

**“Estudio comparativo
de métodos de agrupamiento no supervisado
de latidos de señales ECG”**

Diego Hernán Peluffo Ordóñez

Tesis de Maestría



Universidad Nacional de Colombia
Facultad de Ingeniería y Arquitectura
Departamento de Ingeniería Eléctrica,
Electrónica y Computación
Manizales, 2009

Estudio comparativo de métodos de agrupamiento no supervisado de latidos de señales ECG

Diego Hernán Peluffo Ordóñez

Trabajo de grado para optar al título de
Magíster en Ingeniería - Automatización Industrial

Director
Prof. Germán Castellanos Domínguez

Universidad Nacional de Colombia
Facultad de Ingeniería y Arquitectura
Departamento de Ingeniería Eléctrica, Electrónica y
Computación
Manizales
2009

Comparative study of clustering methods for heartbeats of ECG signals

Diego Hernán Peluffo Ordóñez

Thesis for the degree of
Master in Engineering - Industrial Automation

Supervisor
Prof. Germán Castellanos Domínguez

National University of Colombia - Manizales
Faculty of Engineering and Architecture
Department of Electrical, Electronic and Computing
Engineering
Manizales
2009

A mis padres,
porque detrás de un gran logro
existe una fuente de inspiración de igual proporción

Agradecimientos

Agradezco a Dios por concederme la fuerza y todos los elementos necesarios para resolver mis problemas y aprender de ellos. A mis padres, mi hermano y toda mi familia por su apoyo y comprensión, por su firme convicción de creer en mí y por ser mi interminable fuente de inspiración.

Agradezco al profesor Germán Castellanos por toda su asesoría y apoyo en el desarrollo de este trabajo.

Agradezco también a mi asesor y amigo José Luis Rodríguez (mi compañero de lucha) por su compañía y su invaluable apoyo, de quien aprendí bastante en términos de mi formación académica y personal.

Doy mis más sinceros agradecimientos a Yenny Berrío quien me acompañó en momentos críticos y se convirtió en mi luz y aliento. A ella, de corazón, gracias por su amistad sincera y apoyo incondicional.

Agradezco sentidamente a Mauricio Orozco, quien fue jurado de este trabajo, por sus acertadas observaciones y correcciones.

De igual forma, quiero agradecer a todos mis amigos, especialmente a Oscar Cardona, Victoria Montes, Jorge Gómez y Cristian Castro por su incondicional apoyo y compañía.

Agradezco a los integrantes del Grupo de Control y Procesamiento Digital de Señales y del grupo de estudio de señales biomédicas por el apoyo y las discusiones académicas que contribuyeron a orientar y fortalecer el marco teórico de este estudio.

Y a todas las personas que, de alguna manera, contribuyeron al desarrollo de este trabajo, gracias.

Información del trabajo de tesis

Título:

Estudio comparativo de métodos de agrupamiento no supervisado de latidos de señales ECG.

Desarrollado por:

Diego Hernán Peluffo Ordóñez

Director:

Profesor. César Germán Castellanos Domínguez. Universidad Nacional de Colombia, sede Manizales

Jurados:

José Luis Rodríguez Sotelo. Universidad Nacional de Colombia, sede Manizales

Mauricio Orozco Alzate. Universidad Nacional de Colombia, sede Manizales

Este proyecto fue realizado dentro del marco del proyecto “Análisis de bioseñales en la identificación de arritmias mediante técnicas no supervisadas”, correspondiente a la Convocatoria Nacional de Investigación 2009, modalidad V (apoyo a tesis de posgrado).

Índice general

Información del trabajo de tesis	xi
Índice general	xv
Lista de figuras	xvii
Lista de tablas	xix
Lista de algoritmos	xxi
Resumen	xxiii
Abstract	xxv
Notación	xxvii
I Preliminares	1
1 Introducción	3
1.1 Planteamiento del problema	3
1.2 Organización del documento	4
2 Objetivos	7
2.1 Objetivo general	7
2.2 Objetivos específicos	7
3 Contexto	9
3.1 Contexto Fisiológico	9
3.2 Base de datos de arritmias	12
3.3 Procesamiento digital de señales ECG	12
3.3.1 Estándares del análisis ECG	15
4 Estado del arte de la clasificación de señales ECG	17
4.1 Extracción y selección de características	17
4.2 Clasificación	22
4.2.1 Clasificación supervisada	22
4.2.2 Clasificación no supervisada	24

II	Marco teórico	31
5	Agrupamiento particional	33
5.1	Introducción	33
5.2	Agrupamiento basado en centroides	34
5.2.1	K-medias	35
5.2.2	H-medias	37
5.2.3	Idea general del agrupamiento iterativo	37
5.2.4	H-medias basado en GIM	39
5.2.5	K-medias armónicas	40
5.2.6	H-medias difuso	41
5.2.7	Agrupamiento de máxima esperanza gaussiana	42
5.2.8	Algoritmos híbridos	43
5.3	Criterios de inicialización	44
5.3.1	Criterio máx-mín	44
5.3.2	Criterio J-medias	45
5.4	Análisis por segmentos	46
6	Agrupamiento espectral	49
6.1	Introducción	49
6.2	Preliminares sobre teoría de grafos	50
6.2.1	Medidas sobre los grafos	52
6.3	Agrupamiento espectral normalizado	53
6.3.1	Criterio de particionado multi-clase	54
6.3.2	Representación matricial	54
6.3.3	Algoritmo de agrupamiento	55
6.4	Medidas de afinidad	63
6.5	Estimación del número de grupos	64
6.5.1	Estimación basada en SVD	64
6.5.2	Estimación empleando vectores propios	65
III	Marco experimental	67
7	Descripción de los experimentos	69
7.1	Señales ECG	69
7.2	Conjunto de características	70
7.3	Metodología 1	71
7.4	Metodología 2	71
8	Medidas de desempeño del agrupamiento	73
IV	Discusión de resultados	77
9	Resultados y discusión	79
9.1	Resultados obtenidos con la metodología 1	79

9.1.1	Prueba 1	79
9.1.2	Prueba 2	83
9.1.3	Prueba 3	85
9.2	Resultados obtenidos con la metodología 2	86
V Comentarios finales		91
10 Conclusiones y trabajo futuro		93
10.1	Conclusiones	93
10.2	Trabajo futuro	94
VI Apéndices		95
A Selección de características		97
A.1	Convención	98
A.2	Proyección ortonormal	98
A.3	Análisis de componentes principales	98
A.4	PCA ponderado	101
A.5	Algoritmo $Q-\alpha$	103
A.6	$Q - \alpha$ supervisado	106
A.7	Otra medida de relevancia	107
B Caracterización del complejo QRS usando el modelo de Hermite		111
B.1	Modelo paramétrico de Hermite	111
B.2	Método de reconstrucción y caracterización	112
B.2.1	Extracción del complejo QRS	112
B.2.2	Reconstrucción de la señal	113
B.2.3	Comparación de los espectros	115
B.2.4	Búsqueda de σ óptimo	115
B.3	Características resultantes	116
C Alineamiento temporal no lineal DTW		119
C.1	Algoritmo DTW	120
D Discusión académica		125
Bibliografía		127

Lista de figuras

3.1	Derivaciones para el análisis ECG	10
3.2	Electrocardiograma normal	10
3.3	Diferentes morfologías de latido	11
6.1	Ejemplos de grafos	51
6.2	Grafo ponderado	52
6.3	Diagrama explicativo del algoritmo de agrupamiento espectral	56
7.1	Diagrama de bloques de la metodología 1 (tradicional)	71
7.2	Diagrama explicativo de la metodología 2 (por segmentos)	72
9.1	Características de los registros 207(R, L, V) y 215 (N)	82
9.2	Robustez hacia la inicialización en DBC	84
9.3	Desempeño de GEMC en la metodología 1	86
9.4	Desempeño de NPDBC en la metodología 1	86
9.5	Desempeño de NPDBC con $Q - \alpha$ en la metodología 2	87
9.6	Desempeño de NPDBC con WPCA (vectores propios ponderados) en la metodología 2	87
A.1	Valores de relevancia de los diferentes métodos	109
B.1	Ejemplos de Bases de Hermite obtenidas con $\sigma = 15$ ms (a) $n = 2$ (b) $n = 4$	113
B.2	Primera base de Hermite ($n = 0$) obtenida con diferentes valores de σ	114
B.3	Reconstrucción con las primeras 6 bases de Hermite con $\sigma = 25$ ms (Latido 20 del registro 200)	114
B.4	Espectro de la reconstrucción empleando las 3 primeras bases (Latido 20 del registro 200)	116
B.5	Espectro de la reconstrucción empleando las 6 primeras bases (Latido 20 del registro 200)	117
C.1	Matriz de programación dinámica \mathbf{G} asociada a los vectores x e y	120
C.2	Diferentes producciones para las transiciones	122
C.3	Restricciones globales de DTW	123

Lista de tablas

3.1	Características de algunas arritmias. Información tomada de [1]. * Latidos por minuto.	11
3.2	Grupos de arritmias recomendados por la AAMI	15
4.1	Técnicas de caracterización de señales ECG	21
6.1	Algunas medidas de afinidad	64
7.1	Conjunto de características seleccionadas para el análisis de arritmias	70
8.1	Medidas de desempeño empleadas	75
9.1	Registros empleados en la prueba 1	80
9.2	Resultados obtenidos empleando el algoritmo H-medias en la prueba 1	81
9.3	Resultados obtenidos empleando el algoritmo KHM en la prueba 1	81
9.4	Resultados aplicando GEMC y diferentes criterios de inicialización	83
9.5	Resultados de los métodos DBC empleando la metodología 1	85
9.6	Resultados obtenidos con la metodología 2, empleando NPDBC y $Q - \alpha$ no supervisado	88
9.7	Resultados obtenidos con la metodología 2, empleando NPDBC y WPCA con ponderación de vectores propios	89
A.1	Algunos métodos de ponderación para WPCA	103

Lista de algoritmos

1	K-medias	36
2	H-medias	37
3	Modelo iterativo genérico para CBC	39
4	Criterio J-medias	45
5	Agrupamiento por segmentos	48
6	Agrupamiento espectral normalizado	62
7	Método $Q - \alpha$ estándar	105

Resumen

En este trabajo se presenta un estudio comparativo de diferentes técnicas de agrupamiento no supervisado con enfoque particional y espectral, orientado a la agrupación de patrones representativos de latidos extraídos de registros electrocardiográficos ambulatorios. Debido a la naturaleza de las señales estudiadas y a que, en muchos casos, no es factible el etiquetado de las mismas, se prefieren las técnicas de agrupamiento no supervisado para su análisis. El uso de un modelo genérico de agrupamiento particional y la estimación de parámetros de inicialización adecuados empleando técnicas espectrales, son algunos de los aportes más significativos de esta investigación. Los experimentos se realizan sobre una base de datos de arritmias estándar del MIT (*Massachusetts Institute of Technology*) y la extracción de características se hace con técnicas recomendadas por la literatura.

Otro aporte importante, es el desarrollo de un método de análisis por segmentos que reduce el costo computacional y mejora el desempeño del agrupamiento en comparación con el análisis tradicional, es decir, analizando todo el conjunto de datos en una sola iteración del procedimiento. Adicionalmente, se sugiere un esquema completo de análisis no supervisado de señales ECG, incluyendo etapas de caracterización, selección de características, estimación del número de grupos, inicialización y agrupamiento. También se diseñan medidas adecuadas de desempeño, basadas en la conformación de los grupos, que relacionan el agrupamiento con el número de grupos empleados y costo computacional. Este estudio se realiza teniendo en cuenta las recomendaciones de la AAMI (*Association for the Advanced of Medical Instrumentation*).

Abstract

This work presents a comparative study of different partitional and spectral clustering techniques to cluster heartbeats patterns of long-term ECG signals. Due to the nature of signals and since, in many cases, it is not feasible labeling thereof, clustering is preferred for analysis. The use of a generic model of partitional clustering and the appropriate estimation of initialization parameters via spectral techniques represent some of the most important contributions of this research. The experiments are done with a standard arrhythmia database of MIT (*Massachusetts Institute of Technology*) and the feature extraction is carried out using techniques recommended by literature.

Another important contribution is the design of a sequential analysis method which reduces the computational cost and improves clustering performance compared to traditional analysis that is, analyzing the whole data set in one iteration. Additionally, it suggests a complete system for unsupervised analysis of ECG signals, including feature extraction, feature selection, initialization and clustering stages. Also, some appropriate performance measures based on groups analysis were designed, which relate the clustering performance with the number of resultants groups and computational cost. This study is done taking into account the AAMI standard (*Association for the Advance of Medical Instrumentation*).

Notación

VARIABLES Y FUNCIONES

\mathbf{X}	Matriz de datos de $n \times d$
\mathbf{x}_i	i -ésima observación de \mathbf{X}
\mathbf{Q}	Conjunto de centroides
\mathbf{q}_j	j -ésimo centroide
\mathbf{C}	Conjunto de grupos
\mathbf{P}_k	Conjunto de particiones de k elementos
N_{iter}	Número máximo de iteraciones
$f(\cdot)$	Kernel de una función objetivo
m	Valor de membresía de un elemento a un grupo
w	Peso o factor de ponderación de un elemento
δ, ϵ	Valores de umbral o restricción
$\varphi_{\mathbf{Q}}$	Función de actualización de centroides
$\varphi_{\mathbf{C}}$	Función de asignaciones para conformar una partición
\mathbb{G}	Grafo
λ	Valor propio
\mathbf{R}, \mathbf{U}	Matrices de rotación ortonormal
\mathbf{y}	Señal ECG
\mathbf{b}	Latido

OPERADORES MATEMÁTICOS

$n_e(\cdot)$	Número de elementos contenidos
$\ \cdot\ $	Norma euclidiana
$d(\cdot, \cdot)$	Medida de distancia o disimilitud
$ \cdot $	Valor absoluto
$\ \cdot\ _F$	Norma de Frobenius
$\text{lazos}(\cdot, \cdot)$	Peso total de las conexiones entre sus argumentos
$\text{lazosnor}(\cdot, \cdot)$	Proporción relativa de conexiones entre sus argumentos
$\text{grado}(\cdot)$	Grado de un grafo
$\langle \cdot \rangle$	Indicador binario del valor de verdad de su argumento.
$\mu(\cdot)$	Media aritmética
$\langle \cdot, \cdot \rangle$	Producto escalar
$\text{tr}(\cdot)$	Traza de una matriz
$\text{diag}(\cdot)$	Diagonal de la matriz de su argumento
$\text{Diag}(\cdot)$	Matriz diagonal formada por el vector de su argumento
s. a.	Sujeto a

Abreviaturas

AHA	<i>American Heart Association</i>
AAMI	<i>Association for the Advanced of Medical Instrumentation</i>
MIT	<i>Massachusetts Institute of Technology</i>
MIT/BIH	<i>MIT/Beth Israel Hospital</i>
HRV	<i>Heart rate variability</i>
GIM	Modelo iterativo genérico del agrupamiento
CBC	Agrupamiento basado en centroides
HM	H-medias
KHM	K-medias armónicas
FHM	H-medias difuso
GEMC	Agrupamiento basado en la máxima esperanza gaussiana
MSSC	Agrupamiento basado en la mínima suma de cuadrados
SVD	Descomposición en valores singulares
WT	Transformada de wavelet
PCA	Análisis de componentes principales
cte	Constante

Parte I

Preliminares

1. Introducción

El electrocardiograma (*ECG*) es una de las pruebas diagnósticas más importantes para el análisis del comportamiento eléctrico del corazón [1]. Además, es de uso muy frecuente por tratarse de una técnica no invasiva. El examen electrocardiográfico más común es el ECG de superficie de 12 derivaciones, que se realiza cuando el paciente se encuentra en reposo y se aplica por un breve periodo de tiempo. Comúnmente, los registros o señales ECG son evaluadas por un especialista quien, de acuerdo a su conocimiento y experiencia, emite un juicio sobre la salud cardíaca del paciente. En la actualidad, los especialistas se apoyan en herramientas computacionales, que permiten obtener medidas más precisas sobre las señales de forma que el diagnóstico sea más acertado. Existe otro tipo de prueba diagnóstica, denominada electrocardiografía ambulatoria, que se realiza para evaluar al paciente durante prolongados periodos de tiempo, sin alterar su actividad diaria, lo que permite el examen dinámico del ECG en su ambiente natural. La necesidad de un registro ambulatorio radica en que la detección de algunas patologías transitorias e infrecuentes no puede hacerse con el examen ECG de 12 derivaciones [1]. Los registros obtenidos del examen ambulatorio se les conoce con el nombre de registros *Holter*, en honor a Norman Jefferis Holter, quien desarrolló el primer grabador de señales ECG portátil.

1.1 Planteamiento del problema

Uno de los mayores problemas de la evaluación diagnóstica ambulatoria es la longitud de los registros, lo que hace que la inspección visual sea un trabajo laborioso para los especialistas. Por esta razón se han desarrollado herramientas de asistencia diagnóstica basadas en técnicas de procesamiento digital de señales y reconocimiento de patrones. En este caso, la idea general de la asistencia diagnóstica es disminuir la cantidad de latidos que debe revisar el especialista, de manera que, en lugar de revisar todos los latidos, únicamente deba revisar los latidos representativos de cada grupo obtenido por el sistema de análisis y, al final, sugerir un tipo de patología asociada [2]. No obstante,

el análisis computarizado de los registros Holter también representa un problema en términos de tiempo de procesado. En el proceso de formar subconjuntos de latidos existen dos esquemas: supervisado y no supervisado. En el esquema supervisado es necesario disponer de un conjunto completo de latidos conocidos y etiquetados para clasificar los nuevos latidos del registro a través de una comparación con latidos de las clases conocidas y, de esta manera, obtener un diagnóstico automático, aunque de rendimiento bajo porque una comparación con latidos base o plantilla no genera, en todos los casos, buenos resultados debido a la variabilidad morfológica y de longitud de los latidos. De otro modo, los métodos supervisados requieren entrenamiento por cada paciente o registro, y esto complica la labor del especialista.

Por otra parte, en los métodos no supervisados, no es necesario etiquetar el conjunto de datos, ni se requiere de entrenamiento por cada registro y la tarea es más flexible. Estos aspectos hacen que el análisis no supervisado sea la técnica de uso más frecuente en el diseño de sistemas para examinar los latidos de un registro Holter. En la clasificación no supervisada, las técnicas de agrupamiento o *clustering* han resultado muy útiles, sin embargo aún existe el problema abierto de la selección del método adecuado para agrupar patrones homogéneos, que sea robusto ante factores como la gran cantidad de latidos, costo computacional, clases desequilibradas, inicialización y la alta variabilidad morfológica [3].

En este trabajo de tesis se presenta una evaluación de diferentes métodos de agrupamiento particional y espectral, aplicados a patrones representativos de señales ECG, extraídos de segmentos de registros Holter que presentan una variedad de arritmias cardíacas, con el fin de determinar los que presentan un buen compromiso entre efectividad y tiempo de procesado en el agrupamiento de latidos homogéneos.

1.2 Organización del documento

Este documento está compuesto por 10 capítulos distribuidos en 5 partes, así: preliminares, marco teórico, marco experimental, resultados y discusión, y comentarios finales. El contenido es el siguiente:

- En el capítulo 2 se presentan los objetivos de este estudio.
- En el capítulo 3 se presenta un recorrido conceptual de algunas generalidades de la fisiología del corazón y el procesamiento digital de señales ECG. Adicionalmente,

se describe la base de datos empleada y algunos estándares del análisis ECG.

- El capítulo 4 es un estado del arte de la clasificación de señales ECG, que corresponde a una revisión bibliográfica de algunos trabajos relevantes en el área del procesamiento de señales ECG, específicamente, en procesos de extracción - selección de características y clasificación.
- Los capítulos 5 y 6, corresponden al marco teórico y abarcan los dos enfoques de análisis no supervisado estudiados: particional y espectral, respectivamente. Cada uno de estos capítulos tiene una sección de introducción en la que se presenta un breve contexto y se mencionan las secciones que componen el capítulo.
- En el capítulo 7 se describen las metodologías diseñadas para evaluar los métodos de agrupamiento y en el capítulo 8 se mencionan las medidas de desempeño aplicadas en este estudio. Los resultados de los experimentos se discuten en el capítulo 9.
- Finalmente, en el capítulo 10 se presentan las conclusiones de este estudio y se menciona el posible trabajo futuro.

2. Objetivos

2.1 Objetivo general

Comparar el desempeño de las mejores técnicas de agrupamiento no supervisado asociadas a esquemas de tipo particional y espectral, en la partición del conjunto de patrones descriptivos de la señal ECG, con el fin de encontrar las que presenten equilibrio entre la efectividad en la separación de clases y el costo computacional, considerando que las clases son desequilibradas y puede existir una variedad de morfologías de latido en cada señal.

2.2 Objetivos específicos

- * Implementar algoritmos representativos de agrupamiento particional y espectral para aplicarlos a patrones de señales ECG y evaluar su desempeño, teniendo en cuenta costo computacional y efectividad en la separación de clases.
- * Desarrollar una metodología de comparación de los algoritmos de agrupamiento que permita identificar el algoritmo adecuado para el análisis de señales ECG.
- * Proponer una metodología de análisis de registros Holter basada en métodos no supervisados, orientada a la clasificación de arritmias.

3. Contexto

En este capítulo se presentan algunos conceptos y discusiones preliminares orientadas al soporte diagnóstico de ECG. En la sección 3.1 se presenta un breve recorrido conceptual alrededor de la fisiología del corazón a nivel eléctrico, el trazado del electrocardiograma y los tipos de latido. En las secciones 3.2 y 3.3, se describe la base de datos empleada y se presentan generalidades del procesamiento digital de señales ECG. Adicionalmente, en la sección 3.3.1, se mencionan estándares para el análisis de señales ECG.

3.1 Contexto Fisiológico

El corazón posee un sistema electrogénico especializado que genera de forma rítmica impulsos que se transmiten por toda su superficie y producen la contracción del músculo cardíaco necesaria para el bombeo de la sangre. Este sistema rítmico de conducción es susceptible de lesiones por cardiopatías, especialmente, por la isquemia de los tejidos cardíacos, en consecuencia se presenta un ritmo anormal y el trabajo de bombeo se ve gravemente afectado, incluso podría causar la muerte [4]. En el proceso de bombeo de sangre intervienen cuatro cavidades que posee el corazón, dos superiores y dos inferiores denominadas aurículas y ventrículos, respectivamente. Las aurículas reciben la sangre del sistema venoso y la llevan a los ventrículos, y desde ahí se transporta a la circulación arterial. La aurícula y ventrículo derecho reciben la sangre baja en oxígeno y la bombean al sistema de circulación pulmonar, y las cavidades izquierdas bombean la sangre oxigenada a todo el organismo.

A nivel eléctrico, este proceso genera un trazado que tiene una pauta normal de acuerdo a la configuración de los electrodos. Las configuraciones se conocen con el nombre de derivaciones (ver Figura 3.1). El trazo obtenido se denomina electrocardiograma o señal electrocardiográfica (ECG), en donde, en el caso normal, se presentan tres ondas significativas: onda *P*, complejo *QRS* y onda *T*, como se muestra en la Figura 3.2.

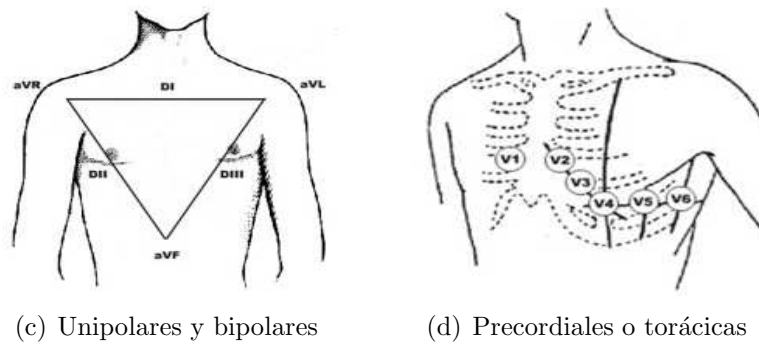


Figura 3.1: Derivaciones para el análisis ECG

La onda P se genera antes de cada contracción auricular, es decir, cuando las aurículas se despolarizan. El complejo QRS ocurre mientras el flujo de despolarización atraviesa los ventrículos. La onda T es de repolarización, por tanto se genera cuando los ventrículos se dilatan y se recuperan de su estado de despolarización. Los conceptos de polarización y despolarización se asocian al estado de los potenciales de membrana, en este caso, los tejidos del corazón.



Figura 3.2: Electrocardiograma normal

Algunos tramos de la señal ECG tienen relación directa con el tipo de anomalía del paciente. Por ejemplo, en la hipertrofia de un ventrículo, la despolarización tarda más en pasar por el ventrículo hipertrófico y por tanto se extiende la longitud del complejo QRS. También en las miocardiopatías ocurren alteraciones en el voltaje del complejo QRS, en las que se nota una considerable disminución del pico máximo y puede significar tendencia a infarto del miocardio [1].

El impulso cardíaco, en el caso normal, debe propagarse a través del haz de His y sus divisiones (rama derecha e izquierda), atravesando las fibras de Purkinje. Sin

embargo, cuando uno de los elementos de este sistema de conducción se altera, el impulso eléctrico puede dejar su trayectoria normal y propagarse a través del propio músculo ventricular, lo que genera un retraso y cambios morfológicos (melladuras) en el complejo QRS que podrían ser atribuidos a contracciones ventriculares prematuras (*V*), o un bloqueo de rama derecha (*R*) o izquierda (*L*). A este tipo de arritmias se les denomina ventriculares y pueden detectarse analizando el ritmo, frecuencia y contorno de los complejos QRS [1]. Las patologías auriculares, como los latidos auriculares prematuros (*A*), se diagnostican analizando la onda P.

La Figura 3.3 muestra algunos tipos de morfología de latido.

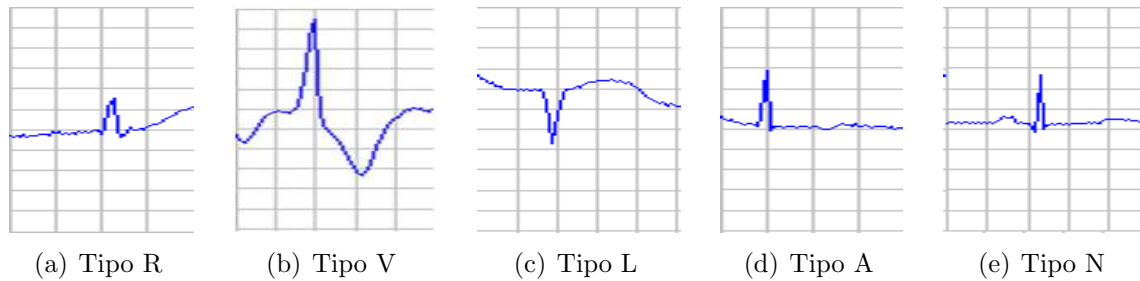


Figura 3.3: Diferentes morfologías de latido

En la Tabla 3.1 se relacionan algunos tipos de patología con sus características morfológicas y de latencia.

Tipo de arritmia o latido	Onda P			Complejo QRS		
	Frecuencia (bpm)*	Ritmo	Contorno	Frecuencia (bpm)	Ritmo	Contorno
Ventricular prematuro	60 a 100	Regular	Normal	60 a 100	Regular	Anormal, > 200 ms
Bloqueo de rama derecha	60 a 100	Regular	Normal	60 a 100	Regular	Anormal, > 200 ms
Bloqueo de rama izquierda	60 a 100	Regular	Normal	60 a 100	Regular	Anormal, > 200 ms
Auricular prematuro	-	-	-	Similar a una taquicardia	Regular	Normal

Tabla 3.1: Características de algunas arritmias. Información tomada de [1]. * Latidos por minuto.

Existen otras características diagnósticas como la respuesta a cierta medicina, respuesta ventricular después de un masaje carotídeo y otros indicadores provenientes de la exploración física.

Además de las patologías cardíacas existen otros factores a nivel fisiológico, asociados a antecedentes clínicos y complejidad del paciente, a los que se le atribuye la variabilidad morfológica de los latidos. Por ejemplo, la diferencia de potencial del punto R al punto más bajo de la señal, normalmente, debe encontrarse entre 0.5 mV y 2 mV, sin embargo esta relación no se cumple constantemente ni siquiera en pacientes normales. Esto se debe a que la corriente eléctrica alrededor del corazón se genera en proporción a la masa muscular [4].

Una causa muy común de la disminución del potencial en el electrocardiograma es la presencia del líquido pericardio. Este líquido es buen conductor y disminuye la amplitud de la pauta normal de la señal ECG. Otra razón por la que puede afectarse la amplitud normal de los potenciales es el padecimiento de enfisema pulmonar, puesto que el paciente diagnosticado con esta enfermedad suele tener un ensanchamiento de la cavidad torácica y los pulmones tienden a envolver el corazón, por tanto se aísla el voltaje entre la superficie del cuerpo y el corazón.

3.2 Base de datos de arritmias

El *MIT* (*Massachusetts Institute of Technology*) tiene para uso público bases de datos de señales ECG, entre ellas la base de datos de arritmias *MIT/BIH* que dispone de registros con diferentes tipos de arritmia (A, L, R, V, entre otros). Las señales fueron adquiridas a través de dos canales (0 y 1), con una frecuencia de muestreo de 360 hz, 11 bits de resolución y 10 mv de rango. Esta base de datos consta de 48 registros obtenidos de diferentes derivaciones. Estos registros se encuentran etiquetados con diversas anotaciones útiles para su análisis y clasificación. En [5] se explica detalladamente la organización de esta base de datos.

3.3 Procesamiento digital de señales ECG

El procesamiento de señales ECG orientado a la asistencia diagnóstica consta de varias etapas, en general, implica pre-procesamiento, reducción de perturbaciones, segmentación, extracción - selección de características y clasificación. El resultado final del sistema de diagnóstico automático depende del funcionamiento de todas las etapas, en otras palabras, si el conjunto de características no representa las propiedades intrínsecas de la señal ni genera buena separabilidad, los resultados de la clasificación no serán correctos [6].

La señal adquirida no debe utilizarse directamente para realizar un proceso de diagnóstico, ya que ésta posee información adicional o distorsionada debido a las variaciones de la línea base, interferencias con la red y otros tipos de ruido [6]. Esta información, ajena a la sustancial, debe minimizarse para optimizar el proceso de análisis y obtener un soporte adecuado para el diagnóstico. Las perturbaciones de la señal se pueden clasificar en dos tipos: artificiales y biológicas. Las perturbaciones artificiales se refieren al equipo electrónico y pueden darse por interferencia con la línea de potencia, ruido por dispositivos electrónicos o artefactos por contacto de electrodos. Mientras que las perturbaciones de tipo biológico se le atribuyen a las señales electromiográficas (*EMG*), las cuales se superponen a la señal requerida y ocultan rasgos característicos. El ruido por desplazamiento de la línea base debido a la respiración y los artefactos por movimiento, también se catalogan como perturbaciones biológicas. Las técnicas de reducción de perturbaciones se orientan al diagnóstico, por tanto, después del proceso de filtración, la señal no debe perder sus características intrínsecas como su morfología y duración [7]. Entre los principales métodos que se emplean para la disminución del ruido se encuentran las técnicas clásicas de filtración, en las que se aplican filtros pasa bajas (FIR e IIR) [8], aunque no son muy utilizadas por su limitada selectividad. También se emplean filtros de media móvil [9], aunque su selectividad también es baja y traslapa los componentes espectrales de alta frecuencia [10]. Otro método común es el de promediado temporal de latidos que hace uso de la regularidad de los latidos y la correlación de la señal con el ruido [6]. En [11] se plantean métodos de aproximación mediante la representación de la señal con funciones ortogonales, utilizando productos escalares y estimación adaptativa basada en el algoritmo de los mínimos cuadrados. Este estudio se extiende en [12]. Un método relativamente nuevo para reducir perturbaciones es el de la transformada de *wavelet* (*WT*), en el que se descompone la señal hasta el nivel que se requiera y al final se calcula la transformada inversa para recuperar la señal sin los elementos atribuidos al ruido, que fueron eliminados de acuerdo a un umbral preestablecido como se estudia en [13], [14] y [15]. La conveniencia de un método de filtración u otro depende de las condiciones iniciales de la señal, como la relación señal a ruido, y de los requerimientos de diseño como costo computacional y fidelidad.

En la etapa de segmentación se identifican latidos y tramos significativos de los mismos para analizar los patrones de la señal, por tanto, es el punto de partida de todo sistema de clasificación. Debido a factores como perturbaciones, dinámica de la señal

y la cantidad de morfologías de latido que se pueden presentar en el registro, es necesario que los algoritmos de segmentación empleen umbralización adaptativa en tiempo y amplitud [16, 17, 18, 19] y, principalmente, ajuste adaptativo de morfología, para la detección de la marca fiducial del pico R, que es el punto relevante para separar los latidos y continuar con el proceso. De acuerdo al tipo de latido o patología que se requiera clasificar, puede ser necesaria la detección de las ondas P, T y/o QRS [20, 21, 22].

