Aplicación de Software para la Clasificación de Señales Electrocardiográficas de Infarto Agudo de Miocardio Implementando la Técnica Medida de Disimilaridad Basada en Compresión

MARCOS AMARIS GONZÁLEZ
Ingeniero de Sistemas

Universidad Industrial de Santander
Facultad de Ingenierías Fisicomecánicas
Escuela de Ingeniería de Sistemas e Informática
Programa de Maestría en Ingeniería e Informática
Bucaramanga, 2012

Aplicación de Software para la Clasificación de Señales Electrocardiográficas de Infarto Agudo de Miocardio Implementando la Técnica Medida de Disimilaridad Basada en Compresión

MARCOS AMARIS GONZÁLEZ
Ingeniero de Sistemas

Trabajo de investigación presentado para optar al Título de Magíster en Ingeniería de Sistemas e Informática

Director:
Ms. Ing. Victor Eduardo Martinez Abaunza

Co-director:
Dr. Pablo Emilio Guillén Rondón

Universidad Industrial de Santander
Facultad de Ingenierías Fisicomecánicas
Escuela de Ingeniería de Sistemas e Informática
Programa de Maestría en Ingeniería e Informática
Bucaramanga, 2012

A Dios.

A mi Madre.

A mi Herman@

AGRADECIMIENTOS

A Dios por darme la madre que tengo, quien me ha ayudado e iluminado incondicionalmente siempre con su apoyo, comprensión y amor.

Mi sincero agradecimiento al profesor Alfonso Mendoza, por la confianza puesta en mi desde un comienzo, y sus valiosos consejos científicos y humanos durante todo mi periodo como estudiante de maestría en la UIS.

Al profesor Victor Eduardo Martinez Abaunza por ser orientador y amigo.

Al profesor Pablo Emilio Guillén Rondón por su apoyo y orientación académica.

A la Fundación Cardiocascular de Colombia, sede Bucaramanga; por colocarse a la orden durante la consecución del presente proyecto y permitir el acceso a la base de datos de señales electrocardiográficas de sujetos Sanos.

A mi familia, pórque a pesar de todo, siempre seremos una familia.

A la gente del GIIB, pórque se investiga, se ríe, se escucha música y se pasa bueno; y, a todas aquellas personas que directa o indirectamente me ayudaron e hicieron cumplir este objetivo.

Índice General

Índice de Figuras

Índice de Tablas

RESUMEN

Título: Aplicación de Software para la Clasificación de Señales Electrocardiográficas de Infarto Agudo de Miocardio Implementando Medida de Disimilaridad Basada en Compresión. *

Autor: Marcos Amaris González.**

Palabras claves: Clasificación, Electrocardiograma, Infarto Agudo de Miocardio, Análisis Wavelet, Medida de Disimilaridad basada en Compresión, Algoritmos de Agrupamiento.

En el presente documento se muestra la continuación del trabajo realizado por el Grupo de Investigación en Ingeniería Biomédica, en la línea de investigación de tratamiento de señales electrofisiológicas, orientadas a la contrucción de aplicaciones para la detección de enfermedades cardíacas en este caso el Infarto Agudo de Miocardio. Esta investigación fue enfocada al uso de dos técnicas de minería de datos, Compression-based Dissimilariy Measure y Symbolic Agreggate AproXimation, para la clasificación de la Variabilidad de la Frecuencia Cardiaca de señales electrocardiográficas digitales por medio de máquinas de aprendizaje no supervisadas.

En este proceso de clasificación de señales electrocardiográficas se realiza un filtrado, la respectiva caracterización de las ondas y se extrae la VFC utilizando herramientas de análisis Wavelet, esto es llamado preprocesamiento; posteriormente se utilizan técnicas de minería de datos para una transformación y clasificación de la VFC de cada señal electrocardiográfica; por la anterior razón se mencionan los resultados de esta investigación como una metodología de descubrimiento de conocimiento en base de datos.

Se presenta una fundamentación teórica de las técnicas de medidas de similaridad basadas en compresión, la base teórica de estas técnicas es la complejidad de Kolmogorov, en este documento se definen conceptos importantes de esta complejidad y ciertas analogías con la teoría de la información de Shannon, también se muestran algunas aplicaciones en máquinas de aprendizajes para la clasificación entre series temporales, imágenes, ADN, video, audio, ente otros.

Se presenta una teoría básica del filtrado y análisis Wavelet en señales electrocardiográficas sobre el ambiente numérico **R**, todos los algoritmos y una interfaz gráfica fueron realizados en el entorno de software para estadística y computación R el cual satisfizo todas las necesidades, y brinda la posibilidad de desarrollar facilmente algoritmos paralelos, debido al modelos de datos por medio de listas.

---

*Proyecto de grado de Maestría

**Facultad de Ingeniería Físico-Mecánicas. Escuela de Ingeniería de sistemas e informática. Director: Victor Eduardo Martinez Abaunza. Codirector: Pablo Emilio Guillén Rondón

ABSTRACT

Title: Software Application for Classification Electrocardiographic Signals of Myocardial Acute Infartion Implementing Compression-based Dissimilirity Measure [*]

Author: Marcos Amaris González.[**]

Keywords: Acute Myocardial Infarction, Wavelet analysis, Clasification, Compression-based Dissimilirity Measure, Clustering Algorithm.

This paper shows the work continuity of the Research Group in Biomedical Engineering (GIIB for its acronym in Spanish) in the area of electrophysiological signal processing, about developing and building applications oriented for the heart disease detection, in this case the Acute Myocardial Infarction. This research was focused on the use of two data mining techniques, Compression-based Disimilarity Measure and Symbolic Agreggate AproXimation, for classification of the Heart Rate Variability of electrocardiographic signals using machine learning unsupervised.

To make possible a good classification of signal electrocardiographic a filter process is performed, the respective characterization of the waves, and extracting of the HRV using wavelet analysis tools; this process is named preprocessing, then data mining techniques are used for the data transformation and data classification; for the last reason the results of this research are mentioned like a methodology for knowledge discovery in databases.

A theoretical foundation of the techniques based similarity measures in compression is presented, the theorical bases of these techniques are based in the Kolmogorov complexity, this paper defines key concepts of this complexity and some analogies with the Shannon's information theory, also a few applications in machine learning for classification between time series, images, DNA, video, audio etc.

Important concepts about the wavelet analysis over electrocardiographic signal with the environment for statistical computing **R**. All the algorithms and graphical user interface were develop in the software environment for statistical computing R, R satisfied all the needs, and provides the ability to easily develop parallel algorithms, for his data models based on lists.

[*] Master Research Work

[**] Faculty of Physical-Mechanical Engineerings. Systems Engineering nd Informatiques Department. Advisor: Victor Eduardo Martinez Abaunza. Co-advisor: Pablo Emilio Guillén Rondón

Introducción

El avance de la tecnología digital en los últimos 50 años ha hecho posible el análisis de señales electrofisiólogicas en tiempo real, en tiempo discreto y se han podido detectar enfermedades con un proceso de filtrado, detección de caracteristicas principales de las señales y posibles diágnosticos. En estos avances tecnológicos, también se han creado técnicas para el análisis y clasificación de series temporales, en la última década fueron creadas varias medidas de similaridad basadas en compresión de datos, estas son técnicas del área de minería que se basan en ciertas analogías de la teoría de la información para hallar similitudes en un conjunto de datos, dentro de estas medidas de similaridad está la técnica CDM (Compression-based Dissimilirity Measure). La idea principal de este proyecto es utilizar la técnica CDM en señales electrocardiográficas con el fin de realizar una clasificación de pacientes con Infarto Agudo de Miocardio; el índice a utilizar para la clasificación es la Variabilidad de la Frecuencia Cardiaca.

Para hallar la VFC en los electrocardiogramas se debe realizar un filtrado con el fin de eliminar el ruido en la señal, luego encontrar los picos R y hallar la distancia entre cada uno; el hallazgo de los picos R se realiza con máximos locales y ventanas deslizantes con base a la frecuencia de muestreo de los electrocardiogramas.

Luego de haber extraído la VFC de cada electrocardiograma se implementa una transformación con la técnica SAX (Symbolic Aggregate approXimation), esta trans-

formación realiza una aproximación numérica basada en la sustitución de números por símbolos, esto es con el fin de conseguir una mejora en la compresión de los archivos y lograr una buena medida de distancia al implementar la técnica CDM. Finalmente se realiza una clasificación utilizando técnicas de árboles jerarquizados que analiza una matriz de disimilaridad y agrupa los objetos analizados.

El presente libro está estructurado de la siguiente manera, en el Capítulo 1 se presenta un marco teórico de las técnicas de medida de similaridad basadas en compresión y ciertas analogías con la teoría de la información de C. Shannon, en el Capítulo 2 se presentan las principales características de un Electrocardiograma, la VFC y lo importante en el estudio de enfermedades cardiovasculares como el infarto agudo de miocardio también se muestran conceptos importantes del procesamiento digital de señales; el Capítulo 3 trata del análisis Wavelet de señales electrocardiográficas sobre el ambiente numérico R, el Capítulo 4 presenta los resultados del proceso de filtrado, caracterización, transformación y clasificación de las señales electrocardiográficas; los resultados se muestran utilizando la metodología de descubrimiento de conocimiento, finalmente en el Capítulo 5 se presentan las conclusiones del presente trabajo de investigación. En el Apéndice A se presentan algunos conceptos básicos de la Computación de Alto Rendimiento sobre el ambiente R y se propone un modelo de comunicación distribuído y paralelo sobre base de datos electrocardiográficas.

1. Medidas de Similaridad Basadas en Compresión

1.1. Introducción

En la última década han surgido variadas técnicas de medidas de distancias de (di)similitud basadas en compresión de datos [2–4]; la idea fundamental de la compresión de datos es utilizada con el fin de usar diversos esquemas de máquinas de aprendizaje, especialmente algoritmos de agrupamiento [5], resultando bastante frutuoso en muchas áreas de investigación tales como bioinformática [6–8,8–10], aprendizaje automático [11], minería de datos, algoritmos de compresión, teoría de la información [12], filtrado de spam [13], Ingeniería Biomédica [14–17], categorización de texto [18–20], métricas de software [21], entre otros [22–24]. Estas medidas de distancias de similaridad basadas en compresión se basan en que al comprimir dos archivos conjuntamente, estos puede que tengan un porcentaje de información mutua entre ellos [25]. Al implementar modelos de máquinas de aprendizaje hace que se reduzca el problema en el proceso de selección de patrones o características comúnes [26].

Estas técnicas son aplicaciones de la teoría de la complejidad de Kolmogorov [27–29], esta última tiene diferentes analogías con operaciones realizadas en la teoría de la información de Shannon y estadística de variables aleatorias o estocástica [30–34]. Estas

técnicas de similaridad hace uso de algoritmos de compresión con el fin de aproximar la complejidad de Kolmogorov y así crear un espacio de información idealizado por medio de una matriz de similaridad entre cada uno de los objetos de estudio para su respectiva y posterior clasificación [35].

La distancia normalizada de información es una medida de distancia universal entre objetos de toda clase. Variadas definiciones y teoremas de la teoría de la complejidad de Kolmogorov, exigen una rigurosidad matemática de alto nivel; con el fin de realizar una clara definición, aquí se proveen conceptos generales, ciertas analogías, notaciones y términos relacionados de gran utilidad en el campo científico e ingenieril.

1.2. Conceptos y Trabajos Relacionados

En ciencias de la computación los conceptos de algoritmo e información son fundamentales; así las medidas de información o algoritmos son cruciales en el sentido de descripción. La navaja de Ockam, principio de economía o principio de parsimonia es un principio filosófico atribuído a Guillermo Ockham (1280 - 1349), según el cual, cuando 2 teorías con igualdad de condiciones tienen las mismas consecuencias, la teoría más simple tiene más probabilidad de ser más cierta que la compleja. La forma moderna de la navaja de ockam es la medida de la complejidad de Kolmogorov [36].

Andréi Nikoláyevich Kolmogórov (25 de abril de 1903 - 20 de octubre de 1987), ver Figura 1.1, fue un matemático ruso que hizo progresos importantes en los campos de la teoría de la probabilidad. En particular, desarrolló una base axiomática que supone el pilar básico de la teoría de las probabilidades a partir de la teoría de conjuntos [31], entre sus más grandes aportes se encuentra la teoría de la complejidad algoritmica de Kolmogorov.

Figura 1.1 Andréi Nikoláyevich Kolmogórov (1903 - 1987).

1.2.1. Complejidad de Kolmogorov

El principio fundamental de la Complejidad de Kolmogorov apareció a través de tres artículos realizados por Solomonoff, Kolmogorov y Chaitin; a través de los artículos [37], [38] y [39] respectivamente; la complejidad de Kolmogorov es también conocida como complejidad algorítmica. Básicamente la complejidad de Kolmogorov K_U de una cadena binaria $x \in \{0,1\}^*$ es la descripción más corta que un programa p necesita para que una Máquina Universal de Turing genere la misma cadena x y se detenga; matemáticamente se describe la complejidad de Kolmogorov como se muestra en la ecuación 1.1

$$K_U(x) = min\{length(p) : U(p) = x\} \tag{1.1}$$

y graficamente se describe según la Figura 1.2

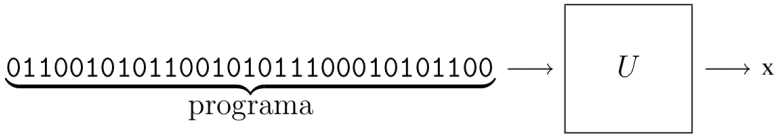

Figura 1.2 MTU U que genera x con el programa p como entrada.

Como una afirmación pragmática que permite evitar las inmensas dificultades involucradas en el diseño de una máquina de Turing que resuelva un problema concreto no trivial: la tesis de Church afirma que al ser capaz de construir un procedimiento, que se pueda llamar finitista o mecánico, y que resuelve un problema de contrucción de una

secuencia aleatoria x, dicho de esta forma un algoritmo representado en algún lenguaje de programación puede suplir estas necesidades de máquina de Turing. El largo de x depende de la máquina de Turing Universal U pero sólo en una constante aditiva. Si K de una secuencia x de largo n es cercana a n, se dice que x es una secuencia aleatoria de grado n, y por ende no se puede comprimir [40, 41].

1.2.2. Condicional de la Complejidad de Kolmogorov

La complejidad de Kolmogorov $K(x)$ de una secuencia x es la descripción de menor tamaño que puede tomar un programa binario para computar la secuencia x en un modelo de computador universal. $K(x)$ representa la mínima cantidad de información requerida para generar la secuencia x por medio de un proceso efectivo [42]. La condicional de complejidad de Kolmogorov $K(x|a)$ se define como el programa más corto que describe x en una máquina universal de Turing; si la cadena a es propuesta como una cinta auxiliar o entrada auxiliar al proceso. $K(x)$ es un caso especial $K(x|a)$ cuando la cinta auxiliar de U está vacía.

La declaración para la complejidad de Kolmogorov no se mantiene exacta, y solamente es verdadera por encima de un factor logarítmico [43], tal como se muestra en la ecuación 1.2

$$K(x, y) = K(x) + K(y|x) + O(log(K(x, y))), \tag{1.2}$$

lo anterior indica que el programa más corto que reproduce x y y, está usando un programa que describe a x y un programa que describe a y dado x más un factor logaritmico.

Las funciones $K(\cdot)$ y $K(\cdot|\cdot)$ son definidas en términos de un modelo de máquina universal (Lenguaje de programación). La expresión matemática de la condicional de

la complejidad de Kolmogorov está dada por la ecuación 1.3

$$d_K(x,y) = \frac{K(x|y) + K(y|x)}{K(xy)} \tag{1.3}$$

esta medida se encuentra muy relacionada con la entropía de Shannon. La complejidad de Kolmogorov de una secuencia aleatoria, puede ser vista como la cantidad de información absoluta y objetiva que es capaz de describir una secuencia cualquiera. No obstante, aún no existe un algoritmo con garantía de finalización que, al ser ejecutado por una máquina de Turing universal y alimentado con cadenas de un alfabeto cualquiera, proporcione la complejidad de la secuencia de simbolos para casos generales; este trabajo no da lugar para una discusión detallada de la no computabilidad de la complejidad de Kolmogorov, sino de ciertas analogías del área de la teoría de la información y su aplicación en información mutua para crear medidas de similaridad entre objetos basadas en la compresión de datos, en la siguiente sección se hará una mejor explicación de esta analogía.

Dado un algoritmo de compresión, entonces se define $C(x)$ como el tamaño en bytes de la secuencia x y se puede aproximar 1.3 con la siguiente expresión:

$$d_C(x,y) = \frac{C(x|y) + C(y|x)}{C(xy)}. \tag{1.4}$$

El principio fundamental de estas medidas de similaridad, es que al tener 2 secuencias x y y, y estas son comprimidas juntas (concatenadas), entonces estas secuencias deben compartir información (di)similar. El algoritmo que realice el mejor radio de compresión respecto los datos de estudio, realizará la mejor aproximación de la medida d_c para d_k.

1.2.3. Entropía e Información Mutua

En muchos trabajos hay diferentes y parecidas definiciones de probabilidad, en esta versión, probabilidad es la cantidad de *E*speranza que tiene un observador específico de

que ocurra un evento determinado. Si hay N distintos eventos posibles determinados cada uno por una variable x, es decir $(x_1, x_2, ..., x_N)$ y los eventos se producen con frecuencias desconocidas n $(n_1, n_2, ..., n_N)$, se dice que la probabilidad de ocurrencia de un evento x_i está dado por el número de ocurrencias durante todos los distintos eventos, y está dada por

$$P(x_i) = \frac{n_i}{\sum_{j=n}^{N} n_j}, \tag{1.5}$$

la expresión anterior tiene una importante propiedad, y es que la suma de probabilidades de los eventos ocurridos debe ser igual a 1, es decir,

$$\sum_{i=1}^{N} P(x_i) = 1. \tag{1.6}$$

Sean dos eventos cualesquiera x y y, otras propiedades básicas de probabilidad son: $P(\sim x) = 1 - P(x)$ y $P(x \cup y) = P(x) + P(y) - P(x \cap y)$, comúnmente se denota $P(x \cap y)$ por $P(x, y)$. Si $P(x, y) = 0$, entonces se dice que x y y son mutuamente exclusivos. Teniendo en cuenta estas propiedades se puede dar un salto a la definición de probabilidad condicional, la cual se denota como $P(x|y)$, y se expresa como la probabilidad de ocurrencia de un evento x dado que el evento y ya ha ocurrido, la probabilidad condicional es evaluada por medio del teorema de Bayes y está obtenida por la ecuación 1.7

$$P(x|y) = \frac{P(x|y)P(x)}{P(y)}. \tag{1.7}$$

En 1948 Claude E. Shannon estableció los principios de la teoría de la información, por lo cual se le conoce como el padre de esta teoría. Shannon expuso el costo computacional en bits (log_2) en la transmisión de información entre una fuente a un receptor por medio de un canal de datos con presencia o ausencia de ruido, ver Figura 1.3.

