathematical function of time as a function of frequency. The basic idea of Fourier transform is the decomposition of a signal into a time series, this is, according to the theory, any signal can be represented by the sum of an infinite number of sine and cosine basis functions, where each function is known as a Fourier coefficient. The discrete version of the Fourier Transform enables the summation to be among a finite number of terms [22, 24].

Therefore, the DFT is used to map time sequences of long time series to frequency domain enabling representation and approximation of a time series by a set of elementary basis function [34] , [39]. It is also useful to characterize the magnitude and phase of a signal.

The exponential representation of DFT could be defined as Eq.10 [34].

$$X(F) = DFT\big(X(t)\big) = \frac{1}{\sqrt{N}} \sum_{i=0}^{N-1} X(i) e^{-\frac{j2\pi Fi}{N}} \qquad F = 0,1, \dots, N-1 \qquad (10)$$

$$e^{-\frac{j2\pi Fi}{N}} = cos\left(\frac{j2\pi Fi}{N}\right) + j\, sin\left(2\frac{j2\pi Fi}{N}\right) \qquad (11)$$

From Eq.10 we can conclude that the Discrete Fourier Transform decompose periodic signals into a time-series in the frequency domain, where imaginary and real parts produce symmetric spectra. In this work, Discrete Cosine Transform (DCT) is used as a method of measuring similarity between two time series, omitting the imaginary part of the spectrum.

The first coefficients of the DFT concentrate and contain most of the time series' information and can capture a good approximation of it. As an example, Figure.5 (a) shows an Arterial Blood Pressure (ABP) signal that has been decomposed into its Fourier basis functions, the first five being shown in Figure.5 (c) to (g), and then Figure.5(b) shows the reconstructed signal when only these five basis functions are considered; As may be noticeable the reconstructed signal presents a good approximation of the original signal.

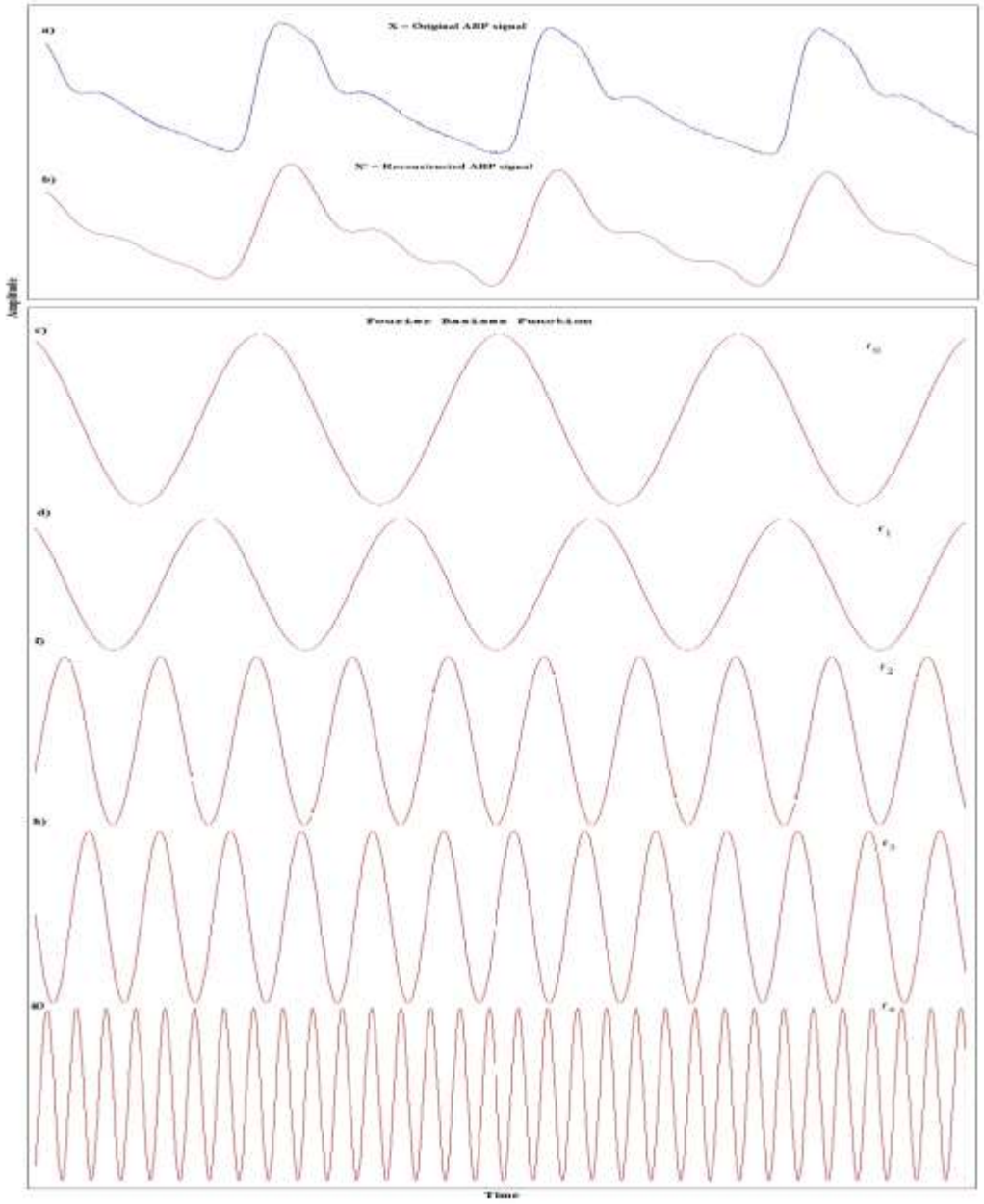


Figure 5: (a)Arterial blood pressure signal and (c) to (g) its basis functions; (b) reconstructed signal using the five basis functions presented from (c) to (g).

According to the Parseval's theorem which specify that the Fourier Transform preserve the Euclidean distance between time series in time and frequency domains, it is possible to use the first coefficients for measuring similarity of two time-series instead of using the original ones [9] [34] [15] [39].

The method of measuring similarity with DCT will be explained in detail in chapter.4.

2.3.3 Discrete Wavelet Transform

Usage of Fourier transforms comprises losing the opportunity to analyze the time domain transformations although having taken care about information preservation in the frequency domain. To represent the behavior of a time series in both domains, Wavelet based functions enable better and higher resolution in both time and frequency domains. In time domain via translations of the mother wavelet and in the frequency domain via dilations in the scale. The wavelet coefficients represent the correlation between the wavelet and a localized section of the time series. The wavelet coefficients are calculated for each wavelet segment, giving a time-scale function relating the wavelets' correlation to the signal. Unlike the Fourier transform, wavelet transforms have an infinite set of possible basis functions and provides a way of analyzing the local behavior of functions [7] [40] [41] [42].

In wavelet transforms lower frequency bands are represented in lower scales and higher frequency bands are represented in higher scales. Although wavelets can be represented in different types such as Daubechies, Symmlets or Haar wavelets. In this thesis the Haar wavelet coefficients are employed and Euclidean distance is used to measure the similarity between Haar wavelet coefficients of both time series.

Haar wavelet transforms are the most popular wavelet transformation that ensures the preservation of the Euclidean distance between any two time-series in the transformed space. For more details about wavelets, see [34].

For instance, Table.1 shows the wavelet transformation of a time series with length of 8 (in general would be with length N). The number of the steps (levels of decomposition) would be j , which can be found by setting $N = 2^j$. In case of this example it would be 3. As it

shows summation and subtraction of each coefficient is calculated it is send to the next step. This process continues until j steps are reached and the whole summation of coefficients representing the time series is achieved (in this case the last line of the table).

Resolution	Sum				Details			
	a1	a2	a3	a4	a5	a6	a7	a8
3	a1+a2	a3+a4	a5+a6	a7+a8	a1-a2	a3-a4	a5+a6	a7+a8
2	a1+a2+a3+a4		a5+a6+a7+a8		(a1+a2)-(a3+a4)		(a5+a6)-(a7+a8)	
1	a1+a2+a3+a4+a5+a6+a7+a8				(a1+a2+a3+a4)-(a5+a6+a7+a8)			

Table 1: Haar wavelet transformation process on the time series with length of 8

The whole process consist of (2^j-1) subtractions plus a summation performed recursively as shown in Figure.6. It is clear that the reconstruction of a time series is possible from this summation and subtraction actions without any loss of information regarding the established levels of decomposition;

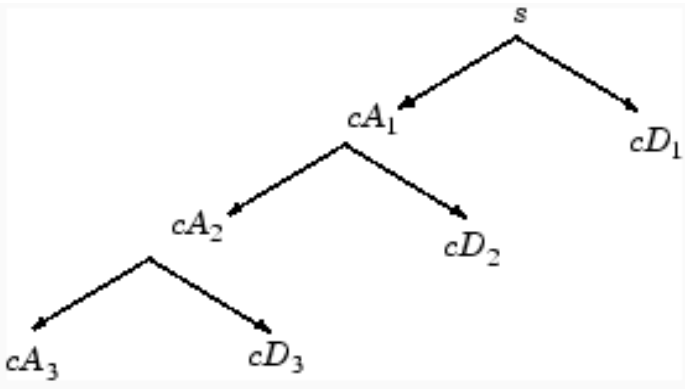


Figure 6: Decomposition of the time series in wavelet

Another explanation point of view for describing the decomposition of a signal with wavelets is to mention that two kind of filters are used. As shown in Figure.6, time series decompose into two sets of coefficients. In the first step approximation coefficients CA_1 , and detail coefficients CD_1 [43] are calculated. The low-pass filter produces the average signal, while the high-pass filter produces the detail signal. The scientific name of each step is octave. The detail signals are kept, but the higher octave averages can be discarded. The

low-pass filter applies a scaling function to a signal, while the high-pass filter applies the wavelet function. Each wavelet packet is also decomposed into two parts using the same approach as in previous octave, giving rise to CA_2, CD_2. This makes wavelet a very complete analysis and the whole binary tree is produced as shown in Figure.7.

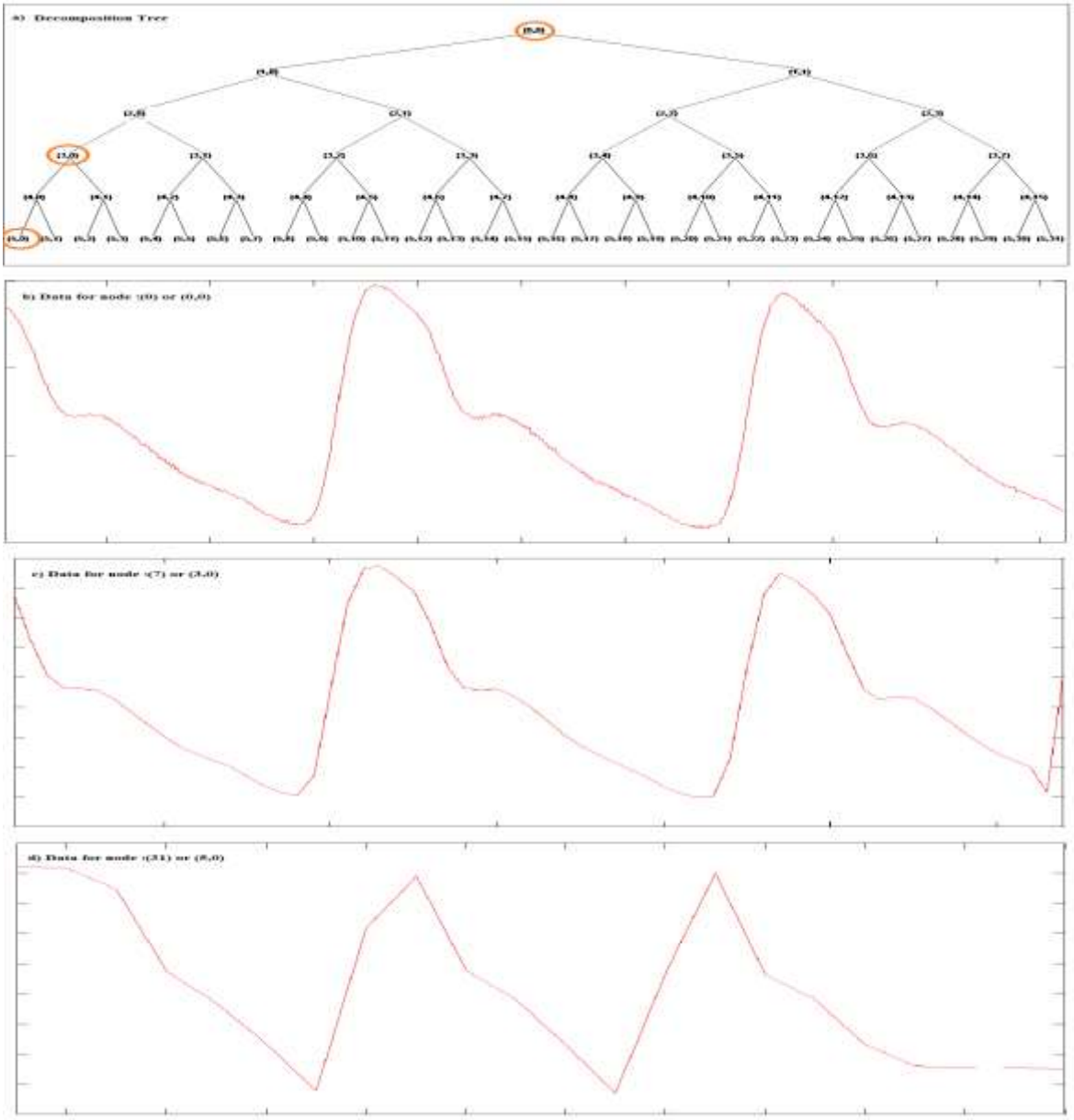


Figure 7: Decomposition Tree of ABP signal with three different resolution [45]

It shows ABP decomposition tree using Haar wavelet. It is also shown that the resolution at different steps depends on the different scale and details that resolution required. For more information about the interpretation of wavelet as filters see [44].

A time series can be decomposed into a linear combination of its basis-functions, So the signal could be approximated by different resolutions through Eq.12, as it may be seen in Figure.8 , $X'(t)$ is an approximation of the time series and its accuracy is dependent on the level of the basis functions (J) that are used to reconstruct the signal .

$$X'(t) = \sum_{j=1}^{J} \varphi_j(t) \tag{12}$$

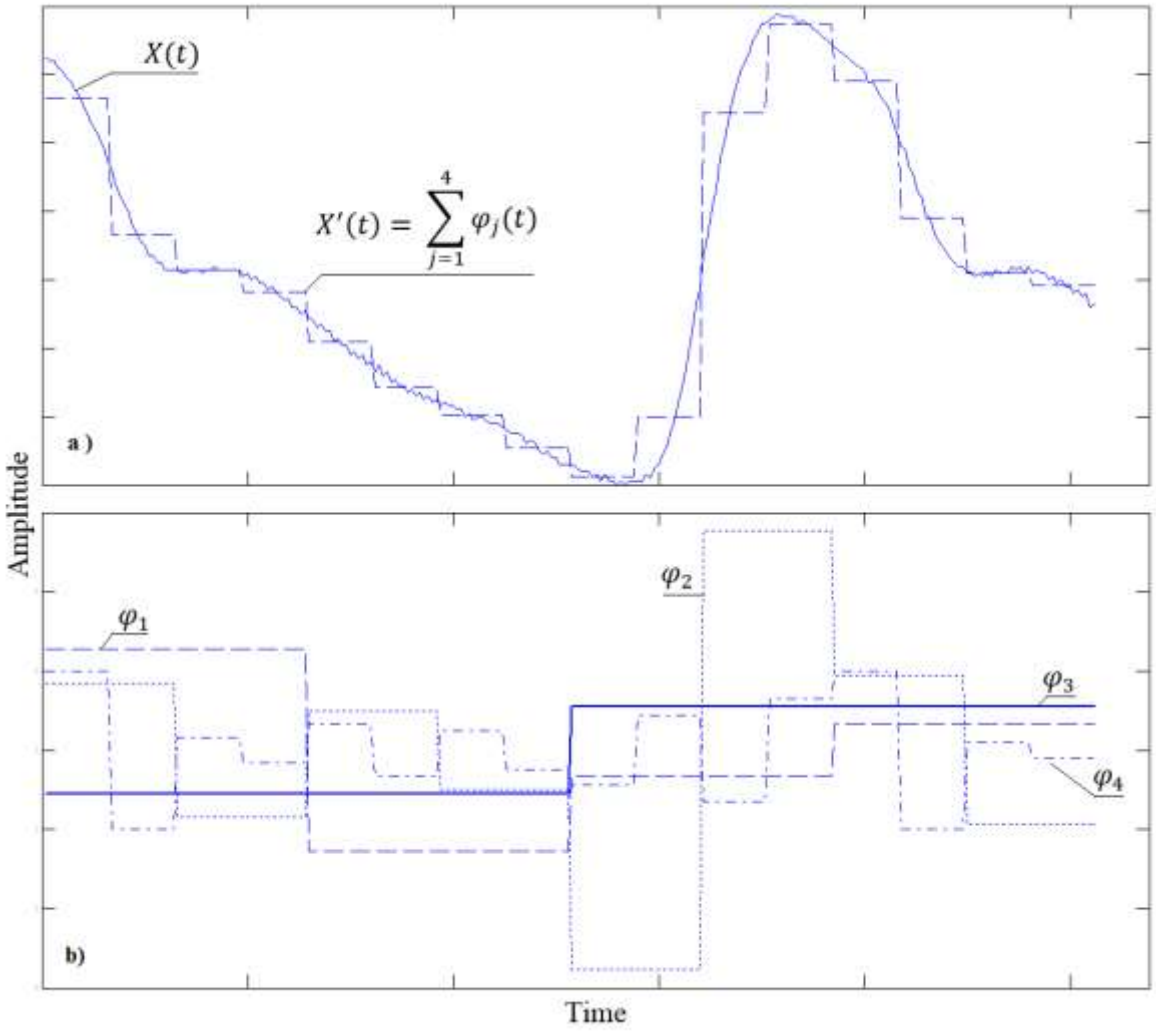


Figure 8: Signal approximation using the Haar wavelet decomposition: a) ABP signal approximation. b) Basis functions φ_j for $j = 1,2,3,4$.