Generalmente, en el procesamiento de señales ECG no se emplean directamente las muestras, sino características de las mismas o señales transformadas, porque esto último resulta ventajoso considerando que la señal puede verse afectada por perturbaciones, como se comentó en párrafos anteriores. Por otra parte, un conjunto de características que se encuentre en un espacio de análisis grande puede presentar problemas asociados a costo computacional e información redundante, por esta razón, si el conjunto de características resultante se encuentra en un espacio, relativamente, de alta dimensión, es necesario realizar una etapa de selección de características efectiva para reducir el número de parámetros y obtener resultados de clasificación mejores o iguales que los obtenidos de emplear todo el conjunto. En suma, las etapas de extracción y selección de características se llevan a cabo para encontrar, idealmente, la mínima información relevante que permita identificar diferentes clases de latidos.

Existen diversos métodos para la extracción de características: métodos heurísticos, métodos que utilizan información estadística de la señal (análisis de componentes principales *PCA* y sus variantes [23, 24]), métodos de representación a través de bases (*Hermite* [25], WT [26], distribuciones de tiempo-frecuencia [27]), métodos de aproximación con polinomios o curvas características [6], métodos basados en muestreo no lineal de la señal [6], entre otros. El conjunto de características de la señal, comúnmente, lo conforman patrones morfológicos de la señal, coeficientes de representación de la señal o medidas del ritmo a partir de la distancia entre los picos R (*HRV: Heart rate variability*) [2].

En el procedimiento de extraer latidos representativos de un registro, se puede optar por técnicas de análisis supervisado o no supervisado. En los métodos supervisados, es necesario disponer de un conjunto de latidos conocidos y etiquetados para clasificar los nuevos latidos de acuerdo al análisis comparativo con cada una de las clases conocidas y, de esta manera, obtener un diagnóstico automático. El mayor problema que presenta este tipo de análisis es la sensibilidad a la variabilidad de la señal debido a

la diversidad de patologías, equipos de adquisición, artefactos, entre otras; además, la detección de un latido específico exige entrenamiento para cada registro o paciente. Cabe anotar que el diagnóstico automático obtenido con estas técnicas no podría ser suficiente para reemplazar, absolutamente y en todos los casos, a la inspección visual realizada por el cardiólogo.

Por otra parte, en el análisis no supervisado, no es necesario etiquetar el conjunto y el proceso de clasificación es más flexible, sin embargo, no puede obtenerse un diagnóstico automático. Las técnicas no supervisadas son las más empleadas porque, considerando una medida de disimilitud adecuada, permiten obtener un pequeño subconjunto de latidos representativos con el fin de reducir la cantidad de latidos que debe revisar el especialista, entonces, la tarea del especialista es diagnosticar adecuadamente un prototipo de cada subconjunto de latidos homogéneos. En la clasificación no supervisada, los métodos de agrupamiento han resultado muy útiles en el análisis exploratorio de datos, sin embargo aún representan un problema combinatorio grande porque debe seleccionarse el método adecuado para agrupar los patrones representativos del registro, considerando la gran cantidad de latidos, costo computacional y clases desequilibradas [3].

3.3.1 Estándares del análisis ECG

La *AAMI* (*Association for the Advanced of Medical Instrumentation*), recomienda algunos estándares para reportar resultados de desempeño de algoritmos de análisis del ritmo cardíaco [28]. Entre las recomendaciones se encuentran: el uso de al menos 5 clases de latidos y la comparación de al menos 12 configuraciones de características de procesamiento obtenidas de uno o varios canales ECG. En la Tabla 3.2 se muestran los conjuntos de arritmias que recomienda la AAMI relacionados con la notación de la base de datos MIT/BIH.

Descripción del latido	N Latidos diferentes de S,V, F o Q	S Ectópico supraventricular	V Ectópico ventricular	F Latidos mezclados	Q Latidos desconocidos
Tipos de latido de la base de datos MIT-BIH	N, L, R, fuga arterial (e), fuga de unión nodal (j)	A, aberración arterial, prematura (a), unión prematura (J), supraventricular prematuro (S)	V, fuga ventricular (E)	Unión de V y N (F), unión de P y N (f)	Generado por marcapasos (P), Sin clase (Q)

Tabla 3.2: Grupos de arritmias recomendados por la AAMI

La *AHA* (*American Heart Association*), acogiendo las recomendaciones de la AAMI, sugiere analizar, además de los latidos normales, las arritmias estándar (A, R, L y V), porque la identificación de este tipo de latidos permite diagnosticar diversas patologías que se derivan de estas arritmias.

4. Estado del arte de la clasificación de señales ECG

A continuación se presenta una revisión bibliográfica de algunos trabajos relevantes en el área del procesamiento de señales ECG, específicamente, en procesos de extracción y selección de características, y clasificación.

4.1 Extracción y selección de características

Los médicos especialistas emplean reglas basadas en características morfológicas y del ritmo de la señal ECG para emitir un concepto sobre la salud cardíaca de un paciente. Partiendo de este hecho, en muchos trabajos se realiza la extracción de características a través de la cuantificación de factores diagnósticos como morfología y ritmo. En [29], se propone realizar la caracterización de la señal ECG a partir de su segmentación. Los parámetros obtenidos corresponden a la detección de puntos significativos (inicio, final, pico) de las ondas P, T y el complejo QRS. El principal objetivo del trabajo fue realizar la extracción en tiempo real y brindar la información al especialista para su posterior análisis. En [30] se desarrolla un método de clasificación de latidos basado en la comparación de los latidos empleando plantillas del QRS de un conjunto descriptor de ECG. Este método considera la máxima correlación cruzada, diferencia de los espectros de frecuencia y características temporales a través del análisis de los intervalos *RR*. En [31] se presenta un algoritmo para la detección de latidos auriculares prematuros (*APB*) empleando etapas de detección y clasificación de latidos basadas en el cálculo de las diferencias ponderadas entre latidos en el intervalo *RR* y la morfología del QRS (diferencia de ancho y área), obteniendo 92.2% de sensibilidad y 96% de especificidad. La extracción del QRS se realizó con el método propuesto en [32].

Otros trabajos utilizan características resultantes de aplicar transformaciones o representaciones de la señal en otro espacio. Por ejemplo, en [33], se realiza la extracción

de características aplicando métodos generalizados que consisten en hacer aproximaciones a tramos de la señal, utilizando seis funciones definidas (constante, recta, triangular, trapecio, exponencial y sinusoidal). Todas las funciones tienen parámetros que se sintonizan a través de un criterio basado en el error cuadrático medio. Se probaron las funciones independientemente y también combinaciones de las mismas. Para medir la efectividad de la técnica, se clasificaron 2 clases de señales (anormales y normales), utilizando un clasificador de tipo árbol de decisión (*CART*) y se comparó el desempeño en porcentaje de clasificación con técnicas como la transformada identidad, transformada de Fourier y WT. Se obtuvieron resultados superiores, aunque no mayores del 90% en porcentaje de clasificación.

En tareas de reconstrucción y caracterización del complejo QRS, el modelo de Hermite constituye una buena alternativa [34, 35, 25]. En [34], se determinó que, en promedio, el 98.6% de la energía del complejo QRS puede representarse utilizando tres coeficientes de Hermite. En [35], se desarrolla un sistema de detección, clasificación e identificación en línea de complejos QRS, en el que se aplica el modelo de Hermite como uno de los métodos para la extracción de características. Para esto, se automatiza el cálculo del parámetro de escala (σ), a partir del algoritmo *Levenberg-Marquardt*. El conjunto de características resultantes lo conformaron, además de los coeficientes de Hermite, algunas características heurísticas, y se obtuvieron errores de clasificación del orden del 3% para un total de 75988 latidos normales y anormales de la base de datos MIT/BIH. Aunque el proceso es relativamente rápido, el cálculo del parámetro σ debe realizarse para cada complejo QRS detectado, lo que retarda el proceso para requerimientos de tiempo real. En otros estudios [36, 37], se emplea la caracterización de Hermite (coeficientes con $n = 5$, σ óptimo) y morfológica (energía del QRS), y análisis no supervisado orientado a la detección de arritmias ventriculares (V, L y R). En [37], se propone una metodología para la reconstrucción y extracción de características del complejo QRS, empleando el modelo paramétrico de Hermite. La reconstrucción de las señales se realiza obteniendo, con métodos iterativos, el valor óptimo del parámetro de escala y el menor número de bases de Hermite que generan una reconstrucción con poca pérdida de información espectral. El valor óptimo del parámetro de escala se obtuvo a partir del menor grado de disimilitud entre la señal original y la reconstruida, empleando el algoritmo *DTW* (*Dynamic Time Warping*) como medida de disimilitud [6]. Para estimar la cantidad mínima de bases necesarias, se analizan los espectros de frecuencia de las señales reconstruidas con diferentes grados, considerando que las

variaciones más significativas del espectro de potencia del QRS se presentan en el rango de 1 – 20 Hz [34], y además, los componentes asociados a dicho rango de frecuencia, no se alteran por elementos de alta frecuencia (superior a los 20 Hz) como la interferencia de la línea de potencia (50 Hz / 60 Hz) y el ruido muscular, ni por componentes de muy baja frecuencia (≤ 1 Hz) [38]. La caracterización se evalúa realizando un proceso de agrupamiento con el algoritmo *K-medias máx-mín*, implementado como se explica en [36]. Las características aplicadas fueron la energía del QRS, la diferencia de los espectros de los complejos originales y un complejo plantilla, y las características de Hermite.

Otra técnica empleada en la extracción de parámetros de señales ECG es la transformada de Karhunen - Løeve (*KLT*), que es lineal, ortogonal y óptima en términos del error cuadrático medio, es decir que concentra la información de la señal en un mínimo número de parámetros. Tiene propiedades adicionales tales como: entropía de representación mínima y coeficientes no correlacionados. Realizando un análisis de la expansión de alguna serie de tiempo con una base ortonormal, se puede llegar al problema de los valores propios o autovalores, donde los vectores propios, asociados a los autovalores de mayor valor, representan la mayor proporción de energía proyectada. Por tanto, la *KLT* como extractor de características de orden N , debe usar los primeros N vectores propios ordenados de manera descendente. La *KLT* ha sido empleada ampliamente en análisis de señales ECG, como en análisis del segmento ST y onda T [39], y compresión de ECG [40].

El problema de la selección de características de un conjunto de datos considerablemente grande es típico en el reconocimiento de patrones y aprendizaje de máquina, y se presenta en diferentes ramas de la ciencia: procesamiento de texto, bio-informática, entre otras. La caracterización de latidos en señales ECG puede resultar en un conjunto de características de alta dimensión, debido a las diversas morfologías y la alta variabilidad de la señal, por esta razón se han realizado diversos trabajos en torno a la selección de características. En [41] se presenta una definición de relevancia basada en las propiedades espectrales del laplaciano de la matriz de características empleando análisis supervisado y no supervisado. El proceso de selección de características se realiza a través de un método iterativo de optimización basado en mínimos cuadrados.

Debido a la naturaleza no estacionaria de la señal ECG, su análisis a partir de la transformada de *wavelet* (WT), ha brindado buenos resultados en todas las tareas

del análisis de la señal, desde preprocesamiento hasta clasificación. En [42], se utiliza WT con una función madre de tipo *Mexican Hat* que presenta un ajuste similar a la señal ECG, de acuerdo al trabajo realizado en [43]. Se utilizaron las escalas diádicas 2^j , $j = 1, 2, 3$, para obtener los parámetros. Las transformaciones forman la entrada a un modelo oculto de Markov (*HMM*) para propósitos de caracterización de la señal en sus principales complejos. A partir de la transformada de *wavelet* se pueden diseñar métodos que pueden aplicarse a muchas tareas en el contexto del procesamiento de señales. Las funciones *wavelet* son ideales para el análisis de cambios súbitos de las señales de corta duración. Una aplicación muy importante es la capacidad de calcular y manipular características. El cálculo de la WT continua, implica que los parámetros de escalamiento y traslación cambien continuamente. Sin embargo, el cálculo de los coeficientes de onda para cada posible escala puede representar un esfuerzo considerable debido a la gran cantidad de datos. Por esta razón, la transformada *wavelet* discreta (*DWT*) es preferida en muchos estudios.

En [44] se realizó un análisis espectral de señales ECG, pletismográficas (*PPG*) y electroencefalográficas (*EEG*), mediante la DWT. Para esto se escogieron 4 niveles de descomposición, considerando los componentes de frecuencia dominantes de la señal. Estos niveles fueron escogidos de tal forma que se conserven, en los coeficientes *wavelet*, las partes de la señal que se correlacionan bien con las frecuencias requeridas en la tarea de clasificación. Entonces, las señales fueron descompuestas en 4 sub-bandas de detalle: D_1 (37.5 - 75 Hz), D_2 (18.75 - 37.5 Hz), D_3 (9.375 - 18.75 Hz), D_4 (4.6875 - 9.375 Hz); y una aproximación final A_4 (0 - 4.6875 Hz). Se escogió la función madre *Daubechies* de orden 2 (db2) porque fue la que presentó la mayor eficiencia en la detección de cambios en las señales consideradas en este estudio. Los coeficientes de la DWT generan una representación compacta de la distribución de energía de la señal en el tiempo y la frecuencia, por esta razón fueron escogidos para realizar la etapa de extracción de características. Las características consideradas fueron: el máximo y mínimo coeficiente de la WT de cada sub-banda, y la media y desviación estándar de los coeficientes de la WT de cada sub-banda.

En [45] se propone un método para el reconocimiento y detección de latidos ECG usando una red neuronal adaptativa basada en la WT (*AWN*). Se realizó la extracción del complejo QRS mediante la detección del pico R y una ventana de 280 ms. Luego se tomaron 100 muestras alrededor de dicho pico (50 puntos antes y 50 puntos después), con una frecuencia de muestreo de 360 Hz. Las características extraídas corresponden a la parte real de las funciones *wavelet* ($\phi_{R_{d,i}}(x_i)$, $i = 1, 2, \dots, 100$).

Después de esta revisión de de algunos trabajos relevantes relacionados con la extracción y selección de características de señales ECG se puede concluir que, en general, las técnicas de representación con bases ortogonales (Hermite, KL), técnicas basadas en la estadística de la señal (PCA), técnicas de representación tiempo-frecuencia (WT) y variantes de las mismas, son herramientas de uso frecuente por su eficacia en tareas de extracción y selección de información relevante de la señal.

La Tabla 4.1 relaciona algunas técnicas de caracterización con la referencia principal y el tipo de latido.

Tipo de latido	Técnica / Característica	Ref.
A	Diferencia ponderada de la distancia RR : $RR_{Diff} = \frac{RR_n - RR_{n-1}}{\frac{1}{5} \sum_{i=n-7}^{n-2} RR_i} \times 100\%$ Diferencia del ancho del QRS: $QRS_{WidthDiff_n} = \frac{ QRS_{Width_n} - QRS_{Width_{REF}} \times 100}{QRS_{Width_{REF}}}$ Diferencia de área del QRS: $QRS_{AreaDiff_n} = \frac{ QRS_{Area_n} - QRS_{Area_{REF}} \times 100}{QRS_{Area_{REF}}}$	[31]
NSVT	Radio de potencia espectral (LF/HF) para cuantificar la HRV: $r_{i,j/m,n} = \frac{LF}{HF} = \frac{P_{i,-5} + P_{j,-5}}{P_{m,-3} + P_{n,-2}}$ $P_{i,j} = \frac{1}{T_j} \sum_{k=(i-1)M}^{iM-1} d_j^2(k), d_j \text{ son los coeficientes de la DWT}$	[46]
N, V, A, L, R	Multiplicación de LSAF de orden 20 y 30 con la señal ECG.	[47]
N, V, A, R, L, F	Espectro de frecuencia del QRS (1-12 Hz)	[48]
SV	Distribución de Husimi del QRS	[49]
PVC	Diferencia del espectro de frecuencia: $FS_{Diff} = \sum_{f=2}^F SHW_f - STW_f , F = 25Hz$	[30]
N, anormal	Índice de irregularidad del QRS (coeficiente de Lipschitz α): $\min_{\alpha} \sum (\lg W_{2^j}(n) - \lg k - j\alpha \log 2)^2,$ k es la constante de Lipschitz ([51]) s. a. $\lg W_{2^j}(n) \leq \lg k + j\alpha \log 2$ Características estadísticas obtenidas mediante DWT (db2): Máximo coeficiente <i>wavelet</i> en cada sub-banda ($D_1 - D_4$). Media de los coeficientes <i>wavelet</i> en cada sub-banda. Mínimo coeficiente <i>wavelet</i> en cada sub-banda. Desviación estándar de los coeficientes <i>wavelet</i> en cada sub-banda.	[50] [44]

Tabla 4.1: Técnicas de caracterización de señales ECG

La Tabla 4.1 se hizo de acuerdo a la siguiente notación:

- RR_i : Vector de distancias RR de un registro.
- RR_{Diff} : Diferencia ponderada de las distancias RR.
- QRS_{Width} : Vector de las longitudes del QRS.
- $QRS_{Width_{REF}}$: Longitud de la plantilla del QRS.
- $QRS_{WidthDiff_n}$: Diferencia de los anchos del QRS con respecto de una referencia.
- $QRS_{Area_{REF}}$: Área de la plantilla del QRS.
- FS_{Diff} : Diferencia de los espectros de frecuencia.
- SHW : Espectro de un latido en particular.
- STW : Espectro de la plantilla.
- NSVT: Taquicardia ventricular no sostenida.

4.2 Clasificación

A continuación, se realiza una revisión de algunos trabajos relacionados con la clasificación de latidos en señales ECG considerando dos enfoques: supervisado y no supervisado.

4.2.1 Clasificación supervisada

En general, los métodos de clasificación supervisada se dividen en métodos estadísticos, sintácticos y de inteligencia artificial.

En muchos estudios, se plantean esquemas de clasificación de latidos empleando redes neuronales artificiales [38, 52]. Los más conocidos son el perceptrón multicapa, redes auto-organizativas de Kohonen, sistemas difusos o neuro-difusos y la combinación de diferentes redes neuronales en sistemas híbridos. El sistema de reconocimiento de latidos típico aplica clasificadores neuronales entrenados con varias redes neuronales y se escoge la mejor. El método más eficiente está basado en la combinación de clasificadores utilizando diferentes estructuras de redes o diferentes métodos de pre-procesamiento (extracción y selección de características) [2]. En [53], se propone un

método para clasificar arritmias utilizando una arquitectura de red neuronal (NN) y agrupamiento no supervisado difuso (FC), denominada $FCNN$ y se compara su desempeño con un clasificador de redes neuronales con arquitectura perceptrón multicapa (MLP) desarrollado con entrenamiento de propagación hacia atrás. Las señales ECG, son tomadas de la base de datos MIT/BIH las cuales se utilizan para entrenar el clasificador para 10 diferentes tipos de arritmias. Los resultados de las pruebas comprueban que el método $FCNN$ puede generalizar mejor y aprender más rápido que la arquitectura clásica MLP . La ventaja del método propuesto está en que disminuye el número de segmentos por grupo en los datos de entrenamiento debido al agrupamiento difuso realizado con el algoritmo c -medias.

El estudio realizado en [54] emplea máquinas de vectores de soporte (SVM) para clasificar 12 tipos de arritmias de la base de datos MIT/BIH. La etapa de extracción de características se realiza empleando dos enfoques: HOS (*High Order Statistics*) y coeficientes de Hermite. Los errores de clasificación para la caracterización de Hermite y HOS fueron de 2.15 % y 3.04 %, respectivamente.

En [55], se presenta un clasificador de latidos específico para un paciente (conocido como clasificador local), combinado con un clasificador global diseñado a partir de una base de datos de entrenamiento de señales ECG. Los dos clasificadores fueron combinados utilizando mezcla de expertos (MOE). El clasificador global, por sí solo, alcanza una efectividad de 62.2% y el clasificador MOE alcanza un 94% de efectividad para diferenciar latidos ectópicos ventriculares (VEB) de otros ($No VEB$).

En [56] se desarrolla una metodología para la detección de latidos de señales ECG utilizando técnicas de preprocesamiento y clasificación supervisada. Este trabajo tiene en cuenta los estándares de la AAMI e involucra etapas de filtración, segmentación, extracción de características y clasificación. La etapa de filtración de perturbaciones de alta y baja frecuencia se lleva a cabo con filtros digitales clásicos. La segmentación de las ondas significativas se realiza a través de la estimación de la ubicación del pico R. El conjunto de características lo conforman parámetros heurísticos obtenidos de la morfología de la señal, intervalos de tiempo (periodo RR), y duración de las ondas P, QRS y T. Para la etapa de clasificación se utilizó un clasificador discriminativo lineal. Este modelo fue determinado a través de la estimación de la máxima verosimilitud de los datos de entrenamiento.

4.2.2 Clasificación no supervisada

La técnica más usual para el análisis no supervisado de latidos en señales ECG es el agrupamiento no supervisado, que consiste en la clasificación de patrones a través de la conformación de grupos de patrones similares. El problema del agrupamiento ha sido abordado por investigadores de muchas disciplinas y en diferentes contextos, comprobando su versatilidad, practicidad y eficiencia en el análisis exploratorio de datos. Sin embargo, agrupar patrones homogéneos aún es un problema combinatorio abierto [57].

El enfoque clásico del agrupamiento no supervisado es el particional o agrupamiento basado en centroides *CBC*, en el que iterativamente se actualizan centroides y se asignan elementos a cada grupo, considerando criterios basados en distancias, disimilitudes y medidas estadísticas, hasta la convergencia de una función objetivo, que debe ser coherente con la función de actualización de los centroides. El algoritmo representativo en el agrupamiento basado en la mínima suma de cuadrados (*MSSC*) es el *K-medias* (MacQueen, 1967), del que se han planteado variantes y mejoras considerando criterios de inicialización como *máx-mín*, y diversas funciones de membresía de los puntos a cada grupo [58]. En [3] se presenta un método heurístico de búsqueda para resolver el problema del MSSC, en el que se actualizan los centroides realizando una evaluación local de la función objetivo, es decir, considerando únicamente una cierta vecindad de los centroides y no todo el conjunto de datos. Los movimientos de los centroides se realizan de forma aleatoria, y cada movimiento se acepta o se rechaza de acuerdo al valor de la función objetivo. A este método se le denomina *J-medias*.

En [58] se presenta una forma generalizada de realizar CBC, estudiando la proporción o grado de pertenencia de un punto a un grupo, dicha proporción se denomina función de membresía. Además, también se considera el grado de influencia o peso de cada punto en el cálculo de los nuevos centroides por cada iteración, que corresponde a una ponderación de los datos. De esta manera, el algoritmo de agrupamiento se realiza estableciendo un agrupamiento inicial, luego se calculan las funciones de membresía y peso de cada punto, a partir de estos valores se actualizan los centroides, y por último se repiten los dos pasos anteriores hasta la convergencia del algoritmo o hasta se que cumplan un número predeterminado de iteraciones.

En [59], se presenta una metodología para el agrupamiento de latidos de señales

ECG de la base de datos MIT/BIH. La primera etapa corresponde a la estimación de la ubicación del pico R, utilizando el algoritmo propuesto en [60], donde se utiliza la primera derivada de la señal ECG, transformaciones no lineales y umbralización adaptativa, para la estimación de la marca fiducial. Posteriormente, se extraen los latidos teniendo en cuenta el criterio del 20% de la distancia entre el pico R actual y el anterior para determinar el punto de inicio del latido, y el punto final, se encuentra con el 80% de la distancia entre el pico siguiente y el actual. Para tener correspondencia entre los latidos y realizar la comparación de los mismos, se aplica un proceso de normalización en amplitud y en tiempo. Para esta tarea, se utiliza la técnica DTW, aplicando restricciones locales y globales para reducir el costo computacional, obteniendo resultados aproximados al procesar sin aplicar las restricciones. La técnica se basa en submuestreo e interpolación uniforme de las señales que se comparan, teniendo en cuenta la morfología de los latidos. La siguiente etapa corresponde a la extracción de características del latido, en la que se prueban 4 técnicas: las muestras de la señal, segmentación de traza (muestreo no lineal de la señal), aproximación poligonal y WT. Se realiza un etiquetado de los latidos con el fin de medir el desempeño de los algoritmos de agrupamiento. Se probaron medidas de similitud para comparar las características de los latidos, como las normas L_1 y L_2 de *Minkowski*, que se usan comúnmente. Se aplica una primera etapa de *pre-agrupamiento* para disminuir la cantidad de latidos en el análisis, dado que un registro *Holter* puede albergar cientos de miles de latidos. Para llevar a cabo esta tarea se aplicó una medida de disimilitud entre latidos con un umbral relativamente bajo. Con esto se descartan los latidos que presenten cierta semejanza con los que ya han sido seleccionados.

Finalmente en la etapa de agrupamiento, se implementan dos algoritmos ampliamente utilizados en la literatura. En primer lugar, se aplica el algoritmo *máx-mín*, que corresponde a un algoritmo particional no paramétrico basado en el criterio voraz que disminuye el costo computacional. El segundo método, corresponde al *K-medias*, que es particional y no paramétrico, y se basa en re-calcular los centroides de forma iterativa. En este caso, se hace necesaria una inicialización completa de los centroides, porque utilizar una inicialización aleatoria, puede dar lugar a una partición inicial inapropiada. Dado que se necesitan re-calcular los centroides, lo cual no es posible en un espacio no euclidiano, se toma otro criterio, correspondiente a la mediana, modificando el algoritmo *K-medias*. Esta modificación recibe el nombre de algoritmo *K-medianas*. Después de realizar diversas combinaciones de los algoritmos de extracción de características y los algoritmos de agrupamiento no supervisado, se concluyó que la combinación que

obtuvo mayor rendimiento en el proceso de agrupamiento, consistió en el método de extracción de características, segmentación de traza y el algoritmo *K-medianas*.

En [61], se aplican los resultados obtenidos en [59] a la clasificación de latidos *VE* (Extrasístoles ventriculares), aplicando un esquema similar. Se encuentra el pico R, de cada latido. Se extrae el latido y se aplica la etapa de extracción de características donde se agregan algunas características propias de los latidos *VE*. La duración del intervalo *RR*, una medida de polaridad que calcula la posición relativa del valor de amplitud promedio del latido entre el máximo y el mínimo valor, y los puntos obtenidos al aplicar segmentación de traza, conforman el grupo de características que es evaluado por el clasificador. Para la inicialización se aplican algunas restricciones que tienen en cuenta la morfología de los latidos *VE*, con el fin de disminuir el costo computacional. Finalmente, se aplica el algoritmo *K-medias* para agrupar los latidos. Se obtienen resultados sobre el 90% en la sensibilidad, lo que resulta comparable con los mejores métodos que se encuentran en la literatura.

En [62], se estudia el desempeño de algunos métodos de segmentación y agrupamiento aplicados a señales de las bases de datos *MIT/BIH* y *ST-T/VALE* de la sociedad Europea. El algoritmo de segmentación propuesto es sensible al ruido, opera en el dominio del tiempo y se fundamenta en el concepto de la longitud de curva. En este procedimiento se extraen parámetros a lo largo del tiempo de la señal ECG, los cuales sirven como punto de decisión para determinar la presencia de un complejo QRS. Aunque el algoritmo presenta buenos resultados en comparación a otros tipos de algoritmos clásicos de segmentación, éste se puede ver afectado por morfologías donde el pico R es pequeño respecto a los otros complejos, debido a la derivación escogida. Una vez se ha estimado la marca fiducial, se procede a extraer el latido estimando los puntos característicos de los otros complejos. En la siguiente etapa se extraen las características de los latidos, aplicando los tres primeros componentes de PCA para caracterizar los complejos. La varianza acumulada de los tres primeros componentes es superior al 90%. Con las características obtenidas se realiza la etapa de agrupamiento, donde se aplica el algoritmo propuesto en [63], que consiste en una modificación del *K-medias*, denominado el *KHM*. El mayor problema del *K-medias*, radica en la sensibilidad a la selección de la partición inicial, convergiendo a un mínimo local de la función objetivo si los centroides no son escogidos de manera adecuada. El algoritmo *KHM*, resuelve dicho problema reemplazando la mínima distancia de una muestra a los centroides por la media armónica de la distancia de la muestra a todos los centros.

El algoritmo presenta mejor desempeño que el *K-medias*, principalmente al reemplazar la media armónica por la estrategia *winner-takes-all*, propia del *K-medias*. Además, se demuestra que el costo computacional del algoritmo es menor que los algoritmos de agrupamiento clásicos. Aunque en el trabajo no se especifica la cantidad de patologías analizadas, se presenta un resultado global, donde el 97% de los latidos de las bases de datos fue agrupado correctamente.

En [64], se realiza un estudio comparativo de medidas de similitud aplicadas al agrupamiento de complejos QRS obtenidos de registros de la base de datos MIT/BIH. Las medidas de similitud son cuatro: Manhattan (L_1), euclidiana (L_2), coeficiente de correlación y el grado relacional *Grey*. El algoritmo de agrupamiento analizado corresponde al llamado método no supervisado de dos pasos, el cual ha reportado mejor desempeño que los algoritmos de agrupamiento jerárquicos. Para evitar la sensibilidad a la partición inicial, se realizaron aleatoriamente 5 particiones iniciales para cada método, seleccionando el promedio de las 5 iteraciones. Se realizó un ajuste de umbral por cada iteración con el fin de obtener mejores resultados por cada vez. Los resultados de la clasificación fueron empleados para medir el desempeño de las medidas de similitud. Como resultado, la medida del grado relacional *Grey* presentó, en el peor caso, un desempeño del 97.55% frente a un 3% por debajo de este porcentaje en las otras medidas. Para las otras cuatro iteraciones el desempeño supera el 99%. Se concluye que, con el algoritmo de agrupamiento propuesto, la medida del grado relacional *Grey* presenta mejores resultados.

El agrupamiento no supervisado basado en teoría espectral es un enfoque relativamente nuevo del análisis no supervisado, sin embargo ya ha sido estudiado en diferentes trabajos que comprueban su eficiencia en tareas de clasificación, sobre todo en los casos en que los grupos no son linealmente separables. Esta técnica de agrupamiento ha sido empleada en una multitud de aplicaciones como diseño de circuitos [65], balanceo eficiente de carga computacional en computación intensiva [66], segmentación de imágenes [67, 68], entre otras. Lo que hace atractivo al análisis espectral aplicado en el agrupamiento de datos es el uso de la descomposición en valores y vectores propios para obtener óptimos locales aproximados a los óptimos globales continuos.

Con base en que los elementos representativos de cada grupo corresponden a los vectores propios asociados a los valores propios de mayor valor, han surgido diversos

estudios empleando técnicas espectrales y *kernels* de algoritmos clásicos como *k-medias* y el método de los k vecinos más cercanos *K-NN*. Por ejemplo, en [69] se presenta un algoritmo de agrupamiento espectral sencillo en el que se aplica el método convencional para estimar los elementos representativos de los grupos a través de la descomposición en valores y vectores propios (descomposición propia) de la matriz de afinidad normalizada. Subsecuentemente, se obtiene una matriz de vectores propios re-normalizada en amplitud, la cual se agrupa en k subconjuntos empleando el algoritmo *K-medias*. Por último, se asignan los elementos a los grupos correspondientes, de acuerdo a la matriz de datos inicial.

Un trabajo posterior [70], presenta de forma más elaborada un método de agrupamiento espectral multi-clase en el que se plantea un problema de optimización para garantizar que el valor de convergencia sea aproximado a los óptimos globales. Para esto se formulan dos sub-problemas de optimización, uno para obtener los óptimos globales en un dominio continuo y no restrictivo, y el otro en el que se obtiene una solución discreta a partir de la solución obtenida en la primera parte. A diferencia del algoritmo explicado en [69], este método no requiere de un algoritmo de agrupamiento adicional sino que genera por sí mismo una matriz binaria que indica la pertenencia de un elemento a un único grupo. En este estudio se emplean transformaciones ortonormales y descomposición en valores singulares (*SVD*) para encontrar la solución discreta óptima, considerando el principio de la invariancia ortonormal y el criterio de la solución propia óptima. Otro estudio [71], reúne elementos de [70] y [69] para explicar teóricamente la relación entre el *kernel K-medias* y el agrupamiento espectral, partiendo de la generalización de la función objetivo del *K-medias* hasta obtener un caso especial de la función objetivo del agrupamiento que sea aplicable al método de particiones normalizadas (*NC: Normalized cuts*). Para esto se asume una matriz positiva definida de similitud y se aplican algoritmos basados en vectores propios, y diversos esquemas de búsqueda local y de optimización para mejorar los resultados del *kernel*, de forma tal que se garantice que la función objetivo del método *NC* sea monótona decreciente.