Para definir información, suponga que tiene n cantidad de símbolos $\{a_1, a_2, \ldots, a_n\}$ de un determinado alfabeto, estos símbolos son enviados por una fuente con probabilidad de ocurrencia $\{p_1, p_2, \ldots, p_n\}$. Si se observa un símbolo determinado de la secuencia a_i en un momento dado de la información se obtendría $log(1/p_i)$ de información

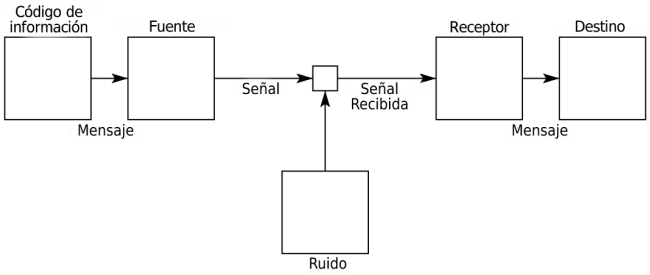

Figura 1.3 Esquema de un sistema de comunicación según C. E. Shannon.

Base logarítmica	Unidad
2	bit
e	nat
10	dit o hartley
256	byte
2^{1000}	kilobit
2^{8000}	kilobyte
2^{10^6}	megabit
$2^{2^9 5^6}$	megabyte
2^{10^9}	gigabit
$2^{2^{12} 5^9}$	gigabyte

Cuadro 1.1 Unidades de información y su base logarítmica.

del símbolo en particular y donde la base del logaritmo es arbitraria y depende de la unidad de información, véase Tabla 1.1.

En un mensaje largo de símbolos aleatorios de tamaño N, frecuentemente se necesita el tamaño total de los símbolos en el mensaje para lograr una información máxima, según la formula de información total, I está dada por la siguiente expresión

$$I = \sum_{i=1}^{n} (N * p_i) * log(1/p_i), \tag{1.8}$$

entonces el promedio de información que se obtiene por símbolo observado sería

$$\begin{aligned} I/N &= (1/N)\sum_{i=1}^{n}(N * p_i) * log(1/p_i) \\ &= \sum_{i=1}^{n} p_i * log(1/p_i), \end{aligned} \tag{1.9}$$

esta última ecuación se conoce como la entropía y la mayoría de las veces se encuentra en la literatura como $H(X)$, y se lee como la cantidad de esperanza o incertidumbre que existe en un proceso de comunicación con la información total esperada, es decir, $H(X) = E(I(X))$; ahora supóngase que se tiene un conjunto de probabilidades (una distribución de probabilides) $P = \{p_1, p_2, ..., p_n\}$, se define entropía de la distribución P por

$$H(P) = \sum_{i=1}^{n} p_i * log(1/p_i). \tag{1.10}$$

Cabe anotar que la ecuación anterior define la entropía de una distribución de probabilidad discreta, si se quisiera extender la ecuación anterior al caso continuo es cuestión de escribirla en forma de integral y la expresión queda así

$$H(P) = \int P(x_i) * log(1/P(x_i))dx. \tag{1.11}$$

Según la ecuación 1.10 la entropía de una distribución de probabilidad es el valor de certeza esperado de ella misma [44, 45], es decir, $H(X)$ es la cantidad de valores finitos que puede tomar dicha variable durante un proceso de comunicación. Un concepto que cabe mencionar es el de entropía relativa [46, 47], dadas dos distribuciones de probabilidad cualesquiera, P_i y Q_i, la entropía relativa entre las dos funciones de probabilidad está dada por

$$d(P||Q) = \sum_{x \in X} P(x) log \frac{P(x)}{Q(x)} \geq 0, \tag{1.12}$$

véase la sumatoria anterior como una medida de distancia de similitud entre dos distribuciones de probabilidad, esta distancia no necesariamente es simétrica, es decir, $d(P||Q) \neq d(Q||P)$.

La entropía también cuenta con una formula de condicional; considere un sistema estocástico con entradas X y salida Y, ambas X y Y están determinadas por valores discretos x_i y y_i. El concepto de entropía condicional está dado por la ecuación 1.13

$$H(X|Y) = H(X,Y) - H(Y), \tag{1.13}$$

la ecuación anterior tiene una importante propiedad, y es que

$$0 \leq H(X|Y) \leq H(X), \tag{1.14}$$

la cantidad expresada como $H(X,Y)$ en 1.13, se refiere a la entropía en común entre X y Y, definida en la ecuación 1.15

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)log(p(x,y)), \tag{1.15}$$

donde $p(x,y)$ representa la función de densidad de probabilidad de la variable aleatoria X y Y.

Teniendo la entropía $H(X)$, la cual representa el grado de incertidumbre antes de haber observado la entrada del sistema; y la entropía condicional $H(X|Y)$ la cual representa el grado de incertidumbre después de haber observado una salida en el sistema; la diferencia entre $H(X) - H(X|Y)$ debe representar la incertidumbre de la posible entrada del sistema, evaluada desde una observación de la salida del sistema, véase Figura 1.4. Esta cantidad es llamada información mutua, evaluada como muestra la ecuación 1.16

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \sum_{x\in X}\sum_{y\in Y} P(x,y)log(\frac{P(x,y)}{P(x)P(y)}) \end{aligned} \tag{1.16}$$

donde $P(x,y)$ es la función de distribución de la probabilidad en común entre x y y, y $P(x)$ y $P(y)$ son las funciones de densidad de probabilidad marginal de x y y

respectivamente. A. N. Kolmogorov se basó en la ecuación 1.16 para dar su enfoque de complejidad algorítmica de la forma como se muestra en la siguiente expresión

$$I(X;Y) = \int \int P_{xy}(dx, dy) log_2 \frac{P_{xy}(dx, dy)}{P_x(dx)P_y(dy)} \tag{1.17}$$

La entropía puede verse como un caso especial de información mutua, cuando $H(X) = I(X;X)$. La información mutua entre X y Y es simétrica, es decir, $I(X;Y) = I(Y;X)$, y no puede ser negativa $I(X;Y) \geq 0$.

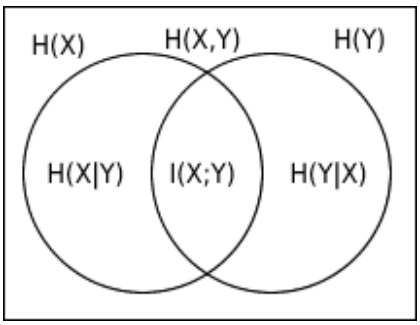

Figura 1.4 Relación entre la información mutua $I(X;Y)$ y la entropía $H(X)$ y $H(Y)$

De la Figura 1.4 se pueden extraer las siguientes igualdades para obtener la información mutua entre dos secuencias:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X,Y) - H(X|Y) - H(Y|X) \end{aligned}$$

1.2.4. Distancia Normalizada de Información

El objetivo de estudio de una distancia de información es encontrar una métrica de distancia absoluta de información entre dos objetos de cualquier clase [48, 49]; cabe anotar que hay una distinción técnica entre los términos métrica y medida. Métrica debe satisfacer los requisitos formales de identidad, simetría y triángulo de desigualdad. Sin

pérdida de generalidad, una distancia solamente necesita operar sobre una secuencia finita de $0'$ y $1'$. Una Función de distancia D con valores reales positivos, definida sobre un producto Cartesiano $X \times X$ de un conjunto X es llamada métrica si para todo $x, y, z \in X$ se mantienen los siguientes axiomas:

- $D(x, y) = 0$ iff $x = y$ (axioma de identidad).
- $D(x, y) = D(y, x)$ (axioma de simetría).
- $D(x, y) + D(y, z) \geq D(x, z)$ (triángulo de desigualdad).

El valor $D(x, y)$ es llamada la distancia entre x y $y \in X$. Definir una distancia de información como la longitud más corta de un programa binario que calcula x desde y, y también consigue calcular y desde x; al ser el más corto, tal programa debe tomar ventaja de cualquier redundancia entre la información necesaria para ir de x a y y viceversa, en términos de esperanza lo anterior se describe matemáticamente según la siguiente expresión

$$E(x, y) = min\{length(p) : U(x, p) = y \wedge U(y, p) = x\}. \tag{1.18}$$

Un límite de cota superior para computar $E(x, y)$ es $K(x|y) + K(y|x)$; también, se puede hallar una distancia máxima entre las dos secuencias x y y; y una suma de distancias entre las dos funciones, quedando

$$E(x, y) = K(x|y) + K(y|x) \tag{1.19}$$

y la distancia máxima entre las dos secuencias x y y se computa de la siguiente manera

$$E(x, y) = max\{K(x|y), K(y|x)\}. \tag{1.20}$$

La distancia normalizada de información (NID según sus siglas e inglés) es una métrica absoluta y universal de similaridad entre dos objetos. Esta es basada en la

complejidad de Kolmogorov y como tal, también es incomputable para casos generales [50], la ecuación matemática de la NID es

$$NID(x,y) = \frac{E(x,y)}{max\{K(x),K(y)\}}. \tag{1.21}$$

NID es la longitud normalizada del programa más corto que puede calcular x conociendo y, y también calcular y conociendo x, obteniendo una métrica absoluta y universal de similaridad entre dos objetos, con las siguientes características:

- $NID(x,y) = 0$ si $x = y$
- $NID(x,y) = 1$ distancia maxima de disimilaridad.

Cómo se dijo anteriormente con el uso de algoritmos de compresión también se puede llegar a diferentes aproximaciones de la expresión 1.21, las cuales resultan útiles y prácticas para hallar (di)similaridad de diferentes clases de objetos en diferentes máquinas de aprendizaje; así surgieron varias técnicas de medidas de similaridad basadas en compresión [51], aquí se explican las más conocidas según la literatura actual.

1.3. Medidas de Similaridad Basadas en Compresión

Se puede utilizar la noción de la complejidad de kolmogorov en el computo de la información mutua entre dos objetos, teniendo en cuenta el gráfico de la Figura 1.4, se describe una función simétrica de la información mutua entre dos conjuntos de la siguiente manera

$$K(x|y) + K(y) = K(y|x) + K(x), \tag{1.22}$$

se factoriza la expresión anterior con el fin de hallar una relación de información entre las dos secuencias, y con el fin de normalizar entre $[0,1]$ el rango de valores, se divide

por la complejidad de Kolmogorov de ambas cadenas concatenadas, es decir $K(xy)$, obtieniendo una relación de las dos secuencias como se muestra a continuación

$$\frac{K(x) - K(x|y)}{K(xy)} = \frac{K(y) - K(y|x)}{K(xy)}, \tag{1.23}$$

y se puede determinar una función de distancia expresada según la ecuación 1.24

$$d_K(x, y) = 1 - \frac{K(x|y) + K(y)}{K(xy)} \tag{1.24}$$

donde 0 es el grado de mayor similitud entre las dos cadenas y 1 cuando muestran una completa disimilaridad entre ellas.

Con el fin de aproximar $K(\cdot)$ y $K(\cdot|\cdot)$ se hace uso de algoritmos y programas de compresión, denotando el número de bytes que representa cada uno de los objetos de estudio como $C(\cdot)$ y $C(\cdot|\cdot)$, los mismos autores de esta medida crearon *GenCompress* [52] un algoritmo para comprimir cadenas de ADN con un excelente radio de compresión, así, definen una métrica de similaridad entre dos secuencias basada en la complejidad de Kolmogorov expresada según la ecuación 1.25

$$d_C(x, y) = 1 - \frac{C(x|y) + C(y)}{C(xy)} \tag{1.25}$$

Otra medida de similaridad bastante conocida y utilizada ampliamente en máquinas de aprendizaje no supervisadas a través de métodos de agrupación es la **NCD** (Normalized Compressión Distance); esta medida opera en un rango de valores entre $[0, 1 + \epsilon]$, y se computa según la siguiente expresión

$$d_C(x, y) = \frac{C(xy) - min\{C(x), C(y)\}}{max\{C(x), C(y)\}} \tag{1.26}$$

la medida NCD descrita según la ecuación 1.26, ha sido utilizada para clasificar cadenas de ADN, lenguajes naturales, ritmos musicales [53], imágenes [54], texto, entre otros. Los autores de ésta técnica también crearon un paquete de herramientas para compresión y aprendizaje automátizado llamado CompLearn el cual se puede encontrar en [55].

Otra medida de distancia de similaridad es llamada **CDM** (Compression-based Dissimilarity Measure), el rango de valores de esta distancia está entre $[\frac{1}{2}, 1]$; donde 0,5 es el mayor grado de similitud entre las secuencias estudiadas y 1 es la completa disimilaridad entre las cadenas [56]. La técnica CDM también aproxima la complejidad de Kolmogorov por medio de algoritmos de compresión y toma pautas teóricas de la técnica NCD. La expresión matemática para computar la distancia CDM está dada por la ecuación 1.27

$$d_C(x, y) = \frac{C(xy)}{C(x) + C(y)}. \tag{1.27}$$

1.4. Clasificación y Agrupamiento

Ha resultado de gran utilidad en diversas áreas la creación distancias normalizadas de información entre objetos con el fin de emplear tareas de máquinas de aprendizaje automátizado, especialmente algoritmos de agrupamiento (Clustering), con el fin de realizar clasificación entre los objetos de estudio. En los últimos años, la clasifiación de objetos con algoritmos de agrupamiento se ha implementado en áreas como la medicina (clasificación de enfermedades), química (agrupamiento de compuestos), estudios sociales (clasificación de estadísticas), entre otros.

Como ya se mencionó en el anterior párrafo, las máquinas de aprendizaje más utilizadas para la clasificación entre objetos usando las técnicas descritas en este capitulo, son los algoritmos de agrupamiento; estos algoritmos utilizan una diferente forma de aprendizaje que emplean aproximaciones no supervisadas, es decir, se trata de construir un clasificador sin información a priori, o sea, a partir de conjuntos de patrones no etiquetados. En la figura 1.5 se muestra un esquema funcional de los algoritmos de agrupamiento, donde a partir de un conjunto de M patrones no etiquetados X_i, $i = 1, 2, ..., M$ encuentra K agrupamientos S_j, $j = 1, 2, ..., K$.

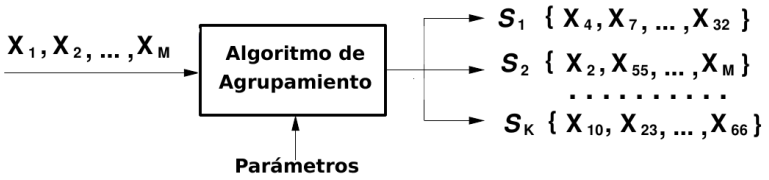

Figura 1.5 Esquema funcional de un algoritmo de agrupamiento

Estos algoritmos se definen como técnicas diseñadas para realizar una clasificación asignando patrones a grupos, de tal forma que cada grupo sea más o menos homogeneo y distinto de los demás en un espacio de representación. El criterio de homogeneidad más simple está basado en una matriz de distancia entre todos los objetos de estudio.

A continuación se muestran dos dendogramas, los cuales son los diagramas más utilizados para graficar la salida de los algoritmos de agrupamiento después de haber analizado las matrices de distancias obtenidas de los objetos analizados. El gráfico de la parte izquierda de la Figura 1.6 fue extraído desde [3], en este se muestra un dendograma donde se clasifican 24 diferentes clases de genomas mitocondriales (mtDNA por sus siglas en inglés) de la clase mammalia; en la parte derecha de la Figura 1.6 se muestra un diagrama de árbol jerárquico copiado desde [57], donde los autores extrajeron aleatoriamente 10 subsecuencias cada una de 2000 muestras de dos base de datos de señales electrocardiográficas y hacen una comparación de la clasificación obetenida con la medida CDM (izq) y la distancia euclidiana (der).

1.5. Recomendaciones

Las medidas de distancias de similaridad basadas en compresión referidas en este documento han demostrado tener amplia aceptación en el campo científico en un sinfín de áreas y aplicaciones, ellas se proponen como un análisis para la clasifiación de datos sin necesidad de parámetros en el proceso; no obstante, se debe escoger siempre un

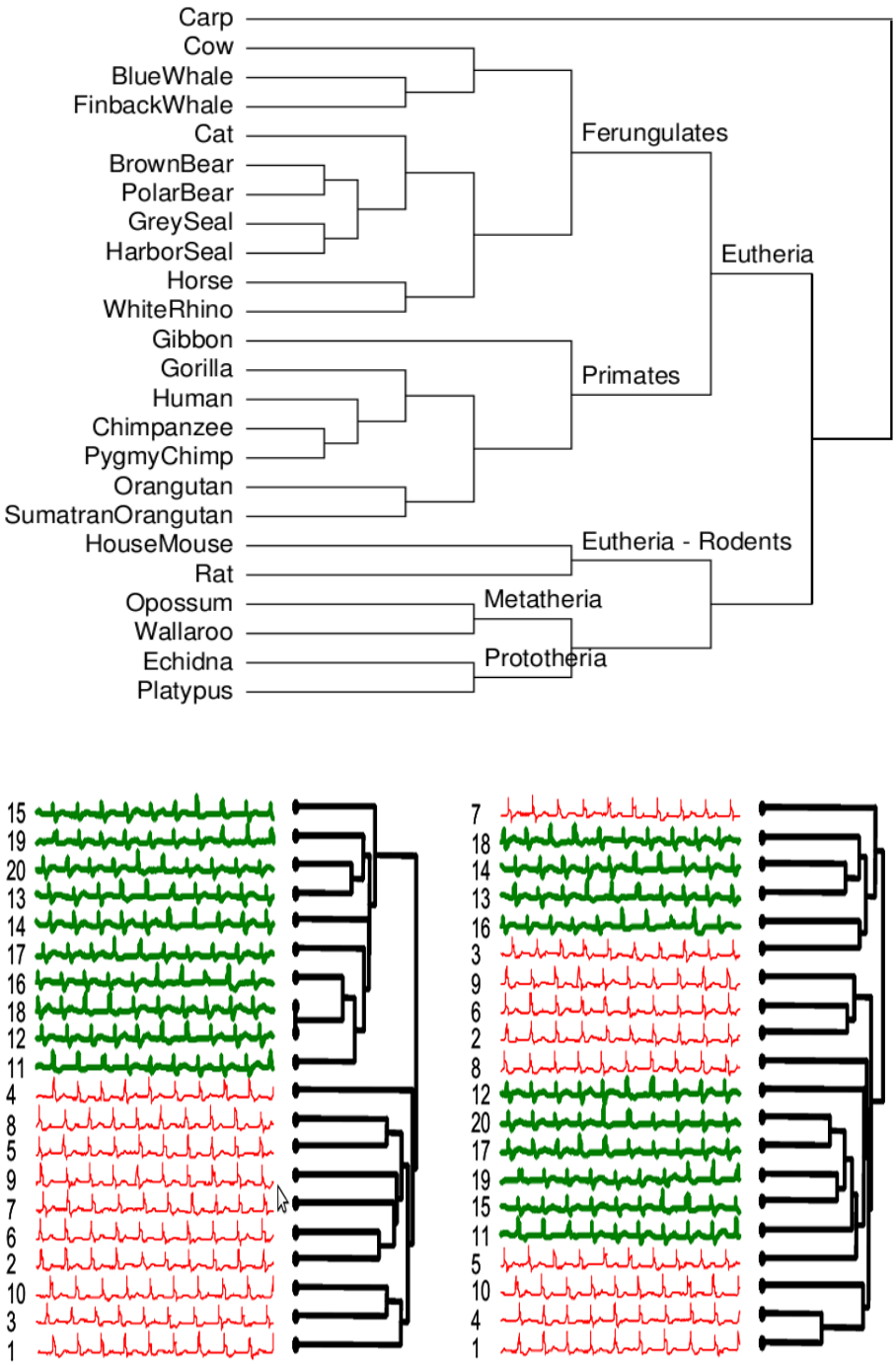

Figura 1.6 Izq: Árbol evolutivo de un conjunto de mtDNA de mamíferos después de haber extraído la matriz de disimilaridad con la técnica NCD. Der: Clasificación de señales electrocardiográficas con la ténica CDM.

campo de carácteristicas en común entre los objetos de estudio, que indique un comportamiento lógico de los datos que se quieren clasificar. Los algoritmos de compresión se clasifican también por su desempeño, aquellos que logren un mejor radio de compresión entre los objetos analizados hará la mejor aproximación de la distancia d_C para d_K.