The basis functions $\varphi_j(t)$ are orthogonal and generated by multiplication of the coefficients $d_j \in \mathbb{R}$ which are scalers with the different orthogonal wavelet basis $\psi_j(t)$ such that (Eq.13)

$$\varphi_j(t) = d_j\, \psi_j(t) \tag{13}$$

The broad trend of the input function is captured in approximation of the original function $\phi(t)$, plus localized changes which are kept as set of detailed functions ranging from coarse to fine $\psi(t)$. If we consider, $\varphi_1(t) = C_{0,0}\phi_{0,0}(t)$ and J as level of decomposition and $j = \log_2 N$; then DWT could be described by Eq.14 and signal X(t) can be approximately represented as a linear combination of N basis functions [19].

$$\tilde{X}_J(t) = C_{0,0}\phi_{0,0}(t) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t) \qquad 1 \le J \le N \tag{14}$$

As a data reduction technique it is possible in Discrete Wavelet Transform approximation to keep the most significant DWT coefficients for measuring similarity of two time-series. T. Rocha et al. [29] proposed interpretable similarity measure to evaluate the similarity between time series by combining the Haar wavelet decomposition with the Karhunen-Loève transforms (KLT) in order to optimum reduce the number of wavelet basis. [26]. In this thesis the same approach has been performed to determine the similarity measurement between two time series and a detailed explanation is included in chapter.4.

The multiresolution aspect of the wavelet transform provides a time-scale decomposition of the signals allowing their visualization and a more accurate clustering of the data into homogeneous groups [3] [19]. Wavelet transform also have some drawbacks. They are only defined for sequences whose length is an integral power of two and also they involve a more complex computational implementation.

Chapter 3

CLUSTERING

3.1 Introduction

Clustering is one of the most frequently used data mining techniques, particularly for similarity search amongst long time series [30]. The objective of cluster analysis is to partition a set of objects and group the subsets into two or more clusters based on the similarity between the analyzed time series. So clustering strategies are challenging tools within several research areas as long as they include similarity search of sequences. The combined methods are also used to recognize dynamic changes in time series. One may list several research works in this area, to be mentioned as examples the approaches provided in [21] [19] [20] [46] [47] [48].

Recalling the aim of this thesis, the comparison of different measuring similarity methods, the experiment of clustering is going to checks the effect of similarity measures in the application of clustering for discovering the accuracy of each method and this could be useful in purposes of Heart rate variability (HRV) diagnosis. All the experiments should be performed with different similarity measurement techniques to identify those who are able to produce more accurate clustering results. Among the published clustering methods the Partitioning Around Medoids (PAM) clustering is going to be used in this thesis.

Next section concentrates on the theory behind clustering which will be implemented in this thesis and detailed in chapter.4.

3.2 Clustering

Clustering is a sort of classification procedure that categorizes all the time series in the study dataset into groups, called *clusters*; the particularity of these classification is that these clusters are not predefined and they are defined by the data itself, based on the similarity between time series. The most important objective is to find the similar clusters that are as distinct as possible from other clusters. In other words, this grouping should maximize inter

cluster variance (maximum similarity inside the cluster) and minimize intra cluster variance (the clusters themselves should be very dissimilar [22].

P. Esling et al. [24] presented a very clear definition of clustering:

Definition: "Given a database of time series $T = (t_1, \ldots, t_n)$ and a similarity measure $\boldsymbol{S}(X,Y)$, find the set of clusters $C = \{c_i\}$ where $c_i = \{ T_j' \mid T_j' \in \boldsymbol{S}_T^n \}$ is a set of subsequences that maximizes inter cluster distance and intra cluster cohesion."

Figure.9 depicts two possible outputs of a clustering algorithm. It can be seen in this figure that the main difficulty in clustering usually is defining the correct number of clusters. It shows two possible outputs from the same clustering system obtained by changing the number of required cluster N=3 and N=8. This difference is because of the way of initializing and selecting the parameters and the level of detail targeted [24].

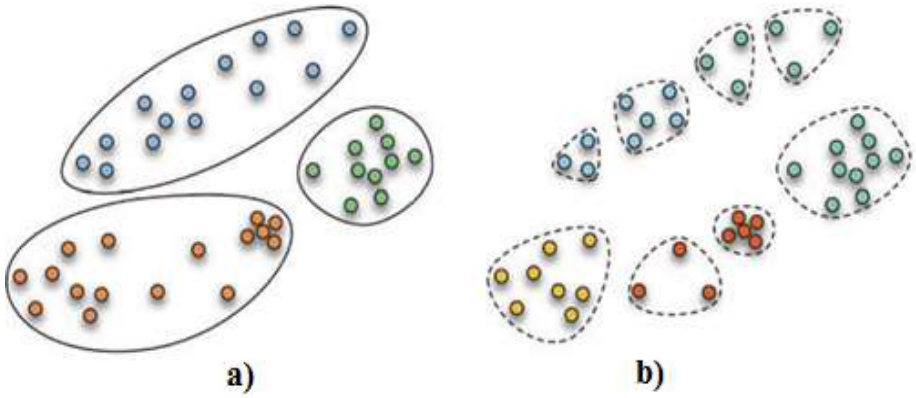


Figure 9: Possible outputs of a clustering algorithm: a) defining 3 clusters, or b) 8 clusters [24]

3.3 Partitioning Around Medoids (PAM)

As explained in previous section the aim of clustering analysis is to partition a set of objects in data base into two or more clusters such that objects within a cluster are similar and objects in different clusters are dissimilar. As also mentioned, for enabling comparison with previously published work in this area, the Partitioning Around Medoids (PAM) is the clustering strategy to be considered.

PAM is based on the search for *k* representative objects, called medoids, among the objects of the dataset. The *medoid* of a cluster is defined as the object for which the average of dissimilarity between objects near that medoid is minimum; in this case a cluster has been identified. If *k* clusters are desired, *k* medoids are found. Once the medoids are found, the data are classified into the cluster of the nearest medoid [49].

Algorithm of Partitioning Around Medoids attempts to minimize the total distance *D* between objects within each cluster. Mathematically *D* can be computed as Eq.15.

$$D = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j \in C_k} d_{ij} \tag{15}$$

where *K* is the total number of clusters, d_{ij} is the distance between objects *i* and *j*, and C_k is the set of all objects in cluster *k* [49].

"The algorithm proceeds through two phases. In the first phase, a representative set of *k* objects is found. The first object selected has the shortest distance to all other objects. That is, it is in the center. An addition *k-1* objects are selected, one at a time, in such a manner that at each step, they decrease *D* as much as possible. In the second phase, possible alternatives to the *k* objects selected in phase one are considered in an iterative manner. At each step, the algorithm searches among the unselected objects for the one that, if exchanged with one of the *k* selected objects, will lower the most the objective function. The exchange is made and the step is repeated. These iterations continue until no exchanges can be found to provide lower values of the objective function. Note that all potential swaps are considered and that the algorithm does not depend on the order of the objects on the database [49]."

Finding dissimilarity (distance) between two time series is fundamental to cluster analysis since the goal is to place similar objects in the same cluster and dissimilar objects in different clusters. The objective of this thesis is comparing different similarity methods so we used different methods in clustering analysis and the output results are compared with predefined datasets. In chapter.4 a more detailed description of the experiments developed under this thesis are explained.

Chapter 4

SIMILARITY MEASURE ANALYSIS

4.1 Introduction

As mentioned before measuring similarity is important for classification, clustering and any analysis. Since clinical signals are random processes with non-stationary characteristics, one is interested in determining disease clustering but having in mind the possibility of personalized variations that may occur between different cardiac cycles of the same patient and among different patients' time series data representation. So, the clustering strategy to be applied in this thesis, should account for possible variations of the time series without loss of general classification on a same cluster.

Therefore, a primary experiment was undertaken to force variations on the template time series and test the influence of increasing variations on the similarity measurements, this is, we wanted to test the sensitivity of the similarity measurement method to different levels of variations.

Secondly, the partitioning around Medoids clustering was assessed for various types of similarity methods, searching for the most robust similarity method that could cluster different CVD among the testing time series. The concept of clustering robustness hereby involved is in the sense of identifying the pair similarity measurement versus PAM that will enable clustering with less sensitivity to possible time series variations. Results are then confronted with the first experiment results to confirm the most suitable strategy for CVD diagnosis [6] [19] [29] [31] [50].

4.2 Implementation of similarity measuring algorithms

Similarity measuring similarity methods apply to pairs of time series. One time series is a template with which we want to measure the similarity to the other time series. The template

time series is forced to different degrees of variations, to be applied on amplitude and on time dimensions.

The template time series corresponded to Arterial Blood Pressure data collected from the public data base PhysioNet [51]. The similarity of these two time series is calculated according to the methods and equations described in chapter.2.

Measuring similarity with Minkowski (p=6) and Euclidean distances' methods, Correlation Coefficient, Mahalanobis distance and Dynamic Time Warping Distance methods is exactly the same as already explained in chapter.2 and the correspondent thereby included. However, the procedure of measuring similarity with the Discrete Fourier Transform and Discrete Wavelet Transform needs deeper explanations below stated.

i. Discrete Fourier Transform

As explained in the chapter.2, The Fourier Transform (FT) decompose time series into imaginary and real parts as two symmetric spectrums. In this work goes for the real part of the signal. Discrete Cosine Transform is the real part of the FT and for a time series with length of N, X(t)= $\{x_1, x_2, \dots, x_N\}$ is derive from a simplified form of Eq.10 that is shown in Eq.16.

$$X'(t) = p(t) \sum_{k=1}^{N} C_k \cos\left\langle \frac{\pi(2k-1)(t-1)}{2N} \right\rangle \qquad t = 1, \dots, N \qquad (16)$$

In Eq.16, the parameters C_k are scale factors of the cosine wave and $p(t)$ is a normalization coefficient factor that could defined as Eq.17:

$$p(t) = \begin{cases} \frac{1}{\sqrt{N}} & , \ t = 1 \\ \sqrt{\frac{2}{N}} & , \ 2 \le t \le N \end{cases} \qquad (17)$$

For measuring similarity of two time series $X(t)$ and $Y(t)$ based on DCT coefficients, the first *m* coefficients could represent a good approximation of time series so this distance could be a good measure of similarity . The template signal, $X(t)$,and the added variation

signal, $Y(t)$,are decomposed into DCT coefficients and the similarity is measured according to Eq.18.

$$D_{DCT}\big(X\,(t), Y(t)\big) = \sqrt{\sum_{k=1}^{m}\Big(C_{k\,X} - C_{k_Y}\Big)^2} \tag{18}$$

This distance could be the same as Euclidean distance if we consider all coefficients (m=N).

In this thesis, the number of coefficients chosen to reconstruct the signal were selected to guarantee a 90 percent of accuracy. First of all, the absolute coefficients are ranked in descend order then the number of basis that could be used in the reconstruction of the signal with the 90 percent approximation are chosen (almost always the first m=4 coefficients).

ii. Discrete Wavelet Transform

Discrete Wavelet Transform decompose time series into the basis functions as explained in chapter.2. On this thesis the similarity measuring is based on the combination of wavelet transform with the Karhunen-Loève transform, which is an optimal dimension reduction method that could guaranty a minimal reconstruction error.

In measuring similarity with DWT, the distance between time series is measured but the reduced number of coefficients are considered according Karhunen-Loève theorem. In this method the time series decompose into the basis functions which are orthogonal to each other. Those are obtained as eigenvectors of the covariance matrix composed of the wavelet basis [52]. The approximation of the signal is obtained by reducing the number of basis that have been employed in the similarity measuring instead of reducing the signal, This reduction is obtained from the first highest *J* eigenvalues of the correspondent covariance matrix [26].

Firstly, the template time series , $X_1(t)$ with the length N, is decomposed into a linear combination of N wavelet basis $\varphi_j(t)$,as Eq.19 .

$$X\,(t) = \sum_{j=1}^{J}\varphi_j(t) \tag{19}$$

Then, the second time series (variation of added signal), $Y(t)$ with the same length of N, also decomposed into the same wavelet basis $\varphi_j(t)$, Eq.20.

$$Y(t) = \sum_{j=1}^{J} \alpha_j \, \varphi_j(t) \tag{20}$$

In Eq.20, the coefficients α_j are always existent and could be derived from Eq. 21 [26].

$$\alpha_j = \frac{\langle Y(t), \varphi_j(t) \rangle}{\langle \varphi_j(t), \varphi_j(t) \rangle} \tag{21}$$

As in Fourier Transforms, the distance of this wavelet coefficients could show similarity of two time series and could be described as Eq.22

$$D_{DWT}\big(X\,(t), Y(t)\big) = \sqrt{\sum_{j=1}^{J} \big(1 - \alpha_j\big)^2} \tag{22}$$

In this distance equation if all set of basis are consider (J=N), the result would be the same as the Euclidean distance. The most important advantage of this method is to reduce data noises and unnecessary parts of the signal.

In this thesis the length of all the signals is set to N=1024 so the appropriate J=4 is chosen for our experiments to achieve the accuracy of 92% in the approximation.

In chapter.2, all the equations for similarity techniques are about the distance between two time series. Eq 23 is used for measuring similarity for all methods in the same exponential scale.

$$S\big(X\,(t), Y(t)\big) = \; e^{-D\left(X\,(t),\; Y(t)\right)} \tag{23}$$

In this equation *S* is the similarity function which in the case of perfect similarity ($D = 0$) would be 1 and in the case of dissimilarity ($D \to \infty$) it would be almost zero.

4.3 Preprocessing of the datasets

In the real life, all the time series collected from sensors or devices are subject to noise and artifacts. The first and most important step of each signal processing experiment is overcoming this problem by means of performing some preprocessing. It could increase the quality of data before running any analysis, data may be transformed into a format that is more easy and effective to be processed. The signals should be prepared by analyzing them carefully to prevent misleading results. Preprocessing includes Noise filtering, normalization, transformation, feature extraction and data selection. Usually noise filtering can be handled by using traditional techniques like digital filters or wavelet thresholding.

Another issue to take into account is concerned the scaling differences between time series. In the case of this thesis, since the range of amplitude values of the raw data (in ECG and ABP signals) varies widely and similarity functions are based on the distance between time series, different ranges of amplitude would produce erroneous results. This problem can be overcome by a linear transformation of the amplitudes.

By performing normalization of all the signals in the dataset, all measured values are adjusted in the common scale. This process called feature scaling or unity-based normalization is typically used to bring all values into the range [0,1]. Eq. 24 describes the rescaling method applied [52].