En [72] se presenta otra forma de plantear y resolver el problema de la discretización de la solución continua. Este trabajo reúne elementos que permiten explicar la diferencia entre los métodos basados en grafos asociados a laplacianos normalizados y sin normalizar, y da una explicación de por qué los vectores propios del laplaciano pueden

contener información de interés para realizar una agrupación correcta. Una de las ventajas de este método es que permite realizar una interpretación directa de la calidad del agrupamiento a partir de un grafo aleatorio, demostrando que el agrupamiento espectral puede ser visto como un agrupamiento de máxima verosimilitud bajo el supuesto de que los datos corresponden a un grafo ponderado con afinidad aleatoria.

En el estudio realizado en [73], se discuten y se proponen alternativas para resolver problemas abiertos del agrupamiento espectral como la selección de la escala de análisis adecuada, manejo de datos multi-escala, la conformación de grupos irregulares y de conocimiento escaso, y la estimación automática del número de grupos. Para esto, los autores proponen un escalamiento local para calcular la afinidad entre cada par de puntos. Este escalamiento mejora el valor y tiempo de convergencia del algoritmo de agrupamiento. Además, se sugiere aprovechar la información contenida en los vectores propios para deducir automáticamente el número de grupos. Al final, el algoritmo genera una buena inicialización para un algoritmo particional como el *K-medias*.

Parte II

Marco teórico

5. Agrupamiento particional

5.1 Introducción

El análisis no supervisado abarca todos los métodos denominados discriminativos, en los que no se requiere de un conocimiento a priori de las clases para la clasificación, por lo general sólo requieren de algún parámetro de inicialización como la cantidad de grupos resultantes o algún otro indicio acerca de la partición inicial. Entonces, la tarea del análisis no supervisado es agrupar patrones homogéneos sin ninguna información acerca de la naturaleza de las clases presentes en el conjunto de datos. Por esta razón, el análisis no supervisado no genera una clasificación automática, sino que genera un subconjunto de datos homogéneos a partir de algún criterio basado en distancias, disimilitudes o medidas estadísticas. De ahí, que el término de clasificación no supervisada se refiere al agrupamiento de los datos en subconjuntos de elementos similares y no algún tipo de clasificación automática. Existen diversas razones por las que los métodos no supervisados son de interés: convergen rápidamente y con buen desempeño en caso de que las características cambien poco en el tiempo, permiten categorizar elementos, son útiles cuando el etiquetado de un conjunto grande de muestras no es factible, entre otras. Sin embargo, la solución generada por un sistema de análisis no supervisado puede verse afectada por factores, como parámetros iniciales no adecuados, que pueden generar una mala convergencia.

El agrupamiento no supervisado o *clustering*, ha mostrado ser útil en el análisis exploratorio de los datos y se han desarrollado diferentes métodos de agrupamiento que atienden a diversos problemas como costo computacional, sensibilidad a la inicialización, clases desbalanceadas, convergencia a un óptimo local, entre otros. Sin embargo, la selección de un método no es una tarea trivial, es necesario considerar la naturaleza de los datos y las condiciones del problema con el fin de agrupar patrones similares, de tal forma que se tenga un buen compromiso entre costo computacional y efectividad en la separabilidad de las clases. Inclusive, en muchas ocasiones resulta

conveniente unir métodos para generar algoritmos híbridos que mejoren la convergencia y reduzcan el costo computacional.

En la literatura se denomina algoritmos particionales de agrupamiento a aquellos en los que se agrupan patrones homogéneos a partir de una división inicial de los datos que se refina, comúnmente, de forma iterativa. La diferencia de un algoritmo particional y otro se da en la medida que se aplique para cuantificar el agrupamiento y la función de actualización de la división o partición.

En este capítulo se estudian los algoritmos representativos del agrupamiento no supervisado de tipo particional: *K-medias* y *H-medias*. También, se desarrolla un modelo iterativo genérico que permite explicar diversos métodos de agrupamiento basados en el principio del algoritmo *H-medias*. En la sección 5.3 se explican algunos algoritmos de inicialización asociados a métodos particionales, y en la sección 5.4 se propone un método de agrupamiento secuencial en donde se analiza el conjunto de datos inicial por segmentos y al final se realiza la unión de grupos bajo un criterio de disimilitud. La notación dada en la sección 5.2 se conservará a lo largo de todo este capítulo.

5.2 Agrupamiento basado en centroides

La técnica clásica del agrupamiento no supervisado, es la basada en centroides, también denominada particional. La idea general del agrupamiento basado en centroides (*CBC*) es minimizar una función objetivo, la cual define cuán buena es la solución del agrupamiento, dicha solución se obtiene iterativamente a través de la actualización de centroides, y la partición resultante por cada iteración corresponde a la asignación de los elementos al subconjunto cuyo centroide sea el más cercano. Las variantes de los algoritmos CBC consisten en cambios de la función objetivo y, por tanto, de la función de actualización de los centroides.

Por ejemplo, en el agrupamiento basado en la mínima suma de cuadrados (MSSC), explicado ampliamente en [3], la función objetivo se puede expresar como:

$$\min_{\rho_k \in \mathbf{P}_k} \sum_{j=1}^k \sum_{\mathbf{x}_l \in \mathbf{C}_l} \|\mathbf{x}_l - \mathbf{q}_j\|^2 \quad (5.1)$$

donde $\|\cdot\|$ denota la norma euclidiana, \mathbf{P}_k denota el conjunto de todas las particiones del

conjunto de datos $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ en el espacio euclídeo R^d , $\mathbf{x}_i = (\mathbf{x}_{1i}, \dots, \mathbf{x}_{di})^\top$ es el vector de atributos de la i -ésima observación, $\mathbf{C} = \{\mathbf{C}_1, \dots, \mathbf{C}_k\}$ representa el conjunto de grupos, $\mathbf{Q} = (\mathbf{q}_1, \dots, \mathbf{q}_k)^\top$ es el conjunto de centroides y \mathbf{q}_j es el centroide correspondiente al grupo j , dado por:

$$\mathbf{q}_j = \frac{1}{n_e(\mathbf{C}_j)} \sum_{l: \mathbf{x}_l \in \mathbf{C}_j} \mathbf{x}_l, \quad j = 1, \dots, k \quad (5.2)$$

donde $n_e(\cdot)$ denota la cantidad de elementos de su grupo argumento y k es el número de grupos.

En adelante, se usará la misma notación.

El objetivo de este método de agrupamiento es encontrar la partición de los datos que minimice la distancia entre los elementos contenidos en cada grupo y su respectivo centroide, es decir, minimizar la varianza intra-clase. Este mismo principio lo aplican los algoritmos básicos de *K-medias* y *H-medias*.

5.2.1 K-medias

En este método se escoge una partición inicial asociada a un conjunto de centroides iniciales, y el movimiento de los mismos, que se da para generar nuevas particiones, se evalúa por cada centroide, por tanto, una vez se haya movido un centroide, se realizan las asignaciones de todos los elementos y se calcula el cambio de la función objetivo. Esta evaluación continua de la función objetivo puede hacer que el algoritmo K-medias presente mejor convergencia que otros algoritmos, dado que se evalúa la actualización de cada centroide de forma independiente, aunque podría representar un costo computacional elevado.

Asumiendo que un elemento \mathbf{x}_i perteneciente a un grupo \mathbf{C}_l en la solución actual es reasignado a algún otro grupo \mathbf{C}_j , la actualización de los centroides puede realizarse aplicando:

$$\mathbf{q}_l \leftarrow \frac{n_l \mathbf{q}_l - \mathbf{x}_i}{n_l - 1}, \quad \mathbf{q}_j \leftarrow \frac{n_j \mathbf{q}_j + \mathbf{x}_i}{n_j + 1} \quad (5.3)$$

donde $n_i = n_e(\mathbf{C}_i)$ y $l \neq j$.

El cambio de la función objetivo generado por cada reasignación es

$$v_{ij} = \frac{n_j}{n_j + 1} \|\mathbf{q}_j - \mathbf{x}_i\|^2 - \frac{n_l}{n_l - 1} \|\mathbf{q}_l - \mathbf{x}_i\|^2, \quad \mathbf{x}_i \in \mathbf{C}_l \quad (5.4)$$

La anterior ecuación aplica en caso de que la función objetivo sea la de MSSC, en el caso general debe tenerse en cuenta la naturaleza de la función objetivo, así:

$$v_{ij} = \frac{n_j}{n_j + 1} f(\mathbf{q}_j, \mathbf{x}_i) - \frac{n_l}{n_l - 1} f(\mathbf{q}_l, \mathbf{x}_i), \quad \mathbf{x}_i \in \mathbf{C}_l \quad (5.5)$$

donde f es el kernel de la función objetivo.

Los cambios de la función objetivo son calculados para todas las posibles reasignaciones y si todos los valores son no negativos ($v_{ij} \geq 0$), el algoritmo converge y la solución lograda corresponde a una partición asociada a un mínimo local de la función objetivo [3]. En el Algoritmo 1 se explica el método iterativo de actualización de las particiones que se emplea en K-medias.

Algoritmo 1 K-medias

1. Inicialización: escoger un valor de k y una partición inicial $\mathbf{C}^{(0)}$ con centroides $\mathbf{Q}^{(0)}$, fijar número máximo de iteraciones N_{iter} , inicializar el contador: $r = 1$

Mientras $r < N_{iter}$

Desde $j = 1$ **hasta** k **hacer**

2. Mover los centroides: $\mathbf{q}_l^{(r)} \leftarrow \frac{n_l \mathbf{q}_l^{(r-1)} - \mathbf{x}_i}{n_l - 1}$, $\mathbf{q}_j^{(r)} \leftarrow \frac{n_j \mathbf{q}_j^{(r-1)} + \mathbf{x}_i}{n_j + 1}$
3. Calcular el cambio de la función objetivo: $v_{ij} = \frac{n_j}{n_j + 1} \|\mathbf{q}_j^{(r)} - \mathbf{x}_i\|^2 - \frac{n_l}{n_l - 1} \|\mathbf{q}_l^{(r)} - \mathbf{x}_i\|^2$,
 $\mathbf{x}_i \in \mathbf{C}_l^{(r)}$

Si $v_{ij} \geq 0$ ($i = 1, \dots, n$ y $j = 1, \dots, k$)

4. El proceso termina y la solución es $\mathbf{C}^{(r)}$

en caso contrario

$r \leftarrow r + 1$

Termina Si

Termina Desde

Termina Mientras

Nótese que en los algoritmos se numeran únicamente los pasos significativos.

5.2.2 H-medias

Entre los diferentes métodos de agrupamiento desarrollados en los últimos años, el algoritmo H-medias es uno de los métodos más conocidos y aceptados en muchas aplicaciones, entre ellas, la clasificación de patrones biomédicos. El algoritmo H-medias es una variante del K-medias, que mejora el costo computacional, calculando los centroides una vez por cada ciclo, es decir, que bajo algún criterio se generan todos los k centroides antes de evaluar el cambio de la función objetivo. La idea general de este método se ilustra en el Algoritmo 2.

Algoritmo 2 H-medias

1. Inicialización: escoger un valor de k y el conjunto de centroides iniciales $\mathbf{Q}^{(0)} = (\mathbf{q}_1^{(0)}, \dots, \mathbf{q}_k^{(0)})^\top$, realizar la asignación inicial $\mathbf{C}^{(0)}$, fijar el número máximo de iteraciones N_{iter} , inicializar el contador: $r = 1$, fijar parámetro de precisión δ

Mientras $r < N_{iter}$

2. Actualizar los centroides: $\mathbf{Q}^{(r)} = \varphi_{\mathbf{Q}}(\mathbf{C}^{(r-1)}, \mathbf{Q}^{(r-1)})$

2. Asignar los elementos: $\mathbf{C}^{(r)} = \varphi_{\mathbf{C}}(\mathbf{X}, \mathbf{Q}^{(r)})$

Si $|\mathbf{d}(\mathbf{q}_j^{(r)}, \mathbf{q}_j^{(r-1)})| < \delta$ ($j = 1, \dots, k$)

3. El proceso termina y la solución es $\mathbf{C}^{(r)}$

en caso contrario

$r \leftarrow r + 1$

Termina Si

Termina Mientras

La función $\mathbf{d}(\cdot, \cdot)$ es una medida de distancia o de disimilitud entre los dos vectores de su argumento, que en este caso se aplica para medir el cambio de los centroides actuales con respecto a los obtenidos en la iteración inmediatamente anterior, de tal forma que cuando el cambio sea menor que cierto umbral δ , el algoritmo converge. Las funciones $\varphi_{\mathbf{Q}}$ y $\varphi_{\mathbf{C}}$ representan, respectivamente, la actualización de los centroides a partir de la partición previa y la actualización de la partición en función de los centroides actuales. En la sección 5.2.4 se extiende la explicación de este método.

5.2.3 Idea general del agrupamiento iterativo

Una forma generalizada de realizar CBC puede obtenerse estudiando la proporción o grado de pertenencia de un elemento a un grupo y la influencia de cada elemento en la actualización de los centroides, como se explica en [58]. El grado de pertenencia

de un elemento a un grupo lo determina una función de membresía, que se denota con $m(\mathbf{q}_j/\mathbf{x}_i)$: grado de pertenencia de \mathbf{x}_i al grupo cuyo centroide es \mathbf{q}_j . El grado de membresía es un valor no negativo y la pertenencia absoluta es 1, por tanto la función m debe satisfacer

$$m(\mathbf{q}_j/\mathbf{x}_i) \geq 0 \quad \text{y} \quad \sum_{j=1}^k m(\mathbf{q}_j/\mathbf{x}_i) = 1.$$

El grado de influencia o peso de cada punto $w(\mathbf{x}_i)$ en el cálculo de los nuevos centroides, es un factor de ponderación de los datos \mathbf{x}_i . Ambas funciones, m y w , están directamente relacionadas con la naturaleza de la función objetivo como se podrá ver en las secciones 5.2.4 a 5.2.8.

La actualización de los centroides se puede escribir como:

$$\mathbf{q}_j = \frac{\sum_{i=1}^n m(\mathbf{q}_j/\mathbf{x}_i)w(\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^n m(\mathbf{q}_j/\mathbf{x}_i)w(\mathbf{x}_i)}, \quad j = 1, \dots, k \quad (5.6)$$

La anterior ecuación es análoga a la expresión usada para el cálculo de un centro de masa, que comúnmente se emplea en geometría: $\mathbf{q} = \sum_i g(\mathbf{r}_i)\mathbf{r}_i / \sum_i g(\mathbf{r}_i)$, donde \mathbf{r}_i es el vector de posición del i -ésimo elemento y $g(\cdot)$ es la función de densidad de masa.

Dado que las funciones de membresía y peso se pueden ajustar a cualquier función objetivo (conservando las restricciones discutidas anteriormente) y que el refinamiento de los centroides se hace de forma iterativa, se puede decir que este método es un modelo iterativo genérico del agrupamiento no supervisado (*GIM*). La heurística de este modelo es la misma del algoritmo H-medias, es decir que se actualizan todos los centroides antes de aplicar el control de convergencia. La aplicación de este modelo en CBC se explica en el Algoritmo 3.

La convergencia de un algoritmo basado en *GIM*, puede evaluarse comparando el valor de la función objetivo obtenido con la partición actual $f(\mathbf{C}^{(r)})$ y el valor obtenido con la partición generada en la iteración inmediatamente anterior $f(\mathbf{C}^{(r-1)})$, a través de criterios de diferencia $|f(\mathbf{C}^{(r)}) - f(\mathbf{C}^{(r-1)})| < \delta$, de cociente $f(\mathbf{C}^{(r)})/f(\mathbf{C}^{(r-1)}) \approx 1$, entre otros.

Algoritmo 3 Modelo iterativo genérico para CBC

1. Inicialización: escoger los centroides iniciales $\mathbf{Q}^0 = (\mathbf{q}_1^{(0)}, \dots, \mathbf{q}_k^{(0)})^\top$, fijar el máximo número de iteraciones N_{iter} , inicializar el contador: $r = 1$
2. Calcular las funciones de membresía $m(\mathbf{q}_j^{(r-1)}/\mathbf{x}_i)$ y peso $w(\mathbf{x}_i)$ para cada punto
3. Actualizar los centroides: $\mathbf{q}_j^{(r)} = \frac{\sum_{i=1}^n m(\mathbf{q}_j^{(r-1)}/\mathbf{x}_i)w(\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^n m(\mathbf{q}_j^{(r-1)}/\mathbf{x}_i)w(\mathbf{x}_i)}$ $j = 1, \dots, k$
4. $r \leftarrow r + 1$ y repetir los pasos 2 y 3 hasta que el algoritmo converja o hasta que $r = N_{iter}$
5. Asignar los elementos: $\mathbf{x}_i \in \mathbf{C}_l^{(r)}$ si $l = \arg \max_j m(\mathbf{q}_j^{(r)}, \mathbf{x}_i)$

A continuación se explican algunos métodos de agrupamiento a partir del modelo iterativo genérico.

5.2.4 H-medias basado en GIM

El objetivo del algoritmo H-medias (*HM*) es encontrar una partición \mathbf{P}_k de \mathbf{X} con k conjuntos disjuntos tal que minimice su función de costo, que corresponde a la varianza intra-clase, es decir:

$$HM(\mathbf{X}, \mathbf{Q}) = \sum_{i=1}^n \min_{j \in \{1, \dots, k\}} \|\mathbf{x}_i - \mathbf{q}_j\|^2 \quad (5.7)$$

donde $\|\cdot\|$ representa la norma euclidiana en el caso de MSSC.

Las funciones de membresía y peso asociadas a cada \mathbf{x}_i son:

$$m_{HM}(\mathbf{q}_j/\mathbf{x}_i) = \begin{cases} 1 & \text{si } l = \arg \min_j \|\mathbf{x}_i - \mathbf{q}_j\|^2 \\ 0 & \text{otro caso} \end{cases} \quad (5.8)$$

y

$$w_{HM}(\mathbf{x}_i) = 1 \quad (5.9)$$

Puede apreciarse que al evaluar las expresiones (5.8) y (5.9) en la función de actualización de centroides mostrada en (5.6), se obtiene la misma ecuación para el cálculo de los centroides que se emplea en MSSC (5.2), con esto se comprueba la generalización del agrupamiento iterativo a través del GIM.

En el caso del H-medias, la función de membresía se denomina *fuerte* porque toma

valores discretos: 0 ó 1. Además, dado que sus pesos son fijos, este algoritmo da igual importancia a cada elemento.

Uno de los mayores problemas del algoritmo básico de H-medias es la sensibilidad a la selección de la partición inicial, lo que puede implicar la convergencia a un mínimo local de la función objetivo distante del mínimo global, si los centroides no son escogidos de manera apropiada. Por tal razón, se han desarrollado diversas alternativas al H-medias empleando GIM, como el *K-medias armónicas*, *H-medias difuso* y el agrupamiento basado en la máxima esperanza Gaussiana.

5.2.5 K-medias armónicas

A diferencia del H-medias, el algoritmo K-medias armónicas (*KHM*) usa la media armónica de la distancia de cada elemento a los centroides [74], en lugar del criterio aplicado en MSSC. Por tanto, la función objetivo es:

$$KHM(\mathbf{X}, \mathbf{Q}) = \sum_{i=1}^n \frac{k}{\sum_{j=1}^k \frac{1}{\|\mathbf{x}_i - \mathbf{q}_j\|^p}} \quad (5.10)$$

donde p ($p \geq 2$) es un parámetro de entrada a ser sintonizado, por ejemplo, usando validación cruzada.

Las funciones de membresía y peso del algoritmo KHM son:

$$m_{KHM}(\mathbf{q}_j / \mathbf{x}_i) = \frac{\|\mathbf{x}_i - \mathbf{q}_j\|^{-p-2}}{\sum_{j=1}^k \|\mathbf{x}_i - \mathbf{q}_j\|^{-p-2}} \quad (5.11)$$

y

$$w_{KHM}(\mathbf{x}_i) = \frac{\sum_{j=1}^k \|\mathbf{x}_i - \mathbf{q}_j\|^{-p-2}}{\left(\sum_{j=1}^k \|\mathbf{x}_i - \mathbf{q}_j\|^{-p} \right)^2} \quad (5.12)$$

La función objetivo de los algoritmos HM y KHM son similares, pero la del KHM tiene una naturaleza *suave* como puede apreciarse en las ecuaciones 5.11 y 5.12.

Una función de membresía suave significa que, no necesariamente, todo \mathbf{x}_i tiene una pertenencia absoluta a un grupo, sino un grado de pertenencia a cada grupo, y las particiones se generan a razón del mayor valor de dicha pertenencia. En este algoritmo, la importancia relativa de cada elemento cambia de acuerdo a la naturaleza de la función objetivo.

5.2.6 H-medias difuso

El algoritmo H-medias difuso (*FHM*), también denominado *c-medias*, es una variación del HM que consiste en usar una función de membresía suave en lugar de pertenencia absoluta, por tanto permite que cada punto tenga un grado de pertenencia a cada grupo [58]. La función objetivo del método FHM se puede escribir como:

$$FHM(\mathbf{X}, \mathbf{Q}) = \sum_{i=1}^n \sum_{j=1}^k u_{ij}^r \|\mathbf{x}_i - \mathbf{q}_j\|^2 \quad (5.13)$$

El factor de ponderación u_{ij} permite hacer difuso el método *HM* básico y debe satisfacer que $\sum_{j=1}^k u_{ij} = 1$ y $u_{ij} > 0$ para que no se afecte la naturaleza de la función objetivo. El parámetro r tiene la restricción $r \geq 1$, dado que si $0 < r < 1$, los aportes de cada factor reducen el acumulado de la función objetivo, lo que puede afectar la convergencia. De igual forma ocurre si $r < 0$ porque puede aumentarse el valor de ponderación de forma considerable. En el caso particular, $r = 1$ y si u_{ij} es de naturaleza fuerte, este método sería igual que el HM. Un valor grande de r hace que el sistema sea más *difuso*.

Dada las condiciones de u_{ij} , este parámetro podría ser la misma función de membresía. La aplicación de FHM usando GIM, se hace a partir de las siguientes funciones de membresía y peso:

$$m_{FHM}(\mathbf{q}_j/\mathbf{x}_i) = \frac{\|\mathbf{x}_i - \mathbf{q}_j\|^{-2/(r-1)}}{\sum_{j=1}^k \|\mathbf{x}_i - \mathbf{q}_j\|^{-2/(r-1)}} \quad (5.14)$$

y

$$w_{FHM}(\mathbf{x}_i) = 1 \quad (5.15)$$

5.2.7 Agrupamiento de máxima esperanza gaussiana

El método de agrupamiento basado en la máxima esperanza gaussiana (*GEMC*), hace parte de los métodos de agrupamiento basados en densidades (*DBC*) y tiene como función objetivo la combinación lineal de distribuciones gaussianas centradas en los centroides de cada grupo, así:

$$GEMC(\mathbf{X}, \mathbf{C}) = - \sum_{i=1}^n \log \left(\sum_{j=1}^k p(\mathbf{x}_i/\mathbf{q}_j)p(\mathbf{q}_j) \right) \quad (5.16)$$

donde $p(\mathbf{x}_i/\mathbf{q}_j)$ es la probabilidad de \mathbf{x}_i dado que es generado por una distribución gaussiana centrada en \mathbf{q}_j y $p(\mathbf{q}_j)$ es la probabilidad a priori del grupo cuyo centroe es \mathbf{q}_j . Se emplea la función logaritmo por facilidad matemática y el signo menos con el fin de que la tarea sea minimizar la función objetivo.

Las funciones correspondientes a la membresía y el peso de cada elemento son, respectivamente,

$$m_{GEMC}(\mathbf{q}_j/\mathbf{x}_i) = \frac{p(\mathbf{x}_i/\mathbf{q}_j)p(\mathbf{q}_j)}{p(\mathbf{x}_i)} \quad (5.17)$$

y

$$w_{GEMC}(\mathbf{x}_i) = 1 \quad (5.18)$$

Nótese que la función de membresía es un valor de probabilidad, por tanto la regla de Bayes puede emplearse para el cálculo de su valor, considerando $p(\mathbf{x}_i)$ como la evidencia:

$$p(\mathbf{x}_i) = \sum_{j=1}^k p(\mathbf{x}_i/\mathbf{q}_j)p(\mathbf{q}_j)$$

El factor $p(\mathbf{x}_i/\mathbf{q}_j)$ puede obtenerse fácilmente con:

$$p(\mathbf{x}_i/\mathbf{q}_j) = f(\mathbf{x}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}_j) = \frac{1}{\det(\boldsymbol{\Sigma}_j)^{\frac{1}{2}}} (2\pi)^{-d/2} e^{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i - \boldsymbol{\mu})^T} \quad (5.19)$$

donde $\boldsymbol{\mu}$ es el centroe ($\boldsymbol{\mu} = \mathbf{q}_j$), d es la dimensión, $\boldsymbol{\Sigma}$ representa la matriz de covarianza y $\det(\cdot)$ denota el determinante de su matriz argumento.

De acuerdo a la regla de Bayes, la matriz $\boldsymbol{\Sigma}_j$ puede ser única ($\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma} = \text{cov}(\mathbf{X})$) o puede calcularse para cada grupo ($\boldsymbol{\Sigma}_j = \text{cov}(\mathbf{C}_j)$, $j = 1, \dots, k$). La segunda opción

es recomendable porque considera la varianza de cada uno de los grupos, además, la matriz Σ_j podría calcularse y refinarse por cada iteración, aunque esto podría aumentar el costo computacional.

Una variante del método *GEMC* puede obtenerse calculando la probabilidad a posteriori $p(\mathbf{x}_i/\mathbf{q}_j)$ de forma no paramétrica, empleando el método de Parzen, que consiste en la superposición de distribuciones gaussianas de un tamaño fijo h centradas en cada \mathbf{x}_i [75]. A este método se le denomina DBC no paramétrico o NPDBC. El valor óptimo de h se puede obtener con validación cruzada.

Matemáticamente, la distribución de probabilidad empleando el método de Parzen es

$$p(\mathbf{x}) = \frac{1}{nh} \sum_{i=1}^n \mathcal{K} \left(\frac{\mathbf{x} - \mathbf{x}_i}{h} \right) \quad (5.20)$$

donde \mathcal{K} es un kernel gaussiano definido como:

$$\mathcal{K}(\mathbf{z}) = \frac{1}{(2\pi)^{-d/2}} \exp^{-\frac{1}{2}\mathbf{z}\mathbf{z}^\top} \quad (5.21)$$

5.2.8 Algoritmos híbridos

Una forma de aprovechar las virtudes de dos métodos a la vez, es conformar un par membresía - peso híbrido, es decir, que se usa la función de membresía correspondiente a un método y la función de pesos correspondiente a otro método.

Por ejemplo, con membresía fuerte y pesos variables:

$$m_{H_1}(\mathbf{q}_j/\mathbf{x}_i) = \begin{cases} 1 & \text{si } l = \arg \min_j \|\mathbf{x}_i - \mathbf{q}_j\|^2 \\ 0 & \text{otro caso} \end{cases} \quad (5.22)$$

y

$$w_{H_1}(\mathbf{x}_i) = \frac{\sum_{j=1}^k \|\mathbf{x}_i - \mathbf{q}_j\|^{-p-2}}{\left(\sum_{j=1}^k \|\mathbf{x}_i - \mathbf{q}_j\|^{-p} \right)^2} \quad (5.23)$$

Otra alternativa se puede plantear con membresía suave y pesos fijos:

$$m_{H_2}(\mathbf{q}_j/\mathbf{x}_i) = \frac{\|\mathbf{x}_i - \mathbf{q}_j\|^{-p-2}}{\sum_{j=1}^k \|\mathbf{x}_i - \mathbf{q}_j\|^{-p-2}} \quad (5.24)$$

y

$$w_{H_2}(\mathbf{x}_i) = 1 \quad (5.25)$$

El híbrido H_1 puede presentar mejor desempeño que el algoritmo HM básico debido a la naturaleza suave de los pesos. Entre tanto, puede ser que el híbrido H_2 , en algunos casos, tenga mejor tiempo de convergencia que KHM y obtenga igual desempeño.

5.3 Criterios de inicialización

Los algoritmos de CBC son sensibles a la inicialización, es decir, que si la partición inicial no es adecuada, el algoritmo puede converger a un mínimo local distante del mínimo global. Una forma de solucionar este problema es garantizar una buena inicialización de los centroides, para esta tarea existen herramientas eficaces como el algoritmo *máx-mín* y *J-medias*.

5.3.1 Criterio máx-mín

El algoritmo máx-mín [59], busca en el conjunto \mathbf{X} , los k elementos más alejados con un criterio de distancia preestablecido, mejorando la cantidad de grupos necesarios para separar las clases y el valor de convergencia. Este algoritmo inicia escogiendo un punto arbitrario de \mathbf{X} como el primer centroide y el resto se escogen siguiendo una estrategia en la que el elemento elegido en la i -ésima iteración es aquél cuyo elemento más cercano entre los $i - 1$ ya elegidos, se encuentra más alejado [59]. Para esto, se selecciona aleatoriamente el primer centroide \mathbf{q}_1 del conjunto \mathbf{X} , y el segundo centroide \mathbf{q}_2 se obtiene como el elemento que presenta la máxima distancia entre \mathbf{q}_1 y los puntos restantes $\{\mathbf{X} - \mathbf{q}_1\}$.

A partir de estos dos centroides se encuentran los centroides restantes aplicando el criterio máx -mín, así:

$$f(\mathbf{x}_l) = \max_{\mathbf{x}_i \in \{\mathbf{X} - \mathbf{Q}\}} \left\{ \min_{\mathbf{q}_j \in \mathbf{Q}} \|\mathbf{x}_i - \mathbf{q}_j\|^2 \right\}, \quad j = 1, \dots, k \quad (5.26)$$

donde $\|\cdot\|$ representa la norma euclidiana.

5.3.2 Criterio J-medias

El algoritmo J-medias consiste en la actualización de los centroides a través de una evaluación local de la función objetivo, es decir, considerando una determinada región alrededor de los centroides en lugar de todo el espacio [3]. Los puntos por fuera de estas regiones son candidatos a ser la ubicación del nuevo centroide, entonces, aleatoriamente se mueve el centroide anterior a un punto disponible y se compara el valor de la función objetivo obtenido de evaluar el centroide anterior con el que se obtiene de evaluar el centroide actual. Por último, se acepta o se rechaza un centroide, considerando que la función objetivo disminuya. En el J-medias tradicional se emplean las condiciones de *MSSC*, es decir, medidas de distancia, por tanto las regiones se establecen con esferas centradas en cada \mathbf{q}_j cuyos radios definen el tamaño de la regiones. Dichos radios deben ser menores que la mitad de la menor distancia entre los centroides ($\varepsilon < \frac{1}{2} \min \|\mathbf{q}_j - \mathbf{q}_i\|$, $i \neq j$), para que no existan intersecciones entre las esferas. El criterio J-medias se describe de forma resumida en el Algoritmo 4.

Algoritmo 4 Criterio J-medias

1. Inicialización: escoger la partición inicial $C^0 = \{C_j^0\}_{j=1}^k$ asociada a Q^0 , $C^1 \leftarrow C^0$.
2. Encontrar los puntos que no corresponden a centroides, elementos por fuera de la esferas de radio ε ($\varepsilon < \frac{1}{2} \min \|\mathbf{q}_j - \mathbf{q}_i\|$ $i \neq j$) centradas en \mathbf{q}_j ($j = 1, \dots, k$).
3. Encontrar la mejor partición C_k^2 y calcular su correspondiente valor de función objetivo f^2 .

Si $f^1 > f^2$

4. El algoritmo converge con la solución Q^1

en caso contrario

5. Cambiar de partición inicial: $C^1 \leftarrow C^2$, $f^1 \leftarrow f^2$ y volver al paso 2.

Termina Si

Si se realizan las asignaciones de los elementos a los centroides Q^1 , se tendría un proceso completo de agrupamiento. Aquí, el J-medias se expresa como un algoritmo de inicialización, bajo el principio de que una unión apropiada de algoritmos de agrupamiento mejora el desempeño individual.

Las variantes de este algoritmo pueden darse en la definición de las regiones, por ejemplo, si en lugar de distancias se empleara momentos estadísticos como la co-

varianza. Así, la función objetivo sería la que se aplica en GEMC (ver (5.16)) y las esferas se definirían con matrices de covarianza. Con esta variante, el método se denomina *J-GEM*. Se denominaría *J-Parzen* en caso de usar estimación no paramétrica de las probabilidades, y *J-H-medias* en el caso de evaluar la función objetivo después de calcular todos los centroides con criterios de MSSC.