La entropía puede verse como un caso especial de información mutua, cuando $H(X) = I(X;X)$ y $K(x)$ como el valor aproximado de la entropía de la distribución de probabilidad de una secuencia aleatoria X. La relación de la complejidad de Kolmogorov y la teoría de la información de Shannon se debe en gran sentido al concepto de información mutua entre variables aleatorias.

Estas y otras técnicas de medidas de similaridad basadas en compresión son implementadas para el análisis de clases de datos que comparten información útil. A través de procesos de compresión se halla una medida de disimilitud entre cada uno de los objetos; en un número que se encuentra normalizado para su posterior clasificación en algoritmos de máquinas de aprendizaje mostrando mejores resultados los algoritmos de agrupamiento. Estas medidas se han implementado en ADN, series temporales, imágenes [58], audio, video [59], entre otras,

EÉstas técnicas basadas en compresión también han sido utilizadas para la detección de anomalías en series temporales, mostrando excelentes resultados en la clasificación de enfermedades y patologías en diferentes clases de señales tales como electroencefalogramas [14].

El profesor Keogh y colaboradores hace poco han creado una aproximación numérica basadas en simbolos llamada Symbolic Aggregate approXimation de siglas SAX [60], esta aproximación lograría una mejor compresión de los datos y como tal una mejor

apoximación de la complejidad de Kolmogorov al momento de comprimir la HRV en señales electrocardiográficas de un periodo de tiempo determinado.

Al observar el dendograma de la parte derecha de la figura 1.6 se observa que se realiza una excelente clasificación en señales electrocardiográficas y como trabajo se propone realizar una implementación de la técnica CDM en la clasificación de enfermedades cardiovasculares como el Infarto Agudo de Miocardio.

2. El Electrocardiagrama y Procesamiento Digital de Señales

2.1. Introducción

Cada corazón tiene su propio ritmo normal provocado por el impecable flujo de impulsos eléctricos que comienzan en el "marcapasos"natural de corazón, el nódulo sinusal. El electrocardiograma es la representación gráfica del impulso eléctrico entre dos puntos del cuerpo y el tiempo. La digitalización de un electrocardiograma y estudio de procesamiento digital de señales conlleva un marco teórico entre el uso de la Variabilidad de la Frecuencia Cardiaca en la clasificación de la enfermedad de Infarto Agudo de miocardio y los principales conceptos del procesamiento de señales digitales.

El sistema cardiaco es bastante complejo y algunas enfermedades coronarias como el Infarto Agudo de Miocardio representan la principal causa de muerte en países desarrollados y la tercera causa de muerte en países en vía de desarrollo, según la Organización Mundial de la salud se calcula que en 2030 morirán cerca de 23,6 millones de personas por enfermedades cardiovasculares, sobre todo por cardiopatías e Infartos Agudo de Miocardio, y se prevé que sigan siendo la principal causa de muerte.

2.2. El Electrocardiograma

A comienzos del siglo XX el experto Willem Einthoven considerado el padre de la electrocardiografía y los doctores G. fahr y A. De Waarts describieron como se podía ver representado el electrocardiagrama como la dirección y el tamaño de los potenciales eléctricos del corazón [61], estos científicos clasificaron cada una de las ondas *PQRST* de las señales emitidas por el corazón cada vez que se presenta un latido, y el proceso realizado para bombear la sangre a las demás partes del cuerpo. El ECG normal se considera una señal periódica, un período de esta señal es un ciclo donde el corazón transfiere sangre a todas partes del cuerpo a través de las arterias [62], ver Figura 2.1.

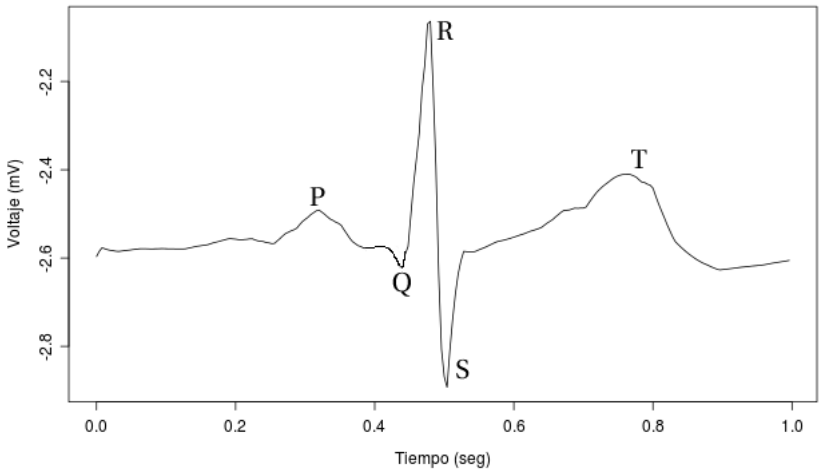

Figura 2.1 Ondas PQRST de un ECG normal propuestas por Einthoven

El sistema cardiaco es bastante complejo y algunas enfermedades coronarias como el Infarto Agudo de Miocardio representan hoy la principal causa de muerte en países desarrollados y la tercera causa de muerte en países en vía de desarrollo [63], según el informe anual del 2004 de la Organización Mundial de la salud esta clase de enfermedades fueron la principal causa de muerte en hombre y mujeres a nivel mundial [64], así todas las contribuciones que se realicen en pro de mejorar la calidad de vida de la humanidad son bienvenidas para la sociedad.

El corazón es el órgano principal del sistema circulatorio; para que el cerebro, los tejidos, células, y demás sistemas en nuestro cuerpo funcionen correctamente necesitan que el corazón les envíe sangre y así suministre el oxigeno que todos estos necesitan. El corazón está formado por cuatro cavidades (ver Figura 2.2) y un conjunto de arterias llamadas arterias coronarias (ver Figura 2.3) que son las encargadas de regar sangre en todo este músculo asegurando así su adecuado funcionamiento [1].

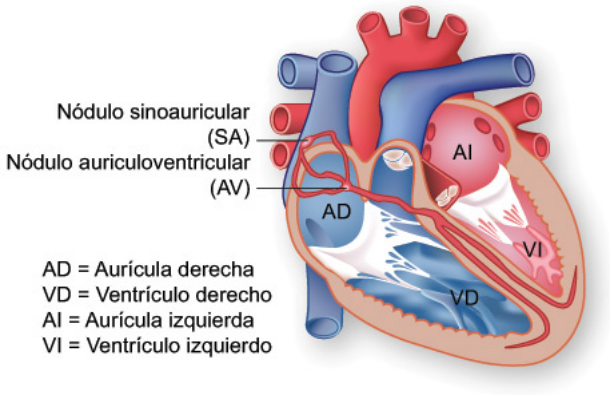

Figura 2.2 Anatomía del corazón.

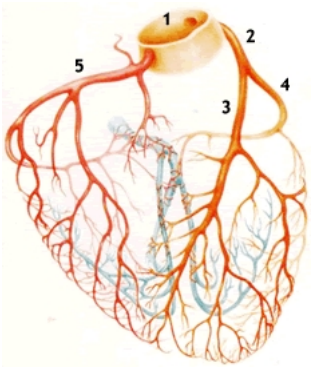

Figura 2.3 Esquema de las arterias coronarias.

En las enfermedades coronarias el ECG juega un papel fundamental, este junto con el cuadro clinico de angina y la elevación de enzimas cardiacas realiza un diagnóstico definitivo de infarto de miocardio. Cuando se produce una obstrucción en estas arterias se suprime el aporte sanguíneo y trae como consecuencia una disminución del flujo sanguíneo a través de las arterias coronarias impidiendo un adecuado suministro de oxígeno y nutrientes al músculo cardiaco ocasionando alteración en el funcionamiento del corazón.

[1] Figura 2.2 tomada del sitio del Instituto del corazón de Texa s y la figura 2.3 del sitio de la Federación argentina de Cardiología

2.2.1. Variabilidad de la Frecuencia Cardiaca - VFC

Puesto que la frecuencia cardiaca es alterada por muchos aspectos externos, permitiendo así que se pueda estudiar la actividad del sistema nervioso autónomo de manera no invasiva. En el ECG se detecta cada onda R y se calcula el tiempo entre ondas R sucesivas o intervalo RR. El intervalo RR mide el período cardiaco, y su inverso mide la frecuencia cardíaca. Para el análisis de la VFC se emplean métodos estadísticos, como la media y la varianza, e indíces como los que se muestran en la Tabla 2.1.

Índice	**Unidades**	**Descripción**
SDNN	mseg	Desviación estándar de la serie RR
SDANN	mseg	Desviación estándar de la media de la serie RR en periodos de 5 minutos
SDNNIDX	mseg	Media de la desviación estándar de la serie RR en periodos de 5 minutos
pNN50	-	Porcentaje de valores de la serie RR que difieren del anterior más de 50 mseg
r-MSSD	mseg	Raíz cuadrada de la media de las diferencias al cuadrado entre los valores de la serie RR adyacentes
TINN	mseg	Base del triángulo al que se ajusta el histograma
HRV index	-	N/M siendo N el número de valores de la serie RR y M el máximo del correspondiente histograma
IRRR	mseg	Diferencias entre los cuartiles tercero y primero de la serie RR
MADRR	mseg	Mediana de las diferencias absolutas entre los valores adyacentes de la serie RR

Cuadro 2.1 Índices estadísticos en el dominio del tiempo de la VFC.

El estudio de la Variabilidad de la serie RR durante mucho tiempo estuvo confinado a los laboratorios, sin embargo, con los avances en la tecnología de los microprocesadores, este análisis puede llevarse a cabo clínicamente, y posiblemente en un futuro cada familia podrá tener un kit de señales fisiológicas eléctricas en sus casas o dispositivos móviles que tomen lectura de esta clase de señales y puedan ser enviadas para su valoración y clasificación por un médico o una computadora en cualquier parte del mundo.

A finales de los 90 se descubrió que al haber indicios de un Infarto Agudo de Miocardio y luego del mismo durante un tiempo de algunos meses (si hay recuperación), existe una disminución de la variabilidad de la frecuencia cardiaca en las personas que

lo sufren [65], véase Figura 2.4.

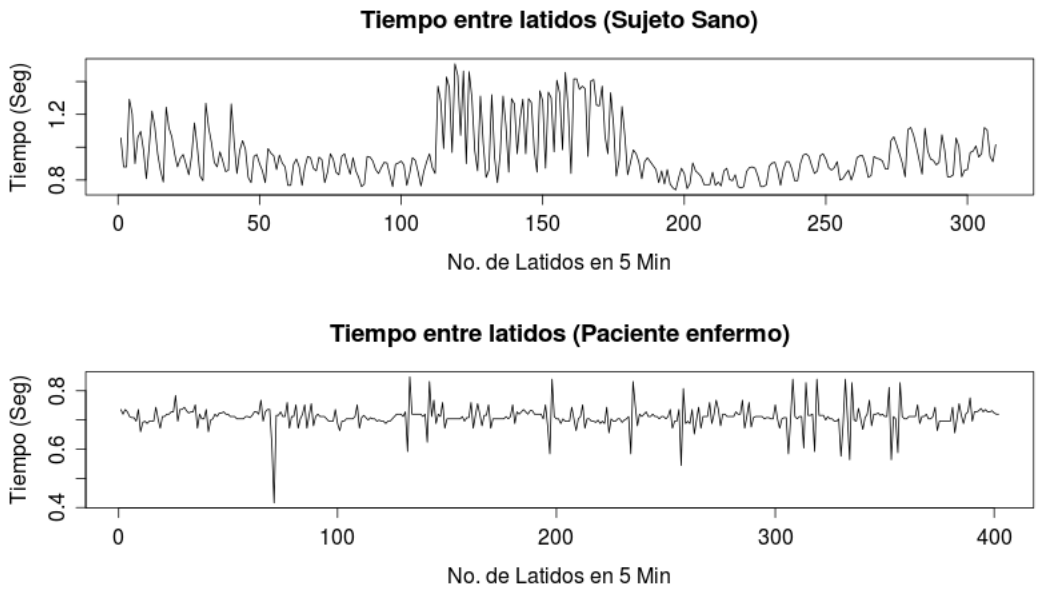

Figura 2.4 Comparación de ritmo cardiacoo en segundos entre una persona sana (arriba) y una persona con Infarto Agudo de Miocardio (abajo).

2.2.2. Infarto Agudo de Miocardio

En 1912 James Bryan Herrick presentó el concepto moderno de trombosis coronaria y reportó el primer caso de Infarto Agudo del Miocardio, en el cual el ECG fue de vital importancia. La electrocardiografía es una de las áreas de la ingeniería biomédica que ha salido mejor favorecida por loos desarrollos tecnológicos, no sólo en equipos de adquisición de señales cardiacas, sino también de aplicaciones de software tendientes a integrar la tecnología de comunicaciones a los procesos de captura, procesamiento y diagnóstico.

En un infarto agudo de miocardio la falta de riego sanguíneo que sufre el corazón al obstruirse las arterias coronarias, las cuales suelen manifiestarse en el ECG por una elevación de las ondas ST e inversión de la onda Q [66], según las manifestaciones anteriores el infarto de miocardio se divide en infarto con Q o sin Q para clasificar los infartos desde el ECG convencional. En ocasiones, esta inversión de T no indica nece-

sariamente la presencia de un infarto agudo de miocardio, sino tan sólo un déficit de riego sanguíneo al corazón sin IAM, por lo cual este proceso patológico suele tener tres fases: Isquemia, Lesión y Necrosis.

La práctica de la medicina ha evolucionado en el sentido de la variedad y la cantidad de información que se maneja. Cada vez se puede tener un volumen más grande de datos con mayor precisión y en menor tiempo. El tratamiento de señales digitales ha sido de mucha utilidad en cuanto a la caracterización de patrones y patologías en muchos tipos de señales fisiológicas, entre los análisis y operaciones más utilizadas es la transformación de Fourier y la transformación Wavelet las cuales se describirán brevemente en la Sección 2.3, y de la transformación Wavelet en señales electrocardiográficas se referirá el Capítulo 3.

2.3. Procesamiento Digital de Señales

La conversión de una señal análoga a digital se realiza a través de un proceso de muestreo, este teorema fue propuesto por Claude Shannon (padre de la Teoría de la Información), formulando la frecuencia de muestreo mínima para garantizar condiciones óptimas de fidelidad al momento de reconstrucción de una señal continua en su representación digital, esta frecuencia es la comúnmente llamada frecuencia de Nyquist [67].

El Teorema del Muestreo de shannon, también conocido como teorema de muestreo de Whittaker-Nyquist-Kotelnikov-Shannon, establece que la frecuencia mínima de muestreo necesaria para evitar el Aliasing, ver Figura 2.5, y hacer una perfecta reconstrucción análoga de una señal $f(x)$ debe ser $fm > 2AB$, con fm frecuencia de muestreo, AB ancho de banda de la señal $f(x)$ a muestrear y $AB = f_{max} - f_{min}$, frecuencia máxima

menos frecuencia mínima; cabe destacar que para señales con $f_{min} = 0$, se puede expresar como $fm > 2f_{max}$.

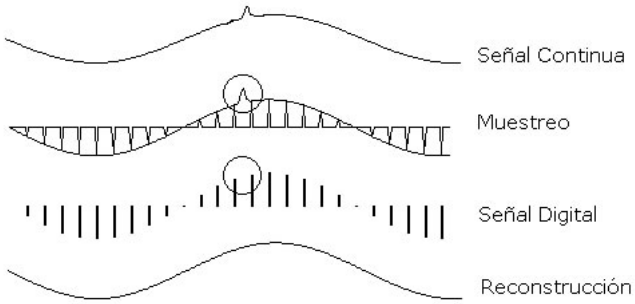

Figura 2.5 Fenómeno de aliasing durante un proceso de muestreo y reconstrucción

Las señales digitales se representan como secuencias de números llamadas muestras, el valor de una muestra de una señal en tiempo discreto se denota cmo $f[n]$, con n un número entero en el intervalo $-\infty$ y ∞. En la mayoría de los casos, las señales de interés son de manera natural funciones discretas de las variables independientes, y es común que estas señales sean finitas, este tipo de señales de extensión finita, a las cuales también se les suele llamar series de tiempo o series temporales, ocurren en muchos ámbitos de la sociedad y desempeñan un papel importante en nuestra vida diaria, los datos de las series de tiempo suelen ser ruidosos y sus representaciones requieren modelos basados en sus propiedades estadísticas.

2.3.1. Análisis de Fourier

El Análisis de Fourier informa acerca de la presencia o ausencia de determinadas frecuencias en la señal que se desea analizar [68]. Una de las herramientas más usadas del análisis matemático son las series de Fourier, llamadas así en nombre del matemático francés Jean Baptiste Joseph Fourier (1768–1830), ver Figura 2.6, las series de Fourier

tienen la forma

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty}[a_n \cos(nx) + b_n \sin(nx)] \tag{2.1}$$

donde a_0, a_n y b_n se denominan coeficientes de Fourier de la función $f(x)$. en términos teóricos la investigación de Fourier dice que toda función que se repite periódicamente puede ser expresada como la suma de senos y/o cosenos de diferentes frecuencias, cada uno multiplicado por un coeficiente diferente [69]

Figura 2.6 Jean Baptiste Fourier (1768-1830).