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{24}$$

Another preprocessing method is removal vertical offsets. Eq.25 shows this preprocessing step when $\bar{X}$ represents the mean value of the time series.

$$X' = X - \bar{X} \tag{25}$$

In the preprocessing of acquired signals in the dataset, if one or more signals have a different sampling frequency, it is recommended to resample them to obtain the same sampling frequency. This is performed order to obtain series of the same length with the same frequency [24].

4.4 Datasets acquisition

As mentioned in chapter.1, all biomedical signals used in this thesis were collected from PhysioNet database [51]. PhysioNet offers free web access to large collections of recorded physiologic signals (PhysioBank) many of them including clinical annotations.

"PhysioBank is a large and growing archive of well-characterized digital recordings of physiological signals and related clinical data for use by the biomedical research community. It currently includes databases of multiparameter cardiopulmonary, neural, and other biomedical signals from healthy subjects and from patients with a variety of illnesses with major public health implications [51]."

As also previously mentioned, two experiments were implemented to measure robustness and sensitivity of similarity measuring methods when applied to biomedical time series, and also, to measure the accuracy of each method. All the algorithms were implemented using the Matlab software [53].

In the experiment of evaluating sensitivity of the similarity measurement methods described in section 4.5, the template signal selected was an Arterial Blood Pressure signal which was collected from MIMIC.II database.

MIMIC.II - Multiparameter Intelligent Monitoring in Intensive Care, is a multi-parametric dataset including time series of vital signs collected from bedside patient monitors in the intensive care units (ICUs) and contains detailed clinical information for many of the patients presented in the Waveform Database [54].

In the experiment of evaluating accuracy of clustering (as described in section 4.6) four groups of ECG signal were collected from the following PhysioNet databases [54]:

Fantasia Database which is related to twenty young (21 - 34 years old) and twenty elderlies (68 - 85 years old) healthy subjects. Both young and elderly were resting while continuous electrocardiograms (ECG) were recorded. All subjects remained in a resting state in sinus rhythm while watching the movie *Fantasia* (Disney, 1940) to help maintain wakefulness. The sampling frequency in this database is 250 Hz but in some of the signals were 333Hz.

The Long-Term AF Database includes 84 long-term ECG recordings of subjects with paroxysmal or sustained atrial fibrillation (AF). The sampling frequency in this database is 128 Hz.

The Long-Term ST Database contains 86 lengthy ECG recordings of 80 human subjects, chosen to exhibit a variety of events of ST segment changes. The sampling frequency in this database is 250 Hz.

The PTB Diagnostic ECG Database contains 549 records from 290 subjects (aged 17 to 87, mean 57.2; 209 men, mean age 55.5, and 81 women, mean age 61.6 with different heart diseases. PTB is an abbreviation for Physikalisch-Technische Bundesanstalt , the National Metrology Institute of Germany, which has provided this digitized ECGs for research. The sampling frequency in this database is 1000 Hz.

4.5 Experiment for evaluating sensitivity of similarity measuring methods

4.5.1 Introduction

A similarity measuring method should be capable of detecting similarity between time series although some variations may occur in signals, this is, it should be invariant to transformations and distortions. It should recognize two similar signal even though they are not mathematically identical. The main goal of this experiment is to measure robustness of each similarity methods when reference signal suffers variations [29].

4.5.2 Variations in time series

The biomedical time series may have different type of variation, in terms of adding noise, scaling or translation in time or amplitude, and also changes in baseline. The similarity measure $S\left(X_1(t), X_2(t)\right)$ in Eq.23 should be robust to any combinations of these transformations. For this experiment, Arterial blood pressure (ABP) signal is used, obtained

from the PhysioNet dataset [54] as explained in section 4.3 . A template signal X(t) representing the 10 sec ABP recording of healthy people is used. This template signal was defined to be the average of 9 cardiac cycle segments randomly collected from the ABP database.

The aim is to evaluate sensitivity and robustness of each similarity method to different distortions. The distortions tested can be represented by the following equations [24]:

(a) Amplitude Scaling: $X_{As}(t) = \beta * X(t)$, (26)

(b) Amplitude Translation: $X_{At}(t) = X(t) + \beta$, (27)

(c) Time Scaling: $X_{Ts}(t) = X(\beta * t)$, (28)

(d) Time Translation: $X_{Tt}(t) = X(t + \beta)$, (29)

(e) Baseline variation : $X_B(t) = rotate(X(t), \theta)$, (30)

(f) Adding Noise: $X_{Wgn}(t) = X(t) + \mathcal{N}(X_i, N)$. (31)

Where β is a constant and Θ is the angle of rotation in the baseline and $\mathcal{N}$ is White Gaussian Noise (WGN). The values of β and Θ employed were based on experiments were small incremental changes were envisaged, therefore 20 possible variations of the series are considered; An example of a possible distortion imposed on a single cardiac cycle of the template ABP signal is depicted in Figure.10.

4.5.3 Experimental results and analysis

According to the equations and methods explained in previous chapters, similarity of signals measured by S_{Ed} (Euclidean distance), S_{DWT} (Discrete Wavelet Transform), S_{FT} (Discrete Fourier Transform), S_{CC} (Correlation Coefficient), S_{Mah}(Mahalanobis distance), S_{Mi} (Minkowski Distance), S_{DTW} (Dynamic Time Warping Distance) were computed considering increasing levels of distortions. Figure.11 presents for each distortion tested a joint representation of all similarity measurement method performance when the 20 levels of distortions are applied.

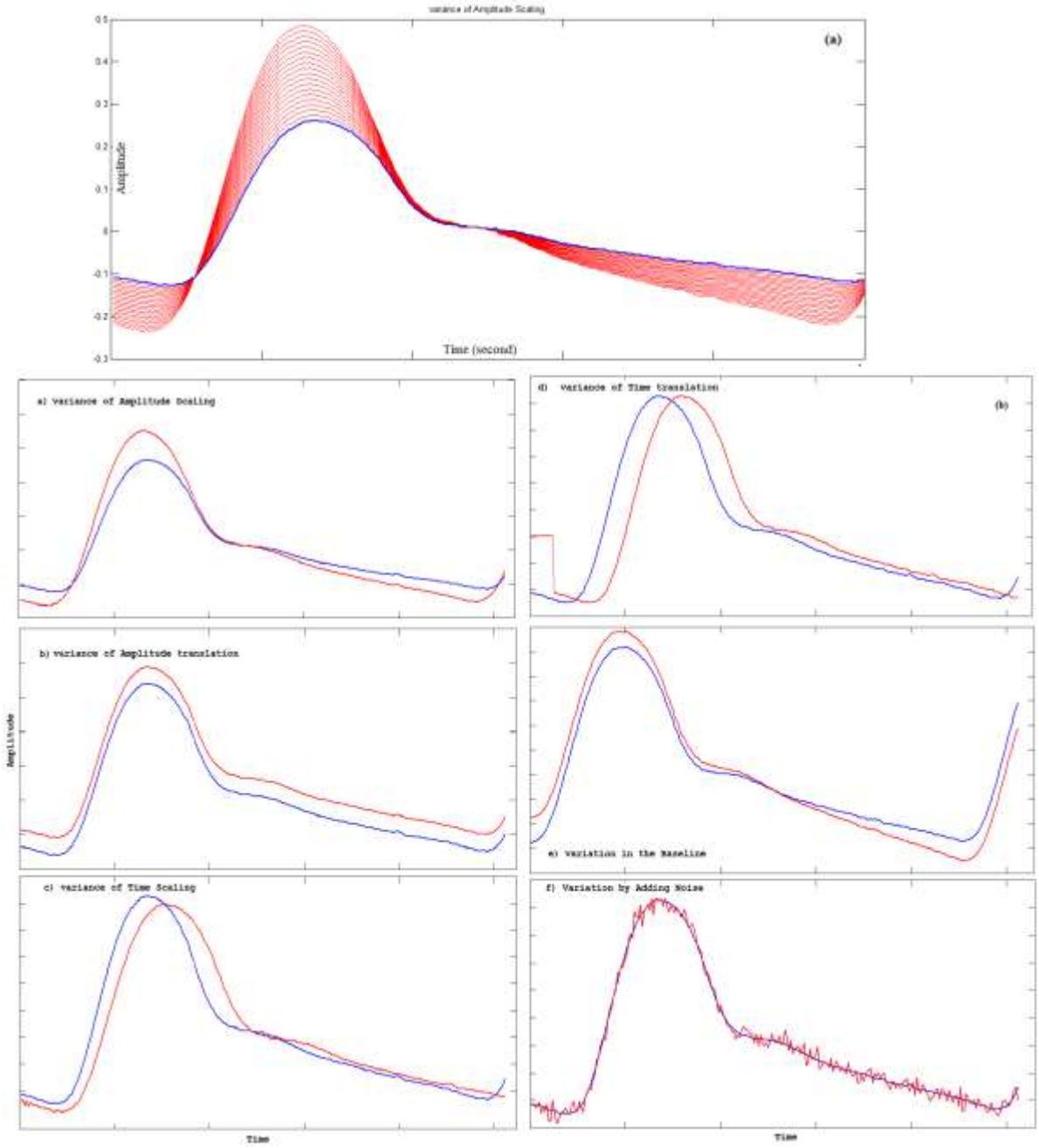


Figure 10: Applying different types of variation on the ABP Time series; a) all 20 steps of amplitude variation together, b) one step at all types of variation

(just on one heart cycle as Template)

As expected and depicted on Figure.11, the similarity measured values decrease when the level of distortion (abscissa values) is increased. The sensitivity to each type of variation is not the same for all the similarity methods as can be seen by the different trend of the curves.

Curves close to similarity value of one represent strong similarity, and those which are maintained close to the horizontal line valued one are less sensitive to variable distortions, therefore they present high robustness.

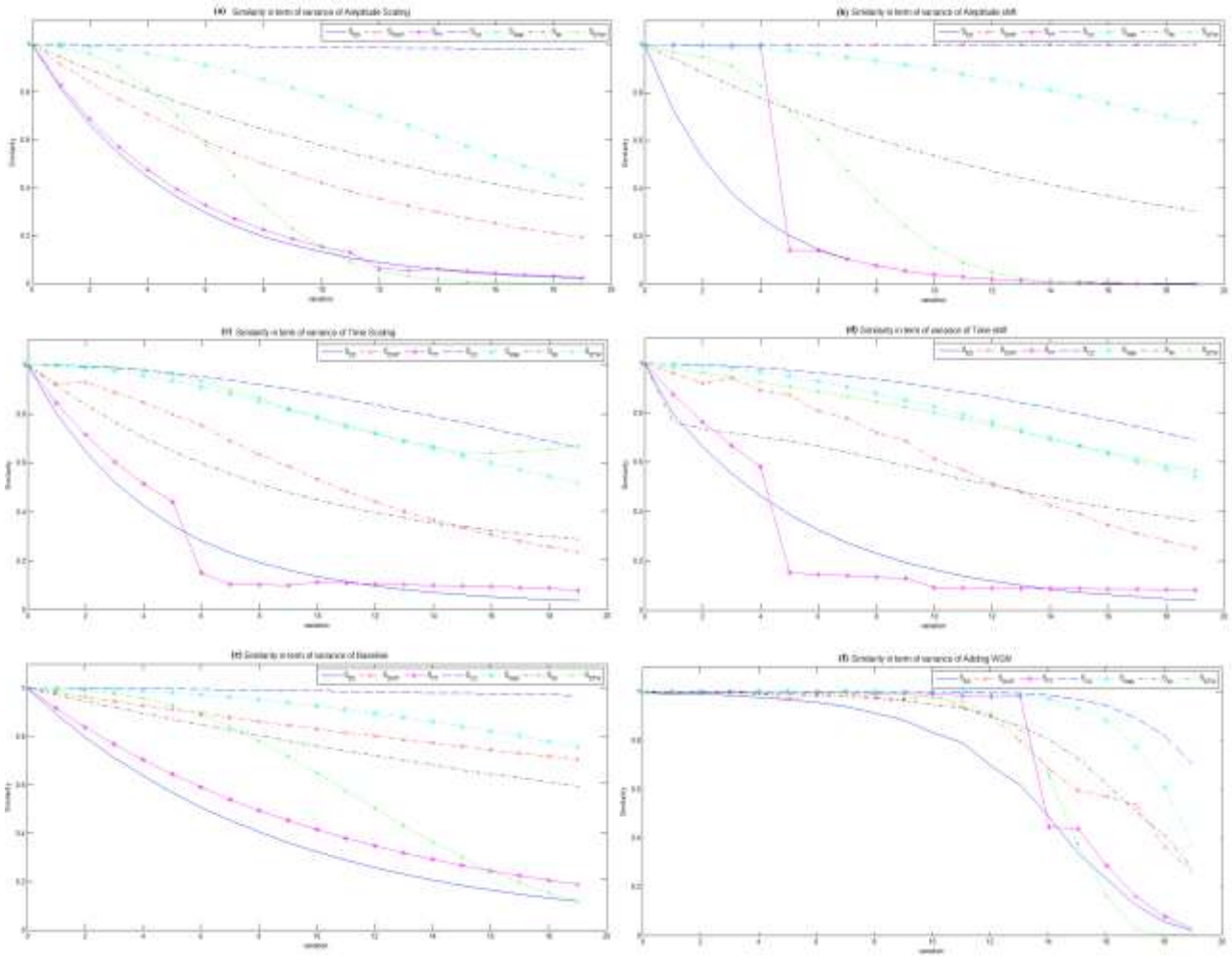


Figure 11 : Similarity measurement methods' sensitivity metrics against degree of distortions for the template signal variations: a) Amplitude Scaling, b) Amplitude Shift, c) Time Scaling, d) Time shift, e) Variation of baseline, f) Variation by Adding white Gaussian Noise. The figures' caption nomenclature stands for: S_{Ed} - *Euclidean distance,* S_{DWT}-*Discrete Wavelet Transform,* S_{FT} -*Discrete Fourier Transform,* S_{CC}-*Correlation Coefficient,* S_{Mah}-*Mahalanobis distance,* S_{Mi}-*Minkowski Distance,* S_{DTW}-*Dynamic Time Warping Distance.*

The similarity measured by Correlation coefficient displays lowest sensitivity to all the signals' variations tested and keeps its trend close to the unity similarity line through all degree of distortion imposed denoting lowest sensitivity in all cases.

The Euclidean distance and Fourier transform showed highest sensitivity to the distortions since as degree of distortions increase the similarity measurement decays rapidly. The Fourier transform being particularly critical when amplitude shift, time scale and time shift variations are tested. The similarity measured by the Euclidean distance demonstrates higher unpredictability results between two time-series whenever variations increase.

The Mahalanobis distance showed a performance closer to the one presented by the Correlation coefficient and much better than the performance of the Euclidean distance or the Minkowski Distance. This is due to the fact that, similarly to the Correlation Coefficient, the Mahalanobis distance equation also takes into consideration the correlation of the data set itself.

The Minkowski Distance proved to be only admissibly insensitive when White Gaussian Noise and variance of the baseline are the data variations considered.

The Discrete Wavelet Transform response is almost insensitive to amplitude shift, closely following the performance of the Correlation coefficient and it has better robustness response, this is, low sensitivity to other variations.

The Dynamic Time Warping Distance presented a performance similar to the Mahalanobis distance when time scaling is considered and is also almost insensitive to White Gaussian Noise when up to 13 degrees of distortion are allowed.

Thinking about the main objective of this thesis, the identification of a similarity method which could enable an efficient clustering of ABP time series, and knowing the individual behavior of the tested similarity methods, next research step will be confirming these results while applying clustering methods to time series (section 4.6).

4.5.4 Conclusion

A comparative study of different time series similarity methods has been performed. Since our target is CVD diagnosis applications, the study considered the well-known and frequently used and referenced PhysioNet data-base, making use of healthy patients' arterial blood pressure signals. We experimentally demonstrated that among the tested time domain

similarity measurement methods the Correlation Coefficient was the most robust method, that is to say, the most insensitive to small distortion, presenting similarity measurements close to unity for amplitude scaling, amplitude shifting, variance of the baseline and additive white Gaussian noise. Concerning the transformed-based similarity methods tested, the Discrete Wavelet Transform performed better than the Discrete Fourier Transform. It is insensitive to amplitude shifts of the signals, almost insensitive to white Gaussian noise up to 14 variations, but for the other variations it is clear a robustness decay as the degree of variation increases.