5.4 Análisis por segmentos

Además de emplear una buena partición inicial para mejorar el tiempo de procesado y valor de convergencia, se puede hacer un análisis por segmentos del conjunto de datos. La forma más intuitiva de realizar un análisis por segmentos consiste en partir el conjunto de datos en N_s subconjuntos (que se denominarán segmentos) y luego aplicar un esquema de análisis no supervisado a cada subconjunto. El conjunto de datos segmentado es $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_{N_s}\}$, donde \mathbf{X}_l es una matriz de $n_l \times d$, $n_l = \text{aprox}(n/N_s)$ es la cantidad de elementos de cada segmento y $\text{aprox}(\cdot)$ representa el entero más aproximado al número de su argumento. Emplear directamente los grupos generados en cada segmento no representa una contribución significativa al procesado de datos y distorsiona el objetivo de la clasificación porque podría generarse una gran cantidad de grupos. Por tanto, es necesario incluir una etapa de unión de grupos. Existen dos formas básicas para esta tarea. La primera consiste en obtener los grupos de todos los segmentos y realizar la unión al final. En la segunda forma, se conforman los grupos de los dos primeros segmentos y se realiza el proceso de unión de ellos, luego se hace el agrupamiento en el siguiente segmento (tercer subconjunto) y al final se unen los grupos del segmento actual con los generados en la unión previa. Este procedimiento se aplica hasta unir los N_s segmentos. Esta forma de unión es ventajosa con respecto a la primera porque representa un acercamiento al análisis secuencial y a las aplicaciones en tiempo real. Por ejemplo, en el análisis de señales ECG, si se adquiere por tramos la señal se podría caracterizar y agrupar los latidos por cada tramo, reduciendo el costo computacional y mejorando el desempeño del sistema de análisis, debido a que agrupar por segmentos es menos sensible a las clases minoritarias que agrupar todo el conjunto de datos. Además, reduce el costo computacional porque, en la mayoría de los casos, la suma de los tiempos de procesado de cada segmento es considerablemente menor que el tiempo que requiere el análisis del conjunto de datos en una iteración, como se verá en los capítulos 7 y 9.

Para aplicar una metodología por segmentos es necesario establecer el tamaño n_l mínimo de los segmentos y un criterio de unión. El parámetro n_l se puede fijar considerando que el número de observaciones debe ser por lo menos 2 veces mayor que el número de características $n_l \geq 2d$, para que se obtengan buenos resultados de clasificación. Adicionalmente, un criterio sencillo para la unión de grupos consiste en estimar una medida de disimilitud o distancia ϑ entre los centroides de cada grupo, de forma que dos grupos se unan si el valor de ϑ es admisible. Intuitivamente, podría pensarse que la forma de hacer la unión de grupos es unir un grupo de un segmento con uno de otro segmento, es decir, que un grupo de un segmento, necesariamente, debe unirse a otro. Esto es de implementación sencilla, sin embargo podría generar resultados bajos en tareas de agrupamiento de conjuntos de datos con clases desbalanceadas, además, los segmentos podrían resultar no homogéneos a razón de la cantidad de clases contenidas en cada uno. Por consiguiente, se deben incluir criterios de exclusión para las clases minoritarias, lo que también resulta útil en la detección de valores atípicos. Este criterio se puede plantear así: si el valor de la función ϑ aplicada a un centroide con respecto a los demás, no es, en ningún caso, un valor admisible, es decir, es mayor que un valor preestablecido ϵ , el grupo asociado a dicho centroide no se une a otro y se mantiene como un grupo independiente en la iteración actual. El parámetro ϵ se puede determinar con un análisis previo sobre los posibles centroides de cada grupo, por ejemplo, empleando disimilitudes como se plantea en [76].

En resumen, la unión de grupos se aplica como sigue. Sea $\mathbf{P}^l = \{\mathbf{C}_1^l, \dots, \mathbf{C}_{k^l}^l\}$ la partición del segmento l , k^l el número de grupos de la misma partición (ver estimación del número de grupos en la sección 6.5), $\mathbf{Q}^l = \{\mathbf{q}_1^l, \dots, \mathbf{q}_{k^l}^l\}$ los centroides de cada grupo, entonces la unión se hace a razón de

$$\vartheta(j^l, j^{l-1}) = \vartheta(\mathbf{q}_{j^l}^l, \mathbf{q}_{j^{l-1}}^{l-1}) \quad (5.27)$$

Con esto, si $\vartheta(j^l, j^{l-1})$ es un valor admisible ($\vartheta(j^l, j^{l-1}) < \epsilon$), el grupo $\mathbf{C}_{j^l}^l$ se une con el grupo $\mathbf{C}_{j^{l-1}}^{l-1}$, de lo contrario se pasa a comparar con los demás grupos. En caso de que un grupo no presente disimilitud admisible con ningún otro, dicho grupo se mantiene como independiente para el análisis del siguiente segmento, con esto se evita la mala agrupación de clases minoritarias y se reduce el costo computacional.

En el Algoritmo 5 se explica, de forma condensada, el procedimiento del análisis por segmentos de un conjunto de datos.

Algoritmo 5 Agrupamiento por segmentos

1. Dividir el conjunto de datos: $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_{N_s}\}$, donde \mathbf{X}_l es una matriz de $n_l \times d$, $n_l = \text{aprox}(n/N_s)$

2. Agrupar los datos de cada segmento en k^l grupos: $\mathbf{P}^l = \{\mathbf{C}_1^l, \dots, \mathbf{C}_{k^l}^l\}$ asociados a los centroides $\mathbf{Q}^l = \{\mathbf{q}_1^l, \dots, \mathbf{q}_{k^l}^l\}$

Desde $l = 2$ **hasta** m **hacer**

3. Comparar los centroides de cada segmento: $\vartheta(j^l, j^{l-1}) = \vartheta(\mathbf{q}_{j^l}^l, \mathbf{q}_{j^{l-1}}^{l-1})$,
 $j^l = 1, \dots, k^l, j^{l-1} = 1, \dots, k^{l-1}$

Si $\vartheta(j^l, j^{l-1}) < \epsilon$

4. El grupo $\mathbf{C}_{j^l}^l$ se une con el grupo $\mathbf{C}_{j^{l-1}}^{l-1}$

en caso contrario

El grupo $\mathbf{C}_{j^l}^l$ se conserva como independiente para el análisis con el siguiente segmento

Termina Si

Termina Desde

6. Agrupamiento espectral

6.1 Introducción

En el área de reconocimiento y clasificación de patrones, los métodos de agrupamiento basados en grafos y análisis espectral son relativamente nuevos, aunque ya han sido aplicados exitosamente en diversos contextos como el diseño de circuitos [65], balance eficiente de carga computacional orientado a programación en paralelo [66], segmentación de imágenes [67, 68], entre otros. Estos métodos son discriminativos, por tanto, no requieren información a priori (suposiciones sobre la estructura de los datos), y las particiones son generadas empleando información obtenida de la misma heurística del método, en este caso, un criterio de decisión global que considera el valor estimado de la probabilidad de que dos elementos pertenezcan a un mismo grupo [70]. Por esta razón, este tipo de análisis se puede explicar fácilmente a partir de la teoría de grafos (como se verá en la sección 6.2), donde dicha probabilidad puede asociarse a una medida de *afinidad* entre los nodos. Comúnmente, este criterio se aplica en un espacio de representación de dimensión menor que la del conjunto de datos, esto implica una reducción de la dimensionalidad en la representación, conservando tan fiel como sea posible las relaciones entre los elementos, y por lo tanto converge en un análisis de vectores y valores propios.

Los métodos de análisis espectral son de gran interés en tareas de clasificación, debido a que la estimación de los óptimos globales en un dominio continuo no restrictivo se obtienen de la descomposición en valores y vectores propios (descomposición propia) [70], basándose en el teorema de Perron - Frobenius en el que se establece que los mayores valores propios, estrictamente reales, de una matriz positiva definida e irreducible definen el radio espectral de la misma [77]. En otras palabras, el espacio generado por los vectores propios está directamente relacionado con la calidad del agrupamiento.

Considerando este principio y la posibilidad de obtener una solución discreta a

través de los vectores propios, surgió un enfoque, relativamente nuevo, del agrupamiento no supervisado, denominado agrupamiento espectral. Sin embargo, obtener dicha solución discreta implica resolver otro problema del agrupamiento, aunque en una dimensión menor a la original [70]. En otras palabras, los vectores propios podrían ser considerados como un nuevo conjunto de datos que pueden ser agrupados empleando algún algoritmo de agrupamiento como *K-medias* [68] (ver sección 5.2.1). Así, se tendría un algoritmo híbrido en donde los algoritmos de análisis espectral generarían los vectores propios y los parámetros de inicialización, y los otros algoritmos particionales se encargarían del agrupamiento en sí, es decir, que el análisis espectral puede apoyar a las técnicas clásicas generando una inicialización que mejore el valor de convergencia. No obstante, de modo similar que los valores fuertes de membresía que se obtienen de algunos algoritmos de agrupamiento basado en centroides (explicados en el capítulo 5), un esquema de agrupamiento espectral genera una matriz binaria que indica la pertenencia de un elemento a un grupo, lo que le permite realizar el proceso de agrupamiento por sí mismo. Una estrategia como ésta no sólo necesita de los vectores propios sino también de los datos obtenidos a través de una transformación ortonormal de la solución generada por la descomposición en vectores propios (solución propia). Esto conduce a que el objetivo del agrupamiento espectral es encontrar la mejor transformación ortonormal que genere una discretización apropiada de la solución continua [68].

En el desarrollo de este capítulo, se explica detalladamente un método de agrupamiento no supervisado denominado agrupamiento espectral normalizado. Para tal propósito, se comienza por mencionar algunos conceptos preliminares sobre teoría de grafos, descritos en la sección 6.2. Luego, en la sección 6.3 se desarrolla el método de agrupamiento, deduciendo un criterio de particionado multi-clase a partir de un análisis topológico (sección 6.3.1) y por último, en la sección 6.3.3, se explica el algoritmo de agrupamiento. Adicionalmente, se incluyen dos secciones, 6.4 y 6.5, en las que se establecen algunas medidas de *afinidad* orientadas al agrupamiento y se describen métodos para la estimación del número de grupos, respectivamente.

6.2 Preliminares sobre teoría de grafos

Una forma sencilla de interpretar los métodos de análisis espectral es a través de una perspectiva geométrica empleando bloques topológicos o grafos. A continuación se

presentan algunas definiciones en torno a teoría de grafos, así como operaciones y propiedades útiles para el análisis posterior.

Los grafos son conjuntos, no vacíos, de objetos denominados nodos o vértices y de líneas denominadas aristas que unen por pares los vértices. Las aristas indican la relación entre los vértices, de forma que si dos vértices están relacionados se traza una línea entre ellos, en caso contrario no debe existir ningún trazo.

En general, un grafo propiamente dicho puede ser descrito por el par ordenado $\mathbb{G} = (\mathbb{V}, \mathbb{E})$, donde \mathbb{V} es el conjunto de vértices y \mathbb{E} es el conjunto de aristas. Por ejemplo, en los grafos mostrados en la Figura 6.1, $\mathbb{V} = \{a, b, c\}$ y $\mathbb{E} = \{ab, bc, ac\}$ para el caso 6.1(a), y para el caso 6.1(b) se tiene $\mathbb{V} = \{a, b, c\}$ y $\mathbb{E} = \{ab, cb\}$. Un grafo propiamente dicho, también denominado no dirigido, es aquél en el que no interesa o no está definido el sentido de la relación, por tanto si $a, b \in \mathbb{V}$ entonces las aristas ab y ba representan el mismo elemento de \mathbb{E} . Los grafos que contienen aristas con orientación definida, se denominan grafos dirigidos (ver 6.1(b)), en este caso, la denominación de las aristas tiene implicación en el sentido de la relación, es decir, $ab \neq ba$. Los grafos, de acuerdo a la trayectoria que forman las relaciones de los vértices, también pueden clasificarse en conexos y no conexos: si dicha trayectoria es cerrada con respecto a un vértice, el grafo se denomina conexo, en caso contrario se denomina no conexo.

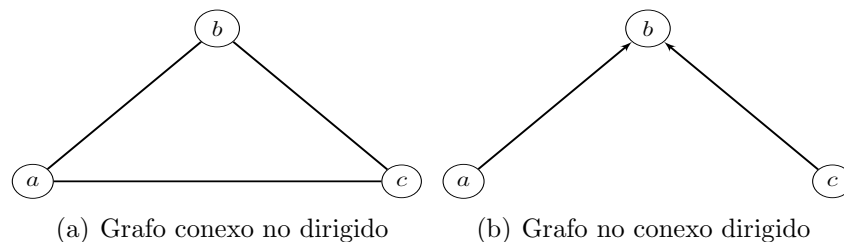


Figura 6.1: Ejemplos de grafos

Una forma generalizada de representar un grafo corresponde a los grafos ponderados en los que, además del orden y el sentido (en el caso de los grafos dirigidos), importa el grado de relación o afinidad entre los nodos, definido a través de una matriz de pesos \mathbf{W} , donde w_{ij} es un valor no negativo que representa el peso de la arista ubicada entre los nodos i y j , como se muestra en la Figura 6.2. Además, en un grafo no dirigido, es evidente que $w_{ij} = w_{ji}$. Por tanto, un grafo ponderado puede ser descrito de la forma $\mathbb{G} = (\mathbb{V}, \mathbb{E}, \mathbf{W})$, donde \mathbf{W} es una matriz simétrica y semidefinida positiva.

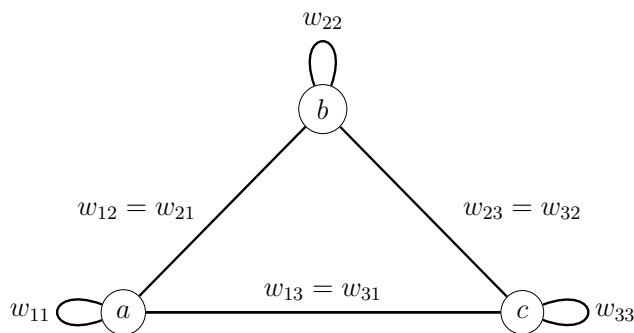


Figura 6.2: Grafo ponderado

6.2.1 Medidas sobre los grafos

Existen diversas medidas, aplicables sobre los grafos, que resultan útiles para clasificarlos y realizar operaciones entre ellos. En este trabajo se tienen en cuenta dos medidas fundamentales, que serán de uso posterior: peso total de un subconjunto de vértices con respecto a otro y grado de un subconjunto.

Dado el grafo $G = (\mathbb{V}, \mathbb{E}, \mathbf{W})$ y los elementos $\mathbb{A}, \mathbb{B} \subset \mathbb{V}$, el peso total de las conexiones entre \mathbb{A} y \mathbb{B} , se puede calcular con

$$\text{lazos}(\mathbb{A}, \mathbb{B}) = \sum_{i \in \mathbb{A}, j \in \mathbb{B}} w_{ij} \quad (6.1)$$

El grado de un grafo no ponderado con respecto a un vértice equivale a la cantidad de aristas adyacentes al mismo. En los grafos ponderados, que es el caso de interés, el grado representa el peso total de las conexiones de un subconjunto con respecto al conjunto de todos los vértices, por tanto el grado de \mathbb{A} podría ser calculado empleando:

$$\text{grado}(\mathbb{A}) = \text{lazos}(\mathbb{A}, \mathbb{V}) \quad (6.2)$$

El grado es comúnmente empleado en la normalización de los pesos de las conexiones. Dicha normalización permite estimar la proporción de un subconjunto de conexiones con respecto a otro, por ejemplo, la proporción de conexiones ponderadas entre \mathbb{A} y \mathbb{B} con respecto al total de conexiones de \mathbb{A} es:

$$\text{lazosnor}(\mathbb{A}, \mathbb{B}) = \frac{\text{lazos}(\mathbb{A}, \mathbb{B})}{\text{grado}(\mathbb{A})} \quad (6.3)$$

6.3 Agrupamiento espectral normalizado

El análisis espectral orientado al agrupamiento de datos se basa en grafos ponderados no dirigidos y conexos $\mathbb{G} = (\mathbb{V}, \mathbb{E}, \mathbf{W})$, en donde $\mathbb{V} = \{1, \dots, n\}$ representa los índices del conjunto de datos a ser agrupado y \mathbf{W} es la matriz de afinidad que, de acuerdo a lo discutido en la sección 6.2.1, se asume semidefinida positiva y simétrica. El objetivo del agrupamiento espectral es descomponer \mathbb{V} en k subconjuntos disjuntos, por tanto, $\mathbb{V} = \cup_{l=1}^k \mathbb{V}_l$ y $\mathbb{V}_l \cap \mathbb{V}_m = \emptyset, \forall l \neq m$, dicha descomposición se realiza empleando información espectral y transformaciones ortonormales, comúnmente.

En el agrupamiento espectral, dos medidas son de particular interés: $\text{lazosnor}(\mathbb{A}, \mathbb{A})$, que define la proporción de conexiones de \mathbb{A} agrupadas correctamente, en otras palabras, indica el grado de conexión intra particiones, y $\text{lazosnor}(\mathbb{A}, \mathbb{A} \setminus \mathbb{V})$, es una medida complementaria a la primera e indica el grado conexión entre particiones. Entonces, un buen agrupamiento se obtiene maximizando el grado de asociación o conexión intra particiones y minimizando el grado de desviación o conexión entre particiones. Estos objetivos se pueden lograr aplicando un criterio de particionado multi-clase que, por lo general, implica medir el total de las conexiones intra y entre particiones. Estas medidas pueden calcularse a través del total de asociaciones normalizadas (kasocn) y el total de desviaciones normalizadas (kdesvn), que corresponden, respectivamente, a:

$$\text{kasocn}(\Gamma_{\mathbb{V}}^k) = \frac{1}{k} \sum_{l=1}^k \text{lazosnor}(\mathbb{V}_l, \mathbb{V}_l) \quad (6.4)$$

y

$$\text{kdesvn}(\Gamma_{\mathbb{V}}^k) = \frac{1}{k} \sum_{l=1}^k \text{lazosnor}(\mathbb{V}_l, \mathbb{V}_l \setminus \mathbb{V}) \quad (6.5)$$

donde $\Gamma_{\mathbb{V}}^k = \{\mathbb{V}_1, \dots, \mathbb{V}_k\}$ es el conjunto de todas las particiones.

Debido a la normalización con respecto al grado, aplicada sobre las medidas de asociación y desviación, se puede verificar que:

$$\text{kasocn}(\Gamma_{\mathbb{V}}^k) + \text{kdesvn}(\Gamma_{\mathbb{V}}^k) = 1 \quad (6.6)$$

6.3.1 Criterio de particionado multi-clase

De la ecuación (6.6) puede concluirse que la maximización de la medida de asociaciones y la minimización de la medida de desviaciones ocurren simultáneamente [70], por tanto no es necesario distinguir entre estas medidas. Con esto, la función objetivo a maximizar podría plantearse como:

$$\varepsilon(\Gamma_{\mathbb{V}}^k) = \text{kasocn}(\Gamma_{\mathbb{V}}^k) \quad (6.7)$$

Entre los diversos criterios que existen para la optimización de las asociaciones, únicamente los criterios de desviaciones mínimas [67] y normalizadas [68] tienen esta propiedad de dualidad. Sin embargo, se prefiere el criterio que emplea desviaciones normalizadas porque es menos sensible al ruido y a la inicialización de la matriz de pesos que el método de desviaciones mínimas.

6.3.2 Representación matricial

En adelante, el conjunto $\Gamma_{\mathbb{V}}^k$ se representará con la matriz de particiones $\mathbf{M} = [\mathbf{M}_1, \dots, \mathbf{M}_k]$ que indica la pertenencia de un elemento a un grupo, de manera que:

$$m_{il} = \langle i \in \mathbb{V}_l \rangle, \quad i \in \mathbb{V}, \quad l = 1, \dots, k \quad (6.8)$$

donde m_{il} es el elemento il de la matriz \mathbf{M} , $\langle \cdot \rangle$ es un indicador binario: toma el valor de 1 si su argumento es verdadero y 0 en caso contrario.

Dado que un vértice es asignado a una única partición, se debe garantizar que $\mathbf{M}\mathbf{1}_k = \mathbf{1}_n$, donde $\mathbf{1}_d$ es un vector d -dimensional en el que todos sus valores son iguales a 1.

Sea la matriz \mathbf{D} el grado de la matriz de pesos definido como:

$$\mathbf{D} = \text{Diag}(\mathbf{W}\mathbf{1}_n) \quad (6.9)$$

donde $\text{Diag}(\cdot)$ denota a una matriz diagonal formada por el vector argumento, entonces las medidas representadas en las ecuaciones (6.1) y (6.3), pueden replantearse como sigue:

$$\text{lazos}(\mathbb{V}_l, \mathbb{V}_l) = \mathbf{M}_l^T \mathbf{W} \mathbf{M}_l \quad (6.10)$$

$$\text{grado}(\mathbb{V}_l) = \mathbf{M}_l^T \mathbf{D} \mathbf{M}_l \quad (6.11)$$

Con lo anterior, el criterio de particionado multi-clase puede expresarse como:

$$\max \varepsilon(M) = \frac{1}{k} \sum_{l=1}^k \frac{\mathbf{M}_l^T \mathbf{W} \mathbf{M}_l}{\mathbf{M}_l^T \mathbf{D} \mathbf{M}_l} \quad (6.12)$$

$$\text{s. t. } \mathbf{M} \in \{0, 1\}^{n \times k}, \quad \mathbf{M} \mathbf{1}_k = \mathbf{1}_n \quad (6.13)$$

Esta formulación se denotará con *NCPM* (*Normalized cuts problem*), la letra “M” se usa para denotar que el problema de optimización está planteado en términos de la matriz \mathbf{M} .

6.3.3 Algoritmo de agrupamiento

El algoritmo de agrupamiento espectral que se estudia en este trabajo, se basa en el método desarrollado en [70] y consiste en resolver el problema de optimización planteado en la ecuación (6.12), en un dominio continuo y sin restricciones, aplicando descomposición en vectores propios, transformaciones ortonormales y un proceso de discretización de la solución, partiendo del hecho de que los vectores propios generan todas las soluciones más aproximadas a las soluciones óptimas continuas. A este algoritmo se le denomina método de desviaciones normalizadas. Este método es robusto a la inicialización y converge más rápido que las técnicas clásicas, sin embargo podría generar un costo computacional más elevado por cada iteración, debido al cálculo de los valores singulares. Las transformaciones ortonormales se aplican sobre la descomposición propia para generar una familia completa de soluciones óptimas globales, que posteriormente se normalizan en longitud para hacer que cada óptimo encontrado corresponda a una partición solución en el dominio continuo. A dicha solución se aplica un proceso iterativo de discretización que es llevado a cabo empleando descomposición en valores singulares (SVD) y seleccionando vectores propios de acuerdo a los valores propios máximos y la cantidad de grupos establecida. El diagrama mostrado en la Figura 6.3 explica de forma gráfica la idea general del algoritmo de agrupamiento.

Idea general del algoritmo

El primer paso en este método es establecer el número de grupos y la matriz de afinidad \mathbf{W} , en las secciones 6.5 y 6.4 se muestran algunas alternativas para estimar el número de

de \mathbf{Z}^* . Para esto, se define \mathbf{Z} como la matriz escalada de particiones [78], que corresponde a un escalamiento de \mathbf{M} de la forma

$$\mathbf{Z} = \mathbf{M}(\mathbf{M}^\top \mathbf{D} \mathbf{M})^{-\frac{1}{2}} \quad (6.14)$$

Dado que $\mathbf{M}^\top \mathbf{D} \mathbf{M}$ es una matriz diagonal, las columnas de \mathbf{Z} corresponden a las columnas de \mathbf{M} escaladas por el inverso de la raíz cuadrado del grado de \mathbf{W} . Con lo anterior, puede establecerse que la condición que deben cumplir las soluciones factibles del problema de optimización es $\mathbf{Z}^\top \mathbf{D} \mathbf{Z} = \mathbf{I}_k$, donde \mathbf{I}_k representa una matriz identidad de $k \times k$. Omitiendo las restricciones iniciales, se puede plantear un nuevo problema de optimización empleando la matriz \mathbf{Z} , así:

$$\max \varepsilon(\mathbf{Z}) = \frac{1}{k} \text{tr}(\mathbf{Z}^\top \mathbf{W} \mathbf{Z}) \quad (6.15)$$

$$\text{s. a. } \mathbf{Z}^\top \mathbf{D} \mathbf{Z} = \mathbf{I}_k \quad (6.16)$$

El hecho de emplear la variable \mathbf{Z} en un dominio continuo y omitir las restricciones iniciales, hace que el problema discreto de optimización se convierta en un problema continuo relativamente sencillo. Resolver este problema implica tener en cuenta la propiedad de invariancia ortonormal y el criterio de la solución óptima basada en vectores propios (solución propia). En adelante, el problema de optimización continuo del método de desviaciones normalizadas se denotará con *NCPZ*.

Proposición 6.1. (Invariancia ortonormal) *Sea \mathbf{R} una matriz de $k \times k$ y \mathbf{Z} una posible solución a NCP, entonces $\mathbf{Z} \mathbf{R} = \mathbf{R}^\top \mathbf{R} = \mathbf{I}_k$ y $\varepsilon(\mathbf{Z} \mathbf{R}) = \varepsilon(\mathbf{Z})$.*

Por tanto una solución factible de *NCPZ*, pese a la rotación arbitraria debida a la transformación aplicada, conserva las mismas propiedades, es decir, sigue siendo igualmente probable a ser considerada como un óptimo global. En la Proposición 6.2 se plantea un nuevo enfoque del *NCPZ* orientado a la representación en términos de \mathbf{Z}^* y se comprueba que entre los óptimos globales se encuentran los vectores propios de la matriz de afinidad normalizada \mathbf{P} definida como:

$$\mathbf{P} = \mathbf{D}^{-1/2} \mathbf{W} \mathbf{D}^{-1/2} \quad (6.17)$$

La matriz \mathbf{P} es estocástica [79], por tanto puede verificarse que $\mathbf{1}_n$ es un vector pro-

pio trivial de \mathbf{P} asociado al mayor valor propio que, en este caso, es 1 debido a la normalización.

Proposición 6.2. (Solución propia óptima) Sea (\mathbf{V}, \mathbf{S}) la descomposición propia de \mathbf{P} :

$$\mathbf{P}\mathbf{V} = \mathbf{V}\mathbf{S}, \mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_n] \text{ y } \mathbf{S} = \text{Diag}(\mathbf{s}) \text{ con valores propios ordenados de forma decreciente } s_1 \geq \dots \geq s_n,$$

obtenida a partir de la solución propia ortonormal $(\tilde{\mathbf{V}}, \mathbf{S})$ de la matriz simétrica $\mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$ donde

$$\mathbf{V} = \mathbf{D}^{-\frac{1}{2}}\tilde{\mathbf{V}} \quad (6.18)$$

$$\mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}\tilde{\mathbf{V}} = \tilde{\mathbf{V}}\mathbf{S}, \quad \tilde{\mathbf{V}}^T\tilde{\mathbf{V}} = \mathbf{I}_n \quad (6.19)$$

entonces cualquier conjunto de k vectores propios es un candidato a ser un óptimo local de $NCPZ$ y los k primeros vectores conforman el óptimo global.

El valor de la función objetivo de los conjuntos candidatos a ser solución de $NCPZ$, se puede escribir como

$$\varepsilon[\mathbf{v}_{\pi_1}, \dots, \mathbf{v}_{\pi_k}] = \frac{1}{k} \sum_{l=1}^k s_{\pi_l} \quad (6.20)$$

donde $\boldsymbol{\pi}$ es un vector de índices compuesto por k enteros distintos de $\{1, \dots, n\}$. Entonces, de acuerdo a la Proposición 6.2, los óptimos globales son obtenidos con $\boldsymbol{\pi} = [1, \dots, k]$ y pueden escribirse como:

$$\mathbf{Z}^* = [\mathbf{v}_1, \dots, \mathbf{v}_k] \quad (6.21)$$

cuya matriz de valores propios y función objetivo son, respectivamente:

$$\boldsymbol{\Lambda}^* = \text{Diag}([s_1, \dots, s_k]) \quad (6.22)$$

$$\varepsilon(\mathbf{Z}^*) = \frac{1}{k} \text{tr}(\boldsymbol{\Lambda}^*) = \max_{\mathbf{Z}^T\mathbf{D}\mathbf{Z}=\mathbf{I}_k} \varepsilon(\mathbf{Z}) \quad (6.23)$$

En conclusión, el óptimo global de $NCPZ$ es un subespacio generado por los vectores propios asociados a los k mayores valores propios de \mathbf{P} a través de matrices ortonor-

males:

$$\{\mathbf{Z}^\top \mathbf{R} : \mathbf{R}^\top \mathbf{R} = \mathbf{I}_k, \mathbf{P}\mathbf{Z}^* = \mathbf{Z}\mathbf{\Lambda}^*\} \quad (6.24)$$

El valor de función objetivo de las soluciones de la forma $\mathbf{Z}^* \mathbf{R}$, representa una cota superior al valor de las soluciones de *NCPM*.

Corolario 6.1. (Acotado superior) Para cualquier k ($k \leq n$) se cumple que

$$\max \varepsilon(\Gamma_{\mathbb{V}}^k) \leq \max_{\mathbf{Z}^\top \mathbf{D}\mathbf{Z} = \mathbf{I}_k} \varepsilon(\mathbf{Z}) = \frac{1}{k} \sum_{l=1}^k s_l \quad (6.25)$$

Además, el valor optimal de *NCPZ* disminuye conforme se aumente el número de grupos considerados.

Corolario 6.2. (Monotonidad decreciente) Para cualquier k ($k \leq n$) se cumple que

$$\max_{\mathbf{Z}^\top \mathbf{D}\mathbf{Z} = \mathbf{I}_{k+1}} \varepsilon(\mathbf{Z}) \leq \max_{\mathbf{Z}^\top \mathbf{D}\mathbf{Z} = \mathbf{I}_k} \varepsilon(\mathbf{Z}) \quad (6.26)$$

Después de obtener la solución en términos de los vectores propios es necesario aplicar una transformación de \mathbf{Z} para volver al problema inicial, por tanto si T es la función de mapeo de \mathbf{M} a \mathbf{Z} , entonces T^{-1} es la normalización de \mathbf{M} , es decir:

$$\mathbf{Z} = T(\mathbf{M}) = \mathbf{M}(\mathbf{M}^\top \mathbf{D}\mathbf{M})^{-\frac{1}{2}} \quad (6.27)$$

y

$$\mathbf{M} = T^{-1}(\mathbf{Z}) = \text{Diag}(\text{diag}^{-\frac{1}{2}}(\mathbf{Z}\mathbf{Z}^\top))\mathbf{Z} \quad (6.28)$$

donde $\text{diag}(\cdot)$ representa la diagonal de la matriz de su argumento. Desde una perspectiva geométrica, considerando las filas de \mathbf{Z} como coordenadas de un espacio k -dimensional, se puede decir que T^{-1} es una normalización en longitud que ubica los puntos dentro de una hiper-esfera unitaria. Con esta normalización, se transforma un óptimo continuo $\mathbf{Z}^* \mathbf{R}$ del espacio de \mathbf{Z} al espacio de \mathbf{M} . Dado que \mathbf{R} es ortonormal se tiene que:

$$T^{-1}(\mathbf{Z}^* \mathbf{R}) = T^{-1}(\mathbf{Z}^*) \mathbf{R} \quad (6.29)$$

Esta simplificación permite caracterizar directamente los óptimos continuos con

$T^{-1}(\mathbf{Z}^*)$ en el espacio de \mathbf{M} , así:

$$\{\tilde{\mathbf{M}}^* \mathbf{R} : \tilde{\mathbf{M}} = T^{-1}(\mathbf{Z}^*), \mathbf{R}^\top \mathbf{R} = \mathbf{I}_k\} \quad (6.30)$$

La segunda etapa para resolver *NCPM* es obtener la solución discreta óptima. Los óptimos de *NCPZ*, en general, no son soluciones óptimas de *NCPM* [70], sin embargo resultan útiles para obtener la solución discreta. Dicha solución no puede ser el absoluto maximal de *NCPM* pero si es aproximado al óptimo global, debido a la naturaleza continua (continuidad) de la función objetivo. Con esto, el objetivo de la discretización es encontrar una solución discreta que satisfaga las condiciones binarias del problema original, mostradas en (6.8) y (6.13), de forma que las soluciones encontradas sean aproximadas al óptimo continuo, como se describió en (6.30).