Una de las técnicas más utilizadas en el tratamiento de señales para la extracción de información, ha sido representar las señales en el dominio del tiempo; esto ha permitido asociar cambios de la señal con los parámetros amplitud-tiempo de la señal; posteriormente, el dominio de la frecuencia vino a complementar esta información [70]. La Transformación de Fourier es una herramienta de gran utilidad para el Tratamiento de señales estacionarias, sobre todo desde el descubrimiento de la Transformación rápida de Fourier [62]. La definición de la transformación de Fourier está dada por la siguiente ecuación:

$$F(y) = \int_{\infty}^{\infty} f(x) e^{-2i\pi xy} dx \tag{2.2}$$

para todo x sea este número real y la variable del dominio directo de la función de entrada y y es la variable a ser representada en el dominio de la frecuencia.

La transformación de Fourier (TF) es una herramienta de gran utilidad para el tratamiento de señales estacionarias; pero para el tratamiento de señales no estacionarias la transformación de Fourier presenta más inconvenientes que ventajas, ya que esta transformación supone que el espectro y la amplitud de las señales son uniformes durante todo el período muestreado, y en la práctica se observa que estos parámetros varían en el tiempo en la mayoría de la señales naturales y sociales [71].

2.3.2. Representaciones Tiempo-Frecuencia

A raíz de las limitaciones de la TF surgió la necesidad de realizar representaciones bidimensionales, estas representaciones han sido descritas tomando como bases la teoría de la transformación de Fourier, las representación tiempo-frecuencia más populares en el tratamiento de señales digitales son la transformación de Fourier de tiempo corto y la transformación Wavelet (STFT y WT respectivamete por sus siglas en inglés).

Transformada de Fourier de Tiempo Corto

La operación que se realiza para obtener la Short Time Fourier Transform, se muestra en la siguiente ecuación

$$STFT(\tau, y) = \int_{-\infty}^{\infty} f(x)w(x-\tau)e^{-iyx}dx \tag{2.3}$$

la anterior ecuación puede ser vista como una representación bidimensional de la función $f(x)$, por medio de la cual se obtienen la componente de frecuencia y en el instante τ que tiene la señal $f(x)$. En cada instante τ se realiza una transformación de Fourier. Con la STFT es imposible mejorar al mismo tiempo la resolución temporal y de frecuencia debido a que la ventana es estática, además carece de varias propiedades de la Transformación de Fourier.

Transformada Wavelet

Desde el inicio de los 80' fue introducida la transformada Wavelet extediéndose a muchas áreas de aplicación por todo el análisis que se puede realizar con ella. La transformada Wavelet continua difiere del método de la STFT en la forma de hacer el ventaneo para estudiar la señal de entrada, haciendo posible que la ventana pueda ser ajustada a lo largo del análisis de la misma, pudiendo realizar un análisis tiempo-escala, la ecuación de la WT se muestra a continuación

$$WT(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} f(x)\psi\left(\frac{t-b}{a}\right) dx; a, b \in \mathbb{R}, a \neq 0 \tag{2.4}$$

donde a es la escala y b la traslación, se profundizará más la teoría de la transformada wavelet en el siguiente capítulo.

3. Procesamiento Wavelet de Electrocardiogramas sobre R

3.1. Introducción

En este capítulo se presenta una introducción al estudio de procesamiento wavelet en señales electrocardiográficas, usando diferentes paquetes y funciones para el análisis de datos con la transformada wavelet sobre la plataforma de cálculo numérico R[1]. La transformada Wavelet ha resultado de gran utilidad y ha sido implementada en muchas áreas de la ingeniería a través de muchos lenguajes de programación. El lenguaje R no ha sido ajeno a este fenómeno y en el presente existen diversas implementaciones que permiten el análisis estadístico, filtrado y creación de gráficos acerca de análisis wavelet. En este trabajo se probaron y usaron los paquetes msprocess, wmtsa, wavethresh, wavelets, waveslim; y se probaron otras implementaciones en señales electrocardiográficas con el fin de validar los algoritmos de filtrado y caracterización de las bases de datos de señales electrocardiográficas.

3.2. El Motor Estadístico R

R es un un motor estadístico y lenguaje de programación funcional orientado a objetos desarrollado por Ross Ihaka y Robert Gentleman [72], es muy utilizado para el

[1] http://cran.r-project.org

análisis de datos en ciencias de la computación por su generación de gráficos de alta calidad y por su calidad de software libre, últimamente ha recibido un sinnúmero de contribuciones en muchas áreas. R es considerado como un dialecto del lenguaje S creado por Laboratorios AT & T Bell. Este fue premiado en 1998 por la ACM (Association for Computing Machinery) manifestando que este lenguaje ha cambiado la forma como personas analizan, visualizan, y manipulan datos.

R puede ser muy útil con sólo usarlo de manera interactiva. Usos avanzados del sistema llevará al usuario a desarrollar sus propias funciones para sistematizar las tareas repetitivas, o incluso para añadir o cambiar algunas funcionalidades de los actuales paquetes, tomando ventaja de ser de código abierto. R se distribuye gratuitamente bajo los términos GNU (General Public Licence) y está disponible para Linux (Ubuntu, Debian, Fedora, Mandrake, RedHat, SuSe etc), Macintosh y windows. En la figura 3.1 se puede observar un entorno multiplataforma de desarrollo llamado RStudio.

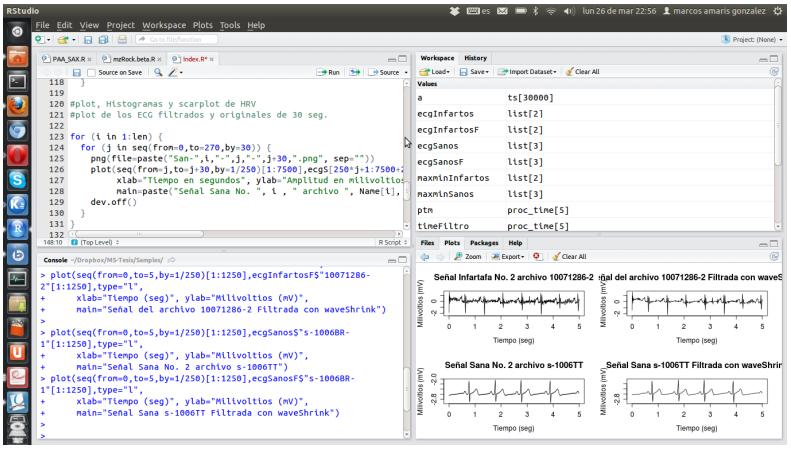

Figura 3.1 Rstudio: IDE multiplataforma de código abierto y gratis para R.

Una de las características más sobresalientes de R es su enorme flexibililidad. Mientras que programas clásicos muestran directamente los resultados de un análisis, R guarda estos resultados como un objeto, de tal manera que se puede hacer un análisis sin necesidad de mostrar su resultado inmediatamente, véase figura 3.2.

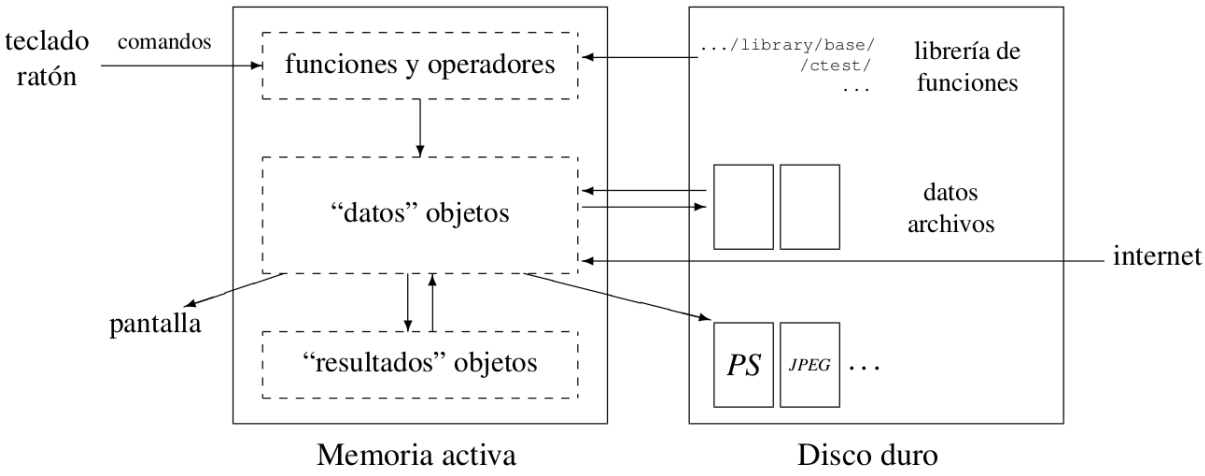

Figura 3.2 Una visión esquemática del funcionamiento de R, tomada de [1]

A continuación se menciona un conjunto de paquetes de R relacionados con el análisis wavelet de señales unidimensionales, bidimensionales y tridimensionales; tales trabajos son importantes y estos exhortan a los investigadores y desarrolladores a reproducir, probar y modificar lo que se crea pertinente por la comunidad científica; esta idea es descrita como investigación reproducible expresada por Buckheit and Donoho en 1995 [73]; estos son: adlift, brainwaver CVThresh, DDHFm, EbayesThresh, msProcesss, nlt, rwt, SpherWave, unbalhaar, waved, wavelets,waveslim, wavthresh, wmtsa etc [74] [75]. En la Tabla 3.1 se presenta una descripción general de los paquetes más utilizados por la comunidad.

Paquete	Versión	Fecha	Funciones	S4	Wavesrink	Alg. Picos	Score
msProcess	1.0.6	2011-02-08	72		Si	Si	47
wavethresh	4.5	2010-03-15	254		Si	Si	144
wmtsa	1.1-1	2011-10-17	53		Si	Si	93
waveslim	1.6.4	2010-06-10	55		Si	No	108
wavelets	0.2-6	2010-04-22	25		No	No	77

Cuadro 3.1 Paquetes Wavelet más utilizados en R

3.3. Breve Teoría Matemática

El término original francés es ondelette, introducido por Jean Morlet y Alex Grossmann, una traducción al español de la palabra es ondeleta u ondícula. El tratamiento realizado con la teansformación wavelet es el Tratamiento de Fourier por medio de pequeñas ondas en un espacio determinado que representa una señal en términos de versiones trasladadas y dilatadas de una onda finita de tiempo llamadas ondeletas, véase Figura 3.3.

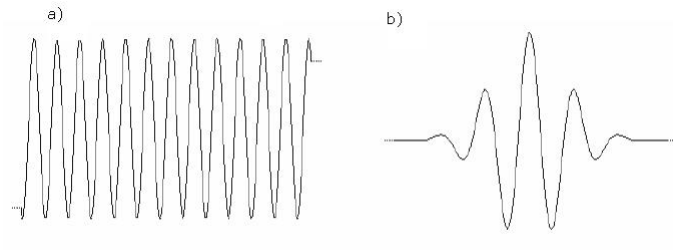

Figura 3.3 a) Onda Senoidal b) Wavelet.

Las wavelets son familias de funciones de análisis que examinan a la señal de interés para obtener sus características de espacio, tamaño y dirección [76]. Principalmente este conjunto de familias de funciones deben cumplir con dos propiedades básicas para poder ser clasificada como wavelet; estas son [77]:

1. El integral de $\psi(t)$ debe ser $= 0$.

$$\int_{-\infty}^{\infty} \psi(t)dt = 0. \tag{3.1}$$

2. El integral del cuadrado de ψ debe ser $= 1$.

$$\int_{-\infty}^{\infty} \psi^2(t)dt = 1. \tag{3.2}$$

Dentro de las wavelets más simple y la más antigua se puede citar la Haar, se

describe con la siguiente ecuación y su gráfica se muestra en la Figura 3.4

$$= \quad 1; \qquad 0 \leq t\frac{1}{2} \tag{3.3}$$

$$\psi(t) \quad = \quad -1; \qquad -\frac{1}{2} \tag{3.4}$$

$$= \quad 0; \qquad \text{otro valor.} \tag{3.5}$$

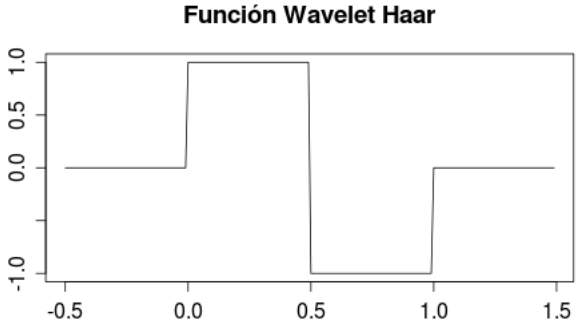

Figura 3.4 Función wavelet Haar.

Al momento de optar por realizar un análisis con la transformada wavelet, se debe realizar un estudio previo para seleccionar el tipo de familias con las que se trabajará, dependiendo del grupo de datos o señales que se vaya a procesar. Una de las ventajas de trabajar con la transformada wavelet para procesamiento de señales biomédicas, tales como el electrocardiograma, es que existen múltiples familias wavelet con las que se puede trabajar.

Otras condiciones que deben cumplir las funciones de familias wavelet determinadas por matemáticos y científicos en esta área son:

1. La función debe tener energía finita

$$E = \int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty. \tag{3.6}$$

2. Si $\hat{\psi}(f)$ es la Transformada de Fourier $\psi(f)$, es decir,

$$\hat{\psi}(f) = \int_{-\infty}^{\infty} \psi(t) e^{-i2\pi ft} dt \tag{3.7}$$

entonces se debe mantener la siguiente condición

$$C_\psi = \int_0^\infty \frac{|\psi(f)|^2}{f} df < \infty \tag{3.8}$$

la ecuación 3.8 es conocida comúnmente como la condición de admisibilidad y C_ψ es llamada la constante de admisibilidad. El valor de C_ψ depende de la selección de la función wavelet.

En la figura 3.5 se muestran respuesta en frecuencia de algunas madres wavelet continuas, las madres wavelet son Haar, Gaussian1, Gaussian2 y Morlet.

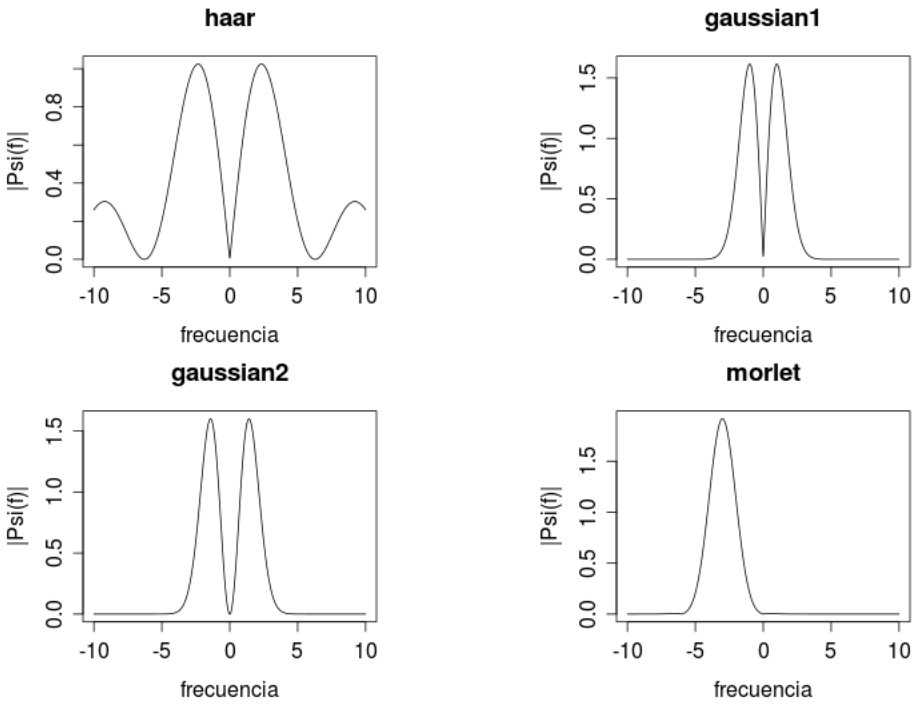

Figura 3.5 Respuesta en frecuencia de filtros de wavelet continua.

.

Las funciones wavelet se caracterizan también por la ortogonalidad, simetría y soporte compacto, dependiendo del objetivo final del procesamiento estas propiedades ayudarán a seleccionar una base wavelet para el análisis de la señal específica, véase Tabla 3.2. Por ejemplo, la propiedad de ortogonalidad indica que el producto interno

de la onda de base es la unidad consigo misma, y de cero con otras wavelet escaladas y trasladadas. Como resultado, una wavelet ortogonal es eficiente para la descomposición de la señal en bandas de subfrecuencias que no se superponen. La propiedad de simetría asegura que una base wavelet puede servir como un filtro de fase lineal. Una wavelet de soporte compacto es una función cuya base es distinto de cero sólo dentro de un intervalo finito. Esto permite que la transformada wavelet represente eficazmente a las señales que tienen características localizadas.

Característica	**Daubechies**	**Symmlet**	**Coiflet**
Nombre corto	Db	Sym	Coif
Orden N	N strictly positive integer	$N = 2, 3, ...$	$N = 1, 2, ..., 5$
Ejemplos	Db1 o haar, Db4, Db15	Sym2, Sym8	Coif2, Coif4
Ortogonal	Si	Si	Si
Biortogonal	Si	Si	Si
soporte compacto	Si	Si	Si
DWT	Posible	Posible	Posible
CWT	Posible	Posible	Posible
Tamaño de Filtros	2N	2N	6N
Simetría	Desde lejos	Desde cerca	Desde cerca
Momentos de Desvanecimiento de ψ	N	N	2N

Cuadro 3.2 Características generales de las familias wavelet más populares

3.3.1. Transformada Wavelet Continua

En el capítulo anterior en la sección 2.4 se muestra la ecuación de la Transformada Wavelet Continua, se observa que el resultado es una señal de dos dimensiones, donde la energía de la señal en la escala a y ubicación b está dada por una función de densidad de energía wavelet bi-dimensional conocida gráficamente como escalograma, véase Figura 3.6.