To conclude, in what consists an election of the most robust similarity method the Correlation Coefficient wins. However, when data reduction is required due to computational burden of the whole system, the Discrete Wavelet Transform as proposed in (Rocha (2014) [29]) is the similarity measurement approach to be elected.

To be also mentioned that the detailed experiments hereby reported are useful to identify the most suitable similarity method to be applied on long time series when the time series main characteristics are known in advance. For instance if a researcher is going to deal with time series that are essentially corrupted by noise, the similarity method to be applied should be the Pearson Correlation Coefficient or the Mahalanobis distance, baut never the Discrete Fourier transform.

Next section addresses the influence of selecting these similarity methods when clustering efficiency is envisaged.

4.6 Experiments for accuracy evaluation of PAM Clustering with various similarity measuring methods

4.6.1 Introduction

As mentioned in the chapter.3, clustering is one of the most frequently used data mining techniques. The objective of cluster analysis is to partition a set of objects into two or more clusters based on the similarity between the analyzed time series. In this thesis Partitioning Around Medoids is employed. PAM is based on the search for k representative objects, called

medoids, among the objects of the dataset. If the average of dissimilarity between objects near a medoid is minimum, a cluster is identified.

4.6.2 Datasets and Clustering performance evaluation metrics

To test the clustering performance 4 groups of ECG signals were generated. All the signals were obtained from the PhysioNet databases [54] as explained in details in section 4.4**.**

For the first Group, 40 related to healthy subjects signals from Fantasia Database were gathered, for the second group 84 signals randomly were pick up from The Long-Term AF Database, for the third group, 85 signals from The Long-Term ST Database were used and for the fourth group 71 signals from The PTB Diagnostic ECG Database were employed as can be summarized in Table.2.

	Number of signals		
PhysioNet Data Base	***Available***	***Randomly selected***	***Signals' characteristics***
Fantasia	40 (2h records)	40	Healthy subjects
MIT-BIH Atrial Fibrillation	25 (10 h records)	84	Atrial Fibrillation
Long-Term ST	86 (21-24 h records)	85	ST level drifted signals
PTB Diagnostic ECG	549 (variable record length)	71	Diagnostic ECG (different pathologies)

Table 2: Dataset acquisition

To be mentioned that, as stated in Table.2 the original data bases included more signals than those employed in this study. The selection of 40 time series of 10 seconds length from the Fantasia data base involved random selection of these 10 seconds records. The same strategy was employed on the selection of the working time series from the other data bases.

Using these time series three testing collection were composed were healthy and diseased patients' records were grouped, as summarized in Table.3.

	Two-class clusters:		
Data Base	***Collection 1***	***Collection 2***	***Collection 3***
Fantasia	30	30	30
MIT-BIH Atrial Fibrillation	45		
Long-Term ST		55	
PTB Diagnostic ECG			50

Table 3: Data Base collections

The first collection included 30 healthy 10 seconds length time series and 45 time series belonging to patients with atrial fibrillation. So, the clustering algorithm should differentiate a specific illness among 75 time series.

The second collection, included 30 time series of healthy patients (not exactly the same 30 time series of collection 1) and 55 time series (also 10 seconds length) from Long-Term ST data base. In this case, clustering strategy was tested against another time of time series characteristics than those encountered in collection 1.

The third collection was also composed of 30 time series of healthy patients but now the 50 PTB time series randomly selected might include different pathologies since the PTB data base is composed of diagnostic ECG signals.

It is expected that PAM clustering will be able to differentiate the healthy from the diseased records within each collection, and, through computation of the clustering performance for each similarity method employed on previous experiments one can conclude about the most effective and robust similarity method to be employed on CVD clustering.

To assess the performance of clustering precision and efficiency of each similarity method should be analytically computed [55]. Gavrilov et la [56] proposed a cluster similarity metric as defined by Eq.32,33 :

$$Sim(G_i, A_j) = 2\,\frac{|G_i \cap A_j|}{|G_i| + |A_j|} \tag{32}$$

It computes a cluster similarity metric based on the G_i "ground-truth", this is, the predefined members of each datasets and A_j representing the clustering results obtained by using PAM with various types of similarity method. Numerator of Eq.32 introduces the number of correct similar time series A_j that are recognized out of a predefine dataset G_i.

$$Sim(G, A) = \frac{\sum_i \max_j Sim(G_i, A_j)}{k} \tag{33}$$

Eq.33 computes the accuracy of the clustering results. *A* cluster out of the *G* groups where k is the number of clusters considered [3], [56]. This metric will be zero if two clustering are completely dissimilar and 1 if they are similar.

4.6.3 Clustering experiment results

The result of Sim in Eq.33 will be 0 if clustering results are completely dissimilar and 1 if the clustering results are similar to the established ground-truth. To clarify, if for instance within collection 1, 15 out of the 30 records (healthy patients) and 45 of the 45 atrial fibrillation records were detected as similar we would have an accuracy of 75%. Here we are designating 'accuracy' as the precision of correctness clustering of the data under analysis.

Each of the datasets is clustered using various type of similarity methods and Eq.33 is computed to obtain the clustering results as stated in Table.4.

The clustering accuracy results in the Table.4 reveals that Discrete Wavelet Transform provide the most accurate clustering on the selected time series for all collections tested with a clustering accuracy ranging from 72.7 to 77.6.

Similarity methods used in clustering	Accuracy of clustering in Percent		
	Collection1	*Collection2*	*Collection3*
Euclidian distance	49.44	49.54	66.13
Auto correlation coefficient	**64.06**	**72.02**	**68.53**
Discrete Cosine Transform	49.44	49.54	**69.26**
Discrete Wavelet Transform	**72.79**	**77.60**	**73.09**
Dynamic time warping	49.44	49.54	66.13
Mahalanobis distance	52.68	63.36	66.13
Minkowski metric(P=6)	**66.85**	**75.70**	58.97

Table 4: Comparison of clustering accuracy with different similarity measuring methods

In an attempt to improve the accuracy of the clustering procedure another experiment was performed. At the pre-processing stage, all the time series were aligned among all collections according to their first peak location (see Figure.12).

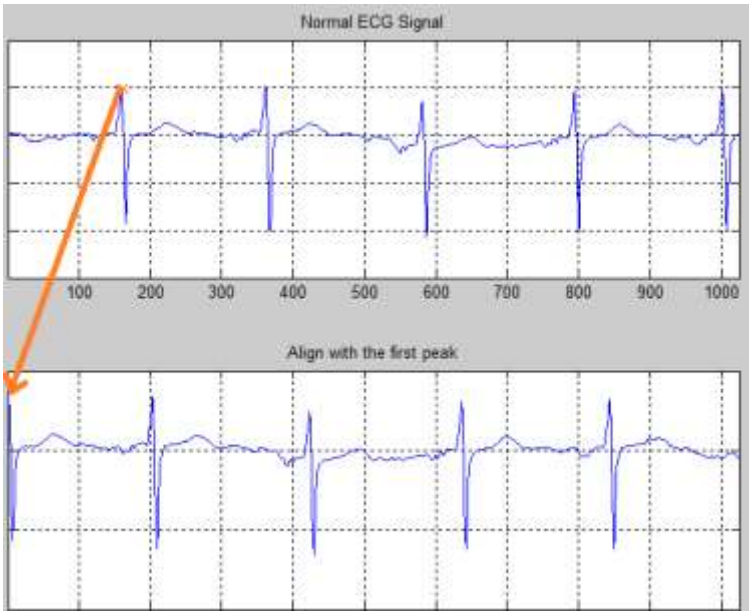


Figure 12: Align signals in datasets according first peak

After this preprocessing step the same algorithms were applied. Table.5 shows the modified results obtained with this additional preprocessing stage.

Similarity methods used in clustering	Accuracy of clustering in Percent		
	Collection1	*Collection2*	*Collection3*
Euclidian distance	**77.22**	**81.60**	69.87
Auto correlation coefficient	66.60	67.05	**74.21**
Discrete Cosine Transform	**77.22**	77.92	**72.97**
Discrete Wavelet Transform	**75.86**	**83.93**	**80.20**
Dynamic time warping	**77.22**	**81.60**	69.87
Mahalanobis distance	**76.13**	**81.60**	69.87
Minkowski metric(P=6)	71.40	65.03	68.12

Table 5: Comparison of clustering accuracy with different similarity measuring methods (time series first peak alignment)

Now we can observe that accuracy has in fact increased, for the Discrete Wavelet Transform the range of accuracy is now between 75.8 and 83.9, but at the same time some other similarity methods became more accurate.

In what concerns collection 1 can see from Table.5 that Dynamic Time Warping, Euclidian Distance and Discrete Cosine Transform present higher accuracy (77.2) than the Discrete Wavelet Transform (75.8); the next higher accuracy is obtained with Mahalanobis distance (76.1). So the majority of the highest rated accuracies were obtained through time domain methods. In fact, if we compare the ECG signals from healthy patients and those with atrial fibrillation (Figure.13) one can see that the signals are not much different, so the time domain similarity measurements easily compute the similarity differences.

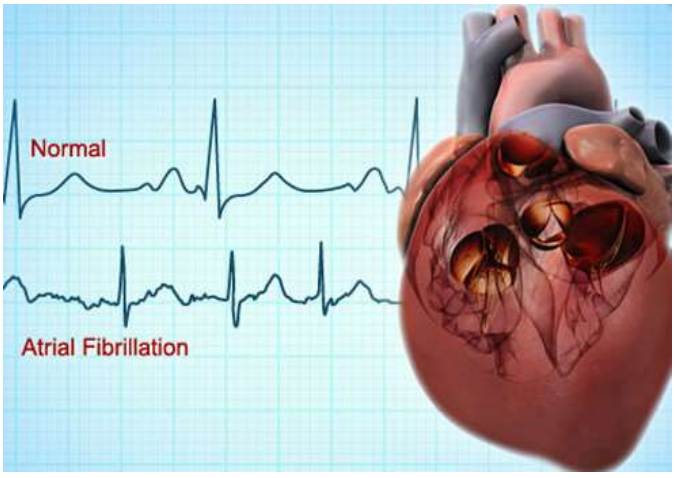


Figure 13: Comparison of normal and atrial fibrillation cardiac cycles ABP signals [57]

In collection 2; the healthy ECG signals is compared with ST variation signals (see Figure.14) that contains variety of events of ST segment changes, including ischemic ST episodes, axis-related non-ischemic ST episodes, episodes of slow ST level drift, and episodes containing mixtures of these phenomena and can see from Table.5 that Discrete Wavelet Transform (83.93) present higher accuracy than Dynamic Time Warping, Euclidian Distance and Mahalanobis distance (81.60), the next higher accuracy is obtained with Discrete Cosine Transform (77.92). It can see in results that those signals which have more distortion distinguished better with DWT.

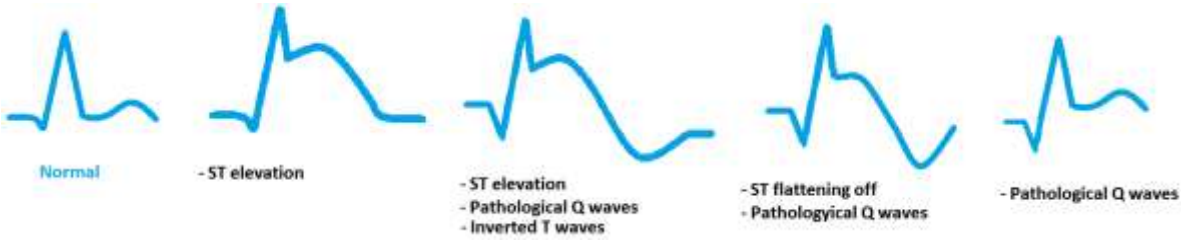


Figure 14: Variation in ST segment [58]

In collection 3; the healthy ECG signals is compared with variety of distortion in signals and can see from Table.5 that Discrete Wavelet Transform (80.20) present higher accuracy than Auto Correlation Transform (74.21), the next higher accuracy is obtained with Discrete Cosine Transform (72.97). In this collection also results shows that those signals which have more distortion distinguished better with DWT.

The above presented clustering results strength the previously obtained results when addressing the efficiency of similarity measurement techniques previously obtained.

4.6.4 Conclusion

In this chapter the similarity measuring method results obtained were validated by a clustering algorithm. The objective of these experiments was to group similar time series according to specifically predefined datasets and to compare the clustering results with predefined groups. In this process by testing different similarity methods inside the PAM clustering the accuracy of each method is measured. It gives us a measure about the level of success and correctness reached by the algorithm.

Analysis of these results reveal that DWT provides the most accurate clustering particularly when the variability of signals occurs (collections 1 and 2). Results obtained for collection 1 evidence that when inside the clustering members exist more similarity among signals (only healthy and atrial fibrillation signals) Euclidian distance related measurements may be more accurate. To be mentioned that if during the preprocessing stage the alignment of the records' first peaks was not performed, the DWT accuracy obtained for collection 1 would be better than any other method. The results improved when an additional preprocessing step is applied.

The datasets were extended as may be seen in Table.6 to enable more experiments for being able to define an accurate final conclusion with the same strategy of evaluating accuracy of clustering. Table.7 depicts the results of all of the experiments and it shows that the DWT provides the best results in this datasets.

	Two-class clusters:						*Three-class clusters:*
Data Base	***Collection 1***	***Collection 2***	***Collection 3***	***Collection 4***	***Collection 5***	***Collection 6***	***Collection 7***
Fantasia	30	30	30				
MIT-BIH Atrial Fibrillation	45				50	50	50
Long-Term ST		55		45		50	50
PTB Diagnostic ECG			50	50	50		50

Table 6: Extended datasets: 6 two-class clustering and 1 three-class clustering

Similarity methods used in clustering	**Accuracy of clustering in Percent**						
	Collection1	***Collection2***	***Collection 3***	***Collection4***	***Collection 5***	***Collection6***	***Collection7***
Euclidian distance	**77.22**	**81.60**	69.87	68.03	**78.57**	65.13	54.84
Auto correlation coefficient	66.60	67.05	**74.21**	**74.89**	67.85	**71.27**	54.34
Discrete Cosine Transform	**77.22**	77.92	**72.97**	66.03	**77.76**	68.60	**57.11**
Discrete Wavelet Transform	**75.86**	**83.93**	**80.20**	**75.86**	**78.75**	**75.73**	**57.54**
Dynamic time warping	**77.22**	**81.60**	69.87	68.03	**78.57**	65.13	54.84
Mahalanobis distance	**76.13**	**81.60**	69.87	68.03	76.74	65.13	54.84
Minkowski metric(P=6)	71.40	65.03	68.12	**71.86**	70.04	**69.93**	**57.26**

Table 7: Comparing accuracy within 7 different datasets

Chapter 5

CONCLUSION AND FUTURE WORKS

5.1 Concluding Remarks

In this thesis a comparative study is performed on the sensitivity and robustness of the various similarity methods in confronting with variation and distortion that may occur in time-series experiments. The aim was to find the appropriate similarity measure for long time–series to achieve the best efficiency in clustering and classification.

We would like to emphasize that the key step in this type of time series data mining endeavor always lies in choosing the right methods dependent on the particular signals and variation existing for that experiment. This means that the similarity measuring method chosen for clustering purposes will depend on the signal itself and the possible variation it suffers. Sometimes the judgment is centered on the signal trend so resolution of approximation is not so important but it may either be concerned with measuring similarity based on the signals' dynamics at specific points in time and in this case more accurate resolution is required.

In case of our specific datasets, we experimentally demonstrate that Discrete Wavelet Transform combined with Karhunen-Loève transforms displayed the most accurate results among the commonly employed time-series similarity measurement methods in terms of accuracy in clustering long time-series. Results also proved that Discrete Wavelet Transform combined with Karhunen-Loève transforms are particularly robust when different types of datasets are considered within the collection under clustering analysis.

It is better to say, even by this achieved results is not reasonable to conclude that one similarity measure is better than the others. We can conclude that measuring the similarity in long time series is dependent on the situation and the goals in the research, presenting different performance in different cases. One particular method could be appropriate for one research and not good for other one.

5.2 Future work

In this work similarity methods are compared to achieve good clustering performance on long time-series. In the future, a deeper study on the length of the time-series capable of maintaining the performance already achieved should be investigated.