Teorema 6.1. (*Discretización óptima*) *Una partición óptima discreta $\tilde{\mathbf{M}}^*$ es aquella que satisface el siguiente problema de optimización, denominado OPD (Optimal discretization problem)*

$$\min \phi(\mathbf{M}, \mathbf{R}) = \|\mathbf{M} - \tilde{\mathbf{M}}^* \mathbf{R}\|_F^2 \quad (6.31)$$

$$s. a. \mathbf{M} \in 0, 1^{n \times k}, \mathbf{M} \mathbf{1}_k = \mathbf{1}_n, \mathbf{R}^\top \mathbf{R} = \mathbf{I}_k \quad (6.32)$$

donde $\|\mathbf{A}\|_F$ denota la norma Frobenius de la matriz \mathbf{A} : $\|\mathbf{A}\|_F = \sqrt{\text{tr}(\mathbf{A}\mathbf{A}^\top)}$. Por tanto, un óptimo local $(\mathbf{M}^*, \mathbf{R}^*)$ del problema de optimización puede ser resuelto iterativamente, dividiendo OPD en dos partes ODPM y ODPR que corresponden al problema de optimización en términos de M y R , así:

El problema de optimización con \mathbf{R}^* conocido (ODPM),

$$\min \phi(\mathbf{M}) = \|\mathbf{M} - \tilde{\mathbf{M}}^* \mathbf{R}^*\|_F^2 \quad (6.33)$$

$$s. a. \mathbf{M} \in 0, 1^{n \times k}, \mathbf{M} \mathbf{1}_k = \mathbf{1}_n \quad (6.34)$$

y con \mathbf{M}^* conocido (ODPR),

$$\min \phi(\mathbf{R}) = \|\mathbf{M}^* - \tilde{\mathbf{M}}^* \mathbf{R}\|_F^2 \quad (6.35)$$

$$s. a. \mathbf{R}^T \mathbf{R} = \mathbf{I}_k \quad (6.36)$$

La solución óptima de ODPM se obtiene seleccionando los k vectores propios asociados a los k mayores valores propios:

$$m_{il}^* = \langle l = \arg \min_j \tilde{m}_{ij} \rangle, \quad j \in \{1, \dots, k\}, \quad i \in \mathbb{V} \quad (6.37)$$

y la solución de ODPR se obtiene a través de vectores singulares,

$$\mathbf{R}^* = \tilde{\mathbf{U}} \mathbf{U}^T \quad (6.38)$$

$$\mathbf{M}^{*T} = \mathbf{U} \mathbf{\Omega} \tilde{\mathbf{U}}^T, \quad \mathbf{\Omega} = \text{Diag}(\boldsymbol{\omega}) \quad (6.39)$$

donde $(\mathbf{U}, \mathbf{\Omega}, \tilde{\mathbf{U}})$ es la descomposición en valores singulares de $\mathbf{X}^{*T} \mathbf{X}^*$, con $\mathbf{U}^T \mathbf{U} = \mathbf{I}_k$, $\tilde{\mathbf{U}}^T \tilde{\mathbf{U}} = \mathbf{I}_k$ y $\omega_1 \geq, \dots, \geq \omega_k$.

Demostración. La función objetivo puede ser extendida de la forma

$$\begin{aligned} \phi(\mathbf{M}, \mathbf{R}) &= \|\mathbf{M}^* - \tilde{\mathbf{M}} \mathbf{R}\|_F^2 = \|\mathbf{M}\|_F^2 + \|\tilde{\mathbf{M}}\|_F^2 - \text{tr}(\tilde{\mathbf{M}} \mathbf{R}^T \tilde{\mathbf{M}} + \mathbf{M}^T \tilde{\mathbf{M}} \mathbf{R}) = \\ &= 2N - 2 \text{tr}(\mathbf{M} \mathbf{R}^T \tilde{\mathbf{M}}^{*T}) \end{aligned}$$

lo que demuestra que minimizar $\phi(\mathbf{M}, \mathbf{R})$ es equivalente a maximizar $\text{tr}(\mathbf{M} \mathbf{R}^T \tilde{\mathbf{M}}^{*T})$. Por tanto, en ODPM, dado que todos los elementos de $\text{diag}(\mathbf{M} \mathbf{R}^T \tilde{\mathbf{M}}^{*T})$ pueden ser optimizados independientemente, se demuestra la ecuación (6.37). Para demostrar la solución de ODPR, se puede calcular un lagrangiano usando como multiplicador la matriz simétrica $\mathbf{\Lambda}$, así:

$$L(\mathbf{R}, \mathbf{\Lambda}) = \text{tr}(\tilde{\mathbf{M}}^* \mathbf{R}^T \mathbf{M}^{*T}) - \frac{1}{2} \text{tr}(\mathbf{\Lambda}^T (\mathbf{R}^T \mathbf{R} - \mathbf{I}_k)) \quad (6.40)$$

donde el óptimo $(\mathbf{R}^*, \mathbf{\Lambda}^*)$ debe satisfacer:

$$L_R = \tilde{\mathbf{M}}^{*T} \mathbf{M}^* - \mathbf{R} \mathbf{\Lambda} = 0 \Rightarrow \mathbf{\Lambda}^* = \mathbf{R}^{*T} \tilde{\mathbf{M}}^{*T} \mathbf{M}^* \quad (6.41)$$

Por último, dado que $\mathbf{\Lambda}^* = \mathbf{U} \mathbf{\Omega} \mathbf{U}^T$, se tiene que $\mathbf{R}^* = \tilde{\mathbf{U}} \mathbf{U}^T$, $\phi(\mathbf{R}) = 2N - 2 \text{tr}(\mathbf{\Omega})$ y el mayor valor de $\text{tr}(\mathbf{\Omega})$ genera la matriz \mathbf{M}^* más aproximada a $\tilde{\mathbf{M}}^* \mathbf{R}^*$.

En resumen, resolver OPD en dos etapas alternantes implica: primero, encontrar el óptimo discreto aproximado $\tilde{\mathbf{M}}^*R$ ($ODPR$), luego determinar la transformación ortonormal más apropiada ($ODPM$), e iterar hasta lo que establezca un determinado criterio de convergencia. Ambas etapas emplean la misma función objetivo ϕ , pero difieren en el parámetro de referencia y las restricciones. Este método iterativo sólo garantiza la convergencia a un óptimo local, pero éste puede aproximarse a un óptimo global a través de una buena estimación inicial.

El Algoritmo 6 reúne los pasos del método de agrupamiento espectral.

Algoritmo 6 Agrupamiento espectral normalizado

1. Inicialización: \mathbf{W} , k

2. Calcular el grado de la matriz \mathbf{W} : $\mathbf{D} = \text{Diag}(\mathbf{W}\mathbf{1}_n)$

3. Encontrar la solución propia óptima \mathbf{Z}^* :

$$\begin{aligned} \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}\tilde{\mathbf{V}}_{[k]} &= \tilde{\mathbf{V}}_{[k]} \text{Diag}(\tilde{\mathbf{s}}_{[k]}), \quad \tilde{\mathbf{V}}_{[k]}^T \tilde{\mathbf{V}}_{[k]} = \mathbf{I}_k \\ \mathbf{Z}^* &= \mathbf{D}^{-\frac{1}{2}}\tilde{\mathbf{V}}_{[k]}, \quad [k] = \{1, \dots, k\} \end{aligned}$$

4. Normalizar \mathbf{Z}^* : $\tilde{\mathbf{M}}^* = \text{Diag}(\text{diag}^{-\frac{1}{2}}(\mathbf{Z}^*\mathbf{Z}^{*T}))\mathbf{Z}^*$

5. Inicializar \mathbf{M}^* calculando \mathbf{R}^* :

$$\begin{aligned} \mathbf{R}_1^* &= [\tilde{m}_{i_1}^*, \dots, \tilde{m}_{i_k}^*], \quad \text{donde } i \text{ es un índice aleatorio } i \in 1, \dots, n \\ \mathbf{c} &= \mathbf{0}_{n \times 1} \\ \text{Desde } l &= 2 \text{ hasta } k \\ \mathbf{c} &= \mathbf{c} + |\tilde{\mathbf{M}}^* \mathbf{R}_{l-1}^*| \\ \mathbf{R}_l^* &= [\tilde{m}_{i_1}^*, \dots, \tilde{m}_{i_k}^*], \quad \text{donde } i = \arg \min \mathbf{c} \\ \text{Termina Desde} & \end{aligned}$$

6. Establecer parámetros de convergencia: δ , $\bar{\phi}^* = 0$.

7. Encontrar la solución discreta óptima $\tilde{\mathbf{M}}^*$:

$$\begin{aligned} \tilde{\mathbf{M}} &= \tilde{\mathbf{M}}^* \mathbf{R}^* \\ m_{ij}^* &= \langle l = \arg \max_j \tilde{m}_{ij} \rangle, \quad j, l \in \{1, \dots, k\}, \quad i \in \mathbb{V} \end{aligned}$$

8. Encontrar la matriz ortonormal óptima \mathbf{R}^* :

$$\mathbf{M}^{*T} \tilde{\mathbf{M}}^* = \mathbf{U} \mathbf{\Omega} \tilde{\mathbf{U}}^T, \quad \bar{\mathbf{\Omega}} = \text{Diag}(\boldsymbol{\omega})$$

$$\phi = \text{tr}(\mathbf{\Omega})$$

$$\text{Si } |\bar{\phi} - \bar{\phi}^*| < \delta$$

El proceso termina y la solución es \mathbf{M}^*

en caso contrario

$$\bar{\phi}^* = \bar{\phi}$$

$$\mathbf{R}^* = \tilde{\mathbf{U}} \mathbf{U}^T$$

Volver al paso 7.

Termina Si

En el paso 3 del Algoritmo 6, $\tilde{\mathbf{V}}_{[k]}$ denota una forma compacta de $[\tilde{\mathbf{v}}_1, \dots, \tilde{\mathbf{v}}_k]$ y puede obtenerse calculando los k primeros vectores propios de la matriz de afinidad normalizada $\mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$. Lo mismo aplica para el vector de valores propios $\tilde{\mathbf{s}}_{[k]}$. En el paso 5, $|\cdot|$ denota el valor absoluto de los elementos de su argumento.

Este método de agrupamiento es robusto a la inicialización aleatoria, debido a la invariancia ortonormal de los óptimos continuos [70]. No obstante, una buena inicialización (\mathbf{W}, k) podría mejorar la velocidad de convergencia del algoritmo.

6.4 Medidas de afinidad

Como se discutió en la sección 6.2, el grado de relación o conexión entre nodos, también denominado grado de afinidad, corresponde a una matriz simétrica y semidefinida positiva. Entonces, una forma aleatoria de establecer la afinidad es $\mathbf{W}_r = \mathbf{X}_r\mathbf{X}_r^\top$, donde \mathbf{X}_r es una matriz aleatoria de $n \times d$. La forma trivial corresponde a una matriz que captura los productos internos entre todas las filas de la matriz de datos, es decir, $\mathbf{W}_t = \mathbf{X}\mathbf{X}^\top$, donde \mathbf{X} corresponde a la matriz de datos. Esta es la medida de afinidad de uso más frecuente.

Dado que la afinidad es una medida de similitud, es decir, tiene un sentido opuesto a la distancia o disimilitud, intuitivamente podrían plantearse alternativas como $w_{ij} = 1/d(\mathbf{x}_i, \mathbf{x}_j)$ para $(i \neq j)$ con $w_{ii} = \text{cte}$, donde $d(\cdot, \cdot)$ es alguna medida de distancia. Esta medida de afinidad presenta una relación inversamente proporcional a la distancia y esto no es conveniente, en todos los casos, porque podría generar cambios bruscos en la matriz \mathbf{W} , por esta razón, se han propuesto versiones suaves empleando una función exponencial, así: $w_{ij} = e^{-d^2(\mathbf{x}_i, \mathbf{x}_j)}$ con $w_{ii} = \text{cte}$. El valor de la constante w_{ii} se establece de acuerdo a las condiciones de análisis, en caso de que no sea de interés la afinidad de un elemento con el mismo, $w_{ii} = 0$.

Una mejora a la anterior medida, propuesta en [73], corresponde a la afinidad exponencial escalada de forma que $w_{ij} = e^{-d^2(\mathbf{x}_i, \mathbf{x}_j)/(\sigma_i\sigma_j)}$, con $w_{ii} = 0$, donde $\sigma_i = d(\mathbf{x}_i, \mathbf{x}_N)$ y \mathbf{x}_N es el N -ésimo vecino más cercano. La selección de N es independiente de la escala y se da en función de la dimensión de los datos [73].

En la Tabla 6.1 se muestran algunas medidas de afinidad.

Medida de afinidad	Expresión
Aleatoria	$\mathbf{W}_r = \mathbf{X}_r \mathbf{X}_r^\top$, donde \mathbf{X}_r es una matriz aleatoria de $n \times d$
Trivial	$\mathbf{W}_t = \mathbf{X} \mathbf{X}^\top$, donde \mathbf{X} corresponde a la matriz de datos.
Exponencial	$w_{ij} = e^{-d^2(\mathbf{x}_i, \mathbf{x}_j)}$ con $w_{ii} = \text{cte}$
Exponencial escalada	$w_{ij} = e^{-d^2(\mathbf{x}_i, \mathbf{x}_j)/(\sigma_i \sigma_j)}$, con $w_{ii} = 0$, donde $\sigma_i = d(\mathbf{x}_i, \mathbf{x}_N)$ es el N -ésimo vecino más cercano [73].

Tabla 6.1: Algunas medidas de afinidad

6.5 Estimación del número de grupos

En el estado del arte del análisis no supervisado, no se encuentran muchos estudios alrededor de la estimación automática del número de grupos, por lo general, este parámetro se establece manualmente. En las técnicas de análisis espectral, se encuentran herramientas que permiten obtener un indicio del número de clases contenidas en el conjunto de datos. En este trabajo se describen dos alternativas para esta tarea, una basada en la descomposición en valores singulares (SVD) y otra empleando los vectores propios de la matriz de afinidad.

6.5.1 Estimación basada en SVD

Muchas de las propiedades de la SVD son útiles en una variedad de problemas y aplicaciones del procesamiento de señales. Por ejemplo, mapear un conjunto de datos a un espacio en donde quedan mejor representados para posteriores tareas de clasificación, mejorando la velocidad y el desempeño de los clasificadores (ver apéndice A). Otra aplicación, no tan común, es la estimación del número de grupos, que se explica a continuación. Sea \mathbf{X} una matriz de datos de $n \times d$, entonces su descomposición en valores singulares es:

$$\mathbf{X} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top = \sum_{i=1}^p \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \quad (6.42)$$

donde $p = \min(n, d)$, $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_m]$ es una matriz ortonormal de $n \times n$, $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_m]$ es una matriz ortonormal de $d \times d$, $\mathbf{\Sigma} = \text{Diag}(\boldsymbol{\sigma})$ es una matriz diagonal y $\boldsymbol{\sigma}$ es el vector de valores singulares. La matriz \mathbf{U} corresponde a los vectores propios de $\mathbf{X} \mathbf{X}^\top$ y \mathbf{V} corresponde a los vectores propios de $\mathbf{X}^\top \mathbf{X}$, y se denominan matrices singulares derecha e izquierda, respectivamente.

La estimación del número de grupos se hace a partir del principio que establece que si una matriz está normalizada con respecto a la norma de Frobenius, esto es,

$$\|\mathbf{X}\|_F^2 = \sum_{i=1}^n \sum_{j=1}^d x_{ij} = n$$

se cumple que

$$\|\mathbf{X}\|_F^2 = \sum_{l=1}^{\rho(\mathbf{X})} \sigma_l^2 \quad (6.43)$$

donde $\rho(\cdot)$ denota el rango de la matriz argumento.

Con esto, el número de grupos se escoge como

$$k = \arg \min \left\{ \alpha n \leq \sum_{l=1}^k \sigma_l^2 \right\} \quad (6.44)$$

donde α es un parámetro a ser sintonizado.

En el estudio realizado en [80], se explica ampliamente esta definición y se discuten otras propiedades útiles de la SVD.

6.5.2 Estimación empleando vectores propios

Una forma intuitiva de encontrar el posible número de grupos contenidos en un conjunto de datos, es analizar los valores propios de la matriz de afinidad. En [69] se demuestra que el primer valor propio (que corresponde al mayor) de la matriz de afinidad normalizada \mathbf{P} (ver ecuación (6.17)) es 1 y tiene una multiplicidad igual al número de grupos, es decir, que el número de grupos corresponde al número de valores propios iguales a 1. Estimar el número de grupos analizando los valores propios de la matriz de afinidad, es apropiado cuando los grupos presentan buena separabilidad porque la multiplicidad de valores propios iguales a 1 coincide con la cantidad de grupos. Sin embargo, si los grupos no son claramente separados, los valores propios se alejan de ser 1 y la elección se hace complicada. En este caso, podría considerarse los valores más aproximados a 1, aunque esto podría derivar en un número muy grande, lo cual no tendría sentido práctico. Además, carece también de sentido conceptual porque un valor propio puede aproximarse a 1 debido a razones diferentes de ser representativo de un grupo, por ejemplo, la presencia de valores atípicos puede causar este efecto.

Una alternativa mejor para la estimación del número de grupos puede lograrse teniendo en cuenta los vectores propios de la matriz de afinidad [73]. Para esto se divide la matriz \mathbf{P} en c submatrices $\{\mathbf{P}^1, \dots, \mathbf{P}^c\}$, y luego se genera una matriz diagonal por bloques de $n \times c$. El valor de c ($c \leq n$) representa la cantidad tentativa de grupos con la que se inicializa la heurística de búsqueda, este parámetro se fija manualmente. La matriz diagonal por bloques $\hat{\mathbf{V}}$ agrupa los vectores propios en la diagonal de la matriz y los demás elementos son 0, así:

$$\hat{\mathbf{V}} = \begin{bmatrix} \mathbf{v}^1 & \mathbf{0}_n & \mathbf{0}_n \\ \mathbf{0}_n & \cdots & \mathbf{0}_n \\ \mathbf{0}_n & \mathbf{0}_n & \mathbf{v}^c \end{bmatrix}_{n \times c} \quad (6.45)$$

donde \mathbf{v}^i es de $1 \times n$ y representa el i -ésimo vector propio de la submatriz \mathbf{P}^i y $\mathbf{0}_n$ es un vector columna n -dimensional en el que todos sus elementos son 0.

Como se mencionó antes, el valor propio igual a 1 se repite con multiplicidad igual a la cantidad de grupos, por lo mismo, la solución propia pudo haber sido generada por cualquier otro conjunto de vectores ortogonales que abarcan el subespacio de las mismas columnas de $\hat{\mathbf{V}}$. Por tanto, $\hat{\mathbf{V}}$ puede ser reemplazado por $\mathbf{V} = \hat{\mathbf{V}}\mathbf{R}$, donde \mathbf{R} es cualquier matriz de rotación ortonormal de $c \times c$. Para esto se debe garantizar que exista una rotación $\hat{\mathbf{R}}$, tal que cada una de sus filas tenga un único elemento diferente de 0.

Sea $\mathbf{Z} = \mathbf{V}\mathbf{R}$ y $\beta_i = \max_j z_{ij}$, entonces para encontrar un espacio en donde los grupos queden bien representados, en términos de separabilidad, se puede plantear la siguiente función de costo:

$$J = \sum_{i=1}^n \sum_{j=1}^c \frac{z_{ij}^2}{\beta_i^2} \quad (6.46)$$

Minimizar esta función de costo sobre todas las posibles rotaciones genera la mejor rotación de los vectores propios, por tanto el número de grupos se escoge como el valor comprendido entre 1 y c que genere el valor mínimo de J . En caso de que más de un valor cumpla esta condición, se escoge el mayor de ellos. El proceso de optimización se lleva a cabo empleando el esquema del gradiente descendiente estocástico, descrito en [73].

Parte III

Marco experimental

7. Descripción de los experimentos

En este estudio se aplicaron diversas pruebas para evaluar los diferentes algoritmos de agrupamiento, en todos los aspectos de interés: efectividad en la separación de las clases, efecto de la inicialización, tiempo de procesado, número de grupos resultantes y, en general, el funcionamiento de cada una de las etapas con respecto al desempeño de todo el proceso de análisis. Todo esto, enmarcado en dos metodologías de análisis de señales ECG, que se describen en las secciones 7.3 y 7.4. En la sección 7.1 se explica el proceso de extracción y normalización de las señales, y en la sección 7.2 se describe el conjunto de características utilizado.

7.1 Señales ECG

Las señales empleadas en este estudio fueron obtenidas de los registros de la base de datos de arritmias del MIT (MIT/BIH), descrita en la sección 3.2. Se utilizó para todos los experimentos el canal 0, que en la mayoría de los casos corresponde a la derivación MLII. Con el fin de evitar el efecto de DC y ajustar la señal de tal forma que la máxima amplitud sea 1, se aplicó una normalización a cada registro \mathbf{y} , de forma que

$$\mathbf{y} \leftarrow \frac{\mathbf{y} - \mu(\mathbf{y})}{\max|\mathbf{y}|} \quad (7.1)$$

Los latidos y complejos QRS fueron extraídos usando las anotaciones de la base de datos que se encuentran sobre el pico R, así:

$$\mathbf{b}_i = \mathbf{y}(p_i - 0.35RR_i : p_i + 0.65RR_i) \quad (7.2)$$

y

$$\mathbf{c}_i = \mathbf{y}(p_i - \alpha Fs : p_i + \beta Fs) \quad (7.3)$$

donde p_i es la ubicación del pico R del i -ésimo latido \mathbf{b}_i del registro, \mathbf{c}_i es el complejo QRS del latido i , \mathbf{RR} es el vector que almacena la distancia entre picos R (HRV),

F_s es la frecuencia de muestreo y $\mathbf{v}(a : b)$ representa todos los elementos del vector \mathbf{v} comprendidos entre las posiciones a y b . Nótese que la extracción de los latidos se hace a partir de una variable dinámica, por tanto la longitud de cada latido extraído es diferente.

Los complejos se extrajeron con un ancho de ventana de 150 ms y con el pico R centrado, es decir, $\alpha = \beta = 0.075$ s.

7.2 Conjunto de características

El conjunto de datos corresponde a características tomadas de trabajos previos que han reportado buen desempeño en la caracterización de la morfología, variabilidad y representación de la señal. Estas características han sido empleadas en aplicaciones para detectar latidos de tipo N, V [61], L [36], R [36, 52], A [81], entre otros. La Tabla 7.1 muestra, en detalle, el conjunto de características, incluyendo su descripción y una posible forma de calcularlas.

#	Tipo	Descripción
1 2 3	HRV [81]	<ul style="list-style-type: none"> • Intervalo RR (f_1) • Intervalo pre-RR (f_2) • Intervalo post-RR (f_3)
4 5 6	Prematuridad [81]	<ul style="list-style-type: none"> • Diferencia entre intervalos RR y pre-RR, $f_4 = f_1 - f_2$ • Diferencia entre intervalos post-RR y RR, $f_5 = f_3 - f_1$ • Latidos A continuos <ul style="list-style-type: none"> ◦ $f_6 = \left(\frac{f_3}{f_1}\right)^2 + \left(\frac{f_2}{f_1}\right)^2 - \left(f_4 + \beta \cdot \frac{1}{3} \sum_{i=1}^3 f_i^2 \cdot \log(f_i)^2\right)$, $0 < \beta < 0.1$
7 8	Morfología [81, 61]	<ul style="list-style-type: none"> • DTW entre onda P actual y onda P promedio • Polaridad del complejo QRS <ul style="list-style-type: none"> ◦ Sea \mathbf{b} las muestras de un latido, entonces, $f_8 = \left \frac{\max(\mathbf{b})}{\min(\mathbf{b})} \right$
9		<ul style="list-style-type: none"> • Energía complejo QRS <ul style="list-style-type: none"> ◦ $f_9 = E_{b_j} = \sum_{i=1}^n b_j(i)^2$
10, ..., 19	Representación [82]	<ul style="list-style-type: none"> • 10 Coeficientes de Hermite: $f_{i=10:19} = C_n^\sigma = \langle \mathbf{x}, \phi_n^\sigma \rangle, \quad \text{donde } \phi_n^\sigma = \frac{e^{-t^2/2\sigma^2}}{\sqrt{2^n n! \sqrt{\pi}}} H_n(t/\sigma),$ <p>H_n es el polinomio de Hermite de grado n, \mathbf{x} es el complejo QRS y σ es el ancho de la ventana</p>
20, ..., 90	[44]	<ul style="list-style-type: none"> • Wavelet Db2 (A4: 20 – 25, D4: 26 – 31, D3: 32 – 41, D2: 43 – 58, D1: 59 – 90) <ul style="list-style-type: none"> ◦ Usando Transformada Wavelet Discreta (DWT)

Tabla 7.1: Conjunto de características seleccionadas para el análisis de arritmias

7.3 Metodología 1

Esta metodología corresponde al enfoque tradicional de análisis, en la que se llevan a cabo, en el orden, las etapas de pre-proceso y extracción de latidos (sección 7.1), extracción de características (sección 7.2), selección de características (apéndice A), estimación del número de grupos (sección 6.5), inicialización de centroides (sección 5.3 y capítulo 6) y agrupamiento no supervisado (capítulo 5), sobre todo el conjunto de datos. El diagrama de bloques de la Figura 7.1 muestra la aplicación de la metodología 1.

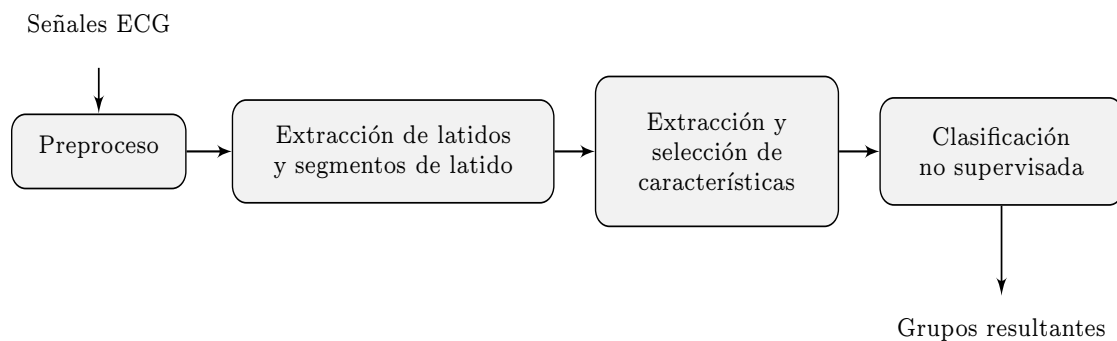


Figura 7.1: Diagrama de bloques de la metodología 1 (tradicional)

La etapa de clasificación no supervisada incluye: estimación del número de grupos, inicialización y agrupamiento. La aplicación de esta etapa es diferente para cada prueba.

7.4 Metodología 2

En esta metodología se propone un análisis a tramos de la señal, en donde, se caracteriza y se agrupa el subconjunto de datos de dicho tramo antes de procesar el siguiente. Para esto se establece el tamaño de la ventana de análisis y un criterio de unión. De acuerdo a la cantidad de características se determinó que el mínimo número de latidos de cada tramo debe ser de por lo menos 300, lo que corresponde, aproximadamente, a ventanas de 5 minutos. El criterio para la unión de grupos se establece a partir de la disimilitud de los latidos asociados a los centroides de cada grupo. La disimilitud es estimada con el algoritmo DTW, explicado en el apéndice C, y se denotará como $dtw(\cdot, \cdot)$. La unión de grupos se aplica con en el principio explicado en la sección 5.4, empleando

N_s segmentos. Sea $\mathbf{P}^i = \{\mathbf{C}_1^i, \dots, \mathbf{C}_{k^i}^i\}$ la partición obtenida con el segmento i , k^i el número de grupos de la misma partición, $\mathbf{Q}^i = \{\mathbf{q}_1^i, \dots, \mathbf{q}_{k^i}^i\}$ los centroides de cada grupo y $\mathbf{b}(\mathbf{q})$ el latido asociado al centroide \mathbf{q} , entonces la unión se hace a razón de:

$$\vartheta(j^i, j^{i+q}) = \text{dtw}(\mathbf{b}(\mathbf{q}_{j^i}^i), \mathbf{b}(\mathbf{q}_{j^{i+1}}^{i+1})), \quad j = 1, \dots, k^i, \quad i = 1, \dots, N_s \quad (7.4)$$

Con esto, si $\vartheta(j^i, j^{i+1})$ es un valor admisible, el grupo $\mathbf{C}_{j^i}^i$ se une con el grupo $\mathbf{C}_{j^{i+1}}^{i+1}$, de lo contrario se pasa a comparar con los demás grupos. En caso de que el centroide asociado a un determinado grupo no presente disimilitud admisible con ningún otro, dicho grupo se mantendrá como independiente para la siguiente iteración. Así, se evita la mala agrupación de latidos de clases minoritarias. En la Figura 7.2 se muestra el diagrama explicativo de la metodología 2.

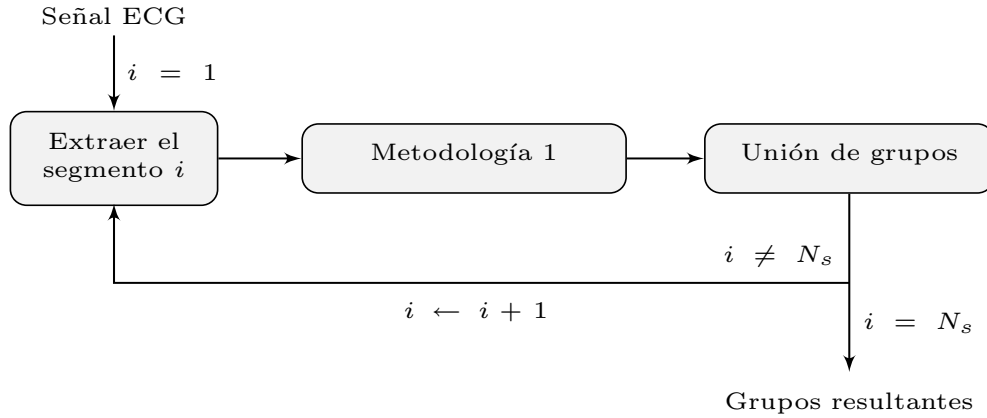


Figura 7.2: Diagrama explicativo de la metodología 2 (por segmentos)

El análisis por cada tramo se hace con el esquema planteado en la sección previa (sección 7.3). En la sección de resultados y discusión (capítulo 9), se mencionan los métodos específicos aplicados en cada prueba.

8. Medidas de desempeño del agrupamiento

Para evaluar los resultados del agrupamiento se usaron medidas de sensibilidad (Se), especificidad (Sp) y porcentaje de clasificación (CP), definidas como:

$$Se = \frac{TN}{TN + FP} \times 100 \quad (8.1)$$

$$Sp = \frac{TP}{TP + FN} \times 100 \quad (8.2)$$

$$CP = \frac{TN + TP}{TN + TP + FN + FP} \quad (8.3)$$

de acuerdo a la siguiente convención:

- TP : Verdaderos positivos. Un latido de la clase de interés (CI) es clasificado correctamente.
- TN : Verdaderos negativos. Un latido diferente de la clase de interés (NCI) es clasificado correctamente.
- FP : Falsos positivos. Un latido NCI es clasificado como CI .
- FN : Falsos negativos. Un latido NCI es clasificado como NCI .

La sensibilidad y especificidad miden, respectivamente, la proporción de latidos NCI y la proporción de latidos CI , clasificados correctamente. Estas medidas se aplican sobre todas las clases presentes en el registro y se usan para medir el desempeño del sistema, pero no tienen implicación en la sintonización de los parámetros del proceso de agrupamiento (todos los criterios de sintonización son estrictamente no supervisados debido a las razones discutidas anteriormente). Dado que no es de utilidad un número de grupos relativamente grande, se aplica un factor de penalización sobre las anteriores

medidas, con el fin de que un agrupamiento correcto realizado con un número de grupos considerablemente grande no registre un desempeño del 100%. El factor de penalización aplicado en este trabajo es de la forma

$$e^{-\eta k_r/k_a} \quad (8.4)$$

donde k_r es la cantidad de grupos resultantes del agrupamiento, k_a es el valor admisible de la cantidad de grupos y η es un factor de ajuste. Con esto, la medida \mathbf{m} (que en este caso, puede ser *Se*, *Sp* o *CP*), se pondera de modo que

$$\mathbf{m} \leftarrow \begin{cases} \mathbf{m} e^{-\eta k_r/k_a} & k_r > k_a \\ \mathbf{m} & k_r \leq k_a \end{cases} \quad (8.5)$$

El término η se fija de acuerdo al rigor de penalización que se determine, teniendo en cuenta que debe ser estrictamente mayor que 0, además, es aconsejable que sea menor que 1 para que la penalización no sea tan rigurosa: $0 < \eta \leq 1$.

Otro índice utilizado en este trabajo, es la medida complementaria al radio de dispersión o desviación del valor de la función objetivo, que se denota con f_1/f_2 . Este índice mide el desempeño del proceso de agrupamiento a través de la relación entre el valor óptimo de la función objetivo f_1 y el valor calculado usando la partición resultante f_2 . La función objetivo evaluada para la estimación de este índice debe ser coherente con el método de agrupamiento empleado. Dado que el mínimo valor de f_2 es f_1 , esto es, $f_2 \geq f_1$, esta medida indica un buen agrupamiento cuando su valor se aproxima a 1.