La CWT transforma la señal de un dominio a otro que depende de 2 variables. La variable de escala a lleva la información de la dilatación y la contracción de la

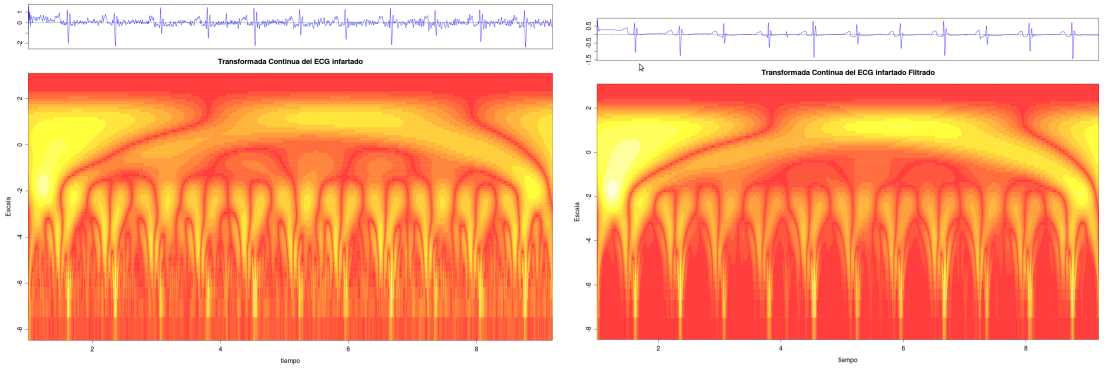

Figura 3.6 Escalogramas de ECG infartado antes y después del proceso de filtrado

señal, pero podría verse desde otro punto de vista donde lo que cambia es la frecuencia y con ello al dilatarse la frecuencia se reduce y al contraerse la frecuencia aumenta. Aquí es justamente donde se aprovecha está característica de las wavelets para obtener la información de la señal y conocer sus componentes de frecuencia. En el dominio del tiempo el análisis es más sencillo ya que la variable de traslación b tiene la información de tiempo. De este modo se completan los datos en forma de una matriz, donde para cada integral que se resuelva se tendrá un punto del plano traslación-escala, que es equivalente a tener la información en el plano tiempo-frecuencia, en la figura 3.7 se observan diferentes escalogramas con las principales wavelet continuas.

3.3.2. Transformada Wavelet Discreta

Los mecanismos para hallar la transformada Wavelet Discreta, son usualmente entendidos por Ingenieros, como un banco de filtros. En el común de los casos, en este proceso se emplean funciones wavelet con carácteristicas de ortonormalidad y ortogonalidad. La DWT se logra a través de una descomposición de la señal original, a esta última se le extraen funciones de detalle y aproximación, véase Figura 3.8; y en la medida en que va aumentando de nivel se va realizando un análisis multiresolución, extrayendo valiosa información del espacio analizado, véase figura 3.9.

El proceso más empleado para hallar la DWT es por medio del algoritmo piramidal

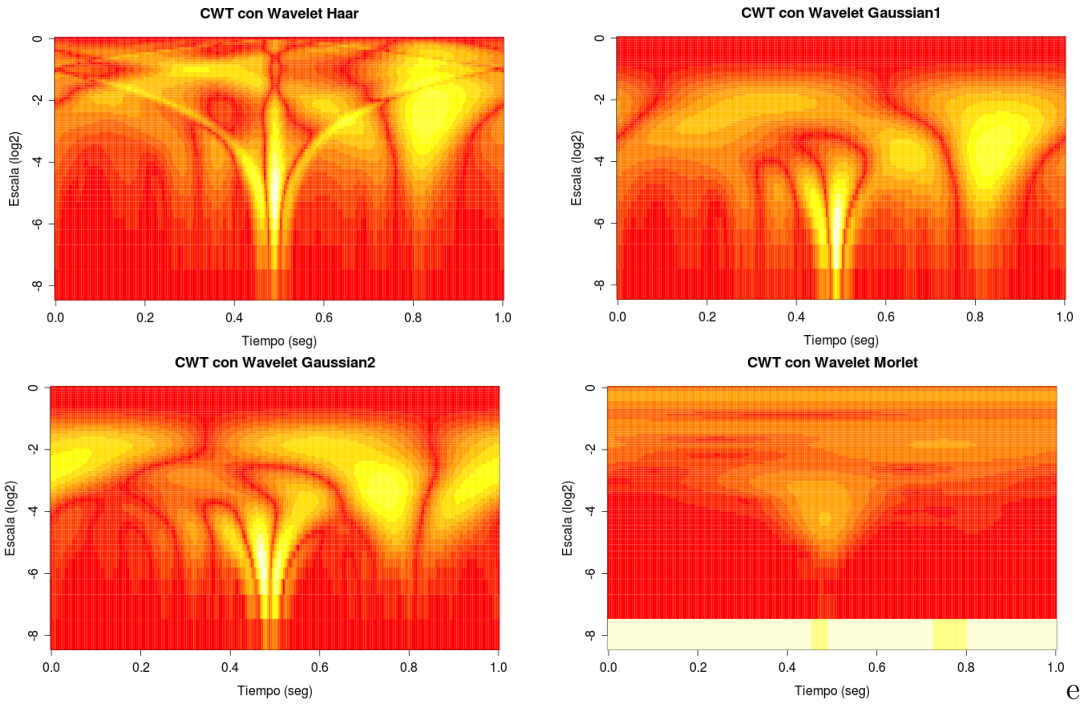

Figura 3.7 Transformadas Wavelet Continua de la Figura 2.1 con la función wavelet Haar, Gauusian1, Gaussian2 y Morlet.

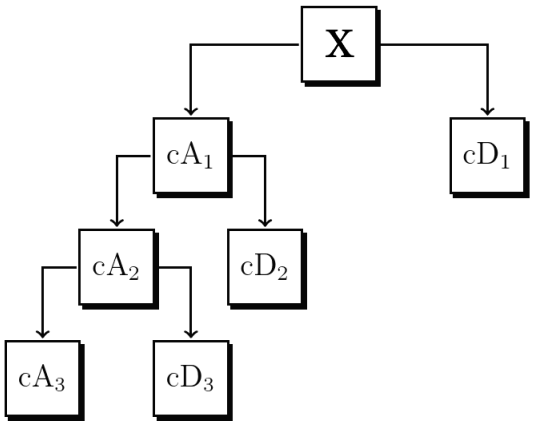

Figura 3.8 Árbol de descomposición wavelet de una señal X.

de Mallat, el cual es utilizado en casi todas las implementaciones en los paquetes de R y en la mayoría de las implementaciones en plataformas numéricas.

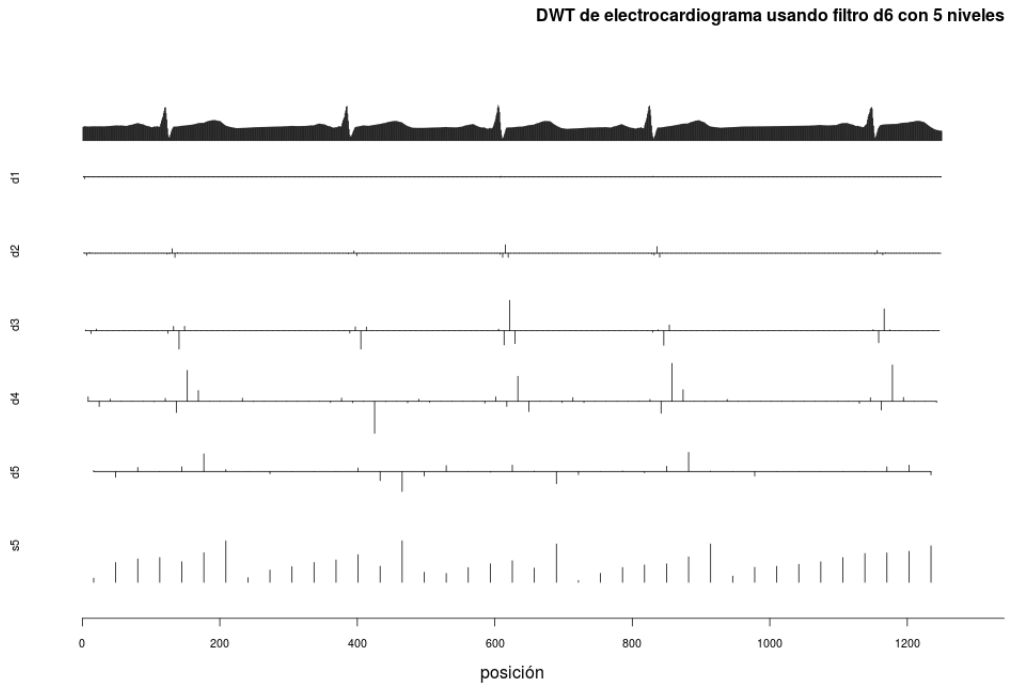

Figura 3.9 DWT de un ECG sano con Filtro d6 de 5 niveles.

3.3.3. Segmentación (Thresholding)

El objetivo del proceso de segmentación es minimizar el error al reconstruir una señal deseada, en general, una de las técnicas para seleccionar la función wavelet y el nivel, se basa en el contenido de energía de la parte de aproximación al realizar el procesamiento de descomposición de la transformación Wavelet [78–80]. Anteriores investigaciones en el área de del procesamiento de señales electrocardiográficas señalan que la familia de wavelet Daubechies debido a su similitud con el ECG realiza un buen desempeño para realizar un proceso de descomposición, segmentación y reconstrucción de la señal, véase Figura 3.10; lo anterior con el fin de realizar una eliminación de ruido y sea óptimo el algoritmo de máximos locales con el cual se hallan las ondas R en cada uno de los ciclos cardiacos.

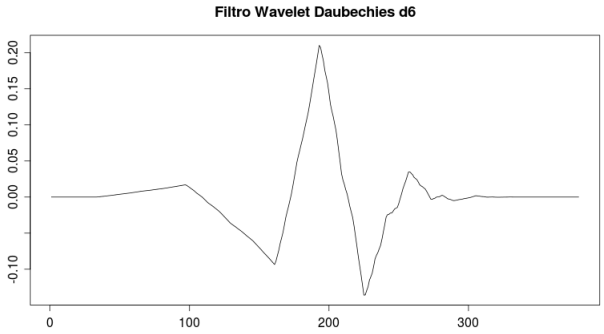

Figura 3.10 Filtro Wavelet Daubechies d6

El nombre de la familia Wavelet Daubechies en `R` está dado por la letra `d` seguido de un número par entre 2 y 20 que representa el orden del filtro. En la figura 3.10 puede verse el filtro wavelet de la familia Daubechies de orden 6, el cual es seleccionado para realizar el proceso de filtrado por medio de la función `wavShrink`; en la tabla 3.3 se observa una comparación de rapidez (tiempo de uso de la CPU, o tiempo de proceso) y score de las principales funciones para realizar la eliminación de ruido.

Atributo	WavSrink	msDenoiseWavelet	Thresholding
user	0.024	0.528	0.596
system	0.004	0.016	0.012
elapsed	0.047	0.591	0.609
score	33	32	25

Cuadro 3.3 Comparativo de las principales funciones para filtrado

Todas las señales digitales son obtenidas con un ruido inmerso en la naturalidad del proceso de adquisición, es decir, $ECG = V + r$ donde V representa la señal de interés (comportamiento del proceso eléctrico del corazón), representado en una señal deterministica desconocida y r el ruido estocástico inmerso en el proceso de adquisición [81]. La función `wavShrink` brinda la posibilidad de obtener la transformada wavelet, proceso de segmentación y reconstrucción. Bajo estas condiciones la función `wavShrink` realiza

un excelente proceso en la eliminación del ruido r esperando una óptima reconstrucción de la señal original, véase Figura 3.11. El algoritmo básico de esta función consiste en tres pasos:

1. Calcula la DWT la señal original con las condiciones deseadas.

2. Aplica el esquema de segmentación a los coeficientes wavelet.

3. Halla la inversa de la DWT, haciendo una reconstrucción de una señal estadísticamente deseada.

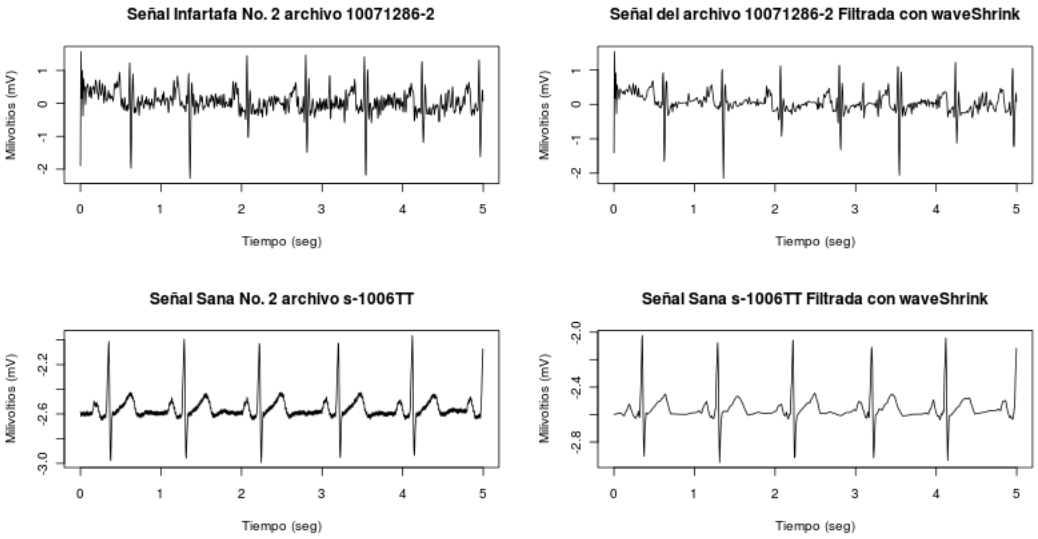

Figura 3.11 Filtros de eliminación de ruido realizados con la función `wavShrink` de WMTSA

4. Análisis de Resultados

4.1. Introducción

En este capítulo se presentan los resultados del proceso que se realizó para llevar a cabo la clasificación de las señales electrocardiográficas por medio de máquinas de aprendizaje utilizando medida de disimilaridad basadas en compresión de datos. La variabilidad de la frecuencia cardíaca es un índice fundamental para la clasificación de personas con enfermedades cardiovasculares, en esta investigación se muestra que se pueden obtener patrones utilizando la técnica SAX y CDM, los cuales analizados por máquinas de aprendizaje llegarán a realizar una buena clasificación entre señales de pacientes enfermos y sujetos sanos.

La minería de datos es la fusión entre la matemática, estadística y computación, el criterio y proceso desempeñado por las técnicas SAX y CDM están ligado con procesos de minería de datos, por tal razón se presentará la descripción de cada una de las fases del proceso KDD (Knowledge Discovery in Databases) en señales electrocadriográficas.

4.2. Proceso KDD en señales el ectrocardiográficas

En esta investigación se establecieron 6 principales fases del proceso KDD, véase Figura 4.1, este proceso es iterativo e interactivo en cada una de sus fases [82]. En esta figura se observa el porcentaje de esfuerzo que requieren las fases del presente proyecto

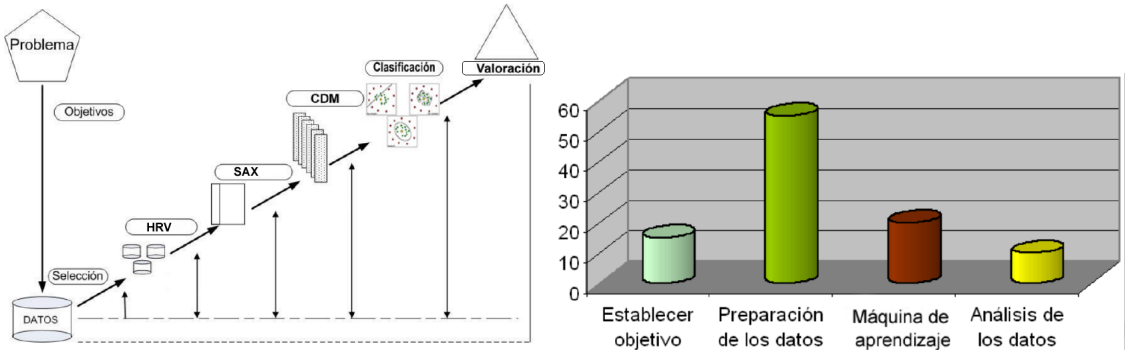

Figura 4.1 Etapas y esfuerzo del proceso KDD propuesto en esta investigación

divididas en 4 etapas, las etapas de objetivos, validación de la máquina de aprendizaje y análisis de los resultados toma menos de un 20 % de esfuerzo cada una, mientras, que la etapa de preparación de datos y preprocesamiento toma más del 50 % de esfuerzo; esta etapa corresponde a las fases de preprocesamiento (filtrado y caracterización de los ECG), transformación de los datos y hallazgo de la medida CDM. Hay una extensa gama de referencias al respecto del descubrimiento de conocimiento en base de datos, no obstante, aquí se describirá el proceso de extracción de conocimiento de un conjunto de señales electrocadriográficas, desde el momento de su adquisición. A continuación se relacionan las acciones y resultados obtenidos durante cada una de las fases del proceso KDD propuesto en esta investigación.

4.2.1. Entendimiento del Dominio del Tema

Cuando se habla de una secuencia de valores observados a lo largo del tiempo, y por tanto ordenados cronológicamente, se denomina, en un sentido amplio, serie temporal. Resulta difícil imaginar una rama de la ciencia en la que no aparezcan datos que puedan ser considerados como series temporales [83].

El electrocardiograma es la representación de la dirección y la magnitud de los impulsos eléctricos producidos en el corazón. El complejo electrocardiográfico normal está formado por una onda P un complejo QRS y una onda T. Una de las técnicas más

utilizadas para diferenciar entre pacientes sanos y enfermos en lecturas de electrocardiogramas es la Variabilidad de la Frecuencia Cardíaca, la VFC se relaciona con las fases más tempranas de la isquemia, con lo cual puede ayudar a una detección precoz de enfermedades coronarias [84]. En 1912 James Bryan Herrick presentó el concepto moderno de trombosis coronaria y reportó el primer caso de Infarto Agudo del Miocardio, en el cual el ECG fue de vital importancia [85].

Para poder realizar una clasificación entre electrocardiogramas sanos y enfermos existen índices estadísticos de la HRV en el dominio del tiempo y frecuencia entre otros. Los métodos en el dominio del tiempo se basan en el cálculo de la frecuencia cardiaca en un instante dado o en los intervalos entre complejos normales sucesivos, véase tabla 2.1. En un registro electrocardiográfico continuo cada complejo QRS es detectado y son determinados los llamados intervalos normales NN.

En estudios de investigación de señales electrofisiológicas la duración de la señal adquirida es tomada según la finalidad de cada investigación. Para el estudio de clasificación de señales electrocadriográficas se sugiere utilizar señales de 5 minutos, con el fin de poder utilizar los indices estadísticos en el dominio del tiempo con ventanas de 300 segundos; hoy en día, aún, están implementando técnicas matemáticas para la extracción de información y análisis de electrocardiogramas, también se han adjudicado muchas patentes de artefactos médicos, cada vez más innovadores [86].