It is also envisaged to enlarge the study to other clustering goals, for instance to identify from the heart rate variability sleep patterns.

Another aspect requiring research is comparing the obtained results with alternative clustering strategies, namely using neural network-based classification methods.

In all cases, a common goal is envisaged, to find similarities (or not) between a patient's signal from his past clinical records and a currently collected signal to conclude about the patient's health evolution aiming at predicting future health trend.

5.3 Publications derived from the thesis

The results obtained from this research work are written in the form of conference papers and submitted to 20th IFAC World Congress, (IFAC WC 2017) [58], and also to BHI2017 - International Conference on Biomedical and Health Informatics [59]. Also a journal article is being prepared enhancing the selection of similarity measurement techniques for long time series clustering purposes.

REFERENCES

[1] E. S. Handbook, "Introduction to Time Series Analysis," NIST/SEMATECH e-Handbook of Statistical Methods, [Online]. Available: http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm. [Accessed 2016 10 15].

[2] Y. Jiang, T. Lan, D. Zhang, "A New Representation and Similarity Measure of Time Series on Data Mining," *Int. Conf. Computational Intelligence and Software Engineering,* p. 1–5, 2009.

[3] K. Kalpakis, D. Gada, V. Puttagunta, "Distance measures for effective clustering of ARIMA time–series," *Proc. IEEE Int. Conf. on Data Mining,* pp. 273-280, 2001.

[4] K. Yang, C. Shahabi, "A PCA-based Similarity Measure for Multivariate Time Series," *Proc. of the 2nd ACM international workshop on Multimedia databases,* pp. 65-74, 2004.

[5] L. Wei, Z. Hua, Q. jianfeng, L. chen, J. Afang, "Based on time series similarity matching algorithm for earthquake prediction research," *3rd Int. Conf. on Advanced Computer Theory and Engineering (ICACTE) ,* p. pp 57, 2010.

[6] S. Lhermitte, J. Verbesselt, W.W. Verstraeten, P. Coppine, "A comparison of time series similarity measures for classification and change detection of ecosystem dynamics," *Remote sensing of environment ,* vol. 115, p. 3129–3152, 2011.

[7] Z. R. Struzik, A. Siebes, "Measuring Time Series' Similarity through Large Singular Features Revealed with Wavelet Transformation," *Database and Expert Systems Applications, Proc.10th Int. Workshop on,* p. 162–166, 1999.

[8] D. Rafiei, "On Similarity-Based Queries for Time Series Data," *15th Inter. Conf. on Data engineering,* pp. 410-417, 1999.

[9] D. Rafiei, A. Mendelzon, "Similarity -Based Queries for Time Series Data," *In Proc. of Int. conference ACM SIGMOD Management of data,* pp. 13-25, 1997.

[10] E. Tsiporkova, E. Kostadinova, V. Boeva, L. Boneva, "An Integrative DTW-based Imputation Method for Gene Expression Time Series Data," *6th IEEE Int. Conf. Intelligent Systems,* pp. 258-263, 2012.

[11] J. Bernatavičien, G. Dzemyda, G. Bazilevičius,V. Medvedev, V. Marcinkevičius, P. Treigys, "Method for Visual Detection of Similarities in Medical Streaming Data," *Int. journal of computers communication &control (IJCCC),* vol. 10, pp. 8-21, 2015.

[12] C.C. Chiu, T.H. Lin, B.Y. LIAU, "Using correlation coefficient in ECG waveform for arrhythmia detection," *Biomed. Eng. Appl. Basis Commun,* vol. 17, pp. 147-152, 2005.

[13] Y. Yeh, "An Analysis of ECG Beats by Using the Mahalanobis Distance Method," *IEEE, 4th Int. Conf. on Innovative Computing, Information and Control,* pp. 1460-1463, 2009.

[14] H.Ding, G.Trajcevski, P.Scheuermann, X.Wang, E.Keogh, "Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures," *PVLDB ,* no. ACM 978-1-60558-306-8/08/08, pp. 1542-1552, 2008.

[15] R.Agrawal, C. Faloutsos, A. Swami, "Efficient Similarity Search in Sequence Databases," *4th Int. Conf. on Foundations of Data Organization and Algorithms,* pp. 69-84, 1993.

[16] F. Castells, P. Laguna, L. Sornmo, A. Bollmann, J.M. Roig, "Principal Component Analysis in ECG Signal Processing," *EURASIP Journal on Applied Signal Processing,* p. 98, 2007.

[17] M. Bandarabadi, M. Karami, A. Afzalian, J. Ghasemi, "ECG denoising using Singular Value Decomposition," *Australian Journal of Basic and Applied Sciences 4(7):2109-2113,* July 2010.

[18] L. Karamitopoulos, G.Evangelidis, "PCA-based Similarity Search: Pre-processing & Distance Measures," *In Proc. 2nd Int. Scientific Conf. The Contribution of Information Technology to Science, Economy, Society and Education,* pp. 318-327, 2007.

[19] A. Antoniadis, X. Brossat, J. Cugliari, J.M. Poggi, "Clustering functional data using wavelets," *Int. Journal of Wavelets, Multiresolution and Information Processing,* p. 30, 2011.

[20] H.G. Koh, W. Loh, S.W. Kim, "An Efficient Subsequence Matching Method Based on Index Interpolation," *Inno. in Applied Artificial Intelligence, Lecture Notes in Computer Science,* vol. 3533, pp. 480-489, 2005.

[21] K. A. S. R. H. D. A. X. Wang, "A Scalable Method for Time Series Clustering," Tech Report, Department of Econometrics and Business Statistics Monash University, Victoria, Australia, 2004.

[22] C. Ratanamahatana, J. Lin, D. Gunopulos, E. Keogh, "MINING TIME SERIES DATA," *Data Mining and Knowledge Discovery Handbook,* pp. 1049-1077, July 2010.

[23] Rob J Hyndman, G. Athanasopoulos , Forecasting: principles and practice, Online book : https://www.otexts.org/fpp/6/1, October 17, 2013 .

[24] P. Esling, C. Ago,, Time-Series Data Mining, ACM Computing Surveys, November 2012. DOI = 10.1145/2379776.2379788..

[25] H. Yin, H. Qi, J. Xu, W.N.N. Hung, X. Song, "Generalized Framework for Similarity Measure of Time Series," *Mathematical Problems in Engineering,* p. 12, 2014.

[26] T. Rocha, S. Paredes, P. Carvalho, J. Henriques, "An Efficient Strategy for Evaluating Similarity between Time Series based on Wavelet / Karhunen-Loève Transforms," *Int. Conf. of the IEEE Engineering in Medicine and Biology,* p. 6216–6219, 2012.

[27] A.C.C. Yang, A.L Goldberger, C.K. Peng, "Information-Based Similarity Index," *http://physionet.org/physiotools/ibs/doc/.*

[28] H. Pree, B. Herwig, K. David, P. Lukowicz, "On general purpose time series similarity measures and their use as kernel functions in support vector machines," *Information Sciences,* vol. 281, p. 478–495, 2014.

[29] T. Rocha, S. Paredes, P. Carvalho, J. Henriques, "Assessing the similarity between time series using a Wavelet transform: application and interpretability aspects," *IEEE-EMBS Int. Conf. on Biomedical and Health Informatics (BHI),* p. 652 – 655, 2014.

[30] D. A. Field, "Postgraduate Statistics: Cluster Analysis," 2000.

[31] X.Dong, Cheng-Kui Gu, Zheng-Ou Wang, "Research on shape-based Time series Similarity measure," *Int. Conf. on Machine Learning and Cybernetics,* p. 1253–1258, 2006.

[32] V. Megalooikonomou, Q. Wang, G. Li, C. Faloutsos, "A Multiresolution Symbolic Representation of Time Series," *21st Int. Conf. on Data Engineering,* pp. 668-679, 2005.

[33] E.J. Keogh, M.J. Pazzani,, "Scaling up Dynamic Time Warping for Datamining Applications," *the sixth ACM SIGKDD international conference on Knowledge discovery and data mining,* pp. 285-289, 2000.

[34] Shasha and Zhu, D. Shasha, Y. Zhu, High performance discovery in time series: techniques and case studies, New York: Springer, ch.2.

[35] F. Iglesias, W. Kastner, " Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns," *Energies (ISSN 1996-1073),* pp. 579-597, 2013.

[36] D. Lane, D. Scott, M. Hebl, R. Guerra, D. Osherson, H. Zimmer, "Introduction to Statistics online edition," [Online]. Available: http://onlinestatbook.com/. [Accessed 1 8 2016].

[37] P. C. Mahalanohis, "On the generalized distance in statistics," *In Proc. National Institute of Science, India,* vol. 2, pp. 49-55, 1936.

[38] K. Sidek, I. Khalil, "Biometric Sample Extraction using Mahalanobis Distance in Cardioid Based Graph using Electrocardiogram Signals," *Annual Int. Conf. of the IEEE EMBS San Diego,* pp. 3396-3399, 2012.

[39] W. Wei, Time Series Analysis: Univariate and Multivariate Methods,2nd edition, Ch. 11, 2006.

[40] V. N. Kopenkov, "Efficient Algorithms of Local Discrete Wavelet Transform with Haar-Like Bases," *Pattern Recognition and Image Analysis,* vol. 18, no. 4, pp. 654-661, 2008.

[41] T. Rocha, S. Paredes, P. Carvalho, J. Henriques, M. Harris, "Wavelet based Time Series Forecast with Application to Acute Hypotensive Episodes Prediction," *Int. Conf. of the IEEE Engineering in Medicine and Biology,* p. 2403–2406, 2010.

[42] I. Popivanov , R.J. Miller, "Similarity Search Over Time-Series Data Using Wavelets," *Proc. IEEE Int. Conf. Data Eng,* pp. 212-221, 2002.

[43] A.R AL-QAWASMI, KH. DAQROUQ, "ECG Signal Enhancement Using Wavelet Transform," *WSEAS TRANSACTIONS on BIOLOGY and BIOMEDICINE ,* p. 7, 2010.

[44] M. Weeks, M. Bayoumi , "Discrete Wavelet Transform: Architectures, Design and Performance Issues," *Journal of VLSI Signal Processing,* vol. 35, p. 155–178, 2003.

[45] Mathworks, "Wavelet Packets," Matlab, [Online]. Available: http://www.mathworks.com/help/wavelet/ug/wavelet-packets.html?refresh=true&s_tid=gn_loc_drop. [Accessed 19 10 2016].

[46] K. a. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, ch. 2, 1990.

[47] X. Wang, K. Smith, R. Hyndman, "Characteristic-Based Clustering for Time Series Data," *Data Mining and Knowledge discovery archive,* vol. 13, no. 3, p. 335–364, 2006.

[48] A. Struyf, M. Hubert, P. J. Rousseeuw, "Clustering in an Object-Oriented Environment," *Journal of statistical software,* 1997.

[49] "NCSS data analysis, Medoid Partitioning," NCSS Statistical software, [Online]. Available: http://www.ncss.com/software/ncss/clustering-in-ncss/. [Accessed 20 july 2016].

[50] MathSoft, S-PLUS 4 Guide to Statistics, Data Analysis Products Division, Ch. 18, Seattle, 1998.

[51] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. Peng , H. Eugene Stanley, "PhysioBank, PhysioToolkit, and PhysioNet:Components of a New Research Resource for Complex Physiologic Signals," PhysioNet, [Online]. Available: http://dx.doi.org/10.1161/01.CIR.101.23.e215. [Accessed 05 09 2016].

[52] T. Rocha, S. Paredes, P. Carvalho, J. Henriques, "Trend Prediction Methodology Based on Time Series Similarity Analysis and Haar Wavelet Decomposition," *2013 2nd Experiment@ International Conference,* p. 122 –127, 2013.

[53] "Matlab," The MathWorks, [Online]. Available: www.mathworks.com/products/matlab/. [Accessed 16 02 2016].

[54] "PhysioNet," Components of a New Research Resource for Complex Physiologic Signals, [Online]. Available: https://physionet.org/physiobank/database. [Accessed 1 March 2016].

[55] K. Kalpakis, D. Gada, V. Puttagunta, "Distance measures for effective clustering of ARIMA time–series," *Technical Report TR–CS–01–14, CSEE, UMBC,* 2001.

[56] M. Gavrilov, D. Anguelov, P. Indyk, R. Motwani, "Mining the stock market: which measure is best?," *KDD Proc. of the 6th ACM SIGKDD Int. conf. on Knowledge discovery and data mining,* pp. 487-496, 2000.

[57] D. L. Kulick, "medicinenet , Atrial Fibrillation Causes, Symptoms, Treatment," [Online]. Available: http://www.medicinenet.com/atrial_fibrillation/article.htm. [Accessed 6 12 2016].

[58] A. Kianimajd . M. G. Ruano, P. Carvalho, J. Henriques, T. Rocha, S. Paredes, "Comparison of different methods of measuring similarity in physiologic time series," in *20th IFAC World Congress, (IFAC WC 2017),* 2016.

[59] A. Kianimajd, M. G. Ruano, P. Carvalho, J. Henriques, T. Rocha, and S. Paredes, "Validation of a similarity measurement method for clustering," in *BHI2017 - International Conference on Biomedical and Health Informatics*, 2016.

ATTACHMENT / APPENDIX

I) IFAC WC 2017 paper

Comparison of different methods of measuring similarity in physiologic time series

A. Kianimajd *, M. G. Ruano. **, P. Carvalho. *, J. Henriques. ****, T. Rocha*****, S. Paredes. ********

* *FCT, University of Algarve, Faro, and, CISUC-University of Coimbra, Portugal (e-mail: adell.kiani@gmail.com).*
** *FCT, University of Algarve, Faro, and, CISUC-University of Coimbra, Portugal (e-mail: mruano@ualg.pt).*
*** *CISUC-University of Coimbra, Portugal (e-mail: carvalho@dei.uc.pt).*
**** *CISUC-University of Coimbra, Portugal (e-mail: jh@dei.uc.pt).*
***** *CISUC-University of Coimbra, Portugal (e-mail: teresa@isec.pt).*
****** *CISUC-University of Coimbra, Portugal (e-mail: sparedes@isec.pt)}*

Abstract: Searching for similarity between time series plays an important role when large amounts of information need to be clustered to integrate intelligent supported personal health care diagnosis systems. Accurate measurement of time series similarity patterns influences the performance of classification, clustering and disease prediction stages. In this paper commonly employed methods of measuring similarity between time series were tested on physiologic data. The similarity methods were applied on longer data segments than the typical cardiac cycle envisaging its use integrated on personalized health care cardiovascular diagnosis systems. Euclidean distance, Discrete Wavelet Transform, Discrete Fourier Transform, Correlation Coefficient, Mahalanobis distance, Minkowski Distance, and Dynamic Time Warping Distance were compared when incremental (20 levels) small variations in amplitude scaling and shift, time scaling and shift, baseline variance and additive Gaussian noise are forced to the tested time series. Concentrating on the performance of the similarity methods in terms of their insensibility to small data variations results demonstrate that the time domain Correlation Coefficient is the most robust method while the Discrete Wavelet Transform is the elected one between the transform-based methods. Selection of a similarity method to be applied should also take into account implementation issues, namely need of data reduction to avoid computational burden.

Keywords: Time series; Similarity measure; Euclidean distance; Discrete Wavelet Transform; Discrete Fourier Transform; Correlation Coefficient; Mahalanobis distance; Minkowski Distance; Dynamic Time Warping Distance;

1. INTRODUCTION

In the last decade there has been intense and significant research on developing and deploying Personal Health Care services in cardiovascular diseases (CVD) management there are still several gaps to be tackled before an automated system can efficiently perform CVD personalized management. Within this context, usage of intelligent algorithms to process data obtained from uncontrolled conditions and to be self-adapting (moving from population-based to patient-specific adaptations) and accurate is still a research challenge. To achieve so, several strategies may be followed, one of them being composed of a prior identification of the personal cardiac signal with a CVD pathology (e.g. by identification of similarity between the personal signal with a reference signal) followed by the automatic classification (i.e. clustering) of the signal under analysis into a specific class of signals (usually disease related). Like in many other application areas the collected cardiac signals may be regarded as long time series i.e. as the simplest representation of temporal data expressing the changes of real values at time or space points, due to sampling at a fixed time interval (Koh et al (2005)).