A partir de lo discutido en la sección 6.3, se puede plantear una medida no supervisada del desempeño del agrupamiento, que cuantifica la coherencia de los grupos empleando un criterio de particionado, así:

$$\varepsilon_M = \frac{1}{k} \sum_{l=1}^k \frac{\mathbf{M}_l^\top \mathbf{W} \mathbf{M}_l}{\mathbf{M}_l^\top \mathbf{D} \mathbf{M}_l} \quad (8.6)$$

donde \mathbf{M} es la matriz conformada por los valores de pertenencia de todos los elementos a cada grupo: $M_{ij} = m(\mathbf{q}_j/\mathbf{x}_i)$ (ver sección 5.2.3), \mathbf{M}_l denota una submatriz de pertenencia asociada al grupo l , \mathbf{W} es la matriz de afinidad y \mathbf{D} es el grado de la matriz \mathbf{W} . De acuerdo a las restricciones mostradas en (6.13), la matriz \mathbf{M} debe ser binaria, por tanto, para los casos en los que se aplique agrupamiento de naturaleza

suave, debe hacerse la siguiente conversión:

$$m_{ij} = \langle \max \arg m(\mathbf{q}_j / \mathbf{x}_i) \rangle, \quad j = 1, \dots, k_r$$

donde $\langle \cdot \rangle$ es 1 si su argumento es verdadero y 0 en otro caso.

Debido a la normalización con respecto de la matriz de afinidad, el máximo valor de ε_M es 1, por tanto indica un buen agrupamiento si su valor es cercano a 1. El problema de esta medida es la sensibilidad a la matriz de afinidad, por eso es aconsejable usar una afinidad recomendada por la literatura, como la exponencial escalada (ver sección 6.4); otra buena alternativa es emplear la matriz de afinidad resultante de la etapa de selección de características (ver apéndice A).

En la Tabla 8.1 se relacionan las medidas de desempeño consideradas en este estudio con su respectiva notación y expresión matemática.

Denominación	Notación	Expresión
Sensibilidad	Se	$\frac{TN}{TN + FP} \times 100$
Especificidad	Sp	$\frac{TP}{TP + FN} \times 100$
Porcentaje de clasificación	CP	$\frac{TN + TP}{TN + TP + FN + FP} \times 100$
Desviación de la función objetivo	f_1/f_2	$\frac{f_1}{f_2}$
Coherencia de los grupos	ε_M	$\frac{1}{k} \sum_{l=1}^k \frac{\mathbf{M}_l^T \mathbf{W} \mathbf{M}_l}{\mathbf{M}_l^T \mathbf{D} \mathbf{M}_l}$

Tabla 8.1: Medidas de desempeño empleadas

Parte IV

Discusión de resultados

9. Resultados y discusión

9.1 Resultados obtenidos con la metodología 1

Esta sección de resultados consta de diferentes pruebas realizadas para analizar el desempeño de los criterios de inicialización y métodos de agrupamiento.

9.1.1 Prueba 1

Esta prueba está enfocada al agrupamiento de arritmias ventriculares de tipo R, L y V, empleando características morfológicas y de representación a través del modelo de Hermite:

- Energía del QRS.
- Parámetro de escala óptimo σ_{opt} .
- Sexto coeficiente de Hermite (C_{σ}^n con $n = 6$).
- Diferencia de \mathbf{c}_i y un complejo plantilla \mathbf{c}_{temp} aplicando (B.11).

Este conjunto de características se explica ampliamente en el Apéndice B.

La etapa de agrupamiento es llevada a cabo con los métodos HM (sección 5.2.4) y KHM (sección 5.2.5).

Los índices de evaluación corresponden a los mencionados en el capítulo 8, calculados de acuerdo a la siguiente convención:

TP: Verdaderos positivos. Un latido normal es clasificado como normal. En caso de que no existan normales, se aplica como latidos V clasificados como V.

TN: Verdaderos negativos. Un latido patológico es clasificado como patológico. En caso de que no existan normales, se aplica como latidos diferentes de V (noV) clasificados como noV.

FP: Falsos positivos. Un latido patológico es clasificado como normal o un latido noV es clasificado como V.

FN: Falsos negativos. Un latido normal es clasificado como patológico o un latido V es clasificados como noV.

La Tabla 9.1 relaciona la cantidad y tipos de latido con los registros empleados en esta prueba.

Registro	Latidos			
	N	R	L	V
118	–	2164	–	16
124	–	1529	–	47
207	–	85	1457	105
214	–	–	200	256
215	3194	–	–	164
217	244	–	–	162
219	2080	–	–	64
221	2029	–	–	396
223	2027	–	–	473
228	1686	–	–	362
230	2253	–	–	1
233	2229	–	–	830
234	2698	–	–	3

Tabla 9.1: Registros empleados en la prueba 1

Las tablas 9.2 y 9.3 muestran los resultados del agrupamiento. Las tres primeras columnas de cada tabla corresponden al desempeño del algoritmo con inicialización aleatoria y las tres columnas restantes corresponden al desempeño del algoritmo aplicando el criterio de inicialización máx-mín. El procedimiento se iteró 10 veces para los dos algoritmos de agrupamiento y se registraron los valores de la media (μ) y desviación estándar (σ) de los índices de desempeño. En las pruebas realizadas sin aplicar el criterio de inicialización, no se consideraron criterios de convergencia del proceso de agrupamiento y se fijó como máximo número de iteraciones $N_{iter} = 20$. Los resultados empleando máx-mín fueron obtenidos con $N_{iter} = 100$ y evaluando la convergencia del algoritmo a través del valor de la función objetivo.

En la Figura 9.1 se muestra como ejemplo el espacio de características del registro 207. Se agregan los latidos de tipo N de otro registro porque el registro 207 no posee latidos de este tipo (ver Tabla 9.1).

Reg.	Inicialización aleatoria ($\mu - \sigma$)			máx- mín ($\mu - \sigma$)		
	$Se(\%)$	$Sp(\%)$	$CP(\%)$	$Se(\%)$	$Sp(\%)$	$CP(\%)$
118	99.26– 0	38.46– 16.97	98.75– 0.12	99.26– 0	51.61– 0	98.85– 0
124	99.48– 0.14	72.34– 6.73	95.77– 0.002	99.54– 0	74.75– 0.12	95.94– 0.002
207	98.4– 0.1	99.1– 0.01	99.5– 0.002	98.9– 0.01	99– 0.002	99.5– 0.001
214	69.5– 0.0474	77.94– 0.09	75.64– 0.009	70.12– 0.021	79.35– 0.09	78.2– 0.008
215	100– 0	100– 0	100– 0	100– 0	100– 0	100– 0
217	100– 0	100– 0	100– 0	100– 0	100– 0	100– 0
219	100– 0	100– 0	100– 0	100– 0	100– 0	100– 0
221	100– 0	100– 0	100– 0	100– 0	100– 0	100– 0
223	100– 0	100– 0	100– 0	100– 0	100– 0	100– 0
228	100– 0	100– 0	100– 0	100– 0	100– 0	100– 0
230	100– 0	100– 0	100– 0	100– 0	100– 0	100– 0
233	100– 0	100– 0	100– 0	100– 0	100– 0	100– 0
234	100– 0	100– 0	100– 0	100– 0	100– 0	100– 0

Tabla 9.2: Resultados obtenidos empleando el algoritmo H-medias en la prueba 1

Reg.	Inicialización aleatoria ($\mu - \sigma$)			máx - mín ($\mu - \sigma$)		
	$Se(\%)$	$Sp(\%)$	$CP(\%)$	$Se(\%)$	$Sp(\%)$	$CP(\%)$
118	99.26– 0	63.19– 3.7	99.6– 0.02	99.26– 0	63, 2– 0.1	99.72– 0.001
124	99.53– 0.1	74.86– 1.2	95.94– 0.0009	99.54– 0	74.87– 0.08	95.94– 0.001
207	98.9– 0.1	99– 1.2	99.5– 0.0009	98.9– 0	99– 0.08	99.5– 0.001
214	69.9– 0.009	80.94– 0.05	78.14– 0.009	70.2– 0.008	80.94– 0.092	78.2– 0.002
215	100– 0	100– 0	100– 0	100– 0	100– 0	100– 0
217	100– 0	100– 0	100– 0	100– 0	100– 0	100– 0
219	100– 0	100– 0	100– 0	100– 0	100– 0	100– 0
221	100– 0	100– 0	100– 0	100– 0	100– 0	100– 0
223	100– 0	100– 0	100– 0	100– 0	100– 0	100– 0
228	100– 0	100– 0	100– 0	100– 0	100– 0	100– 0
230	100– 0	100– 0	100– 0	100– 0	100– 0	100– 0
233	100– 0–	100– 0	100– 0	100– 0	100– 0	100– 0
234	100– 0	100– 0	100– 0	100– 0	100– 0	100– 0

Tabla 9.3: Resultados obtenidos empleando el algoritmo KHM en la prueba 1

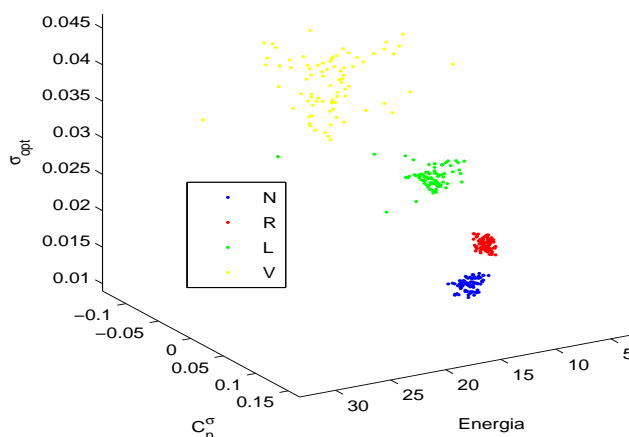


Figura 9.1: Características de los registros 207(R, L, V) y 215 (N)

Las características morfológicas y de representación del complejo QRS brindan buena separabilidad de las arritmias ventriculares tratadas en este trabajo, debido a la naturaleza fisiológica de estas patologías y su relación directa con la generación del trazado del QRS. En este caso, la energía, las características de Hermite (C_n^σ , σ_{opt}) y las diferencias espectrales empleando plantillas del QRS fueron características adecuadas para esta tarea de clasificación.

En general, los dos métodos de agrupamiento aplicados generan resultados similares, de acuerdo a las Tablas 9.2 y 9.3, cuando se aplica el criterio de inicialización máx-mín. Esto se debe a que la principal fortaleza del algoritmo K-medias armónicas sobre el H-medias es la robustez a la inicialización debido a la media armónica y la naturaleza suave de sus funciones de membresía y peso. Sin embargo, se puede apreciar que sin utilizar inicialización, el algoritmo H-medias baja el rendimiento, mientras que el algoritmo K-medias armónicas, prácticamente, mantiene su desempeño. De esta manera, se podría concluir que es posible obtener iguales resultados con los algoritmos, aunque sacrificando costo computacional debido a que el algoritmo KHM puede obviar este paso, mientras que el H-medias necesariamente debe incluir la inicialización para garantizar buenos resultados. No obstante, con un criterio de inicialización apropiado se tiene una probabilidad mayor de obtener un buen valor de convergencia.

Los resultados de desempeño son similares para todos los casos (9 últimas filas, registros 215 al 234), aunque en los registros 118, 124 y 214 se aprecia que la especificidad

fue baja. Esto ocurre porque las características consideradas no generan, en todos los casos, buena separabilidad de los latidos V con respecto de los otros latidos considerados en este estudio. En los registros 118 y 124, el valor de Se es alto, en contraste con el valor bajo de Sp , debido a que la cantidad de latidos V de estos registros es mucho menor en comparación a la cantidad de los otros latidos (ver Tabla 9.1), por tanto, un solo latido mal agrupado tendría implicaciones considerables en el valor de Sp .

9.1.2 Prueba 2

En esta prueba se evalúan distintos criterios de inicialización sobre el registro 207. Se escogió este registro porque posee 4 de las clases de latidos de interés (A, V, L y R), por tanto, es un registro representativo de la base de datos. En la Tabla 9.4 se muestra el desempeño del algoritmo GEMC inicializado con criterios máx-mín, J-medias y J-GEM. Los valores registrados en esta tabla corresponden a la media y desviación estándar de los índices de Sp y Se obtenidos de evaluar todo el conjunto en 10 iteraciones del procedimiento, considerando la convención dada en el capítulo 8. En la primera columna se mencionan los métodos aplicados, en la segunda columna se muestran los resultados obtenidos iterando el algoritmo sin control de convergencia y la tercera columna corresponde a los resultados obtenidos evaluando la función objetivo. En esta prueba se aplica uno de los algoritmos de DBC, partiendo del hecho de que los algoritmos de naturaleza suave presentan mejor desempeño y son menos sensibles a una mala inicialización, como se discutió en la prueba 1 (sección 9.1.1).

Método	Iteraciones			Función Objetivo			
	Iter.	Se $\mu - \sigma$	Sp $\mu - \sigma$	δ	Iter. Resultantes aprox(μ) - σ	Se $\mu - \sigma$	Sp $\mu - \sigma$
máx - mín	10^1	0.52 - 0.016	0.87 - 0.02	10^{-1}	10 - 0.707	0.56 - 0.022	0.88 - 0.027
	10^2	0.96 - 0.027	0.99 - 0.032	10^{-2}	15 - 1.01	0.63 - 0.015	0.88 - 0.017
	10^3	0.94 - 0.03	0.99 - 0.025	10^{-3}	25 - 1.13	0.95 - 0.032	0.99 - 0.015
J-medias	10^1	0.63 - 0.01	0.91 - 0.008	10^{-1}	5 - 0.48	0.72 - 0.014	0.85 - 0.01
	10^2	0.95 - 0.012	0.99 - 0.009	10^{-2}	11 - 0.53	0.95 - 0.018	0.99 - 0.01
	10^3	0.95 - 0.013	0.99 - 0.008	10^{-3}	14 - 0.39	0.95 - 0.016	0.99 - 0.01
J-GEM	10^1	0.64 - 0.009	0.91 - 0.004	10^{-1}	5 - 0.32	0.72 - 0.006	0.87 - 0.002
	10^2	0.95 - 0.008	0.99 - 0.004	10^{-2}	11 - 0.31	0.95 - 0.004	0.99 - 0.002
	10^3	0.95 - 0.007	0.99 - 0.004	10^{-3}	12 - 0.31	0.95 - 0.004	0.99 - 0.002

Tabla 9.4: Resultados aplicando GEMC y diferentes criterios de inicialización

En general, se puede decir que los algoritmos de inicialización mejoran el desempeño del agrupamiento, brindando una partición inicial apropiada. Los algoritmos basados

en el criterio J-medias presentaron mejores resultados que los obtenidos con máx-mín, esto se debe a la evaluación sistemática de la función objetivo que se realiza en el algoritmo J-medias. Además, dicha función objetivo se puede definir de manera adecuada al tipo de agrupamiento específico que se emplee (DBC ó MSSC). En cambio, el máx-mín usa un criterio de distancia preestablecido y no tiene en cuenta la naturaleza de la función objetivo del proceso de agrupamiento. Adicionalmente, el algoritmo J-medias reduce el costo computacional porque calcula el valor de la función objetivo localmente (ver sección 5.3.2).

Como se mencionó anteriormente, los métodos de naturaleza suave, como el KHM y DBC, son menos sensibles a la inicialización porque los valores de membresía tienen mayor posibilidad de cambio que en los métodos de naturaleza fuerte. En la gráfica mostrada en la Figura 9.2 se aprecia la robustez hacia la inicialización de centroides del método GEMC. En esta gráfica, los puntos verdes son los centroides iniciales y los puntos negros son los centroides finales. El espacio de características corresponde a un conjunto de datos artificiales de 3 clases y 150 observaciones, 50 por cada clase. El desempeño del agrupamiento obtenido en esta prueba, sin control de convergencia (esto es, sin evaluar la función objetivo), por cada iteración fue de $CP = 89\%$ y el número de iteraciones fue $N_{iter} = 20$.

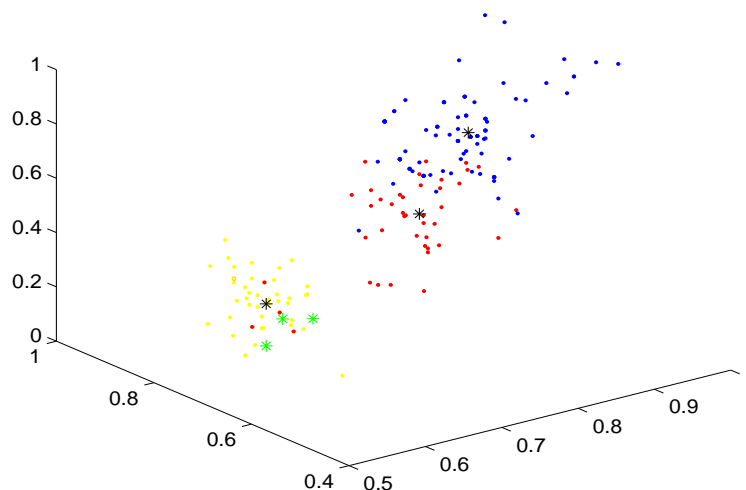


Figura 9.2: Robustez hacia la inicialización en DBC

9.1.3 Prueba 3

En esta prueba, se aplica la metodología 1 con agrupamiento basado en densidades (GEMC y NPDBC usando el método de Parzen) e inicialización J-H-medias. El desempeño del proceso de agrupamiento se midió en términos de Se y Sp por cada clase, y el índice f_1/f_2 . En este caso, las clases corresponden al grupo de arritmias recomendado por la AAMI (ver Tabla 3.2). En la Tabla 9.5 se muestran los resultados obtenidos. La primera columna muestra el número del registro, la segunda columna el desempeño obtenido con NPDBC y en la tercera columna el desempeño del método GEMC. N_{hb} indica el número de latidos que tiene el registro, este valor se especifica por cada clase. En la fila de Σ se registra el desempeño total, $\sum |y_i|$ representa el total de latidos de la clase relacionada en la respectiva columna y $\mu(\cdot)$ es la media aritmética.

Reg.	NPDBC usando el método de Parzen						GEMC							
	N_{hb}	N	S	V	F	Q	f_1/f_2	N_{hb}	N	S	V	F	Q	f_1/f_2
100	2237	33	1	0	0	0	1	2237	33	1	0	0	0	0.99
	Se	100	100	100				Se	100	93.94	100			
	Sp	100	100	100				Sp	94.12	100	100			
106	1506	0	520	0	0	0	0.99	1506	0	520	0	0	0	0.98
	Se	97.81	100					Se	99.67	96.53				
	Sp	100	97.81					Sp	96.53	99.67				
107	0	0	59	0	2076	0	1	0	0	59	0	2076	0	1
	Se		100			99.95		Se		100		100		
	Sp		99.95			100		Sp		100		100		
111	2121	0	1	0	0	0	1	2121	0	1	0	0	0	1
	Se	100	100					Se	100	100				
	Sp	100	100					Sp	100	100				
113	1787	6	0	0	0	0	0.95	1787	6	0	0	0	0	0.95
	Se	100	83.33					Se	100	83.33				
	Sp	83.33	100					Sp	83.33	100				
119	1541	0	444	0	0	0	1	1541	0	444	0	0	0	1
	Se	100	100					Se	100	100				
	Sp	100	100					Sp	100	100				
123	1513	0	3	0	0	0	1	1513	0	3	0	0	0	1
	Se	100	100					Se	100	100				
	Sp	100	100					Sp	100	100				
207	1542	106	210	0	0	0	0.98	1542	106	210	0	0	0	0.85
	Se	97.34	97.14	98.1				Se	97.42	0	98.1			
	Sp	97.78	98.63	98.97				Sp	65.4	100	98.06			
215	3194	2	164	1	0	0	0.85	3194	2	164	1	0	0	0.85
	Se	100	0	99.39	0			Se	100	0	99.39	0		
	Sp	98.18	100	100	100			Sp	98.18	100	100	100		
217	244	0	162	260	1540	0	0.99	244	0	162	260	1540	0	0.98
	Se	100	93.83	95.38	99.81			Se	99.59	88.27	92.66	99.87		
	Sp	99.34	99.76	99.9	99.25			Sp	98.62	99.9	99.64	99.25		
220	1951	94	0	0	0	0	0.97	1951	94	0	0	0	0	0.84
	Se	98.82	92.55					Se	99.74	46.81				
	Sp	92.55	98.82					Sp	46.81	99.74				
221	2029	0	396	0	0	0	1	2029	0	396	0	0	0	1
	Se	99.9	99.75					Se	99.9	99.75				
	Sp	99.75	99.9					Sp	99.75	99.9				
230	2253	0	1	0	0	0	1	2253	0	1	0	0	0	1
	Se	100	100					Se	100	100				
	Sp	100	100					Sp	100	100				
234	2698	50	3	0	0	0	0.95	2698	50	3	0	0	0	0.95
	Se	99.96	76	100				Se	99.96	78	100			
	Sp	77.36	99.96	100				Sp	79.25	99.96	100			
Σ	$\sum y_i $	24616	291	1964	261	3616	$\mu(\cdot)$							
	$\mu(Se)$	99.53	74.84	99.26	47.7	99.88			99.75	50.35	98.5	46.33	99.93	
	$\mu(Sp)$	96.02	99.57	99.7	99.95	99.62	0.98		89.39	99.95	99.7	99.82	99.62	0.96

Tabla 9.5: Resultados de los métodos DBC empleando la metodología 1

Las figuras 9.3 y 9.4, muestran los resultados de la metodología 1 de forma gráfica. Estas gráficas muestran los valores medios, dispersión, máximos y mínimos de las

medidas Sp y Se de los dos métodos de agrupamiento evaluados.

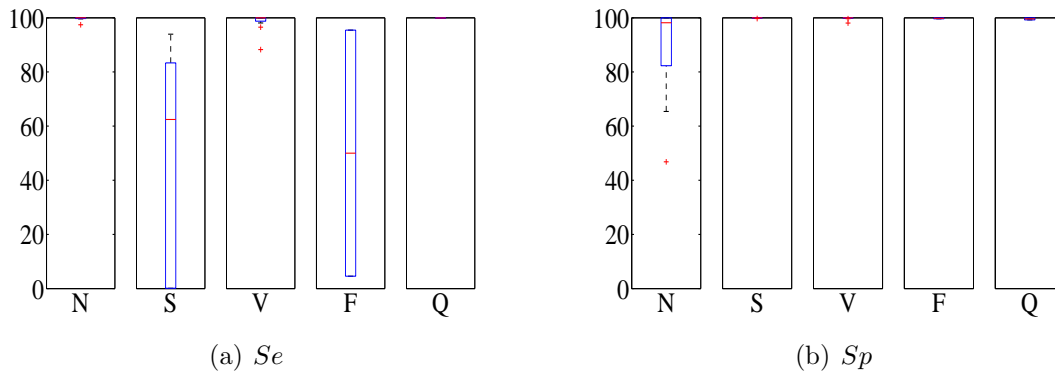


Figura 9.3: Desempeño de GEMC en la metodología 1

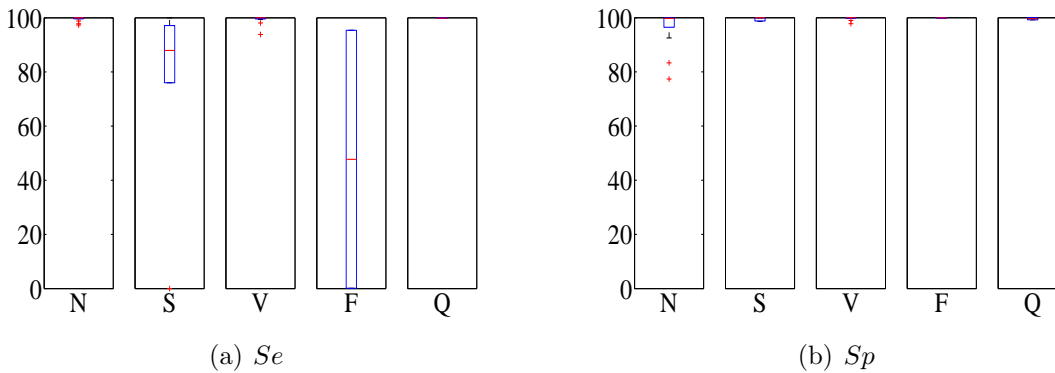


Figura 9.4: Desempeño de NPDBC en la metodología 1

Los métodos DBC presentan buenos resultados porque usan información estadística, como el segundo momento estadístico y la probabilidad a posteriori, para el refinamiento de la partición. Además, son menos sensibles a la inicialización que los métodos clásicos.

9.2 Resultados obtenidos con la metodología 2

En esta prueba se aplica la metodología 2 empleando, en la etapa de selección de características, el algoritmo $Q - \alpha$ no supervisado (sección A.5) y una medida de relevancia basada en la ponderación de los elementos de los vectores propios (sección A.7). El proceso de agrupamiento se aplicó a los datos transformados: $\mathbf{Z} = \mathbf{X}\mathbf{U}$, donde \mathbf{U} es la matriz de vectores propios de la matriz de covarianza de WPCA, dada

en la ecuación (A.13), con $w_i = \alpha_i$ para el caso de $Q - \alpha$ y $w_i = \rho_i$ (criterio del 90 %) para el caso de WPCA con ponderación de vectores propios (ver sección A.4). Se empleó el algoritmo NPDBC con estimación de Parzen como método de agrupamiento. Los criterios de evaluación del desempeño del agrupamiento fueron los mismos que se aplicaron en la prueba de la sección 9.1.3, adicionando, el número de grupos resultantes k_{end} , el número de características relevantes p y el índice ε_M . El valor de ε_M se calculó empleando la matriz resultante de la selección de características, en el caso de $Q - \alpha$, y para la otra prueba se usó la medida de afinidad exponencial escalada. Las medidas Sp y Se fueron penalizadas aplicando la ecuación (8.5), con $\eta = 0.05$ y $k_a = 12$. Para generar los centroides iniciales se aplicó el algoritmo descrito en la sección 6.3.3, con un máximo de iteraciones igual a 2 y la misma matriz de afinidad aplicada para obtener ε_M . En las tablas 9.6 y 9.7, se muestran los resultados obtenidos de esta prueba. Las figuras 9.5 y 9.6 muestran los mismos resultados de forma gráfica.

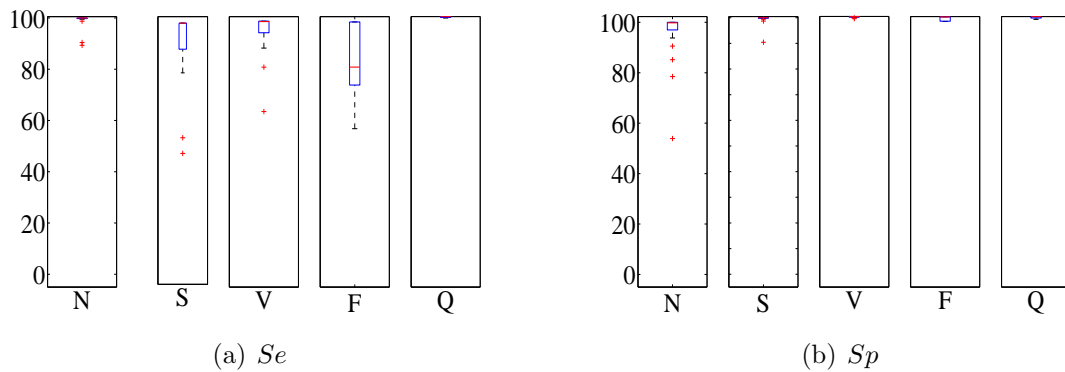


Figura 9.5: Desempeño de NPDBC con $Q - \alpha$ en la metodología 2

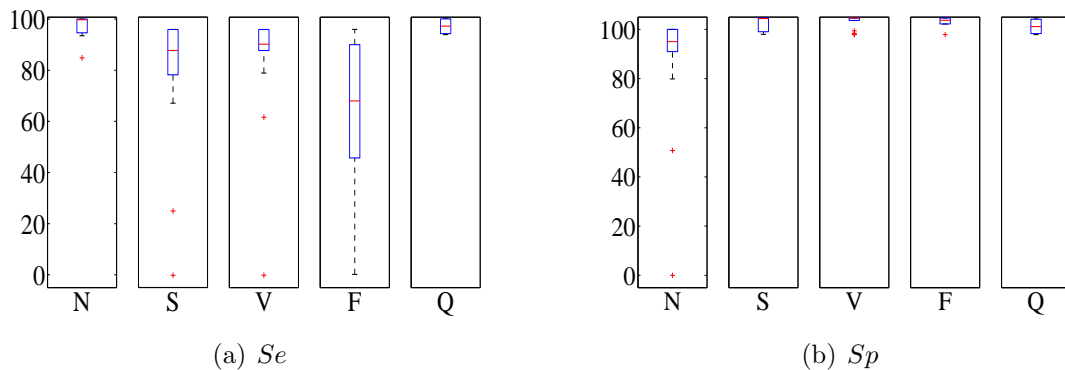


Figura 9.6: Desempeño de NPDBC con WPCA (vectores propios ponderados) en la metodología 2

Reg.		N	S	V	F	Q	Tiempo (s)	f_1	f_2	f_1/f_2	p	k_{end}	ε_M
100	Latidos	2237	33	1	0	0	39.64	-1634.54	-1634.54	1	6	12	0.93
	Se	100	100	100									
	Sp	100	100	100									
101	Latidos	1858	3	0	0	2	32.11	-1265.34	-1265.34	1	5	12	0.91
	Se	100	100	100									
	Sp	100	100	100									
103	Latidos	2080	2	0	0	0	35.9	-1407.01	-1407.01	1	5	7	0.93
	Se	100	100										
	Sp	100	100										
106	Latidos	1506	0	520	0	0	34.75	-962.07	-967.27	0.99	5	8	0.93
	Se	99.67		99.81									
	Sp	99.81		99.67									
107	Latidos	0	0	59	0	2076	36.97	-967.79	-967.79	1	5	8	0.91
	Se			100		100							
	Sp			100		100							
111	Latidos	2121	0	1	0	0	36.62	-1474.27	-1474.27	1	4	4	0.95
	Se	100		100									
	Sp	100		100									
112	Latidos	2535	2	0	0	0	44.25	-1793.16	-1793.16	1	5	7	0.91
	Se	100	100										
	Sp	100	100										
113	Latidos	1787	6	0	0	0	30.68	-1211.32	-1211.32	1	5	11	0.91
	Se	100	100										
	Sp	100	100										
118	Latidos	2164	96	16	0	0	39.28	-1727.78	-1922.98	0.89	5	6	0.71
	Se	99.95	96.84	64.29									
	Sp	98.17	99.72	99.96									
119	Latidos	1541	0	444	0	0	34.78	-754.8	-754.8	1	6	5	0.93
	Se	100		100									
	Sp	100		100									
121	Latidos	1859	1	1	0	0	31.37	-1284.33	-1284.33	1	5	10	0.92
	Se	100	100	100									
	Sp	100	100	100									
123	Latidos	1513	0	3	0	0	25.68	-979.14	-979.14	1	5	12	0.94
	Se	100		100									
	Sp	100		100									
124	Latidos	1534	31	47	5	0	27.76	-1213.84	-1445.01	0.84	6	8	0.71
	Se	99.93	80	89.36	75								
	Sp	85.19	99.94	100	100								
201	Latidos	1633	128	198	2	0	33.85	-1474.16	-1638.04	0.9	7	8	0.8
	Se	99.51	88	96.97	100								
	Sp	94.44	99.4	100	100								
207	Latidos	1542	106	210	0	0	21.7	-1777.18	-1777.18	0.99	6	12	0.9
	Se	99.81	100	99.52									
	Sp	99.68	100	99.82									
210	Latidos	2421	22	195	10	0	46.78	-1666.93	-1742.51	0.79	6	9	0.62
	Se	99.83	47.62	92.78	80								
	Sp	93.78	99.92	99.51	99.89								
213	Latidos	2639	28	220	362	0	57.95	-1916.98	-2081.31	0.92	7	13	0.79
	Se	99.54	96.3	81.82	57.73								
	Sp	78.49	99.97	99.21	98.27								
214	Latidos	2000	0	256	1	2	39.39	-1215.36	-1215.36	0.99	6	10	0.9
	Se	100		99.22	100	100							
	Sp	99.22		100	100	100							
217	Latidos	244	0	162	260	1540	39.41	-1340.65	-1341.88	0.94	7	10	0.84
	Se	89.34		94.9	84.23	99.48							
	Sp	98.31		99.61	98.19	98.94							
220	Latidos	1951	94	0	0	0	35.66	-1331.81	-1331.81	0.99	5	8	0.88
	Se	100	97.83										
	Sp	97.83	100										
221	Latidos	2029	0	396	0	0	42.69	-1534.08	-1534.08	0.99	6	9	0.87
	Se	99.85		99.49									
	Sp	99.49		99.85									
222	Latidos	2272	209	0	0	0	44.56	-1961.96	-2746.74	0.71	7	5	0.52
	Se	98.72	53.92										
	Sp	53.92	98.72										
230	Latidos	2253	0	1	0	0	41.2	-1658.06	-1658.06	1	6	8	0.93
	Se	100		100									
	Sp	100		100									
231	Latidos	1566	1	2	0	0	34.25	-1399.04	-1399.04	1	6	9	0.9
	Se	100	100	100									
	Sp	100	100	100									
232	Latidos	397	1381	0	0	0	31.17	-1575.12	-1.7698	0.9	7	7	0.66
	Se	90.43	100										
	Sp	100	90.43										
234	Latidos	2698	50	3	0	0	48.71	-1873.5	-1943.55	0.93	7	6	0.69
	Se	100	90	100									
	Sp	90.57	100	100									
Σ	$\mu(Se)$	99.06	91.21	95.7	82.83	99.87	$\mu(\cdot)$			$\mu(\cdot)$	$\mu(\cdot)$	$\mu(\cdot)$	$\mu(\cdot)$
	$\mu(Sp)$	95.56	99.3	99.88	99.39	99.74							