El Grupo de Investigación en Ingeniería Biomédica (GIIB), cuenta con una base de datos de señales electrocardiográficas de cinco minutos cada una, estas señales cumplen el protocolo de Infarto Agudo Miocardio creado por la AAMI (The Association Advancement for Medical Instrument); se espera que haciendo uso de técnicas de minería de datos por computadora se realice una buena clasificación de señales electrocardio-

gráficas infartadas y señales sanas.

4.2.2. Selección y Adición

Para el desarrollo de este trabajo se cuenta con dos (2) bases de datos de señales electrocardiográficas, la base de datos de propiedad del GIIB, tomadas a una tasa de 250 muestras por segundo; cada una de las señales fue adquirida en sus 12 derivaciones por un tiempo mayor a 5 minutos y almacenadas en archivos de extensión '.txt '. La segunda base de datos pertenece y fue cedida por la Fundación Cardiovascular (FCV) de Colombia, sede Bucaramanga para la presente investigación; estas señales fueron tomadas por el sistema de adquisición WinDAQ y contiene 26 señales adquiridas a 250 muestras por segundo tomadas por un tiempo promedio de 30 minutos, estas señales solamente presentan la derivación V2 y están almacenadas en archivos de extensión '.wdq '.

Todas las señales electrocardiográficas se deben ordenar de tal forma que conformen tablas basadas en columnas, donde cada columna sea un ECG de la misma derivación. Por lo cual se extrae la derivación V2 de cada una de las señales de la base de datos infartadas y de la base de datos cedida por la FCV, debido a que hay señales de un tiempo mayor a 30 minutos se recomienda dividirlas y utilizar muestras 300 segundos, es decir, 5 minutos; todos los electrocardiogramas se almacenaron en archivos .txt, y se crearon archivos .RData para que sean cargados con mayor rapidez a la plataforma de calculo numérico R.

4.2.3. Preprocesamiento

Una base de datos grande es fundamental para la realización del proceso KDD, y la calidad de la extracción de características de las bases de datos es aún más importante durante el proceso del descubrimiento de conocimiento. A cada una de las señales elec-

trocardiográfica se le debe realizar un proceso de filtrado con el fin de eliminar el ruido presente en cada señal y hallar los picos de la onda R.

Como se comentó en el capítulo 3 el proceso de filtrado con la transformada Wavelet resulta particularmente útil para el análisis de señales con componentes armónicas de alta frecuencia durante períodos muy cortos, y armónicas de baja frecuencia durante largos períodos, como las señales electrocardiográficas [76,87]. La transformada wavelet provee una buena técnica para la eliminación de ruido en una señal, a través del método de fijación del umbral (thresholding), ya que preserva las características de la señal original, en particular la de los complejos QRS [88].

Existen varias familias de Wavelets de diferentes características de las cuales se debe seleccionar una para realizar un proceso de filtrado, el lenguage R presenta multiples paquetes para implementar un análisis wavelet, en la Figura 3.11 se muestran dos segmentos de 5 segundos de la señal de sujeto sana `s-1006TT` y de paciente infartado `10071286-2`, y a la derecha de cada una también se muestran estas señales sin ruido luego de un proceso de filtrado de altas frecuencias con la función `wavShrink` del paquete `wmtsa`.

La mejor ventaja ofrecida por el análisis de señales con Wavelets, es la de poder realizar análisis locales, es decir, analiza áreas localizadas en señales largas, con una función de ventana que se va desplazando a lo largo de la la señal estudiada. En la Figura 4.2 se muestra el resultado del uso del algoritmo `msExtrema` para el hallazgo de picos o máximos locales del paquete `msProcess`, utilizado para hallar los picos de las ondas R en cada uno de los ciclos cardiacos.

En el algoritmo de hallazgo de los picos R, se utilizaron algoritmos de máximos

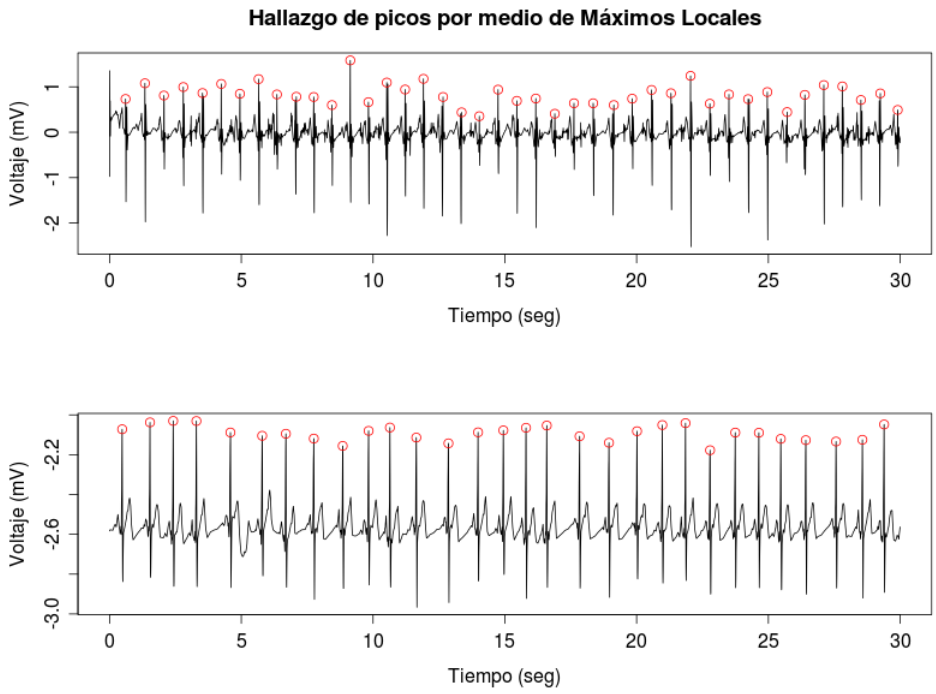

Figura 4.2 Hallazgo de picos por medio de máximos locales

locales con ventanas desplazadas con una tasa constante relacionada con la duración de la señal normal electrocardiográfica. Al tener los máximos de las ondas de las ondas R de cada electrocardiograma, se halla el tiempo entre cada uno de los intervalos RR, y así hallar la variabilidad de la frecuenica cardiaca instantánea en cada ciclo cardiaco; el tiempo entre cada uno de los latidos de un electrocardiograma se encuentra siguiendo la ecuación 4.1 y la VFC instantánea se computa hallando la inversa de la frecuencia y multiplicándolo por 60.

$$T_{RR} = R_n - R_{n-1} \text{ donde } n = 2, ..., N. \tag{4.1}$$

Un corazón normal se contrae aproximádamente 100.000 veces por día, a un ritmo variable de generalmente entre 60 y 100 latidos por minuto. Las ritmos lentos del corazón son típicamente de menos de 60 latidos por minuto. A un ritmo lento del corazón se le llama bradicardia. Si el ritmo es rápido (más de 100 latidos por minuto), se le llama taquicardia.

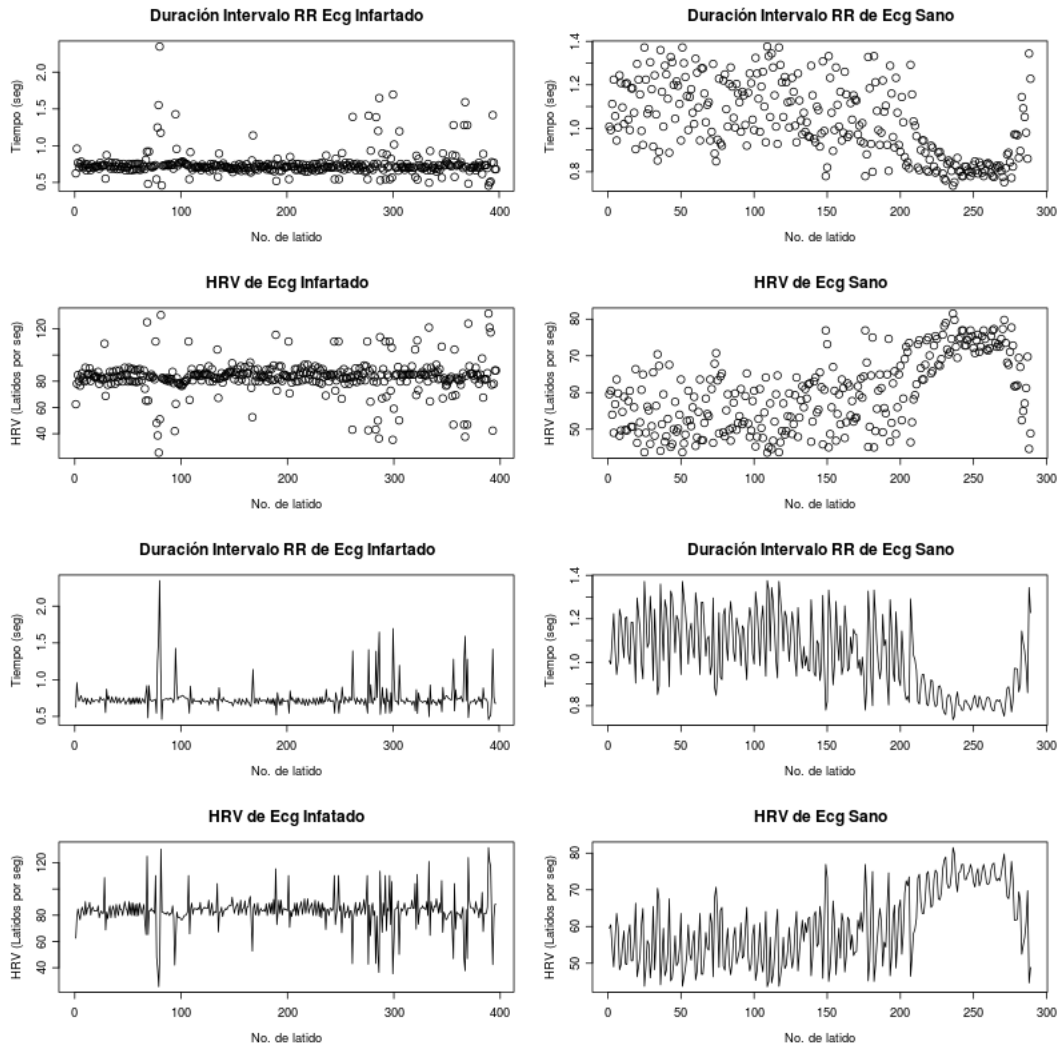

Figura 4.3 Comparación Intervalo RR y HRV entre ECG sano e infartado

En la Figura 4.3 se observa una comparación de la duración de tiempo entre latidos de una señal electrocardiográficas de paciente infartado y una señal electrocardiográfica de sujeto sano, se puede observar que el comportamiento de las gráficas de la derecha (la sana) es mucho más caótico que las gráficas del lado izquierdo (la infartada).

4.2.4. Transformación

Hasta el momento se ha reducido el espacio de búsqueda considerablemente, al extraer la variabilidad de la frecuencia cardiaca y los principales índices estadísticos en el dominio del tiempo, no obstante, para usar técnicas de minería de datos se debe realizar una transformación de los datos e indices estadísticos, es decir, realizar una reducción de espacio implementando técnicas de normalización y aproximación. "SAX es la primera representación simbólica para series temporales que permite realizar una reducción de dimensionalidad e indexado con una medida de distancia de baja banda. En las clásicas tareas de minería de datos tales como técnicas de agrupación y clasificación SAX, es una buena representación."[1]

Los algoritmos para el proceso de aproximación agregando simbolos (SAX) fue utilizado el código basado en el trabajo original de Lin, J., Keogh, E., Lonardi, S. & Chiu, B. [89]; y el código para implementar la medida de disimilaridad basada en compresión (CDM) fue desarrollado observando el algoritmo en el trabajo [57] del profesor Eamonn keogh.

Con la técnica SAX se convierte una serie temporal o cadena de datos numéricos a una cadena de símbolos, véase Figura 4.4 extraída de [90], se define un espacio simbólico y en él se define una métrica, se determinan similaridades o disimilaridades de las señales mediante una medida de distancia, esto también sirve para predecir eventos o

[1] Sitio Web Profesor Eamonn Keogh, visitado el 27 de abril de 2012

anomalías entre señales [89]. implementando la técnica CDM, véase ecuación 1.27; se cuantifica analógamente similaridad o disimilaridad entre cada uno de los registros por medio de una matriz de disimilaridad.

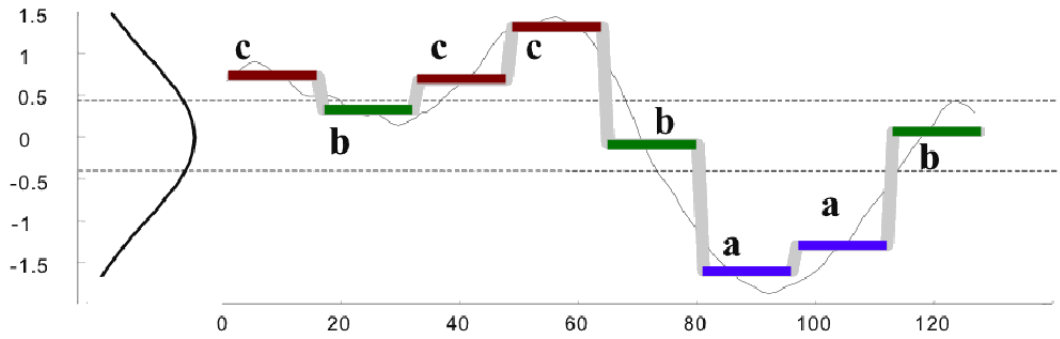

Figura 4.4 Técnica de aproximación basada en simbolos SAX

Con estos nuevos índices estadísticos, se llevan a una máquinas de aprendizaje y se hace una clasificación, como una herramienta de predicción entre sujetos sanos y pacientes con enfermedades coronarias (IAM, arritmias etc).

4.2.5. Clasificación

Desde la invención de la computadora el hombre ha intentado de crear máquinas inteligentes, que puedan realizar labores cotidianas de hombres regulares y de hombres expertos en un tema específico. En esta labor nació el área de inteligencia artificial en las ciencias computacionales, en esta área se han creado algoritmos de máquinas de aprendizaje supervisado y máquinas de aprendizaje no supervisado; la mayor diferencia entre estos dos grupos de algoritmos es que el primero necesita una base de conocimiento o un entrenamiento previo al momento de realizar un proceso de clasificación, la técnica de agrupación (Clustering en inglés) es una herramienta de clasificación de aprendizaje no supervisado muy utilizada en el área de minería de datos.

La técnica de agrupación realiza una clasificación asignando patrones a grupos de tal

forma que cada grupo sea más o menos homogéneo y distinto a los demás, ella permite agrupar los datos en árboles jerarquizados por medio de una medida de distancia de similaridad o disimilaridad, estas técnicas parten de una medida de proximidad entre individuos y a partir de ahí, busca los grupos de individuos más parecidos entre sí, según una matriz de similaridad o disimilaridad. Para graficar el resultado de estos algoritmos se utilizan dendogramas, véase Figura 4.5. Se pueden obtener patrones basados en la técnica CDM los cuales análizados por máquinas de aprendizaje pueden llegar a realizar una buena clasificación entre sujetos sanos y pacientes enfermos [14].

Se hicieron tres pruebas de validación, la primera consistió en utilizar un conjunto de 20 señales de cada una de las bases de datos, en la segunda prueba se utilizaron 50 señales de cada una de las bases de datos y en la tercera prueba se utilizaron las 90 de sujetos sanos y 90 señales de pacientes con infarto del miocardio. Para establecer los criterios de efectividad para el Modelo propuesto se recurre al cálculo de la sensibilidad, capacidad de encontrar todos los miembros de una población anormal; la especificidad, capacidad de encontrar todos los miembros de una población normal. Para cada una de las pruebas se realizaron operaciones de especificidad (ecuación 4.2) y sensibilidad (ecuación 4.3) insertando los valores en la tabla 4.1 con el propósito de determinar el número de señales electrocadriográficas correcta o incorrectamente clasificadas.

$$Especificidad = \frac{Verdaderos\_Negativos}{Verdaderos\_Negativos + Falsos\_Positivos} * 100 \qquad (4.2)$$

$$Sensibilidad = \frac{Verdaderos\_Positivos}{Verdaderos\_Positivos + Falsos\_Negativos} * 100 \qquad (4.3)$$

4.2.6. Valoración e Interpretación

Tal como se muestra en la Figura 4.1 durante todo el proceso la etapa de preprocesamiento es la de mayor porcentaje de esfuerzo y en donde el proyecto demanda de más tiempo para la extracción de información útil para el proceso de clasificación.

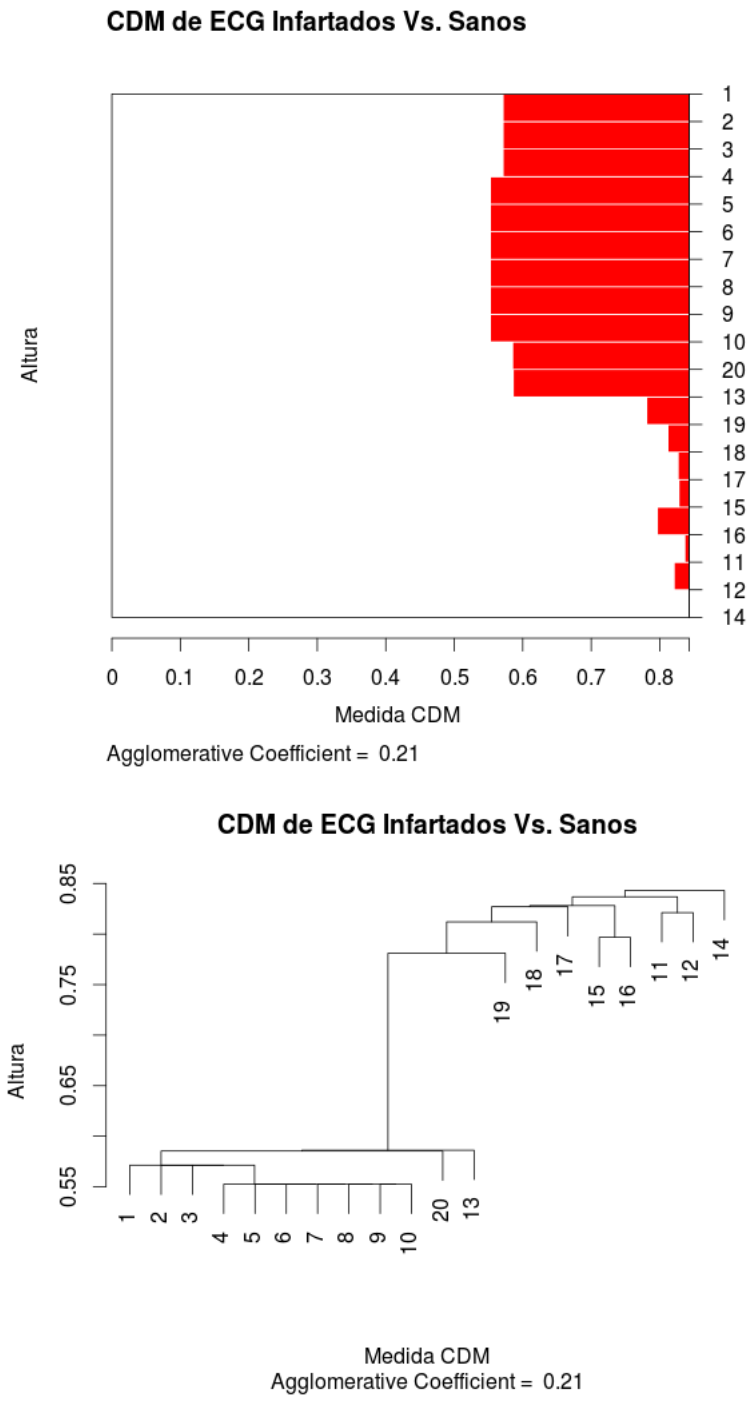

Figura 4.5 Valores de la medida CMD (Arriba). Dendograma de pacientes enfermos Vs. sujetos Sanos usando CDM (Abajo)

Cantid. ECG	VP	VN	FP	FN	Especificidad	Sensibilidad
10	5	4	1	0	80 %	100 %
20	10	8	2	0	80 %	100 %
40	20	18	2	0	80 %	100 %
50	25	23	2	0	92 %	100 %
100	50	48	2	0	96 %	100 %

Cuadro 4.1 Valores de especificidad y sensibilidad de las diferentes pruebas.