Time series similarity measurement is a method of measuring the degree of similarity between two-time series. If we can work with a highly efficient and effective method of measuring similarity and find the relationship among the time series, it will greatly increase precision of the analysis in time series databases and could help improving accuracy and efficiency in classification, prediction and cluster analysis (Jiang et al (2009) Kalpakis et al (2001)).

Several works have been published about similarity measuring methods. Application of similarity matching algorithm is commonly encountered in various multimedia, medical and financial applications (Yang et al (2004.b)). It is one of the main research subjects in earthquake prediction research (Wei et al (2010)), in change detection of vegetation indices in the land ecosystem research (Lhermitte et al (2011)), in stock prices data and money exchange rate analysis (Struzik et al (1999), Rafiei (1999), Rafiei (1997)), bioinformatics (Tsiporkova et al (2012)) and in medical streaming data (Bernatavičien et al (2015)), arrhythmia detection (Chien et al (2005), Yeh (2009)) and several other sciences.

Each of these publications is based on different approaches for similarity search both in terms of working in time (directly

with the data) or transform domain. There are many similarity and distance measuring methods, namely Dynamic Time Warping (DTW) distance (Tsiporkova et al (2012)), Mahalanobis distance (Yeh (2009)), transforming and Dimension reduction techniques like discrete Fourier transform (Agrawal et al (1993)) or Karhunen-Loève transform (Castells (2007)), Singular Value Decomposition transform (Bandarabadi et al (2010)), principal component analysis (Yang et al (2004.b), Karamitopoulos et al (2007)), and discrete wavelet transform (DWT) (Antoniadis et al (2011)).

This paper compares different methods of measuring similarity in long time series and presents our analysis in terms of accuracy and precision when various forced time series variations are imposed.

The organization of the rest of the paper is as follows. A summary of the similarity measurement methods employed on this study is presented in section 2. Description of the experimental tests performed is included on section 3 and the obtained results are presented in section 4. Conclusions and topics to be addressed in the near future are pointed out in section 5.

2. TIME SERIES BACKGOUND

2.1 Time series representation

As previously mentioned time series data represent the variations observed on temporal data expressed as real values in time or space, resultant of a fixed time interval sampling, (Koh et al (2005)). How to effectively manage and use vast amounts of data series, the effective discovery and understanding of the data sequence and knowledge behind the law, has been increasingly adopted as data mining researcher's topic (Wei et al (2010)).

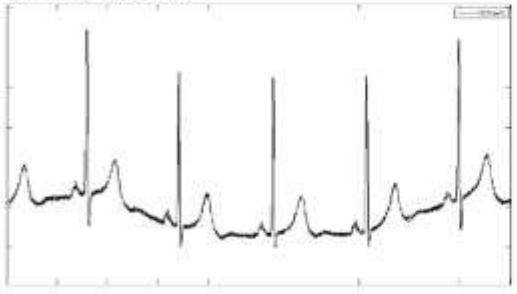


Fig. 1. Example of a time series: an electrocardiogram signal.

Dealing with time series quite often means having to overcome some problems, such as large volumes of data, non-finite or even discrete numerical range, non-constant sampling rate, various noise interference forms (Jiang et al (2009)). So, before applying any analysis technique, pre-processing is required, namely normalization and noise removal.

A brief description of the time domain and transformed based similarity methods used in this paper is below included.

2.2 Time domain similarity measurement methods

The distance between two N sized time series X(t) = {x(1), x(2), ..., x(N)} and Y(t) = {y(1), y(2),..., y(N)} is the length of the path connecting pair of points. This distance is a measure of similarity. Greater distance indicates less similarity and vice versa (Yang et al (2004.a)). The most commonly used and simplest time domain distance measure in classification approaches is derived from the Minkowski distance, represented in (1), where it is described as a general equation for both the Euclidean distance (D_{Ed}) and the Manhattan distance (D_{Man}) (Lhermitte et al (2011))

$$D_{Minkowski}\big(X(t),Y(t)\big) = \left(\sum_{t=1}^{N}|x_t - y_t|^p\right)^{\frac{1}{p}} \quad (1)$$

In the case of p=1, Eq.1 represent the Manhattan distance. If p=2, the Euclidean distance is easy calculated among time series of the same length see (2) (Lhermitte et al (2011), Pree et al (2014).

$$D_{Euclidean}\big(X(t),Y(t)\big) = \sqrt{\sum_{t=1}^{N}|x_t\text{-}y_t|^2}\,. \quad (2)$$

However, the Euclidean distance has limitations. It does not allow different sequence's length, different sampling rates, shifting in time axis (even though these time series are similar to each other). These drawbacks make the Euclidean distance difficult for direct use.

To cope with these problems, modifications have been introduced based on the principle of DTW (Dong et al (2006), Megaloikonomou et al (2005)) as expressed in (3), where two time-series X(t) and Y(t) are 'stretched' or 'compressed' to allow comparison between them. We construct a $n \times m$ warping matrix. The cell (i, j) is correspondent to the alignment of element x_i with y_i and D is the distance function. A detailed explanation of DTW algorithm can be found in (Shasha et al (2004)). (3)

$$D^2_{DTW}(X,Y) = D^2_{i=1:end}(X_i,Y_i) + \min\begin{cases} D^2_{DTW}\big(X_i,\text{Rest}(Y)\big) \\ D^2_{DTW}\big(\text{Rest}(X),Y_i\big) \\ D^2_{DTW}\big(\text{Rest}(X),\text{Rest}(Y)\big) \end{cases}$$

DTW produces a more intuitive similarity measure, allowing similar shapes to match even if they are out of phase and are not perfectly synchronized in the time axis as schematically presented in Fig. 2.

The Pearson Correlation Coefficient is a well-known similarity measure that is invariant to shifting and scaling being expressed by (4) ((Shasha et al (2004))

$$r_{CC}(X(t),Y(t)) = \frac{\sum_{t=1}^{N}(X(t)-\mu_X)(Y(t)-\mu_Y)}{\sqrt{\sum_{t=1}^{N}(X(t)-\mu_X)^2}\sqrt{\sum_{t=1}^{N}(Y(t)-\mu_Y)^2}} \quad (4)$$

where N is length of the time series and μ is the average of each time series (Chien et al (2005)) The Pearson Correlation Coefficient ranges is $-1 < r < +1$ where +1 indicates a perfectly matched between two time-series and 0 indicates that

there is no association between the two variables. A value less than 0 indicates a negative association.

The Mahalanobis distance defined as a dissimilarity measure between two time-series with the same distribution and covariance matrix S is defined on (5) (Mahalanohis (1986)).

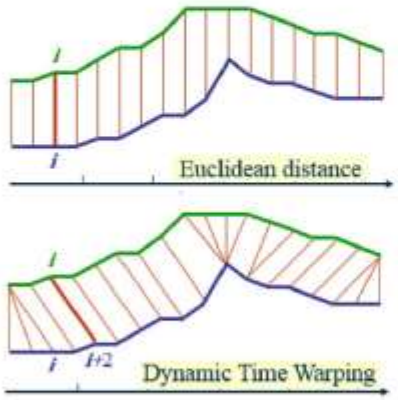


Fig. 2. Dynamic time-warping Vs Euclidean distance.

$$D_{\text{Mahalanobis}}\big(X(t),Y(t)\big) = \sqrt{(X-Y)^T S^{-1}(X-Y)} \qquad (5)$$

The advantage of using Mahalanobis distance is that it takes into consideration the correlations, S, between the time series by which different patterns can be identified and analysed with respect to a based or reference point (Sidek et al (2012))

2.3 Transform-based similarity methods

Usually time series are so long that data reduction techniques can be used to reduce the size of the data without substantial loss of information. The Discrete Fourier Transform (DFT) is a classic data reduction technique and based on that the Discrete Wavelet Transform (DWT) is developed. The DFT is used to map long time series into frequency domain to enable representation of the time series by a set of elementary function called basis (in these case sine and cosine functions) (Shasha et al (2004), Wei (2006)) as given in (6) at its exponential form:

$$X(F) = DFT\big(X(t)\big) = \frac{1}{\sqrt{N}} \sum_{i=0}^{N-1} X(i) e^{-\frac{j2\pi F_i}{N}} \qquad (6)$$

Where F represents the frequency and N the length of the time series. The first few coefficients of the DFT concentrate and contain most information of the time series and can capture good approximation of it. According to Parseval theorem which specifies that the Fourier Transform preserves the Euclidean distance between time series in time and frequency domains, it is possible to use the first few coefficients for measuring similarity of two time-series instead of the original ones (Agrawal et al (1993), Wei et al (2010), Shasha et al (2004)). Fourier transform could change time series from time domain to frequency domain, at the expense of unclear time representation, beside all information being preserved. To represent the behaviour of a time series in both domains, Wavelet-based functions are also employed with better and higher resolution in both time and frequency domains. Unlike the Fourier transform, wavelet transforms have a huge set of possible basis functions and provides a way of analysing the local behaviour of functions (Struzik et al (1999), Kopenkov (2008), Rocha et al (2014), Popivanov et al (2002)). For more details, see (Shasha et al (2004)).

So, time series can be decomposed into linear combinations of the basis-functions. The trend of the input function is captured in approximation to the original function $\phi(t)$, while localized changes are kept as sets of detailed functions, ranging from coarse to fine $\psi(t)$ (Antoniadis et al (2011)). DWT is computed as in (7).

$$\tilde{X}_J(t) = C_{0,0}\phi_{0,0}(t) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t) \qquad (7)$$

Exploring the data reduction ability of DWT for measuring the similarity between two time-series (Rocha (2014)) proposed an interpretable similarity measure by combining the Haar wavelet decomposition with the Karhunen-Loève transforms in order to optimally reduce the number of wavelet basis (Rocha (2012)). The multiresolution aspect of the wavelet transform provides a time-scale decomposition of the signals allowing to visualize and to more accurate clustering the data into homogeneous groups (Antoniadis et al (2011), Kalpakis et al (2001)).

3. EXPERIMENTS

The goal of this work is measuring and comparing the insensitivity of each similarity method to the imposed time series variation thinking on the future application of the best performed method for clustering purposes.

Regarding the usage of a time series sampling frequency of 250Hz and performing an approximation of the series' data points to 2^{11} (please see section 3.1), and, after testing how many DFT coefficients would be required to attain a 90% approximation, we decided to use the first 4 coefficients of DFT (recall (6)). Also, with the same approximation requirement, we considered 11 decomposition levels of DWT (recall (7)). The similarity is hereby computed as the difference between time series under comparison, and the difference is expressed in the range 0 to 1, where 1 means 100% agreement between the time series.

3.1 Time series employed

Dealing with time series involves quite often a pre-processing stage such as normalization and/or noise removal since time series are typically large volumes of data, non-finite or even discrete numerical type, non-constant sampling rate, noise interference forms (Jiang et al (2009)). In our study, we collected our time-series from a public database, and we resampled the signals to adjust all records to the same sampling frequency (250Hz, and used the closest approximation to a power of 2 as data points) and we also normalized data between 0 and 1 considering feature scaling.

In this paper the time series were collected from the public database PhysioNet (Goldberger et al (2000)) that is available in: (https://physionet.org/physiobank/database/#ecg).

To enlarge the applicability of this study besides coronary artery disease and heart failure patient (Rocha et al (2012)), the similarity measurement methods were tested using Arterial Blood Pressure (ABP) signals from Fantasia Database related to healthy subjects, and similarity measurements were taken from longer time series than usual in CVD assessment, i.e., we decided to test time series segments longer than the cardiac cycle, this is, we decided to test 10 sec length data segments.

A template signal X(t) representing 10 sec of healthy people ABP recording was employed. This template corresponded to the mean of randomly selected 10sec segments.

3.2 Time series variations imposed

Biomedical time series may have different type of variation such as additive noise, scaling or translation in time or amplitude, changes in baseline. So selection of a particular similarity method should be based on its sensitivity against variations and to the specific application. Six types of time series distortions were tested using the following equations:

(a) Amplitude Scaling: $X_{As}(t) = \beta * X(t)$,

(b) Amplitude shift: $X_{At}(t) = X(t) + \beta$,

(c) Time scaling: $X_{Ts}(t) = X(\beta * t)$,

(d) Time shift: $X_{Tt}(t) = X(t + \beta)$,

(e) Baseline variation: $X_B(t) = rotate(X(t), \theta)$,

(f) Adding WGN: $X_{Wgn}(t) = X(t) + Z_i \sim \mathcal{N}(X_i, N)$.

Where β is a constant and Θ is the angle of rotation in the baseline and $\mathcal{N}$ is White Gaussian Noise (WGN). The values of β and Θ employed were based on experiments were small incremental changes were envisaged, therefore we decided to consider 20 possible variations of the series as can be seen in Fig. 3, where only the time shift distortion was tested with 21 possible variations.

Signals' similarity were measured by S_{ED} (Euclidean distance), S_{DWT} (Discrete Wavelet Transform), S_{FT} (Discrete Fourier Transform), S_{CC} (Correlation Coefficient), S_{Mah} (Mahalanobis distance), S_{Mi} (Minkowski Distance), S_{DTW} (Dynamic Time Warping Distance) in terms of the above time series variations and results are presented in Fig.3.

4. ANALYSIS OF THE RESULTS

Analysis of performance is always constrained by the goal of the research in course. In our case we were interested in testing the robustness of the similarity methods, where robustness is understood as insensitivity to small variations. Recall that we are testing 10 cardiac cycles instead of single cardiac cycles. So, all experiments developed and the analysis of results below performed are framed by this robustness goal.

As expected and depicted in Fig.3, the similarity measured values decrease when the level of variation is increased (as abscissa values increase). The sensitivity to each type of variation is not the same for all the similarity methods as can be seen by the trend of the curves. Curves close to similarity value of one represent strong similarity, therefore high robustness.

The similarity measured by Correlation coefficient displays lowest sensitivity to all the signals' variations tested.

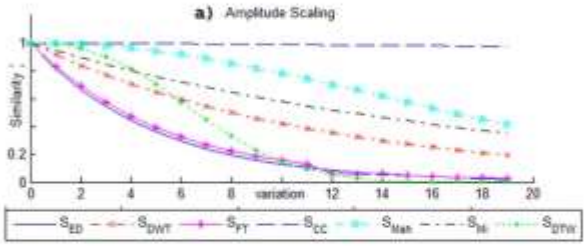
The Euclidean distance and Fourier transform showed highest sensitivity to the variations. The Fourier transform being particularly critical when amplitude shift, time scale and time shift are tested. The similarity measured by the Euclidean distance demonstrates higher unpredictability results between two time-series whenever variations increase.

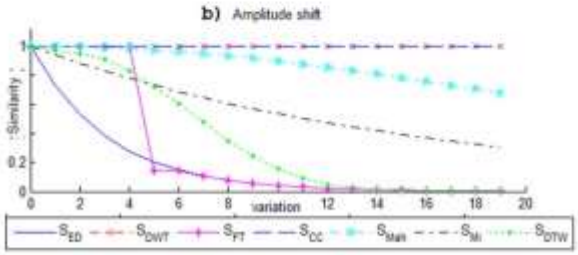
The Mahalanobis distance showed a performance closer to the one presented by the Correlation coefficient and much better than the performance of the Euclidean distance or the Minkowski Distance. This is due to the fact that, similarly to the Correlation Coefficient, the Mahalanobis also takes into consideration the correlation of the data set itself.

The Minkowski Distance proved to be only admissibly insensitive when White Gaussian Noise and variance of the baseline are the data variations considered.

The Correlation coefficient keeps close to the unity similarity line through all degree of variations imposed denoting lowest sensitivity to data variations in all cases.

The Discrete Wavelet Transform response is almost insensitive to amplitude shift, closely following the performance of the Correlation coefficient and it has better robustness response, this is, low sensitivity to other variations.
The Dynamic Time Warping Distance presented a performance similar to the Mahalanobis distance when time scaling is considered and is also almost insensitive to White Gaussian Noise when up to 13 degrees of variation are allowed.