Tabla 9.6: Resultados obtenidos con la metodología 2, empleando NPDBC y $Q - \alpha$ no supervisado

Reg.		N	S	V	F	Q	Tiempo (s)	f_1	f_2	f_1/f_2	p	k_{end}	ε_M
100	Latidos	2237	33	1	0	0	39.53	-1791.07	-1791.07	1	10	15	0.9
	Se	94	94	94									
	Sp	94	94	94									
101	Latidos	1858	3	0	0	2	32.4	-1268.76	-1268.76	1	9	14	0.89
	Se	94.3	94.3			94.3							
	Sp	94.3	94.3			94.30							
103	Latidos	2080	2	0	0	0	36.3	-1443.28	-1443.28	1	10	9	0.9
	Se	100	100										
	Sp	100	100										
106	Latidos	1506	0	520	0	0	35.14	-951.91	-9547.05	0.99	10	9	0.87
	Se	99.73		100									
	Sp	100		99.73									
107	Latidos	0	0	59	0	2076	37.6	-1046.24	-1046.24	1	9	10	0.9
	Se			100		100							
	Sp			100		100							
111	Latidos	2121	0	1	0	0	36.81	-1474.3	-1474.3	1	8	9	0.9
	Se	100		100									
	Sp	100		100									
112	Latidos	2535	2	0	0	0	44.65	-1793.13	-1793.13	1	8	10	0.92
	Se	100	100										
	Sp	100	100										
113	Latidos	1787	6	0	0	0	30.82	-1211.39	-1211.39	1	9	10	0.91
	Se	100	100										
	Sp	100	100										
118	Latidos	2164	96	16	0	0	39.92	-1709.49	-1997.6	0.86	10	14	0.69
	Se	99.91	97.89	64.29									
	Sp	98.17	99.72	99.96									
119	Latidos	1541	0	444	0	0	34.88	-783.16	-783.16	1	9	9	0.89
	Se	100		100									
	Sp	100		100									
121	Latidos	1859	1	1	0	0	31.5	-1280.23	-1280.23	1	10	16	0.88
	Se	93.6	93.6	93.6									
	Sp	93.6	93.6	93.6									
123	Latidos	1513	0	3	0	0	25.68	-1014.66	-1014.66	1	8	9	0.92
	Se	100		100									
	Sp	100		100									
124	Latidos	1534	31	47	5	0	28.16	-1213.01	-1461.5	0.83	9	11	0.66
	Se	99.93	80	89.36	75								
	Sp	85.19	99.94	100	100								
201	Latidos	1633	128	198	2	0	33.88	-1471.78	-1839.72	0.8	10	11	0.63
	Se	99.51	80.83	94.44	0								
	Sp	89.69	99.4	100	100								
207	Latidos	1542	106	210	0	0	21.85	-196.2	-196.2	1	10	15	0.9
	Se	94	94	94									
	Sp	94	94	99									
210	Latidos	2421	22	195	10	0	46.83	-1773.79	-2073.34	0.85	8	16	0.58
	Se	99.75	33.33	91.24	44.44								
	Sp	91.07	99.96	99.27	99.89								
213	Latidos	2639	28	220	362	0	57.81	-1885.29	-2049.23	0.92	9	12	0.83
	Se	99.47	100	82.27	57.43								
	Sp	79.83	99.97	99.07	98.23								
214	Latidos	2000	0	256	1	2	39.32	-1396.67	-1396.67	1	10	16	0.9
	Se	93.5		93.5	93.5	93.5							
	Sp	93.5		93.5	93.5	93.5							
217	Latidos	244	0	162	260	1540	39.29	-1372.59	-1373.85	0.99	9	12	0.9
	Se	84.84		91.98	87.69	99.28							
	Sp	98.32		99.66	97.63	98.95							
220	Latidos	1951	94	0	0	0	36.17	-1327.62	-1327.62	0.99	10	9	0.9
	Se	99.9	96.81										
	Sp	96.81	99.9										
221	Latidos	2029	0	396	0	0	43.42	-1534.98	-1534.98	0.99	10	9	0.91
	Se	100		99.49									
	Sp	99.49		100									
222	Latidos	2272	209	0	0	0	44.84	-1956.25	-2934.38	0.67	9	7	0.49
	Se	98.72	50.75										
	Sp	50.75	98.72										
230	Latidos	2253	0	1	0	0	42.35	-1745.38	-2618.07	0.67	9	10	0.5
	Se	100		0									
	Sp	0		100									
231	Latidos	1566	1	2	0	0	34.82	-1402.27	-1402.27	1	9	13	0.92
	Se	95	95	95									
	Sp	95	95	95									
232	Latidos	397	1381	0	0	0	31.25	-674.52	-691.669	0.98	8	10	0.91
	Se	94.71	100										
	Sp	100	94.71										
234	Latidos	2698	50	3	0	0	49	-1999.52	-1999.52	1	1	12	0.93
	Se	100	90	100									
	Sp	90.57	100	100									
Σ	$\mu(Se)$	97.64	88.27	88.59	59.68	96.77	$\mu(\cdot)$			$\mu(\cdot)$	$\mu(\cdot)$	$\mu(\cdot)$	$\mu(\cdot)$
	$\mu(Sp)$	89.77	97.84	98.57	98.21	96.68							

Tabla 9.7: Resultados obtenidos con la metodología 2, empleando NPDBC y WPCA con ponderación de vectores propios

Se puede apreciar que el análisis por segmentos mejora el desempeño del proceso de agrupamiento, en comparación con el análisis tradicional (sección 9.1.3), porque reduce la mala agrupación de las clases minoritarias. En efecto, si se estima un número de grupos k para todo el conjunto de datos, existirán k grupos que podrían no ser suficientes para separar las clases presentes en el conjunto, por tanto las clases minoritarias se agruparían equivocadamente. Por otra parte, estimar el número de grupos y agrupar por cada segmento, es menos sensible a las clases minoritarias, porque en un segmento es menor la probabilidad de agrupar mal un latido de alguna clase minoritaria.

Un algoritmo de agrupamiento espectral, es una buena opción para la inicialización de un esquema particional aunque de costo computacional elevado. Por esta razón, el algoritmo sólo se itera dos veces, considerando que como se discutió en la sección 6.3.3, converge rápidamente y no es tan sensible a la inicialización.

En la selección de características, los dos algoritmos presentan resultados similares, sin tener en cuenta la penalización debida al número de grupos. Sin embargo, el algoritmo $Q - \alpha$ es preferible porque genera un número de características menor y esto podría disminuir el costo computacional. Además, también se aprecia que la estimación del número de grupos y la medida de ε_M es mejor empleando los valores resultantes de $Q - \alpha$. La matriz de afinidad resultante del algoritmo $Q - \alpha$ fue obtenida de un refinamiento iterativo, es decir, que tiene en cuenta información intrínseca de los datos. Por esta razón, funciona mejor que la matriz de afinidad exponencial escalada.

La medida no supervisada ε_M demostró ser un buen indicador del agrupamiento resultante porque penaliza el número de grupos (ver Corolario 6.2) y genera un valor cercano a 1 cuando el agrupamiento es correcto, aunque es sensible a la matriz de afinidad. Una buena alternativa para la matriz de afinidad es tomar la matriz resultante de la selección de características. En caso de que no se aplique $Q - \alpha$, la matriz de afinidad de tipo exponencial escalada es una medida aconsejable.

En la Tabla 9.7, el registro 100 es uno de los registros, en donde se aprecia el efecto del factor de penalización. Nótese que $f_1/f_2 = 1$, es decir, que los valores de Sp y Se debieron estar en el rango de 98 a 100%, pero esto no ocurre porque el número de grupos resultantes fue mayor que el número admisible.

Parte V

Comentarios finales

10. Conclusiones y trabajo futuro

En este estudio comparativo de métodos de agrupamiento de latidos de señales ECG, se estudiaron técnicas no supervisadas, entre otras razones, porque no requieren del etiquetado de los latidos, ni de entrenamiento por cada registro. Además, los métodos de agrupamiento no supervisado son los de uso más frecuente en el diseño de dispositivos de análisis de ECG.

10.1 Conclusiones

Considerando lo expuesto en este estudio se puede concluir lo siguiente:

- Un análisis secuencial, en donde se divida la señal en segmentos y se agrupen patrones homogéneos por cada segmento teniendo en cuenta las disimilitudes de los centroides y criterios de exclusión, mejora el desempeño de todo el proceso de agrupamiento, en comparación con el análisis tradicional, porque reduce la mala agrupación de las clases minoritarias. Además, en la mayoría de los casos, la suma de los tiempos de procesado de cada segmento es considerablemente menor que el tiempo de análisis de todo el registro en una iteración, por tanto, se puede decir que reduce el costo computacional.
- Una alternativa para garantizar una mejor convergencia del agrupamiento, es incluir un criterio de inicialización que genere una partición inicial adecuada. Para esto, los criterios basados en la evaluación local y sistemática de la función objetivo, representan una buena alternativa porque presentan un buen compromiso entre costo computacional y desempeño.
- Los métodos de agrupamiento particional de naturaleza suave, presentan mejor desempeño que los de naturaleza fuerte, porque los valores de membresía tienen mayor posibilidad de cambio en caso de una asignación errada, además, son menos sensibles a la inicialización que los métodos clásicos.

- El agrupamiento espectral es un método de agrupamiento que presenta buen desempeño, aún cuando los grupos no son linealmente separables, debido a que emplea información de los vectores propios y transformaciones de los datos, aunque es de costo computacional elevado.
- Un método de análisis basado en agrupamiento no supervisado de tipo espectral y particional, teniendo en cuenta el costo computacional, puede lograrse de forma que los algoritmos de análisis espectral se encarguen de la inicialización, estimación del número de grupos y selección de características, y los algoritmos particionales realicen el agrupamiento de los datos.
- La medida no supervisada asociada a la coherencia de los grupos, obtenida a través de un criterio de particionado espectral, es un buen índice del desempeño del proceso de agrupamiento porque penaliza el número de grupos y, con una matriz de afinidad apropiada, genera un valor cercano a 1 cuando el agrupamiento es correcto.
- Las medidas supervisadas aplicadas en el agrupamiento no supervisado, deben tener cuenta el número de grupos resultantes, de forma que el desempeño de un agrupamiento correcto realizado con un número de grupos relativamente grande no sea de 100 %.

10.2 Trabajo futuro

Como trabajo futuro se pretende desarrollar opciones eficaces para aproximarse a la solución del problema del agrupamiento de patrones homogéneos de latidos, considerando clases desbalanceadas, variabilidad morfológica y costo computacional. Además, teniendo en cuenta los resultados y conclusiones de este estudio, desarrollar un sistema completo de diagnóstico automático de señales ECG desde los enfoques particional y espectral, incluyendo etapas óptimas de preproceso, selección de características, estimación del número de grupos, inicialización y clasificación no supervisada. Todo esto, orientado a aplicaciones en tiempo real a nivel de microcontrolador, que impliquen bajo costo computacional y sean de implementación factible.

Parte VI
Apéndices

A. Selección de características

En el área de reconocimiento de patrones, es común encontrar problemas en los que al momento de extraer patrones descriptivos de observaciones o muestras, que posteriormente serán clasificadas, no se tiene información a priori sobre la cantidad necesaria de dichos patrones, ni de la relevancia en la clasificación de los mismos. Por esta razón, los procesos de caracterización generan matrices de datos de alta dimensión, lo que puede representar un problema para la subsecuente tarea de clasificación en términos de desempeño debido a la información redundante y, además, podría tener implicaciones en el costo computacional del algoritmo. A este problema se le denomina selección de características o atributos. Este problema es típico en el reconocimiento de patrones y aprendizaje de máquina, y se presenta en diferentes ramas de la ciencia (procesamiento de texto, bio-informática, procesamiento de señales biomédicas, etc).

Existen diversas alternativas para esta tarea y la escogencia de un método u otro depende de las condiciones del problema y de la naturaleza de los datos [83]. El análisis de componentes principales (PCA) y sus variantes, como WPCA, representan una buena opción, entre otras razones por su naturaleza no paramétrica, facilidad de implementación y versatilidad. Diversos estudios han comprobado su aplicabilidad como técnica de mapeo, extracción de características y reducción de dimensionalidad en diferentes contextos como procesamiento de señales biomédicas [23], [24], detección de rostros [84], entre otros.

A continuación se describen algunos métodos clásicos de selección de características de inferencia supervisada y no supervisado derivados del algoritmo PCA clásico, desde un enfoque unificado obtenido de la proyección ortonormal y la descomposición en valores propios. Para esto se deduce una función objetivo de PCA a partir del error cuadrático medio de los datos y su proyección sobre una base ortonormal. Luego se generaliza este concepto para plantear una expresión asociada al algoritmo fundamental de WPCA. Adicionalmente, se estudia a fondo el algoritmo $Q - \alpha$, propuesto en [41], para casos supervisados y no supervisados, y se explica su relación con PCA.

A.1 Convención

En este estudio, \mathbf{x}_i denota la i -ésima observación de q atributos o características y $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ denota la matriz de datos.

A.2 Proyección ortonormal

Un vector \mathbf{x} de dimensión q puede escribirse como una combinación lineal de los elementos de una base ortonormal, así:

$$\mathbf{x} = \sum_{i=1}^q c_i \mathbf{u}_i \quad (\text{A.1})$$

donde $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_q)$ representa la base ortonormal y $\mathbf{c} = (c_1, \dots, c_q)$ son los pesos de la combinación lineal.

En general, cualquier proyección ortonormal $\tilde{\mathbf{x}}$ se realiza en un espacio p -dimensional ($p < q$), que mejor represente a \mathbf{x} :

$$\tilde{\mathbf{x}} = \sum_{i=1}^p c_i \mathbf{u}_i \quad (\text{A.2})$$

El error cuadrático medio de la proyección ortonormal con respecto de la señal original, puede estimarse de la forma:

$$\overline{e^2} = \mathbf{E} \{ (\mathbf{x} - \tilde{\mathbf{x}})^\top (\mathbf{x} - \tilde{\mathbf{x}}) \} \quad (\text{A.3})$$

Reemplazando las expresiones (A.1) y (A.2) en (A.3), el error se puede re-escribir como:

$$\overline{e^2} = \mathbf{E} \left\{ \left(\sum_{i=1}^q c_i \mathbf{u}_i - \sum_{i=1}^p c_i \mathbf{u}_i \right)^\top \left(\sum_{i=1}^q c_i \mathbf{u}_i - \sum_{i=1}^p c_i \mathbf{u}_i \right) \right\} = \mathbf{E} \left\{ \left(\sum_{i=p+1}^q c_i \mathbf{u}_i \right)^\top \left(\sum_{i=p+1}^q c_i \mathbf{u}_i \right) \right\} \quad (\text{A.4})$$

A.3 Análisis de componentes principales

En estadística, la aplicación más común de PCA es la reducción de la dimensionalidad de un conjunto de datos. El principio fundamental de este método es determinar el

número de elementos descriptivos subyacentes tras un conjunto de datos que contengan información de la variabilidad de dichos datos [83]. En otras palabras, en PCA se busca la proyección en la que los datos queden mejor representados en términos de mínimos cuadrados, dicha proyección corresponde a la varianza acumulada de cada observación. Una de las ventajas de PCA es que reduce la dimensionalidad de un conjunto de datos, reteniendo aquellos atributos o características del conjunto de datos que contribuyen más a su varianza.

PCA construye una transformación lineal de los datos originales de manera que se genere un nuevo sistema de coordenadas en donde la mayor varianza del conjunto de datos es capturada en el primer eje (denominado primer componente principal), la segunda varianza más grande en el segundo eje, y así sucesivamente; donde la medida de varianza la define una estimación de la matriz de covarianza de los datos [83]. Por tanto, el objetivo de PCA es minimizar el error cuadrático medio de la proyección de los datos sobre los vectores propios de la matriz de covarianza, sujeto a una condición de ortonormalidad. Minimizar dicho error (ver (A.4)), es equivalente a maximizar el complemento del mismo, es decir:

$$\mathbf{E} \left\{ \left(\sum_{i=1}^p c_i \mathbf{u}_i \right)^\top \left(\sum_{i=1}^p c_i \mathbf{u}_i \right) \right\} = \mathbf{E} \left\{ \sum_{i=1}^p c_i^2 \right\} = \text{tr}(\mathbf{c}_p^\top \mathbf{c}_p) = \text{tr}(\mathbf{C}_p) \quad (\text{A.5})$$

donde \mathbf{c}_p es un vector compuesto por los p primeros elementos de \mathbf{c} y $\text{tr}(\cdot)$ representa la traza de su matriz argumento.

Se puede apreciar que \mathbf{C}_p es una matriz de $p \times p$ simétrica y semi-positiva definida. En el caso de PCA, para realizar la proyección de todos los vectores \mathbf{x}_i , la matriz \mathbf{C} es de $q \times q$ y corresponde a la matriz de covarianza, que puede ser estimada como:

$$\mathbf{C}_{PCA} = \frac{1}{n} \mathbf{X}^\top \mathbf{X} \quad (\text{A.6})$$

La anterior ecuación se aplica después de centrar los datos en la media de cada observación, es decir:

$$\mathbf{x}_i \leftarrow \mathbf{x}_i - \mu(\mathbf{x}_i), \quad i = 1, \dots, n \quad (\text{A.7})$$

donde $\mu(\cdot)$ representa la media.

Con lo anterior y considerando el criterio de invariancia ortonormal [70], puede plantearse el siguiente problema de optimización:

$$\max \frac{1}{n} \text{tr}(\mathbf{U}^\top \mathbf{X}^\top \mathbf{X} \mathbf{U}) = \sum_{j=1}^p \lambda_j \quad (\text{A.8})$$

$$\text{s. a. } \mathbf{U}^\top \mathbf{U} = \mathbf{I}_p \quad (\text{A.9})$$

donde \mathbf{I}_d representa una matriz identidad de dimensión d y λ_j los valores propios de \mathbf{C}_{PCA} .

Debido a la simetría de la matriz \mathbf{C}_{PCA} , existe una base completa de vectores propios de la misma y por tanto la transformación lineal que mapea los datos a esta base es, justamente, la representación de los datos que se utiliza para la reducción de la dimensionalidad. Los elementos de la base ortogonal se denominan componentes principales y la proyección de los datos se obtiene con:

$$\mathbf{Z}_{PCA} = \mathbf{X} \mathbf{U} \quad (\text{A.10})$$

En la ecuación (A.8) se aprecia que el valor de la función objetivo se asocia directamente a la suma de los valores propios de la matriz de covarianza, por tanto la solución de este problema de optimización conduce al absurdo de tomar todas las características, por esta razón es necesario aplicar un criterio adicional sobre los componentes principales para la selección de características.

En el algoritmo general de PCA, se considera \mathbf{U} como la matriz de vectores propios de \mathbf{C}_{PCA} ordenados de forma descendente, es decir,

$$\begin{aligned} [\mathbf{U}, \mathbf{\Lambda}] &= \text{eig}(\mathbf{C}_{PCA}), \quad \mathbf{\Lambda} = \text{Diag}(\boldsymbol{\lambda}), \\ \boldsymbol{\lambda} &= (\lambda_1, \dots, \lambda_n) \text{ y } \mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_n) \text{ con } \lambda_1 > \dots > \lambda_n \end{aligned} \quad (\text{A.11})$$

donde $\text{eig}(\cdot)$ representa la descomposición en valores y vectores propios, y $\text{Diag}(\cdot)$ denota una matriz diagonal formada por el vector de su argumento.

Por último, se escogen los p primeros componentes principales como los elementos

relevantes, es decir, los que mejor representan a \mathbf{X} . El valor de p , puede definirse a través de algún criterio de varianza acumulada o evaluando iterativamente los resultados de un clasificador. El criterio de varianza acumulada se aplica sobre el valor de la función objetivo normalizado, así:

$$\mathbf{z}^* = \text{diag}(\mathbf{U}^\top \mathbf{X}^\top \mathbf{X} \mathbf{U}) / \text{tr}(\mathbf{U}^\top \mathbf{X}^\top \mathbf{X} \mathbf{U}) \quad (\text{A.12})$$

donde $\text{diag}(\cdot)$ representa la diagonal de su matriz argumento.

\mathbf{z}^* es un indicador de la variabilidad de cada componente, entonces, para un criterio del $N\%$ se consideran los p elementos de \mathbf{Z} que correspondan a un valor de varianza acumulada del $N\%$, es decir, $\sum_{i=1}^p z_i^* \approx N/100$.

A.4 PCA ponderado

La diferencia de este método y el algoritmo básico de PCA radica en la estimación de la matriz de covarianza. En PCA ponderado (*WPCA*), se emplea una matriz de covarianza denominada ponderada para la cual existen dos maneras fundamentales de estimación [85]. La primera corresponde al caso de la ponderación de características, es decir:

$$\mathbf{C}_{WPCA} = \mathbf{W}^\top \mathbf{X}^\top \mathbf{X} \mathbf{W} \quad (\text{A.13})$$

donde $\mathbf{W} = \text{Diag}(\mathbf{w})$ es una matriz diagonal de pesos. De este modo se puede cambiar la importancia relativa de cada característica en la representación de \mathbf{X} .

Por tanto, el problema de optimización para WPCA con ponderación de características podría plantearse como:

$$\max \text{tr}(\mathbf{U}^\top \mathbf{W}^\top \mathbf{X}^\top \mathbf{X} \mathbf{W} \mathbf{U}) = \sum_{j=1}^p \lambda_j \quad (\text{A.14})$$

$$\text{s. a. } \mathbf{U}^\top \mathbf{U} = \mathbf{I}_p, \quad \mathbf{W}^\top \mathbf{W} = \mathbf{I}_p \quad (\text{A.15})$$

donde λ_j corresponde a los valores propios de $\mathbf{X}^\top \mathbf{X}$.

En este caso, la proyección sobre los componentes principales se obtiene con: $\mathbf{Z}_{WPCA} = \mathbf{X}\mathbf{W}\mathbf{U}$.

La segunda forma básica de WPCA consiste en la ponderación de las observaciones o muestras, donde la matriz de covarianza se puede escribir como:

$$\mathbf{C}_{WPCA}^s = \mathbf{X}^T \mathbf{W}^T \mathbf{W} \mathbf{X} \quad (\text{A.16})$$

Nótese, que estimar la covarianza ponderada empleando (A.16), es equivalente a ponderar las observaciones y luego calcular la matriz de covarianza usando (A.6). En efecto, si $\mathbf{X}_w = \mathbf{W}\mathbf{X}$ representa los datos ponderados entonces la proyección se realiza sobre los vectores propios de $\mathbf{X}_w^T \mathbf{X}_w$ que corresponde a \mathbf{C}_{WPCA}^s .

El problema de optimización de WPCA empleando la ponderación de observaciones es:

$$\max \text{tr}(\mathbf{U}^T \mathbf{X}^T \mathbf{W}^T \mathbf{W} \mathbf{X} \mathbf{U}) = \sum_{j=1}^p \lambda_j \quad (\text{A.17})$$

$$\text{s. a. } \mathbf{U}^T \mathbf{U} = \mathbf{I}_p, \quad \mathbf{W}^T \mathbf{W} = \mathbf{I}_p \quad (\text{A.18})$$

Con esto, la proyección sobre los componentes principales se determina como $\mathbf{Z}_{WPCA} = \mathbf{X}\mathbf{U}\mathbf{W}$ [84]. Existen diversas formas de estimar los pesos. En la Tabla A.1 se muestran algunos métodos no supervisados de ponderación.

En general, la ponderación w_i se aplica partiendo de la idea de que no siempre los elementos que generan una buena representación, generan también una buena separabilidad [84]. Además, un elemento relevante en la representación, deberá seguir siendo relevante pese a la ponderación; mientras que un elemento que no sea relevante en la representación pero que si lo sea en la clasificación, podría llegar a ser considerado dentro de los componentes principales después de dicha ponderación.

Método	Expresión
Pre-normalización de PCA	$w_i = \frac{1}{\sqrt{\frac{1}{q} \sum_{j=1}^q X_{ij}^2}}$
Ponderación empleando los valores propios [84]	$w_i = \lambda_i^{-1/2}$

Tabla A.1: Algunos métodos de ponderación para WPCA

A.5 Algoritmo Q- α

En [41] se presenta una definición de relevancia en términos de una matriz de *afinidad* que captura los productos internos de las observaciones y un vector de ponderación. Este concepto se basa en la coherencia de los subconjuntos o grupos resultantes de un proceso de agrupamiento, empleando propiedades espectrales y análisis topológico derivado de la teoría de grafos, donde la matriz de datos representa los vértices de los grafos ponderados no dirigidos y la matriz de afinidad indica los pesos de cada arista del grafo [70]. A este método, los autores lo denominan $Q - \alpha$ debido a que α es el vector de ponderación y Q es la matriz ortonormal de rotación. En este estudio la matriz ortonormal se denota con U .

Sea M una matriz de $q \times n$, definida como $M = \mathbf{X}^T = (\mathbf{m}_1, \dots, \mathbf{m}_q)^T$ y preprocesada de forma que los vectores \mathbf{m}_i tengan media cero y norma unitaria, entonces la matriz de afinidad se puede obtener con:

$$\mathbf{C}_\alpha = \sum_{i=1}^q \alpha_i \mathbf{m}_i \mathbf{m}_i^T = \mathbf{M}^T \text{Diag}(\alpha) \mathbf{M} \quad (\text{A.19})$$

Desde el punto de vista estadístico, la idea general del algoritmo $Q - \alpha$ es la misma de PCA, parte de un principio de variabilidad acumulada basada en mínimos cuadrados, con la diferencia que, en este caso, la variabilidad se mide con la matriz \mathbf{C}_α y no con la covarianza. El factor de escalamiento α_i permite ajustar la importancia relativa de cada observación. Intuitivamente, a partir de estas premisas se podría plantear una función objetivo como la mostrada en (A.8), sin embargo, dado que en este algoritmo se parte de una transformación ortonormal arbitraria, $\mathbf{Z} : \mathbf{Z} = \mathbf{X}U$ no representa la proyección sobre los componentes principales y por tanto la selección de características relevantes se relaciona directamente con el valor de α_i . Entonces, considerando que la solución del problema de optimización consiste en encontrar el vector α , resulta conveniente

re-formular el problema de optimización planteado en (A.8) como la siguiente forma cuadrática:

$$\max_{\boldsymbol{\alpha}} \text{tr}(\mathbf{U}^{\top} \mathbf{C}_{\alpha} \mathbf{C}_{\alpha} \mathbf{U}) = \sum_{j=1}^p \lambda_j^2 \quad (\text{A.20})$$

$$\text{s. a. } \mathbf{U}^{\top} \mathbf{U} = \mathbf{I}_p, \quad \boldsymbol{\alpha}^{\top} \boldsymbol{\alpha} = 1 \quad (\text{A.21})$$

donde $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)$ son los valores de propios de \mathbf{C}_{α} y \mathbf{U} es una matriz ortonormal arbitraria.

Duplicar el término \mathbf{C}_{α} en la formulación planteada en la sección A.3, no cambia la naturaleza de la función objetivo porque \mathbf{C}_{α} es un matriz simétrica y semi-positiva definida, además, en términos matemáticos, esto resulta ventajoso porque permite plantear una forma cuadrática con respecto a la variable de interés, así:

$$\max_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^{\top} \mathbf{G} \boldsymbol{\alpha} \quad (\text{A.22})$$

$$\text{s. a. } \boldsymbol{\alpha}^{\top} \boldsymbol{\alpha} = 1 \quad (\text{A.23})$$

donde \mathbf{G} una matriz auxiliar cuyos elementos son $G_{ij} = (\mathbf{m}_i^{\top} \mathbf{m}_j) \mathbf{m}_i^{\top} \mathbf{U} \mathbf{U}^{\top} \mathbf{m}_j$.

Esta última ecuación corresponde a la función objetivo de la versión no supervisada de $Q - \alpha$. Dado que, en principio, la matriz \mathbf{G} es obtenida a partir de una transformación ortonormal arbitraria, es necesario plantear un método iterativo en el que se sintonicen la matriz \mathbf{U} y el vector $\boldsymbol{\alpha}$. Del problema de optimización planteado en la ecuación (A.20) se aprecia que mientras el vector $\boldsymbol{\alpha}$ apunta a la dirección de las características relevantes, la matriz \mathbf{U} indica su rotación; por tanto la sintonización de estos parámetros es mutuamente dependiente y debe realizarse de forma alternante, como se aprecia en el Algoritmo 7.

Algoritmo 7 Método Q – α estándar

1. Inicialización: $\mathbf{M} = \mathbf{X}^\top$, $\mathbf{U}^{(0)}$ de $k \times n$ ($\mathbf{U}^{(0)\top}\mathbf{U}^{(0)} = \mathbf{I}_n$), $\mathbf{m}_i \leftarrow (\mathbf{m}_i - \mu(\mathbf{m}_i))/\|\mathbf{m}_i\|$.
2. Formar \mathbf{G} : $G_{ij} = (\mathbf{m}_i^\top \mathbf{m}_j) \mathbf{m}_i^\top \mathbf{U} \mathbf{U}^\top \mathbf{m}_j$
3. Calcular $\boldsymbol{\alpha}$ como el vector propio asociado al mayor valor propio de \mathbf{G} .
4. Calcular $\mathbf{C}_\alpha = \mathbf{M}^\top \text{Diag}(\boldsymbol{\alpha}) \mathbf{M}$
5. Obtener la transformación ortonormal: $\mathbf{Z}^{(r)} = \mathbf{C}_\alpha^{(r)} \mathbf{U}^{(r-1)}$
6. Descomposición QR: $[\mathbf{U}^{(r)}, \mathbf{R}] = \text{qr}(\mathbf{Z}^{(r)})$
7. Incrementar r : $r \leftarrow r + 1$ y retornar al paso 2

En [41], también se estudian dos alternativas del Q – α no supervisado: La primera empleando una normalización a través del laplaciano de la matriz \mathbf{C}_α , y la segunda se basa en el criterio de la aceleración de Ritz y descomposición en valores singulares cuando se asume un vector $\boldsymbol{\alpha}$ inicial. Adicionalmente, demuestran la convergencia del algoritmo. No obstante, un indicador de la convergencia del algoritmo puede establecerse con el cambio del vector $\boldsymbol{\alpha}$, es decir, la diferencia del vector obtenido en la iteración actual y el inmediatamente anterior: $\|\boldsymbol{\alpha}^{(r)} - \boldsymbol{\alpha}^{(r-1)}\| < \delta$, donde δ define la precisión.

En el paso 1 del algoritmo 7 se establece una matriz ortonormal inicial de $n \times k$, donde $k : k < n$ puede ser arbitrario, dado que no tiene implicación en la función objetivo a razón de que esta matriz se convierte en una matriz cuadrada de $n \times n$, a partir la segunda iteración. Al final, dicha matriz corresponderá a la base que genera la transformación lineal de los datos.

Es fácil de comprobar que la solución del problema de optimización planteado en (A.22) corresponde al vector propio asociado al mayor valor propio de \mathbf{G} . De la definición de valores y vectores propios se tiene que:

$$\mathbf{G}\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha} \Rightarrow \boldsymbol{\alpha}^{-1}\mathbf{G}\boldsymbol{\alpha} = \boldsymbol{\alpha}^\top\mathbf{G}\boldsymbol{\alpha} = \lambda \quad (\text{A.24})$$

Por tanto, para que $\boldsymbol{\alpha}^\top\mathbf{G}\boldsymbol{\alpha}$ sea máximo, $\boldsymbol{\alpha}$ debe estar asociado al mayor valor propio de \mathbf{G} .

En los pasos 5 y 6, se realiza una proyección ortonormal de \mathbf{C}_α y se aplica des-

composición QR, respectivamente, para obtener la matriz \mathbf{U} refinada de la siguiente iteración.