En la Figura 4.3 se muestra que la HRV de sujetos sanos es mucho más caótica que la de un paciente enfermo de Infarto Agudo de Miocardio.

Al extraer los índices estadísticos de la VFC en los dominios del tiempo se observa que hay diferencias entre las poblaciones de señales electrocadriográficas. En la Figura 4.5 se observa que se puede realizar una clasificación con árboles jerarquizados como técnica de máquina de aprendizaje no supervisado, esta clasificación se logra sin ninguna clase de parámetros, conjunto de entrenamiento o instrucción específica para el proceso de clasificación.

4.3. Interfaz Gráfica sobre R

Para efectos de la presente investigación se desarrollo una interfaz gráfica sobre R, ver figura 4.6, utilizando el paquete `svWidget` y `rpanel`; se creó una ventana con 6 menús los cuales son:

1. 5f Cargar DB: En este menú principal se encuentran tres opciones, `Cargar .txt Ecg's`, `Cargar .txt Ecg's` y `Salir`; con el primero importa las bases de datos señales electrocardiográficas desde los archivos '.txt'y los carga sobre la plataforma R en 2 listas llamadas EcgInfartos y EcgSanos con sus respectivos nombres. El segundo carga las bases de datos desde archivos .RData'. El tercer botón cierra la ventana principal.

Figura 4.6 Interfaz gráfica desarrollada sobre R para el proceso KDD

2. **Preprocesamiento:** En el menú de Preprocesamiento se realiza el proceso de filtrado, hallazgo de los máximos locales, duración de los Intervalos RR, Variabilidad de la Frecuencia Cardiaca e indices estadísticos en el dominio temporal.

3. **Transformación:** En esta parte el usuario transforma los datos para utilizar las técnicas de minería de datos y realizar una calsificación de los datos cargados, en este submenú el usuario Normaliza, aproxima y halla la medida CDM de la HRV de cada uno de los electrocardiogramas de la base de datos.

4. **Clasificación:** En esta parte el usuario implementa diferentes algoritmos de máquinas de aprendizaje no supervisado para realizar la clasificación de los datos transformados en una matriz de medida de disimilaridad basada en compresión de datos.

5. **Graficos:** En esta parte el usuario utiliza los submenu para visualizar los datos de la aplicación por medio de gráficas de lineas, puntos; de la cual el usuario puede extraer información valiosa del proceso de minería de datos en señales electrocadriográficas.

6. **Ayuda:** En este menú se presentan los créditos del presente trabajo de investigación, se imprimen en la consola los autores y el email del autor principal del proyecto.

5. Conclusiones y Trabajos Futuros

Los algoritmos empleados y el procesamiento planteado lograron realizar una buena clasificación de señales electrocardiográficas de Infarto Agudo de Miocardio con aquellas que se presentan sanas, el proceso de minería de datos y clasificación depende en gran manera de la etapa de extracción de características y preprocesamiento; en esta investigación esta etapa consta de un proceso de filtrado y extracción de la variabilidad de la frecuencia cardiaca.

La transformada Wavelet y su representación tiempo-escala es una de las relaciones más importantes en el filtrado y procesamiento de señales electrofisiológicas, ya que permite determinar el nivel de descomposición adecuado para obtener las bandas de frecuencia que se necesiten. Esta transformada permite filtrar ciertas bandas de frecuencia sin necesidad de alterar otras componentes frecuenciales de la señal. En el proceso de filtrado y caracerización se probaron varios paquetes de análisis Wavelet en R, siendo la función `wavShrink` del paquetes `wmtsa` la que brinda el algoritmo de descomposición, segmentación y reconstrucción más velozy con más opciones.

El lenguaje R ha sido un lenguaje muy utilizado en el área de la minería de datos y análisis estadístico en series temporales; y fue muy útil durante el desarrollo del presente trabajo de investigación, satisfaciendo todas las expectativas y necesidades, siendo un pequeño problema la selección de un paquete y función en particular habiendo variedad

de opciones.

De esta manera se cuenta con una herramienta computacional en la plataforma de Calculo numérico R que favorecerá la detección oportuna de personas con enfermedades coronarias tales como el IAM.

La función `lapply` utilizada en la mayoría de la ejecución de las funciones brinda facilidad en la eliminación de ciclos, ejecutando una única función a una lista de datos y regresa el resultado de cada una de esas listas, también en una lista de datos. Con este tipo de diseño de algoritmos es fácil realizar una implementación de tecnologías enfocadas hacia la computación de alto rendimiento, para el manejo de grandes volúmenes de información.

Bibliografía

[1] E. Paradis and J. A. Ahumada, "R para Principiantes," *Evolution*, vol. 42, pp. 1–61, 2003.

[2] X. Chen, S. Kwong, and M. Li, "A compression algorithm for DNA sequences based on approximate matching," in *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB)* (R. Shamir, S. Miyano, S. Istrail, P. Pevzner, and M. Waterman, eds.), (Tokyo, Japan), p. 107, Association for Computing Machinery, April 8–11 2000.

[3] R. Cilibrasi and P. M. B. Vitányi, "Clustering by compression," *IEEE Transactions on Information Theory*, vol. 51, pp. 1523–1545, 2005.

[4] M. Li, X. Chen, X. Li, B. Ma, and P. Vitányi, "The similarity metric," *IEEE Transaction on Information Theory*, vol. 50, no. 12, pp. 3250–3264, 2004.

[5] D. Benedetto, E. Caglioti, and V. Loreto, "Language Trees and Zipping," *Physical Review Letters*, vol. 88, no. 4, pp. 048702+, 2002.

[6] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang, "An information-based sequence distance and its application to whole mitochondrial genome phylogeny," *Bioinformatics*, vol. 17, no. 2, pp. 149–154, 2001.

[7] Z. Dawy, B. Göbel, P. Hanus, and J. Mueller, "Genomic analysis using methods from information theory," in *IEEE Information Theory Workshop*, pp. 55–59, 2004.

[8] A. Rokas, B. I. Williams, N. King, and S. B. Carroll, "Genome-scale approaches to resolving incongruence in molecular phylogenies," *Nature*, vol. 425, pp. 798–804, Octubre 2003.

[9] J. Mellville, J. Riley, and J. Hirst, "Similarity by compresion," *J. chem. Inf. Model*, vol. 47, pp. 25–33, 2007.

[10] P. Ferragina, R. Giancarlo, V. Greco, G. Manzini, and G. Valiente, "Compression-based classification of biological sequences and structures via the Universal Similarity Metric: experimental assessment," *BMC Bioinformatics*, vol. 8, no. 1, p. 252, 2007.

[11] D. Sculley and C. Brodley, "Compression and Machine Learning: A New Perspective on Feature Space Vectors," *Data Compression Conference*, vol. 0, pp. 332–332, 2006.

[12] P. Vitanyi, "Universal Similarity," 2005.

[13] A. Bratko and B. Filipič, "Spam filtering using compression models," tech. rep., 2005.

[14] P. Guillén, H. Spinetti, J. Sanz, M. Barrera, N. Angulo, and V. Malavé, "Máquinas de aprendizaje para clasificar señales electroencefalográficas," in *Cuarto Congreso Colombiano de Computación 4CCC*, UNAB, 2009.

[15] C. C. Santos, J. Bernardes, P. Vitanyi, and L. Antunes, "Clustering Fetal Heart Rate Tracings by Compression," *Computer-Based Medical Systems, IEEE Symposium on*, vol. 0, pp. 685–690, 2006.

[16] M. D. Esposti, C. Farinelli, A. Tolomelli, and M. Manca, "A similarity measure for biological signals: new applications to HRV analysis," *JP J Biostat*, vol. 1, no. 1, pp. 53–78, 2007.

[17] D. Galas, M. Nykter, G. Carter, N. Price, and I. Shmulevich, "Biological Information as Set-Based Complexity," *Information Theory, IEEE Transactions on*, vol. 56, pp. 667–677, Feb. 2010.

[18] I. H. Witten, Z. Bray, M. Mahoui, and B. Teahan, "Text Mining: A new frontier for lossless compression," in *In Data Compression Conference*, pp. 198–207, IEEE Press, 1999.

[19] E. Frank, C. Chui, and I. H. Witten, "Text categorization using compression models," in *In Proceedings of DCC-00, IEEE Data Compression Conference, Snowbird, US*, pp. 200–209, IEEE Computer Society Press, 2000.

[20] C. Bennet, M. Li, and B. Ma, "Chain letters and evolutionary histories," *Scientific American*, vol. 288, pp. 76–81, Junio 2003.

[21] T. Arbuckle, "Measure software - and its evolution - using information content," in *Proceedings of the joint international and annual ERCIM workshops on Principles of software evolution (IWPSE) and software evolution (Evol) workshops*, IWPSE-Evol '09, (New York, NY, USA), pp. 129–134, ACM, 2009.

[22] X. Chen, M. Li, B. Mckinnon, and A. Seker, "A Theory of Uncheatable Program Plagiarism Detection and Its Practical Implementation," Mayo 2002.

[23] X. Chen, B. Francia, M. Li, B. Mckinnon, and A. Seker, "Shared Information and Program Plagiarism Detection," *IEEE TRANS. INFORM. TH*, vol. 50, no. 7, pp. 1545–1551, 2004.

[24] T. Arbuckle, A. Balaban, D. K. Peters, and M. Lawford, "Software documents: Comparison and measurement," in *SEKE07: Proceedings of the 18th Int. Conf. on Software Engineering and Knowledge Engineering*, pp. 740–745, July 2007.

[25] T. Pham, "GeoEntropy: A measure of complexity and similarity," *Pattern Recognition*, vol. 43, pp. 887–896, 2010.

[26] D. Li and S. Simske, "Training Set Compression by Incremental Clustering," *Journal of Pattern Recognition Research*, vol. 6, pp. 56–64, febrero 2011.

[27] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 3 ed., 1997.

[28] V. Nannen, "A Short Introduction to Kolmogorov Complexity," *CoRR*, vol. abs/1005.2400, 2010.

[29] P. D. Grünwald and P. M. B. Vitányi, "Kolmogorov Complexity and Information Theory.With an Interpretation in Terms of Questions and Answers," *J. of Logic, Lang. and Inf.*, vol. 12, pp. 497–529, September 2003.

[30] C. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, 1948.

[31] A. Kolmogorov and S. Fomin, *Elementos de la teoría de funciones y del análisis funcional*. MIR, 1975.

[32] V. V'Yugin, "Algorithmic Complexity and Stochastic Properties of Finite Binary Sequences," 1999.

[33] T. L. Griffiths and J. B. Tenenbaum, "Probability, Algorithmic Complexity, and Subjective Randomness," in *In Proceedings of the 25th Annual Conference of the Cognitive Science Society*, Erlbaum, 2003.

[34] V. G. Vovk and G. R. Shafer, "Kolmogorov's Contributions to the Foundations of Probability," *Problems of Information Transmission*, vol. 39, pp. 21–31, 2003.

[35] E. Pekalska and R. Duin, "Dissimilarity representations allow for building good classifiers," *Pattern Recogn. Lett.*, vol. 23, no. 8, pp. 943–956, 2002.

[36] D. Hankerson, G. A. Harris, and P. D. Johnson, *Introduction to Information Theory and Data Compression*. Champan and Hall, 2 ed., 2003.

[37] R. Solomonoff, "A formal theory of inductive inference. Part I,," *Information and Control*, vol. 7, no. 1, pp. 1–22, 1964.

[38] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Problems of Information Transmission*, vol. 1, no. 1, pp. 1–7, 1965.

[39] G. J. Chaitin, "On the Length of Programs for Computing Finite Binary Sequences," *Journal of the ACM*, vol. 13, pp. 547–569, 1966.

[40] T. Jiang, M. Li, and P. Vitányi, "The Incompressibility Method," in *SOFSEM 2000: Theory and Practice of Informatics* (V. Hlavác, K. Jeffery, and J. Wiedermann, eds.), vol. 1963 of *Lecture Notes in Computer Science*, pp. 206–257, Springer Berlin / Heidelberg, 2000. 10.1007/3-540-44411-4_3.

[41] M. R. Garey and D. S. Johnson, *Computers and Intractability; A Guide to the Theory of NP-Completeness*. New York, NY, USA: W. H. Freeman & Co., 1990.

[42] X. Zhang, Y. Hao, X.-Y. Zhu, and M. Li, "New information distance measure and its application in question answering system," *J. Comput. Sci. Technol.*, vol. 23, pp. 557–572, July 2008.

[43] A. A. Muchnik, "Conditional complexity and codes," *Theor. Comput. Sci.*, vol. 271, pp. 97–109, January 2002.

[44] R. M. Gray, *Entropy and information theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1990.

[45] L. Paninski, "Estimation of entropy and mutual information," *Neural Comput.*, vol. 15, pp. 1191–1253, June 2003.

[46] A. Kaltchenko, "Algorithms for Estimating Information Distance with Application to Bioinformatics and Linguistics," Apr. 2004.

[47] J. Ziv and N. Merhav, "A measure of relative entropy between individual sequences with application to universal classification," *IEEE Trans. Inform. Theory.*, vol. 34, pp. 1270 – 1279, Julio 1993.

[48] C. H. Bennett, P. Gacs, P. Gács, S. Member, M. Li, P. M. B. Vitanyi, and W. H. Zurek, "Information Distance," *IEEE Transactions on Information Theory*, vol. 44, pp. 1407–1423, 1998.

[49] M. Li, "Information Distance and Its Applications," in *Implementation and Application of Automata* (O. H. Ibarra and H.-C. Yen, eds.), vol. 4094 of *Lecture Notes in Computer Science*, pp. 1–9, Springer Berlin / Heidelberg, 2006. 10.1007/11812128_1.

[50] S. A. Terwijn, L. Torenvliet, and P. M. Vitányi, "Nonapproximability of the normalized information distance," *Journal of Computer and System Sciences*, vol. In Press, Corrected Proof, pp.–, 2010.

[51] R. Cilibrasi and P. M. B. Vitanyi, "The Google Similarity Distance," *IEEE Trans. Knowledge and Data Engineering*, vol. 19, p. 370, 2007.

[52] L. Chen, "GenCompress," 1991. `http://www.cs.cityu.edu.hk/ cssamk/gencomp/GenCompress1.htm` Visitado el 20 Febrero de 2011.

[53] R. Cilibrasi, P. Vitanyi, and R. de Wolf, "Algorithmic clustering of music," in *Web Delivering of Music, 2004. WEDELMUSIC 2004. Proceedings of the Fourth International Conference on*, pp. 110–117, Sept. 2004.

[54] B. Hescott and D. Koulomzin, "On clustering images using compression," tech. rep., Boston university, CS Deparment, 2007.

[55] R. Cilibrasi, "The CompLearn ToolKit," 2003. `http://sourceforge.net/projects/complearn/` Visitado 1 Marzo de 2011.

[56] C. Ratanamahatana, E. Keogh, A. J. Bagnall, and S. Lonardi, "A Novel Bit Level Time Series Representation with Implications for Similarity Search and Clustering," in *In Proc. 9th Pacific-Asian Int. Conf. on Knowledge Discovery and Data Mining (PAKDD'05*, pp. 771–777, Springer, 2005.

[57] E. Keogh, S. Lonardi, and C. Ratanamahatana, "Towards parameter-free data mining," in *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, (New York, NY, USA), pp. 206–215, ACM, 2004.

[58] N. Nikvand and Z. Wang, "Generic image similarity based on Kolmogorov complexity," in *ICIP*, pp. 309–312, 2010.

[59] K. Kaabneh, A. Abdullah, and Z. Al-Halalemah, "Video Classification Using Normalized Information Distance," in *Proceedings of the conference on Geometric Modeling and Imaging: New Trends*, (Washington, DC, USA), pp. 34–40, IEEE Computer Society, 2006.

[60] E. Keogh, "SAX," 2002. `http://www.cs.ucr.edu/ eamonn/SAX.htm` Visitado Octubre de 2010.

[61] A. Castellanos, L. Rueda, A. Gualdron, and R. Menco, "Software Educativo para el Aprendizaje de la Electrocardiografía EKG-TUTOR," Octubre 2002. Escuela Ingeniería de Sistemas es Informatica - UIS.

[62] S. Mitra, *Procesamiento de Señales Digitales: Un enfoque basado en computadora.* Mc. Graw Hill, 2007.

[63] *Infarto agudo de miocardio: Clínica y tratamiento*, vol. 28, Ambito Farmaceutico, Educación sanitaria, Marzo 2009.

[64] O. M. de la Salud, *The World Health Report 2009 World Health-Changing History.* 1211 Geneva 27, Switzerland: WHO, 2009.

[65] R. Kleiger, J. Miller, J. Bigger, and A. Moss, "Decreased heart rate variability and its association with increased mortality after acute myocardial infarction," *Am J Cardiol*, vol. 59, no. 4, pp. 256–262, 1987.