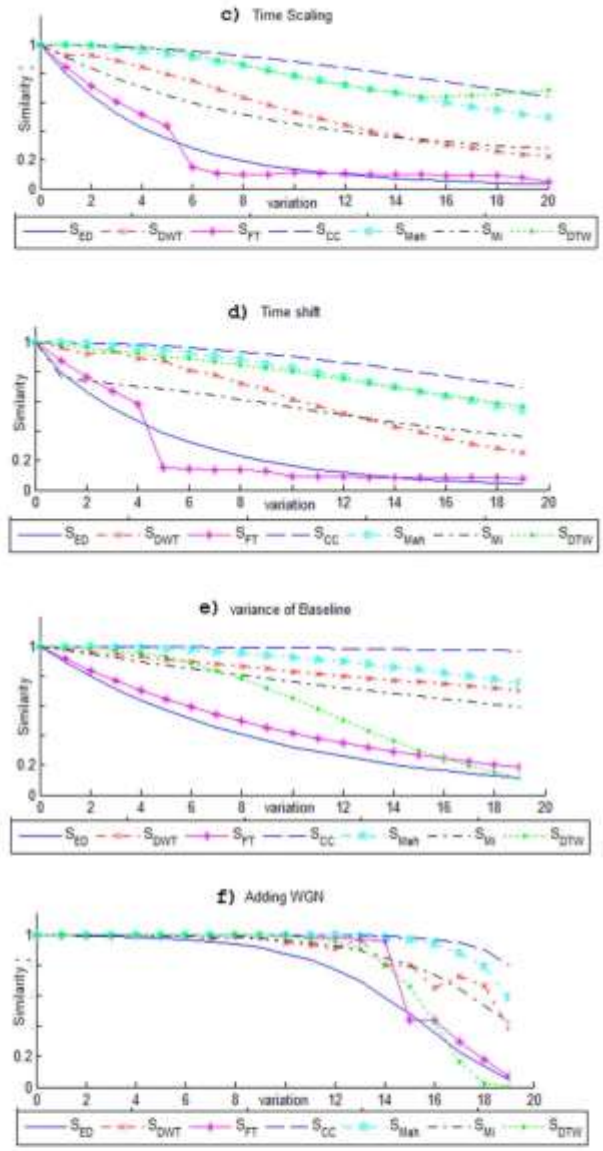


Fig. 3. Similarity sensitivity metrics against degree of variations for a) Amplitude Scaling, b) Amplitude Shift, c) Time Scaling, d) Time shift, e) Variation of baseline, f) Variation by Adding white Gaussian Noise. The figures' caption nomenclature stands for: S_{Ed} - Euclidean distance, S_{DWT}-Discrete Wavelet Transform, S_{FT}-Discrete Fourier Transform, S_{CC}-Correlation Coefficient, S_{Mah}-Mahalanobis distance, S_{Mi}-Minkowski Distance, S_{DTW}-Dynamic Time Warping Distance.

5. CONCLUSIONS AND FUTURE WORK

A comparative study of different time series similarity methods has been performed. Since our target is CVD diagnosis applications, the study considered the well-known and frequently used and referenced PhysioNet data-base, making use of healthy patients' arterial blood pressure signals. We also assumed that testing these similarity methods on data segments longer than the average cardiac cycle (1 sec) would enable a broader application of this study and provide less time consuming evaluation of physiologic time series when a CVD personal health care system is aimed. So, behind this particular study we envisage using these results on our strategy of clustering and disease classification thus considering as the best performed similarity measurement method the one presenting less sensitivity to small variations on the data signal within a range of 20 degrees of increasing variations studied on six types of possible signal variations. Tests and analysis of the results were based on the obtained similarity measurements with seven different similarity measurement methods, five addressing data in time domain and the other two on transformed-based domain.

The case-study time series were collected from the PhysioNet database. We experimentally demonstrated that among the tested time domain similarity measurement methods the Correlation Coefficient was the most robust method, that is to say, the most insensitive to small variations, presenting similarity measurements close to unity for amplitude scaling, amplitude shifting, variance of the baseline and additive white Gaussian noise. Concerning the transformed-based similarity methods tested, the Discrete Wavelet Transform performed better than the Discrete Fourier Transform. It is insensitive to amplitude shifts of the signals, almost insensitive to white Gaussian noise up to 14 variations, but for the other variations it is clear a robustness decay as the degree of variation increases.

To conclude, in what consists an election of the most robust similarity method the Correlation Coefficient wins. However, when data reduction is required due to computational burden of the whole system, the Discrete Wavelet Transform as proposed in (Rocha (2014)) is the similarity measurement approach to be elected.

Future work will address the influence of selecting these similarity methods when clustering efficiency is envisaged.

REFERENCES

Agrawal, R., Faloutsos, C., Swami, A. (1993). Efficient Similarity Search in Sequence Databases. 4th International Conference, FODO '93, LNCS 730. .Available at: DOI:101007/3-540-57301-1_5.

Antoniadis, A., Brossat, X., Cugliari, J., Poggi, J. (2011). Clustering functional data using wavelets. International Journal of Wavelets, Multiresolution and Information Processing. Available at: DOI:10.1142/S0219691313500033.

Bandarabadi, M., Karami-Mollaei, M., Afzalian, A., Ghasemi, J. (2010). ECG Denoising Using Singular Value Decomposition. Australian Journal of Basic and Applied Sciences 4(7), p2109-2113.

Bernatavičien, J., Dzemyda, G., Bazilevičius,G., Medvedev, V., Marcinkevičius, V., Treigys, P. (2015). Method for Visual Detection of Similarities in Medical Streaming Data. International journal of computers communication &control. 10(1), p8-21.

Castells, F., Laguna, P., Sornmo, L., Bollmann, A., Millet, J.R. (2007). Principal Component Analysis in ECG Signal

Processing. EURASIP Journal on Advances in Signal Processing. Available at: Doi:10.1155/2007/74580

Chien, C., Hong, T., Yi, B. (2005). Using correlation coefficient in ECG waveform for arrhythmia detection. Biomed Eng Appl Basis Comm. 17, p147-152.

Dong, X., Gu, C., Wang, Z. (2006). Research on shape-based Time series Similarity measure. Proc. of the Fifth International Conference on Machine Learning and Cybernetics. Available at: DOI:10.1109/ICMLC.2006.258648

Goldberger,A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. Ch., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C., Eugene Stanley,H.(2000).PhysioBank, PhysioToolkit, and PhysioNet:Components of a New Research Resource for Complex Physiologic Signals,

Jiang, Y., Lan, T., Zhang, D. (2009). A New Representation and Similarity Measure of Time Series on Data Mining. International Conference on Computational Intelligence and Software Engineering. Available at: DOI:10.1109/CISE.2009.5364532

Kalpakis, K., Gada, D., Puttagunta, P. (2001). Distance measures for effective clustering of ARIMA time–series. Proceedings IEEE International Conference on Data Mining. Available at: DOI:10.1109/ICDM.2001.989529

Karamitopoulos, L., Evangelidis, E. (2007). PCA-based Similarity Search: Pre-processing & Distance Measures. Proc. 2nd International Scientific Conference: The Contribution of Information Technology to Science, Economy, Society and Education, p318-327.

Koh, H., Loh, W., Kim, S. (2005). An Efficient Subsequence Matching Method Based on Index Interpolation. Innovations in Applied Artificial Intelligence in Lecture Notes in Computer Science, (3533), p 480-489.

Kopenkov V. (2008). Efficient Algorithms of Local Discrete Wavelet Transform with Haar-Like Bases. Pattern Recognition and Image Analysis. 18 (4). p 654-661.

Lhermitte, S., Verbesselt, J., Verstraeten, W.W., Coppine, P. (2011). A comparison of time series similarity measures for classification and change detection of ecosystem dynamics. Remote sensing of environment. Available at: DOI: 10.1016/j.rse.2011.06.020

Mahalanohis, P. (1936). On the generalized distance in statistics. Proceedings National Institute of Science, India. 2(1). p. 49-55.

Megaloikonomou, V., Wang, Q., Li, G., Faloutsos, C. (2005). A Multiresolution Symbolic Representation of Time Series. 21st International Conference on Data Engineering. Available at: Doi:10.1109/ICDE.2005.10

Popivanov, I., Miller, R. (2002). Similarity Search Over Time-Series Data Using Wavelets. Proc. IEEE Int'l Conf. Data Eng. (ICDE). p212-221.

Pree, H. Herwig, B., Gruber, T., Sick, B., David, K., Lukowicz, P. (2014). On general purpose time series similarity measures and their use as kernel functions in support vector machines. Information Sciences (281) p 478–495.

Rafiei, D., Mendelzon, A. (1997). Similarity -Based Queries for Time Series Data. Proc. of International conference ACM SIGMOD Management of data. p 13-25.

Rafiei, D. (1999). On Similarity-Based Queries for Time Series Data. Proc. of International conference on Data engineering. p 410-417.

Rocha, T., Paredes, S., Carvalho, P., Henriques, J. (2012). An Efficient Strategy for Evaluating Similarity between Time-Series based on Wavelet / Karhunen-Loève Transforms. 34th Annual International Conference of the IEEE EMBS. DOI: 10.1109/embc.2012.6347414

Rocha, T., Paredes, S., Carvalho, P., Henriques, J. (2014). Assessing the similarity between time series using a Wavelet transform: application and interpretability aspects. IEEE-EMBS International Conference on BHI. Available at: DOI:10.1109/BHI.2014.6864448

Shasha, D., Zhu, Y. (2004). High performance discovery in time series: techniques and case studies. PhD thesis. Available at: https://cs.nyu.edu/media/publications/zhu_yunyue.pdf

Sidek, K., Khalil, I. (2012). Biometric Sample Extraction using Mahalanobis Distance in Cardioid Based Graph using Electrocardiogram Signals. 34th Annual International Conference of the IEEE EMBS.

Struzik, Z.R., Siebes A. (1999). Measuring Time Series' Similarity through Large Singular Features Revealed with Wavelet Transformation. Tenth International Workshop on Database and Expert Systems Applications. Available at: DOI:10.1109/DEXA.1999.795160

Tsiporkova, E., Kostadinova, E., Boeva, V., Boneva, L. (2012). An Integrative DTW-based Imputation Method for Gene Expression Time Series Data. 6th IEEE International Conference Intelligent Systems. Available at: DOI:10.1109/IS.2012.6335145

Wei, L., Hua, Z., Lin Chen, Q., Afang, J. (2010). Based on time series similarity matching algorithm for earthquake prediction research. 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE). Available at: DOI:10.1109/ICACTE.2010.5579640

Wei, W.S. (2006). Fourier Analysis in Chap 11 in Time Series Analysis: Univariate and Multivariate Methods. Pearson Addison Wesley.

Yang, A., Goldberger, A., Peng, C. (2004.a). Information-Based Similarity Index. Available at: http://physionet.org/physiotools/ibs/doc/ [Accessed 17 Feb. 2016]

Yang, K., Shahabi, C. (2004.b). A PCA-based Similarity Measure for Multivariate Time Series. Proc. of the 2nd ACM international workshop on Multimedia databases, p 65-74. Available at: Doi:10.1145/1032604.1032616

Yeh, Y. (2009). An Analysis of ECG Beats by Using the Mahalanobis Distance Method. Fourth International Conference on Innovative Computing, Information and Control. Available at: DOI:10.1109/ICICIC.2009.75

II) BHI2017 paper

Validation of a similarity measurement method for clustering cardiac signals *

A. Kianimajd, M. G. Ruano, P. Carvalho, J. Henriques, T. Rocha, and S. Paredes, *Members, IEEE*

***Abstract*— Development of personalized cardiovascular management systems involves automatic identification of the current data as a normal or pathological; considering cardiac data as time–series (bio signals), the illness identification may be performed by seeking similarity between the current patient's time-series data and a reference signal, as more accurately as possible and then proceeding to illness stratification (clustering) or even prognostic indication. In this paper we analyze the performance of seven of the most common methods of time-series similarity measurement when the Partitioning Around Medoids strategy of clustering is considered. Three different electrocardiogram collections of data were used for testing, each one including both pathological (were different pathologies were included for each collection) and non-pathological time series. Results demonstrate that usage of the reduced basis Discrete Wavelet Transform resulting from the combination of Haar wavelet decomposition with the Karhunen-Loève transforms enables clustering different pathological-dependent and healthy cardiac datasets with better performance than the other methods, presenting an accuracy ranging from 75% to 85% when partitioning around Medoids clustering is used.**

I. NTRODUCTION

Cardiovascular diseases (CVD) are still a major cause of chronic disease human mortality, being responsible for huge numbers of disability adjusted life years and presenting major impact on health expenditure. Also, CVD may develop fast and sometimes silently due to other comorbidities, which leads to many relevant CVD research queries whenever preventive medicine is envisaged. Although in the last decade there has been intense and significant research on developing and deploying Personal Health Care (PHC) services in CVD management there are still several gaps to be tackled before an automated system can efficiently perform CVD personalized management. Within this context, usage of intelligent algorithms to process data obtained from uncontrolled conditions and to be self-adapting (moving from population-based to patient-specific adaptations) and accurate is still a research challenge. To achieve so, several strategies may be followed, one of them assuming the cardiac signal as a time-series. As so, the strategy would be composed of a prior identification of the personal cardiac signal with a CVD pathology (e.g. by identification of similarity between the personal time-series with a reference signal) followed by the automatic classification (i.e. clustering) of the time-series under analysis into a specific class of signals (usually disease related). This procedure may be found in many application areas and particularly at the bioengineering field; clustering is in calling the attention of data mining researchers within application areas such as bioinformatics [1], medical streaming data [2] and arrhythmia detection [3, 4].

These publications are based on different approaches for similarity search, either in terms of time and transformed-base domains or both. Several similarity and distance measuring methods may be found in literature, to be mentioned Dynamic Time Warping (DTW) distance [1] and Mahalanobis distance [3] in time domain, within transforming and dimension reduction techniques, the Discrete Fourier Transform [4], Karhunen-Loève transform [5], Singular Value Decomposition Transform [6], Principal Component Analysis [7], and Discrete Wavelet Transform (DWT) [8] are worth being mentioned.

Many of the clinically collected cardiac signals may be regarded as time series, i.e., as the simplest representation of temporal data expressing the changes of real values at time or space points, due to sampling at a fixed time interval. Whenever general trend of the cardiac cycles are to be analyzed, several cardiac cycles are dealt simultaneously, allowing non-stationarity features to be taken into account. This approach leads us to long time-series analysis, enabling an efficient computational illness clustering regarding the PHC management system.

However, usage of long time-series approaches requires a highly efficient similarity method so that a highly precise stratification phase may happen, and if that is the case, subsequent accurate prediction [9, 10] arises. So, the aim of this paper is to validate the best performed and accurate similarity measure to be applied on datasets of cardiac long time series signals, taking into account previous team research on accuracy and precision evaluation of different methods of measuring similarity between long time series when different possible variations are considered [11]. Present publication is devoted to the description, testing and conclusion of performing clustering by Partitioning Around Medoids (PAM) [8] strategy, when the previously identified [11] best behaved similarity method (in the context of a clustering approach) is employed. In [11] and [5] we tested the methods on Arterial Blood Pressure (ABP) time-series, while on this paper we enlarge the study to electrocardiogram (ECG) signals.

This paper is organized as follows: section II and III synthetize the similarity methods tested and the PAM clustering strategy employed, respectively; section IV specifies the data sets used; the experiments undertaken and the obtained results are presented in section V and finally the

* Research supported by H2020 – 692023.

A. Kianimajd is with University of Algarve, Faro, Portugal (e-mail: Adell.kianimajd@gmail.com).
M. G. Ruano is with University of Algarve, Faro, Portugal (corresponding author phone: +351915392142, e-mail: mruano@ualg.pt) and CISUC, University of Coimbra, Portugal
P. Carvalho is with CISUC, University of Coimbra, Portugal (email: carvalho@dei.uc.pt)
J. Henriques is with CISUC, University of Coimbra, Portugal (email: jh@dei.uc.pt)
T. Rocha is with CISUC, University of Coimbra, Portugal (email: teresa@isec.pt)
S. Paredes is with CISUC, University of Coimbra, Portugal (email: sparedes@isec.pt)

conclusions and issues to be addressed in near future are exposed in section VI.

II. Similarity Measuring Methods

As previously mentioned searching for time-series similarity may be performed in time or transformed (frequency in general) domains or even including both domains. The similarity measuring methods we tackled are described in the sections below.

The distance between two N sized time series $X(t) = \{x(1), x(2), \ldots, x(N)\}$ and $Y(t) = \{y(1), y(2), \ldots, y(N)\}$ is the length of the path connecting pair of points. This distance is a measure of similarity. Greater distance indicates less similarity and vice versa [11].

A. Time Domain Methods

The most commonly used and simplest time domain distance measure in classification approaches is derived from the Minkowski distance, represented in (1), where it is described as a general equation for both the Euclidean distance (D_{Ed}) (P=2) and the Manhattan distance (D_{Man}) (P=1) [11].

$$D_{Minkowski}(X(t),Y(t)) = \left(\sum_{t=1}^{N} |X(t) - Y(t)|^p \right)^{\frac{1}{p}} \quad (1)$$

The Euclidean distance is easily calculated among time series with the same length [11, 12, 13] . However, the Euclidean distance has limitations. It does not allow different sequence's length, different sampling rates, shifting in time axis (even though these time series are similar to each other). These drawbacks make the Euclidean distance difficult for direct use. To cope with these problems, modifications have been introduced based on the principle of DTW [14, 15] where two time-series X and Y are 'stretched' or 'compressed' to allow comparison between them. We construct a $n \times m$ warping matrix. The cell (i, j) is correspondent to the alignment of element x_i with y_i and D is the distance function. A detailed explanation of DTW algorithm can be found in [16]. DTW produces a more intuitive similarity measure, allowing similar shapes to match even if they are out of phase and are not perfectly synchronized in the time axis.

The Pearson Correlation Coefficient is a well-known similarity measure that is invariant to shifting and scaling, being expressed by (2) [16], where N is the length of the time series and $\overline{X}$ and $\overline{Y}$ are the averages of each time series. The Pearson Correlation Coefficient's range is $-1 \le r_{CC} \le +1$, where +1 indicates a perfectly match between two time-series and 0 indicates that there is no association between them. A value less than 0 indicates a negative association

$$r_{CC}(X,Y) = \frac{\sum_{i=1}^{N} (X(t) - \overline{X})(Y(t) - \overline{Y})}{\sqrt{\sum_{i=1}^{N} (X(t) - \overline{X})^2} \sqrt{\sum_{i=1}^{N} (Y(t) - \overline{Y})^2}} \quad (2)$$

The Mahalanobis distance defined as a dissimilarity measure between two time-series with the same distribution and covariance matrix S is defined on (3) [17].

$$D_{Mahalanobis}(X(t),Y(t)) = \sqrt{(X-Y)^T S^{-1} (X-Y)} \quad (3)$$

The advantage of using Mahalanobis distance is that it takes into consideration the correlations, S, between the time series by which different patterns can be identified and analyzed with respect to a based or reference point.

B. Transformed-based methods

Usually time series are so long that data reduction techniques can be used to reduce the size of the data without substantial loss of information a very convenient strategy when long time-series are being processed. The Discrete Fourier Transform (DFT) is a classic data reduction technique and based on that the Discrete Wavelet Transform (DWT) is developed. The DFT is used to map long time series into frequency domain to enable representation of the time series by a set of elementary function called basis (in these case sine and cosine functions) [16, 18] as given in (4) at its exponential form:

$$X(F) = DFT(X(t)) = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} X(t) e^{-\frac{j 2\pi F t}{N}} \qquad F = 0, 1, \ldots, N-1 \quad (4)$$

Where F represents the frequency and N the length of the time series. The first few coefficients of the DFT concentrate and contain most information of the time series and can capture good approximation of it. According to Parseval theorem which specifies that the Fourier Transform preserves the Euclidean distance between time series in time and frequency domains, it is possible to use the first few coefficients for measuring similarity of two time-series instead of the original ones [16, 4, 18]. Fourier transform could change time series from time domain to frequency domain, at the expense of unclear time representation, beside all information being preserved. To represent the behavior of a time series in both domains, Wavelet-based functions are also employed with better and higher resolution in both time and frequency domains. Unlike the Fourier transform, wavelet transforms have a huge set of possible basis functions and provides a way of analyzing the local behavior of functions [19, 20, 21]. For more details, see [16].

So, time series can be decomposed into linear combinations of the basis-functions. The trend of the input function is captured in approximation to the original function $\phi(t)$, while localized changes are kept as sets of detailed functions, ranging from coarse to fine $\psi(t)$ [8]. DWT is computed as in (5).

$$\overline{X}_j(t) = C_{0,0}\phi_{0,0}(t) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t) \quad (5)$$

Exploring the data reduction ability of DWT for measuring the similarity between two time-series T. Rocha et al. [13] proposed an interpretable similarity measure by combining the Haar wavelet decomposition with the Karhunen-Loève transforms in order to optimally reduce the number of wavelet basis [5]. The multiresolution aspect of the wavelet transform provides a time-scale decomposition of the signals allowing to visualize and to more accurate clustering the data into homogeneous groups [8, 10].

III. PAM BASED CLUSTERING

Clustering is one of the most frequently used data mining techniques. The objective of cluster analysis is to partition a set of objects into two or more clusters based on the similarity between time series. On this paper we focus on PAM clustering method which is based on the search for k representative objects, called Medoids, among the objects of the dataset. If the average of dissimilarity between objects near a Medoid is minimum, it is classified as cluster [8]. PAM clustering is hereby performed to analyze the ability of different similarity methods to distinguish signals in the dataset so that precision and efficiency of each similarity methods [10] is assessed. It attempts to minimize the total distance D between objects within each cluster. D could be computed as (6):

$$D = \sum_{k=1}^{K} \sum_{i \epsilon C_k} \sum_{j \epsilon C_k} d_{ij} \qquad (6)$$

where K is the number of clusters, d_{ij} is the distance between objects i and j, and C_k is the set of all objects in cluster k. For more details about algorithmic issues please see [23].

Finding dissimilarity (distance) between two time series is fundamental to cluster analysis since the goal is to place similar objects in the same cluster and dissimilar objects in different clusters. The goal of present work is the identification of the similarity method which would produce more precise clustering when compared with predefined datasets.

IV. LONG TIME SERIES EMPLOYED

Based on the public database PhysioNet [23] we created our four data bases of ECG signals. Each signal in our data base has 10 sec length, randomly selected from the correspondent PhysioNet data base. The following data bases were considered: Fantasia, MIT-BIH Atrial Fibrillation, Long-Term ST and PTB Diagnostic ECG. From these, PhysioNet presented, respectively 40 (2h records), 25 (10 h records), 86 (21-24 h records) and 549 (variable record length), within which 40, 84, 85 and 71 were randomly selected. According these numbers we can see that our data base may include time series of ECG's belonging to the same patient, but, most important, include different pathology samples.

IV. EXPERIMENTS AND ANALYSIS OF RESULTS

Dealing with time series involves quite often a pre-processing stage such as normalization and/or noise removal since time series are typically large volumes of data, non-finite or even discrete numerical type, non-constant sampling rate and noise interference forms [9]. In our study, for each 10 sec signal of our working data base we applied pre-processing. Since the range of amplitude values of the data varied widely and similarity functions do not work properly in such situations, a linear transformation of the amplitudes was performed. By normalization of all the signals in the dataset, all measured values were adjusted in the common scale. This feature scaling process (typically named unity-based normalization) is used to bring all values into the range [0,1] [24]. Another preprocessing applied was the removal of vertical offsets of the time series. We also resampled the signals to adjust all records to the same sampling frequency of 250Hz, and used the closest approximation to a power of 2 as data points. Also data was normalized between 0 and 1 by means of feature scaling. At last every time series was aligned in the datasets according to their first peaks aiming at increased accuracy of the results.

Following previous research, different types of time series variations have been applied on the template time series of ABP signals and the achieved robustness's may be consulted in [11]. At present the conclusions drawn for ABP signals were considered to perform PAM clustering on ECGs.

Within the four working data bases created three collections were formed, each one including 30 signals from Fantasia data base and collection 1 included 45 signals from MIT-BIH Atrial Fibrillation, collection 2 included 55 signals from Long-Term ST data base and finally collection 3 included 50 signals from PTB Diagnostic ECG.

The accuracy of clustering may be evaluated through metrics. Gavrilov et la [25] proposed a cluster similarity metric as defined by (7) and (8)

$$Sim(G_i, A_j) = 2\,\frac{|G_i \cap A_j|}{|G_i| + |A_j|} \qquad (7)$$

where G_i are predefined members of each datasets to be considered as "ground-truth" and A_j are clustering results obtained by using PAM with various types of similarity methods. In our case we considered two G_i members: healthy and diseased. Numerator of (8) introduces the number of similar time series A_j that are recognized within G_i when k clusters are considered. This metric will be zero if two clustering are completely dissimilar and 1 if they are similar [10, 26].

$$Sim(G, A) = \frac{\sum_i \max_j Sim(G_i, A_j)}{k} \qquad (8)$$

The cluster similarity metrics obtained with PAM clustering of these collections are expressed on Table I where a *Sim* value (expressed as percentage) of 0 indicates that the clustering results are completely dissimilar while a value of 100 evidences clustering results similar to the established ground-truth. To clarify, if for instance within collection 1 15 out of the 30 records (healthy patients) and 45 of the 45 atrial fibrillation records were detected as similar we would have an accuracy of 75%. Here we are designating 'accuracy' as the precision of correctness clustering of the data under analysis.

TABLE I. COMPARISON OF CLUSTERING 'ACCURACY' WITH DIFFERENT SIMILARITY METHODS

Similarity methods used in clustering	Accuracy of clustering in Percent		
	Collection1	*Collection2*	*Collection3*
Euclidian distance	**77.22**	**81.60**	69.87
Auto correlation coefficient	66.60	67.05	**74.21**
Discrete Cosine Transform	**77.22**	77.92	**72.97**
Discrete Wavelet Transform	**75.86**	**83.93**	**80.20**
Dynamic time warping	**77.22**	**81.60**	69.87
Mahalanobis distance	**76.13**	**81.60**	69.87
Minkowski metric(P=6)	71.40	65.03	68.12

Analysis of these results reveal that DWT provides the most accurate clustering particularly when the variability of signals occur (collections 1 and 2). Results obtained for collection 1 evidence that when inside the clustering members exist more similarity among signals (only healthy and atrial fibrillation signals) Euclidian distance related measurements may be more accurate. To be mentioned that if during the pre-processing stage the alignment of the records' first peaks was not performed, the DWT accuracy obtained for collection 1 would be better than any other methods. Also, these clustering results strength the previously obtained results [5, 11] confirming the election of DWT as the similarity measurement to apply on CVD analysis.

V. Conclusion

This paper describes comparative studies performed to assess accuracy of seven commonly employed similarity methods by comparison of clustering accuracy metrics obtained for those similarity methods when applied on three different collections of data. Results show that, besides previously published work [5, 11], when ECG long term data series are considered usage of DWT combined with Karhunen-Loève transforms for clustering purposes is the most accurate among the commonly employed time-series similarity measurement methods being particularly robust when different types of CVD are considered within the collection under analysis. Therefore, this study enables a generalized conclusion that DWT is a reliable similarity measurement method for biomedical and in particular ECG time series screening for clustering purposes. Future work includes extending analysis to other intelligent-based methods of clustering.

References

[1] E. Tsiporkova, E. Kostadinova, L. Boneva, "An Integrative DTW-based Imputation Method for Gene Expression Time Series Data," *6th IEEE Int. Conf. Intelligent Systems*, pp. 258-263, 2012.

[2] J. Bernatavičien, G. Dzemyda,V. Medvedev, P. Treigys, "Method for Visual Detection of Similarities in Medical Streaming Data," *Int. journal of computers communication &control (IJCCC)*, vol. 10, pp. 8-21, 2015.

[3] Y. Yeh, "An Analysis of ECG Beats by Using the Mahalanobis Distance Method," *IEEE, 4th Int. Conf. on Innovative Computing, Information and Control*, pp. 1460-1463, 2009.

[4] R.Agrawal, C. Faloutsos, A. Swami, "Efficient Similarity Search in Sequence Databases," *4th Int. Conf. on Foundations of Data Organization and Algorithms*, pp. 69-84, 1993.

[5] T. Rocha, S. Paredes, P. Carvalho, J. Henriques, "An Efficient Strategy for Evaluating Similarity between Time Series based on Wavelet / Karhunen-Loève Transforms," *Int. Conf. of the IEEE Engineering in Medicine and Biology*, p. 6216–6219, 2012.

[6] M. Bandarabadi, M. Karami, J. Ghasemi, "ECG denoising using Singular Value Decomposition," *Australian Journal of Basic and Applied Sciences 4(7):2109-2113*, July 2010.

[7] K. Yang, C. Shahabi, "A PCA-based Similarity Measure for Multivariate Time Series," *Proc. of the 2nd ACM international workshop on Multimedia databases*, pp. 65-74, 2004.

[8] A. Antoniadis, X. Brossat, J.M. Poggi, "Clustering functional data using wavelets," *Int. Journal of Wavelets, Multiresolution and Information Processing*, p. 30, 2011.

[9] Y. Jiang, T. Lan, D. Zhang, "A New Representation and Similarity Measure of Time Series on Data Mining," *Int. Conf. Computational Intelligence and Software Engineering*, p. 1–5, 2009.

[10] K. Kalpakis, D. Gada, V. Puttagunta, "Distance measures for effective clustering of ARIMA time-series," *Technical Report TR-CS-01-14, CSEE, UMBC*, 2001.

[11] A. Kianimajd, M. G. Ruano, P. Carvalho, J. Henriques, T. Rocha, and S. Paredes, "Comparison of different methods of measuring similarity in physiologic time series," submitted to *20th IFAC World Congress, (IFAC WC 2017)*.

[12] S. Lhermitte, J. Verbesselt , P. Coppine, "A comparison of time series similarity measures for classification and change detection of ecosystem dynamics," *Remote sensing of environment*, vol. 115, p. 3129–3152, 2011.

[13] H. Pree, B. Herwig, P. Lukowicz, "On general purpose time series similarity measures and their use as kernel functions in support vector machines," *Information Sciences*, vol. 281, p. 478–495, 2014.

[14] T. Rocha, S. Paredes, P. Carvalho, J. Henriques, "Assessing the similarity between time series using a Wavelet transform: application and interpretability aspects," *IEEE-EMBS Int. Conf. on Biomedical and Health Informatics (BHI)*, p. 652 – 655, 2014.

[15] L. Wei, Z. Hua, Q. jianfeng, J. Afang, "Based on time series similarity matching algorithm for earthquake prediction research," *3rd Int. Conf. on Advanced Computer Theory and Engineering (ICACTE)*, p. pp 57, 2010.

[16] V. Megalooikonomou, Q. Wang, G. Li, C. Faloutsos, "A Multiresolution Symbolic Representation of Time Series," *21st Int. Conf. on Data Engineering*, pp. 668-679, 2005.

[17] Shasha and Zhu, D. Shasha, Y. Zhu, High performance discovery in time series: techniques and case studies, New York: Springer, ch.2.

[18] P. C. Mahalanobis, "On the generalized distance in statistics," *In Proc. National Institute of Science, India*, vol. 2, pp. 49-55, 1936.

[19] W. Wei, Time Series Analysis: Univariate and Multivariate Methods,2nd edition, Ch. 11, 2006.

[20] Z. R. Struzik, A. Siebes, "Measuring Time Series' Similarity through Large Singular Features Revealed with Wavelet Transformation," *Database and Expert Systems Applications, Proc.10th Int. Workshop on*, p. 162–166, 1999.

[21] V. N. Kopenkov, "Efficient Algorithms of Local Discrete Wavelet Transform with Haar-Like Bases," *Pattern Recognition and Image Analysis*, vol. 18, no. 4, pp. 654-661, 2008.

[22] T. Rocha, S. Paredes, P. Carvalho, J. Henriques, M. Harris, "Wavelet based Time Series Forecast with Application to Acute Hypotensive Episodes Prediction," *Int. Conf. of the IEEE Engineering in Medicine and Biology*, p. 2403–2406, 2010.

[23] "NCSS data analysis, Medoid Partitioning," NCSS Statistical software, [Online]. Available: http://www.ncss.com/software/ncss/clustering-in-ncss/. [Accessed 20 july 2016].

[24] T. Rocha, S. Paredes, P. Carvalho, J. Henriques, "Trend Prediction Methodology Based on Time Series Similarity Analysis and Haar Wavelet Decomposition," *2013 2nd Experiment@ International Conference*, p. 122 –127, 2013.

[25] https://physionet.org/physiobank/database.

[26] M. Gavrilov, D. Anguelov, R. Motwani, "Mining the stock market: which measure is best?," *KDD Proc. of the 6th ACM SIGKDD Int. conf. on Knowledge discovery and data mining*, pp. 487-496, 2000.