[66] G. Baselli, S. Cerutti, S. Civardi, F. Lombardi, A. Malliani, M. Merri, M. Pagani, and G. Rizzo, "Heart rate variability signal processing: A quantitative approach as an aid to diagnosis in cardiovascular pathologies," *International Journal of Bio-Medical Computing*, vol. 20, no. 1-2, pp. 51–70, 1987.

[67] W. R., "Sampling Theory and Spline Interpolation," Septiembre 2003.

[68] W. Goodman, *Introduction to Fourier Optics*. McGraw-Hill, 2 ed., 1996.

[69] K. Tang, *Mathematical Methods for Engineers and Scientists V. 3*. Springer, 2007.

[70] P. M. Y. A., "The Influence of Fast Fourier Transform on the signal-spectrum Estimation," *Radiophysics and Quantum Electronics*, vol. 45, no. 3, pp. 239–245, 2002.

[71] A. Lara, "Sobre la Transformación Tiempo-Frecuencia y la aplicación del proceso de Convolución a la dinámica de sistemas físicos," *Acústica*, vol. 38, no. 1-2, pp. 7–13, 2006.

[72] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[73] J. Buckheit, J. B. Buckheit, D. L. Donoho, and D. L. Donoho, "WaveLab and Reproducible Research," pp. 55–81, Springer-Verlag, 1995.

[74] G. P. Nason, *Wavelet Methods in Statistics with R*. New York: Springer, 2008. ISBN 978-0-387-75960-9.

[75] L. Gong, W. Constantine, Y. A. Chen, and M. L. Gong, "Type Package Title Protein Mass Spectra Processing Version 1.0.6Date 2011-02-07," 2011.

[76] S. Mallat, *A Wavelet Tour of Signal Processing, 3rd ed., Third Edition: The Sparse Way*. Academic Press, 3 ed., December 2008.

[77] D. B. Percival and A. T. Walden, *Wavelet Methods for Time Series Analysis (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press, Feb. 2000.

[78] G. D. Clifford, F. Azuaje, and P. McSharry, *Advanced Methods And Tools for ECG Data Analysis*. Norwood, MA, USA: Artech House, Inc., 2006.

[79] F. Sachse, *Computacional Cardiology: Modeling of Anatomy, Electrophysiology, and Mechanics*. Springer, 2004.

[80] W. Ardila and L. Aristizábal, "Caracterización Mediante Wavelet de electrocardiogramas para Efectos y Clasificación de Cardiopatías," *Scientia et Technica*, vol. 32, pp. 155–158, Diciembre 2006.

[81] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[82] J. C. Riquelme, R. Ruiz, and K. Gilbert, "Mineria de Datos: Conceptos y Tendencias," *Inteligencia Artificial, Revista Iberoamericana de IA*, vol. 10, no. 29, pp. 11–18, 2006.

[83] A. Aguirre, *Introducción al tratamiento de series temporales. Aplicación a las ciencias de la salud*. Díaz de santos S.A., 1994.

[84] J.Vila, *Análisis de la variabilidad de señales fisiológicas de señales fisiológicas. Integración en un sistema de monitorización inteligente*. PhD thesis, Universidad de Santiago de Compostela, Departamento de Electrónica y Computación. Santiago de Compostela, 1997.

[85] J. Herrick, "Clinical features of sudden obstruction of the coronary arteries," *JAMA*, vol. 59, pp. 2015–2019, 1912. Reproduced in JAMA 1983;250:1757-65. PMID 6350634.

[86] H. Vullings, H. Verbruggen, and M. Verhaegen, *Advances in Intelligent Data Analysis Reasoning about Data*, vol. 1280, ch. ECG segmentation using time-warping, pp. 275–285. Springer Berlin / Heidelberg, 1997.

[87] N. M. y. D. Lütfiye, "Optimum Wavelet Transform-based ECG Compression and Dissimilarity Measure based Noise Performance Analysis," in *15th European Signal Processing Conference (EUSIPCO 2007)*, pp. 155–158, Poznan, Poland, Septiembre 2007.

[88] P. Addison, "Wavelet transforms and the ECG: a review," *Physiological Measurement*, vol. 26, pp. 155–199, Agosto 2005.

[89] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A Symbolic Representation of Time Series, with Implications for Streaming Algorithms," in *In Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp. 2–11, ACM Press, 2003.

[90] E. Keogh and J. Lin, "Hot sax: Efficiently finding the most unusual time series subsequence," in *ICDM 2005*, pp. 226–233, 2005.

[91] V. C. Barbosa, *An introduction to distributed algorithms*. Cambridge, MA, USA: MIT Press, 1996.

[92] G. Coulouris, T. Kindberg, and J. Dollimore, *Sistemas Distribuidos: Conceptos y Diseño.* Pearson Educación, 2001.

[93] P. D. C. H., "Grid computing: promesa de los sistemas distribuidos.," *Revista Sistemas ACIS*, vol. Edición 98, pp. 45–57, Octubre-Diciembre 2006.

[94] V. S. Sunderam, "PVM: A Framework for Parallel Distributed Computing," *Concurrency: Practice and Experience*, vol. 2, pp. 315–339, 1990.

[95] I. Foster, *Designing and Building Parallel Programs: Concepts and Tools for Parallel Software Engineering*. Boston, MA, USA: Addison-Wesley Longman Publishing Co. Inc, 1995.

[96] J. B. B. M. F. K., *Introduction to Grid Computing*. IBM RedBooks, 2006.

[97] J. B. et ál. et ál., *Introduction to Grid Computing with Globus*. IBM RedBooks, 2003.

[98] I. Foster, "The Anatomy of the Grid: Enabling Scalable Virtual Organizations," in *Euro-Par '01: Proceedings of the 7th International Euro-Par Conference Manchester on Parallel Processing*, (London, UK), pp. 1–4, Springer-Verlag, 2001.

[99] M. Schmidberger, M. Morgan, D. Eddelbuettel, H. Yu, and L. Tierney, "State-of-the-Art in Parallel Computing with R," *Journal of Statistical Software*, vol. 31, Agosto 2009.

[100] A. Goldman and C. Queiroz, "A model for parallel job scheduling on dynamical computer Grids," *Concurrency and Computation: Practice and Experience*, vol. 16, pp. 461–468, 2004.

[101] L. G. Valiant, "A bridging model for parallel computation," *Commun. ACM*, vol. 33, pp. 103–111, Aug. 1990.

[102] D. Skillicorn and D. Talia, "Models and languages for parallel computation," *ACM Comput. Surv*, vol. 30, no. 2, pp. 123–169, 1998.

A. Computación de Alto Rendimienro en Lenguage R

A.1. Introducción

Un computador paralelo es una máquina que tiene 2 o más procesadores y por ende puede ejecutar más de un proceso al mismo tiempo. Un sistema distribuido es aquel en que sus partes están geograficamente distribuidas compartiendo funciones de software y hardware para un fin en común. Existen ciertas características y condiciones que deben cumplir los sistemas distribuidos y paralelos [91]. Hace pocos años surgió la Computación de Alto Rendimienro como u nservicio ofrecido a través de Internet, esto ha devenido en que la accesibilidad ha aumentado y que su uso sea mayor y más amplio. El lenguaje R no ha sido ajeno a este fenómeno, obteniendo la contribución de varios paquetes para hacer posible la paralelización de algoritmos y brindar acceso a servicios de computación de alto rendimiento, e.g. clodnumbers.com [1]. En este capítulo se presentan conceptos básicos de los Sistemas Distribuidos y la Computación en Paralelo, se mencionan algunos paquetes y funciones de la Plataforma de Cálculo Numérico R, y al final se propone un modelo teórico del análisis de señales electrocardiográficas en paralelo usando R.

[1] HPC para R, cloudnumbers.com/r-for-hpc, visitado el 28 de abril de 2012

A.2. Computación de Alto Rendimiento - CAR

Desde el ábaco, la calculadora, los computadores personales y las supercomputadoras, el objetivo siempre ha sido hacer el mayor número de operaciones con mayor número de datos en menos tiempo; así, con el uso de sistemas de computación en paralelo tales como clusters, arquitecturas mutiprocesador y supercomputadoras se puuede obtener un menor tiempo en la ejecución de algoritmos.

Para comenzar a realizar una definición de CAR se realiza la definición de Sistema Distribuido y sus principales propiedades; los sistemas distribuidos son una colección de objetos de hardware y/o software separados geograficamente, los cuales están conectados por una red, trabajando por una misma causa y comunicándose entre ellos y otros sistemas a través del paso de mensajes u otras señales de sistemas operativos [92–94]. Entre las propiedades que se esperan de estos sistemas están:

- Concurrencia: Es la propiedad que permite que múltiples procesos sean ejecutados al mismo tiempo, y que potencialmente puedan interactuar entre sí. Los procesos concurrentes solamente existe cuando son ejecutado en diferentes procesadores.

- Escalabilidad: Es la propiedad que indica a un sistema la habilidad para crecer o extender el número de las partes o funciones sin perder la calidad de su funcionamiento.

- Modularidad: La modularidad es la capacidad que tiene un sistema de ser estudiado, visto o entendido como la unión de varias partes que interactúan entre sí y que trabajan para alcanzar un objetivo común.

- Eficiencia: La idea base de los sistemas distribuidos es hacer abarcar cada vez más información en el menor tiempo posible. El mayor esfuerzo de los investigadores en el área de la computación de alto rendimiento y sistemas distribuidos es crear

algoritmos y herramientas que permitan una mayor velocidad en un sistema distribuido.

La Computación de Alto Rendimiento (HPC por sus siglas en inglés) es una herramienta en el desarrollo de grandes problemas [95]. Para lograr este objetivo, la computación de alto rendimiento se apoya en tecnologías computacionales como los clusters, supercomputadoras o mediante el uso de la computación en paralelo.

Teóricamente, si se dobla el número de procesadores, el tiempo de ejecución en secciones paralelas deberia reducirse a la mitad. Esto teniendo en cuenta que todo programa consta de una o más porciones que no se pueden paralelizar y una o más porciones paralelizables; la ley de Amdahl es un modelo matemático que permite conocer la relación entre la aceleración A esperada de la implementación paralela P de un algoritmo y la implementación serial del mismo algoritmo, el modelo matemático A.1 se observa en la ecuación de la ley de Amdahl. La ley de Amdahl se mide en unidades genéricas, es decir, los resultados no son porcentajes, ni unidades de tiempo. En términos simples, el algoritmo es el que decide la mejora de velocidad, no el número de procesadores.

$$\frac{1}{(1-P)+\frac{P}{A}} \tag{A.1}$$

Otra caracteristica y clasificación fundamental en la computación en paralelo es la memoria, teniendo tres clasificaciones que son, Memoria compartida, Memoria distribuida y Memoria compartida distribuida; los sistemas distribuidos también mantienen otras clasificaciones según su arquitectura, véase Tabla A.1. Esta clasificación fue creada por Michael J. Flynn en 1972; en la computación en paralelo el modelo de arquitectura más utilizado es el Single Instruction Multiple Data, en este modelo una misma función es utilizada paralelamente con un conjunto de datos diferentes.

	Una Instrucción	Multiple Instrucción
Un dato	SISD	MISD
Multiples Datos	SIMD	MIMD

Cuadro A.1 Taxonomía de Flynn

A.2.1. Cluster (Máquinas Paralelas)

En ciencias computacionales la palabra Cluster se define como un conjunto de dispositivos de hardware (procesadores, memorias, almacenamiento) en lo posible heterogéneos conectados por medio de una red, los cuales trabajan de forma paralela y se comunican entre ellos a través del paso de mensajes comportándose como si fueran una única Computadora. Así con la implementación de cluster se tiene acceso a la construcción de supercomputadores de bajo costo y de excelente producción.

La computación paralela es una técnica de programación en la que muchas instrucciones se ejecutan simultáneamente. Se basa en el principio de que los problemas grandes se pueden dividir en partes más pequeñas que pueden resolverse de forma concurrente. Para aprovechar la paralelización de datos y funciones en un sistema distribuido Cluster se utilizan librerías desarrolladas en los diferentes lenguajes de programación, a continuación se realiza una descripción general de las librerías MPI y OpenMP:

MPI Message Passage Interface, interface de paso de mensajes en español; esta librería es un estándar para la implementación de aplicaciones y algoritmos en ambientes paralelos, MPI crea una comunicación entre máquinas conectadas en un ambiente paralelo y así gestiona los trabajos en cada una de ellas. Su principal característica es que no precisa de memoria compartida.

OpenMP Es una librería para la programación paralela en un ambiente de memoria compartida. OpenMP es un modelo portable y escalable, que ofrece a los programadores

una simple y flexible interfaz para el desarrollo de aplicaciones paralelas en arquitecturas con multiples procesadores, es decir, arquitecturas multicore.

A.2.2. Grid Computing

La necesidad de aprovechar los recursos disponibles conectados a Internet y simplificar su utilización dio lugar a una tecnología llamada Grid Computing o computación en Grid [96] [97], se han propuesto varias definiciones para esta palabra como mallas o grilla, sin embargo es mejor no traducirla.

Grid Computing es un nuevo modelo computacional, el cual permite compartir todos los recursos conectados a através de una red (cpu, memoria, almacanamiento, sensores remotos, clusters, supercomputadoras, aplicaciones etc), esto haciendo que la unión de todos los elementos puedan comportarse como una única máquina, por medio de una organización lógica llamada "Organización Virtual" [98], en la cual habrán privilegios y permisos para acceder a los recursos. Se suele determinar un sistema Grid como una infraestructura con múltiples capas y componentes que interactúan entre sí. Esta infraestructura debe proporcionar a los usuarios un servicio seguro a todos los niveles como: capacidad de cómputo, integridad de datos, seguridad de acceso.

A.3. Paquetes y Funciones en R para Computación de Alto Rendimiento

La práctica de la medicina ha evolucionado en el sentido de la variedad y la cantidad de información que se maneja. Cada vez podemos manejar un volumen más grande de datos con mayor precisión y en menor tiempo. El tratamiento de señales digitales es una de las áreas donde se presenta esta demanda de tiempo y cada vez un mayor volumen de datos.

El lenguaje R es un proyecto de código abierto, especializado para el cálculo intensivo de datos, análisis de datos estadístico y gráficos. Con Lenguaje R se puede fácilmente crear interfaces con los lenguajes C, C++, Java, Python, entre otros. Hace pocos años se han venido recibiendo y desarrollando contribuciones en el uso de herramientas para aplicaciones paralelas [95] sobre R, R ha incorporado paquetes para trabajar con diferentes librerías tales como MPI y PVM. Los paquetes más utilizados para trabajar con procesamiento en paralelo en R son `Rmpi`, `snow`, `papply`, `multicore` [99]. En la página del Proyecto R, se puede encontrar una larga lista de paquetes, agrupados por tópicos, que son de gran utilidad al hacer uso de la CAR sobre R [2].

Con el fin de realizar algoritmos más rápidos y masivamente paralelos se han creado modelos de algoritmos para computación paralela [100]; entre estos modelos se encuentra el BSP (Bulk Synchronous Parallel) propuesto en 1990 por Leslie Valiant [101,102], el BSP ha sido implementado en los principales lenguajes de programación para la computación distribuida-paralela, tales como C, Fortran, Python.

R cuenta con una gran variedad de tipos de objetos, entre estos, boleanos, vectores, matrices, cadenas, listas, dataframes etc. Al tener un conjunto de lista de datos del mismo tipo y querer computar ese conjunto de datos con una misma función se puede utilizar la función `apply` con el fin de eliminar ciclos y lograr optimizar los recursos de hardware necesitados para ejecutar tal acción. R brinda la posibilidad de reemplazar estas funciones por `papply` (Parallel Apply) y `mclapply` con el fin de brindar un paralelismo en un modelo de memoria distribuida y compartida respectivamente.

Teniendo en cuenta la ley de Amdahl es posible realizar un cálculo ideal de las op-

[2] HPC y Computación en Paralelo con R, Visitado el 30 Abril de 2012

eraciones al momento de implementar algoritmos paralelos, esta operación se realiza teniendo en cuenta los tiempos de ejecución o las secciones donde se implementarían operaciones paralelas, en la Tabla A.2 se observa el tiempo de ejecución de los algoritmos implementados variando el número de señales electrocardiográficas a analizar.

Algoritmo	1	5	20	50	100	200	300
Recuperación Baselinea (seg)	0.308	1.405	4.906	11.712	23.257	46.878	69.533
Reducción Ruido (seg)	0.194	0.214	0.820	1.452	2.740	5.520	8.590
Maximos Locales (seg)	4.381	19.729	77.678	201.492	406.679	807.683	1179.361
Normalización Z (seg)	0.004	0.006	0.011	0.025	0.049	0.081	0.108
SAX (seg)	0.007	0.027	0.099	0.264	0.362	0.716	1.103
TOTAL (segundos)	4.894	21.381	83,514	214,945	433,087	860,878	1258,695

Cuadro A.2 Tiempo de ejecución de los algoritmos

En la Figura A.1 se observa un modelo distribuido de tratamiento de señales electrocardiográficas, utilizando la bases de datos distribuida accesable desde internet junto con una serie de servicios de CAR (Supercomputadores y Clusters) para el cálculo de las funciones paralizables, a través del modelo de única función a multiples datos.

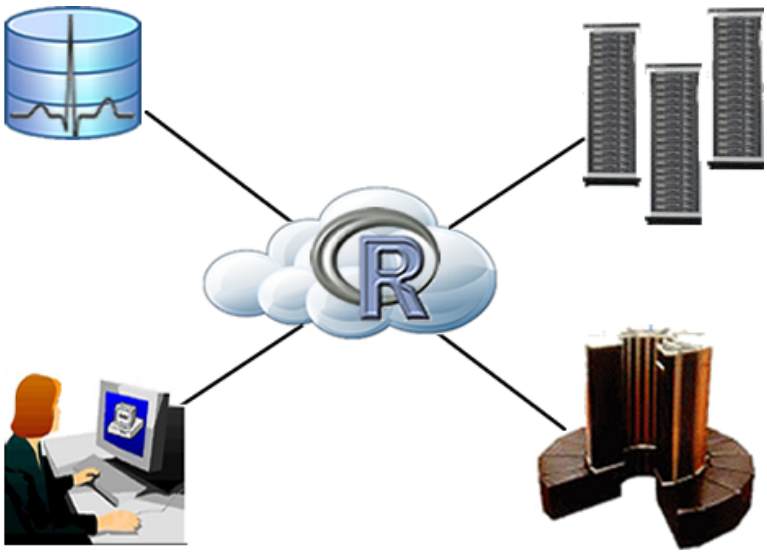

Figura A.1 Modelo Distribuido-Paralelo del análisis de ECG con **R**

La función con prioridad a paralelizar es la de máximos locales por ser la de mayor exigencia de cálculo y por ende la que mayor tiempo lleva en su ejecución tal como se muestra en la Tabla A.2.