Por último, las p características relevantes se seleccionan como los elementos de \mathbf{M} que cumplan $\sum_{i=1}^p \alpha_i^2 \approx N/100$, para un criterio del $N\%$ de varianza acumulada.

A.6 $Q - \alpha$ supervisado

En la versión supervisada del algoritmo $Q - \alpha$, la matriz \mathbf{C} puede indicar la afinidad intra y entre clases empleando las etiquetas, es decir:

$$\mathbf{C}_\alpha^{gh} = \sum_{i=1}^q \alpha_i \mathbf{m}_i^g \mathbf{m}_i^{h\top} \quad (\text{A.25})$$

donde g y h son indicadores de las clases.

Entonces, si se tienen k clases se deben formar $n_\alpha : n_\alpha = \binom{k}{2} + k$ matrices. Estas matrices se analizan en el siguiente problema de optimización:

$$\max_{\alpha, \mathbf{U}^{gh}} \sum_{l=1}^k \text{tr}(\mathbf{U}^{ll\top} \mathbf{C}_\alpha^{ll} \mathbf{C}_\alpha^{ll} \mathbf{U}^{ll}) - \gamma \sum_{g \neq h} \text{tr}(\mathbf{U}^{gh\top} \mathbf{C}_\alpha^{gh} \mathbf{C}_\alpha^{gh} \mathbf{U}^{gh}) \quad (\text{A.26})$$

$$\text{s. a. } \mathbf{U}^{gh\top} \mathbf{U}^{gh} = \mathbf{I}_p, \quad \alpha^\top \alpha = 1 \quad (\text{A.27})$$

El anterior problema de optimización captura los dos objetivos que generan un buen agrupamiento, desde el punto de vista topológico: Maximiza la afinidad intra clase (parte izquierda de la función) y minimiza la afinidad entre clases (parte izquierda de la función), todo sujeto a una condición de ortonormalidad. El parámetro de regularización γ se puede determinar por validación cruzada $\gamma = 0.5$.

La solución de (A.26) puede ser obtenida a través del algoritmo 7, generando un vector α por cada matriz, es decir, $\Delta = (\alpha_1, \dots, \alpha_{n_\alpha})$. Dada la naturaleza aditiva de

la función objetivo, el vector resultante puede ser obtenido así:

$$\boldsymbol{\alpha} = \sum_{l=1}^k \boldsymbol{\alpha}^{ll} - \gamma \sum_{g \neq h} \boldsymbol{\alpha}^{gh} \quad (\text{A.28})$$

Al final, dado que $\boldsymbol{\alpha}$ debe ser unitario ($\boldsymbol{\alpha}^\top \boldsymbol{\alpha} = 1$), es necesario aplicar la siguiente normalización en amplitud:

$$\boldsymbol{\alpha} \leftarrow \frac{1}{\|\boldsymbol{\alpha}\|} \boldsymbol{\alpha} \quad (\text{A.29})$$

donde $\|\cdot\|$ denota la norma euclídea.

De otro modo, el problema de optimización planteado en (A.26) se puede interpretar como:

$$\max_{\boldsymbol{\alpha}} \sum_{l=1}^k \text{tr}(\boldsymbol{\alpha}^\top \mathbf{G}^{ll} \boldsymbol{\alpha}) - \gamma \sum_{g \neq h} \text{tr}(\boldsymbol{\alpha}^\top \mathbf{G}^{gh} \boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^\top \boldsymbol{\mathcal{G}} \boldsymbol{\alpha} \quad (\text{A.30})$$

donde $G_{ij}^{gh} = (\mathbf{m}_i^g \mathbf{m}_j^g) \mathbf{m}_i^h \mathbf{m}_j^h \mathbf{U}^{gh} \mathbf{U}^{gh \top}$ y $\boldsymbol{\mathcal{G}} = \sum_{l=1}^k \mathbf{G}^{ll} - \gamma \sum_{g \neq h} \mathbf{G}^{gh}$.

De este modo, se puede aplicar directamente el algoritmo 7, empleando la matriz auxiliar $\boldsymbol{\mathcal{G}}$.

A.7 Otra medida de relevancia

Antes se discutió que el problema de optimización asociado a PCA, podría plantearse como maximizar el complemento del error cuadrático medio de los datos originales y su proyección ortonormal (ver ecuación (A.5)), es decir:

$$\mathbf{E} \left\{ \left(\sum_{i=1}^p c_{li} v_{li} \right)^\top \left(\sum_{i=1}^p c_{li} v_{li} \right) \right\} = \mathbf{E} \left\{ \sum_{i=1}^p \lambda_i u_{li}^2 \right\} \quad (\text{A.31})$$

donde \mathbf{V} y \mathbf{U} son las matrices resultantes de la descomposición en valores singulares de \mathbf{X} : $\mathbf{X} = \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^\top$, donde $\boldsymbol{\Sigma} = \text{Diag}(\boldsymbol{\sigma})$ y $c_{li} = \sigma_i u_{li}$.

Lo anterior corresponde a la combinación lineal de los elementos del vector propio i al cuadrado donde los valores propios son los pesos. La demostración conduce a

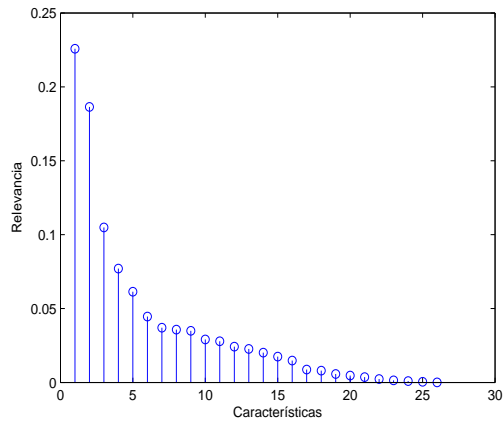
que una medida de relevancia puede escribirse como $\sum_{i=1}^p \lambda_i(\mathbf{u}_i^\top)^2$, donde $(\mathbf{u}_i)^2$ es un vector donde cada elemento es el cuadrado del elemento asociado de \mathbf{u}_i . Esto comprueba que cada elemento de los componentes principales representa el aporte de cada característica. Adicionalmente, los valores propios se asocian a la varianza acumulada de cada componente. Con esto, la medida de relevancia se puede establecer como:

$$\boldsymbol{\rho} = \sum_{i=1}^q \lambda_i(\mathbf{u}_i^\top)^2 \quad (\text{A.32})$$

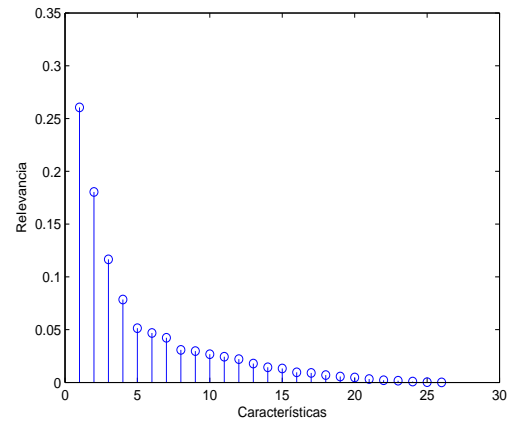
El j -ésimo elemento de $\boldsymbol{\rho}$ (ρ_j) indica la relevancia de la j -ésima característica. Este vector también puede ser utilizado para rotar los datos en un esquema de WPCA, de forma que $w_i = \rho_i$.

En la Figura A.1 se muestran los valores de relevancia obtenidos de cada uno de los métodos mencionados al analizar una base de datos experimental de 23 características.

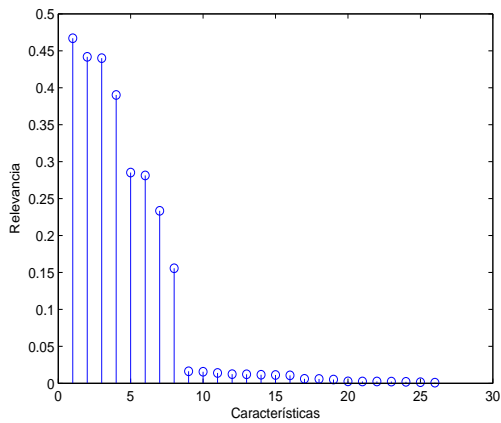
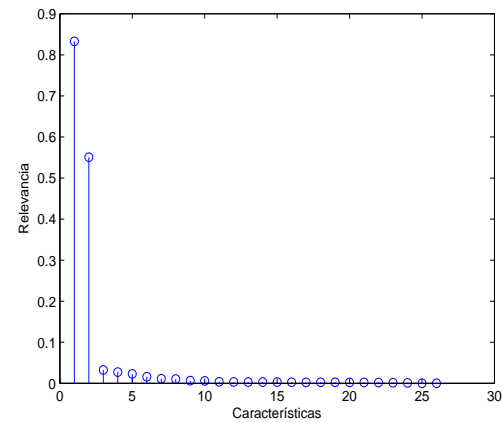
Se puede apreciar que el método $Q - \alpha$ no supervisado es el más estricto o *fuerte* al momento de seleccionar las características, como se muestra en la Figura A.1, esto se debe a que la relevancia la mide el primer (correspondiente al mayor) vector propio de la matriz auxiliar, y por esta razón la variabilidad tiende a concentrarse radicalmente en los primeros elementos. La selección de un número de característica podría ser una ventaja del método $Q - \alpha$, si los resultados de clasificación son mejores o por lo menos iguales a los que se obtienen con las características seleccionadas por otros métodos.



(a) PCA clásico



(b) WPCA

(c) $Q - \alpha$ supervisado(d) $Q - \alpha$ no supervisado**Figura A.1:** Valores de relevancia de los diferentes métodos

B. Caracterización del complejo QRS usando el modelo de Hermite

En este Apéndice se presenta una metodología para la reconstrucción y caracterización de los complejos QRS empleando el modelo paramétrico de Hermite. Los complejos son extraídos de la base de datos MIT/BIH, con un tamaño de ventana fijo y empleando las anotaciones de los registros. La reconstrucción se realiza aplicando el valor óptimo del parámetro de escala de las bases de Hermite obtenido mediante la minimización de la disimilitud de la señal original y la reconstruida. Se emplea DTW como medida de disimilitud (ver Apéndice C). Adicionalmente, se explica un método para obtener la cantidad mínima de bases que generan una reconstrucción con alta confiabilidad, basado en la comparación de los espectros de frecuencia en el rango de 1 – 20 Hz.

B.1 Modelo paramétrico de Hermite

El polinomio de Hermite H de orden n es una solución de la ecuación diferencial

$$\varphi''(z) - 2z\varphi'(z) + 2n\varphi(z) = 0$$

donde n es un entero no negativo. Con esto se puede definir los polinomios de Hermite como:

$$H_n(z) = (-1)^n e^{z^2} \frac{d^n}{dz^n} e^{-z^2} \quad (\text{B.1})$$

Los polinomios de Hermite forman un conjunto ortonormal con respecto a la función peso e^{-z^2} , es decir que:

$$\frac{1}{\sqrt{2^n n! \sqrt{\pi}}} \langle e^{-z^2} H_n(z), H_m(z) \rangle = \delta_{m,n}$$

donde $\delta_{m,n}$ es el delta de Kronecker y $\langle \cdot, \cdot \rangle$ es el producto interno.

Haciendo $z = \frac{t}{\sigma}$ es posible construir bases de la forma:

$$\phi_n^\sigma(t) = \frac{e^{-t^2/2\sigma^2}}{\sqrt{2^n \sigma n! \sqrt{\pi}}} H_n(t/\sigma) \quad (\text{B.2})$$

donde σ es el parámetro de escala (ver Fig. B.2).

A (B.2) se le conoce como modelo paramétrico de Hermite.

Los coeficientes de Hermite asociados a las señal $x(t)$ se obtienen con:

$$C_n^\sigma = \frac{1}{F_s} \int_{t=-\infty}^{\infty} x(t) \phi_n^\sigma(t) dt \quad (\text{B.3})$$

Por último, la reconstrucción de la señal sería:

$$x(t) = \sum_{n=0}^{\infty} C_n^\sigma \phi_n^\sigma(t) \quad (\text{B.4})$$

B.2 Método de reconstrucción y caracterización

B.2.1 Extracción del complejo QRS

La extracción de los complejos QRS se realiza empleando las anotaciones de la base de datos del MIT que se encuentran sobre el pico R, así:

$$\mathbf{c}_i = \mathbf{y}(p_i - \alpha F_s : p_i + \beta F_s) = \mathbf{x}_i \quad (\text{B.5})$$

donde p_i es la ubicación del pico R del i -ésimo latido del registro \mathbf{y} y F_s es la frecuencia de muestreo.

La extracción se hace de manera que el ancho de la ventana sea de 200 ms y el pico R estuviese centrado. Por tanto $\alpha = \beta = 0.1$ s. Adicionalmente, se debe aplicar una normalización de manera que:

$$\mathbf{x} \leftarrow \frac{\mathbf{x} - \mu(\mathbf{x})}{\max |\mathbf{x}|} \quad (\text{B.6})$$

B.2.2 Reconstrucción de la señal

Para la reconstrucción de la señal se puede aplicar la ecuación recursiva de los polinomios de Hermite, como sigue:

$$H_n(z) = 2zH_{n-1}(z) - 2(n-1)H_{n-2}(z) \quad (\text{B.7})$$

con $H_0 = 1$ y $H_1 = 2z$.

Las bases se extienden en el intervalo $(-t_0, t_0)$ con $t_0 = 100$ ms. Luego, para ajustar la longitud de la base a la de la señal se emplea el vector de tiempo

$$t = -t_0 : \frac{2t_0}{L_{QRS}-1} : t_0$$

donde L_{QRS} es la longitud de los complejos (200 ms).

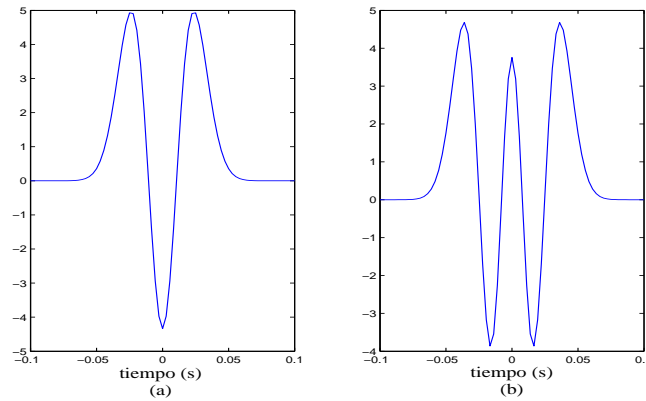


Figura B.1: Ejemplos de Bases de Hermite obtenidas con $\sigma = 15$ ms (a) $n = 2$ (b) $n = 4$

El parámetro de escala σ se agrega para que la ventana se ajuste al ancho del QRS, como se aprecia en la Figura B.2.

Los coeficientes de Hermite se obtienen con la forma discreta de (B.3), asumiendo que los términos por fuera del intervalo $(-t_0, t_0)$ son cero:

$$C_n^\sigma = \frac{1}{F_s} \sum_{i=-t_0}^{t_0} x(i) \cdot \phi_n^\sigma(i) = \frac{1}{F_s} \langle \mathbf{x}, \phi_n^\sigma \rangle \quad (\text{B.8})$$

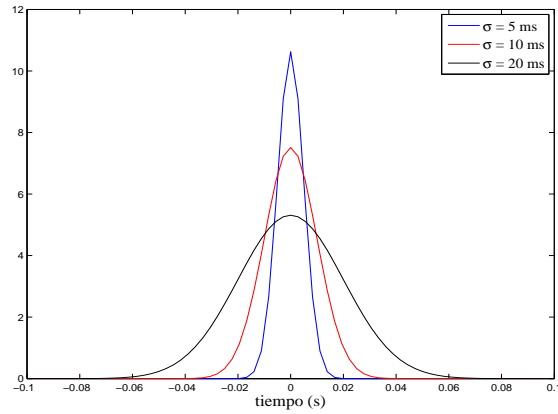


Figura B.2: Primera base de Hermite ($n = 0$) obtenida con diferentes valores de σ

Con esto, la reconstrucción se puede escribir como

$$x(t) = \sum_{n=0}^{N-1} C_n^\sigma \phi_n^\sigma(t) + \xi(t) = \hat{x}_N^\sigma(t) + \xi(t) \quad (\text{B.9})$$

donde $\hat{x}_N^\sigma(t)$ es la señal reconstruida empleando las N primeras bases y $\xi(t)$ es el factor de truncamiento.

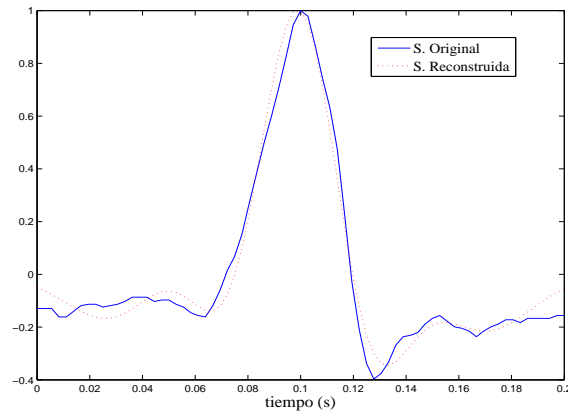


Figura B.3: Reconstrucción con las primeras 6 bases de Hermite con $\sigma = 25$ ms (Latido 20 del registro 200)

La señal discreta $\hat{\mathbf{x}}_N^\sigma$ se normalizó aplicando (B.6).

B.2.3 Comparación de los espectros

En esta etapa se estudia el cambio del espectro de la señal reconstruida considerando diferentes valores de N ($N \in (3, 20)$). Para esto se compara el espectro de frecuencia de la señal original con el espectro de la señal reconstruida ($\hat{\mathbf{x}}_n^\sigma$) en el rango de 1 – 20 Hz con el fin de obtener el orden adecuado, es decir, la cantidad mínima de bases (N_{min}) que deben considerarse para que la pérdida de información sea despreciable en comparación con los resultados de la reconstrucción. En esta etapa se debe emplear un valor de σ fijo, por ejemplo, $\sigma = 25$ ms es un valor aconsejable.

La densidad espectral de potencia se estima empleando periodograma [86]:

$$S(e^{j\omega}) = \frac{1}{n} \left\| \sum_{l=1}^n x_l e^{j\omega l} \right\|^2 \quad (\text{B.10})$$

La diferencia de los espectros se realiza aplicando:

$$diff_N = \frac{1}{F} \sum_{f=2}^F |S_f(\mathbf{x}) - S_f(\hat{\mathbf{x}}_N^\sigma)| \quad (\text{B.11})$$

donde $F = 20$ Hz y $N \in (3, 20)$.

B.2.4 Búsqueda de σ óptimo

El valor óptimo del parámetro de escala (σ_{opt}) se obtiene aplicando una medida de disimilitud entre la señal original y su reconstrucción. En este caso, se emplea el método básico de alineamiento temporal no lineal DTW. El algoritmo DTW se empleó como se sugiere en [6], sin aplicar restricciones globales.

Considerando que un valor de σ menor a 5 ms o mayor a 100 ms no es necesario para la reconstrucción de los complejos QRS, la obtención de σ_{opt} se puede plantear como el siguiente problema de optimización:

$$\begin{aligned} \min_{\sigma} \text{dtw}(\mathbf{x}_i, \hat{\mathbf{x}}_{N_{min}}^\sigma) \\ S.T. \sigma \in (5, 100)\text{ms} \end{aligned} \quad (\text{B.12})$$

B.3 Características resultantes

Con esta metodología se puede formar el siguiente conjunto de características:

- Energía del QRS:
La morfología de las arritmias ventriculares sugiere que la energía es una característica adecuada:

$$E(\mathbf{x}) = \sum_{i=1}^{L_{QRS}} x_i^2$$

- σ_{opt}
- C_σ^n con $n = 6$.
- Diferencia de \mathbf{x}_i y un complejo plantilla \mathbf{x}_{temp} aplicando (B.11), donde:

$$\mathbf{x}_{temp} = \mu(\mathbf{x}_i) \forall i$$

En todos los casos, se comprueba que, con $n \geq 6$ la diferencia del espectro de la señal original y el espectro de la señal reconstruida con $\sigma = 25$ ms es pequeña (ver Figura B.5), y la reconstrucción es muy aproximada (ver Figura B.9).

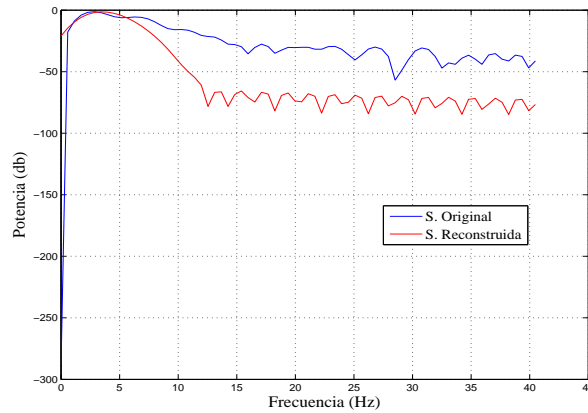


Figura B.4: Espectro de la reconstrucción empleando las 3 primeras bases (Latido 20 del registro 200)

El valor admisible para la diferencia de los espectros de la señal reconstruida y la original (B.11) se encuentra en el rango $\max|diff_n| \leq 5$.

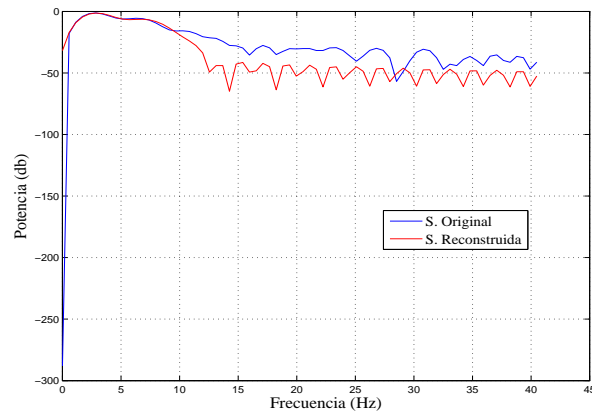


Figura B.5: Espectro de la reconstrucción empleando las 6 primeras bases (Latido 20 del registro 200)

Con esta metodología se logra reducir el espacio de búsqueda del valor óptimo del parámetro de escala σ_{opt} del modelo paramétrico de Hermite a través de la minimización de la disimilitud de la señal reconstruida y la original y, a la vez, se reduce el costo computacional estableciendo la cantidad mínima de bases necesarias para la reconstrucción, que genere la menor pérdida de información espectral. La medida de disimilitud DTW genera buenos resultados porque considera factores morfológicos.

C. Alineamiento temporal no lineal

DTW

El método DTW (*Dynamic time warping*), es un algoritmo de alineamiento temporal no lineal, que se usa comúnmente en la normalización en longitud y en la estimación de la disimilitud entre dos señales. El algoritmo DTW se aplicó en muchas secciones a lo largo de este estudio, por eso se ha destinado este Apéndice para dar una breve descripción.

En el agrupamiento no supervisado y, en general, en los métodos de clasificación de patrones, la elección de la función de disimilitud no es menos importante que la elección de características. Las opciones más inmediatas son la norma L_1 y L_2 . Sin embargo, en algunas ocasiones las señales o vectores representativos son de longitudes diferentes y estas medidas no pueden aplicarse directamente. Por ejemplo, en el análisis de señales ECG, el proceso de extracción de latidos genera señales de diferentes longitudes porque se extraen a partir de la distancia entre picos, que es una variable dinámica. Entonces, en estos casos, debe hacerse una normalización temporal previa, que consiste en ajustar las longitudes de las señales a una mismo valor, para que los vectores puedan ser comparados a través de alguna de las medidas mencionadas. La normalización puede llevarse a cabo con métodos lineales y no lineales. En la normalización temporal lineal se escala el vector de menor longitud para que los dos vectores a comparar tengan la misma longitud. Así, para calcular la disimilitud entre dos señales se deben repetir o interpolar muestras de la señal de longitud menor hasta que coincida con la longitud de la otra señal. De esta manera, se puede interpretar los vectores como elementos de un espacio vectorial y se puede aplicar medidas como la distancia euclídea. Aunque, en el análisis de algunas señales, como las de ECG, al aplicar una normalización temporal lineal se produce un desplazamiento que puede generar resultados inadecuados para el cálculo de un valor de disimilitud, debido al mal ajuste [6].

El método de alineamiento temporal no lineal DTW, mejora las deficiencias de la normalización temporal lineal y se basa en ajustar únicamente la longitud de los tramos de las señales que resulte más conveniente desde el punto de vista de su comparación, sin alterar el resto. Todo esto se logra en un esquema de programación dinámica. Por esta razón el algoritmo DTW es de uso frecuente en la clasificación de muestras de longitudes diferentes. La programación dinámica se puede implementar con una multitud de variaciones de acuerdo a su aplicación. Esta técnica se ha aplicado en reconocimiento automático del habla [87], tanto a palabras aisladas como a palabras conectadas [88]. También se ha empleado como medida de disimilitud para el agrupamiento de latidos en señales ECG [89].

C.1 Algoritmo DTW

El valor de la disimilitud se obtiene en el último elemento de una matriz, denominada matriz de programación dinámica \mathbf{G} , que se denota con $G[n_x - 1, n_y - 1]$, donde n_x y n_y son las longitudes de los vectores \mathbf{x} e \mathbf{y} a comparar. Este valor debe normalizarse respecto a la longitud del camino de alineamiento para evitar el efecto de la longitud. La matriz \mathbf{G} tiene tantas filas como la longitud del primer vector y tantas columnas como la longitud del segundo vector, como se muestra en la Figura C.1. Esta matriz se construye de forma que cada uno de sus elementos o nodos, represente el costo acumulado del alineamiento entre los vectores hasta llegar a ese punto.

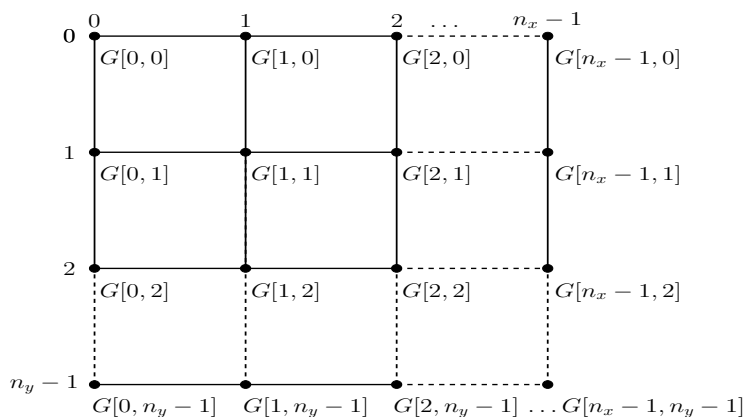


Figura C.1: Matriz de programación dinámica \mathbf{G} asociada a los vectores x e y

La matriz de programación dinámica es el plano discreto en donde se traza el camino de alineamiento. El objetivo del algoritmo consiste en encontrar el camino de

alineamiento de costo mínimo entre los dos vectores. Este camino se definirá como el conjunto ordenado de nodos que se encuentran desde el nodo inicial (i_1, j_1) hasta el nodo final (i_N, j_N) , de la forma: $(i_1, j_1), (i_2, j_2) \dots (i_N, j_N)$.

Para establecer el costo del camino, es necesario definir el costo debido a la transición entre cada par de nodos. La función de costo asociada a la k -ésima transición se denota como $d[(i_{k-1}, j_{k-1})|(i_k, j_k)]$. El costo debe ser un valor positivo y se puede medir a través de la norma L_1 o L_2 .

El camino de alineamiento debe satisfacer las siguientes condiciones:

- *Monotonicidad:* El camino debe tener el sentido de izquierda a derecha, y de arriba a abajo de la matriz \mathbf{G} , es decir, se debe cumplir que $i_k \leq i_{k+1}$ y $j_k \leq j_{k+1}$.
- *Continuidad:* El paso de un nodo al siguiente debe hacerse de modo que no hayan nodos intermedios: $i_k - i_{k-1} \leq 1$ y $j_k - j_{k-1} \leq 1$.
- *Frontera:* $i_1 = j_1 = 0$, $i_N = n_x - 1$ y $j_N = n_y - 1$. Con esta condición se establece que el principio y el final de los vectores deben coincidir con ciertos puntos independientemente del camino seguido.
- *Ventana:* $|i_k - j_k| = \Delta$, donde Δ es un entero positivo.
- *Pendiente:* Se aplican restricciones sobre las pendientes del camino para evitar la búsqueda en zonas donde es difícil realizar un buen ajuste.

El costo asociado a un camino completo puede calcularse con la suma del costo de cada una de las transiciones:

$$D = \sum_{k=1}^K d[(i_{k-1}, j_{k-1})|(i_k, j_k)] \quad (\text{C.1})$$

donde K representa la cantidad de nodos del camino.

Adicionalmente, es necesario aplicar un factor de ponderación a las transiciones. Este factor se relaciona con el conjunto de transiciones o producciones y representa la longitud correspondiente a la transición [6]. El costo acumulado ponderado es de la

forma:

$$D = \sum_{k=1}^K d[(i_{k-1}, j_{k-1})|(i_k, j_k)]w(a_k, b_k) \quad (\text{C.2})$$

donde $w(a_k, b_k)$ representa el factor de ponderación debido a la transición de (i_k, j_k) a (i_{k-1}, j_{k-1}) .

Finalmente, se realiza una normalización con respecto de la longitud, así:

$$D = \frac{\sum_{k=1}^K d[(i_{k-1}, j_{k-1})|(i_k, j_k)]w(a_k, b_k)}{\sum_{k=1}^K w(a_k, b_k)} \quad (\text{C.3})$$

Haciendo $N = \sum_{k=1}^K w(a_k, b_k)$, se tiene que

$$D = \frac{G[n_x - 1, n_y - 1]}{N} \quad (\text{C.4})$$

La expresión recursiva para la obtención de cada elemento de la matriz \mathbf{G} es:

$$G[i, j] = \min_{a,b} \sum_P \sum_{k=1}^K d[(i_{k-1}, j_{k-1})|(i_k, j_k)]w(a_k, b_k) \quad (\text{C.5})$$

donde $G[i - a, j - b]$ representa el costo mínimo acumulado considerando el conjunto de producciones P . En la Figura C.2 se muestran algunos conjuntos de producciones de uso común en programación dinámica, indicando el factor de ponderación correspondiente de cada transición [87].

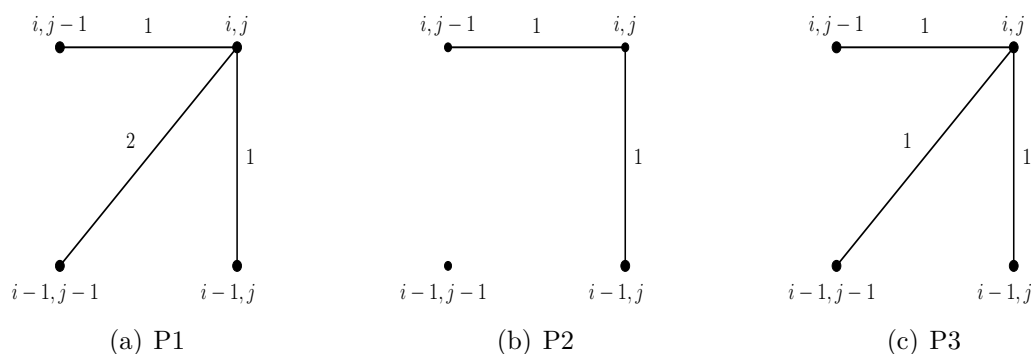


Figura C.2: Diferentes producciones para las transiciones

Utilizando el conjunto de producciones P1, la matriz \mathbf{G} se obtiene con:

$$G[i, j] = \min \begin{cases} G[i, j - 1] + d[(i, j)|(i, j - 1)] \\ G[i - 1, j] + d[(i, j)|(i - 1, j)] \\ G[i - 1, j - 1] + 2d[(i, j)|(i - 1, j - 1)] \end{cases} \quad (\text{C.6})$$

donde el nodo inicial es $G[0, 0] = |x[0] - y[0]|$. Con esto la disimilitud entre los dos vectores es:

$$D = \frac{G[n_x - 1, n_y - 1]}{n_x - 1 + n_y - 1} \quad (\text{C.7})$$

Por lo general, se aplican restricciones globales al camino de alineamiento para reducir el costo computacional (ver Figura C.3). Para esto se establece una región válida \mathfrak{R} , de forma que los elementos que se encuentren fuera de esta región se les asigna un costo muy grande en comparación con los posibles costos del resto, para que el camino no los tenga en cuenta. Esto es $G[i, j] = \infty$, si $(i, j) \notin \mathfrak{R}$.

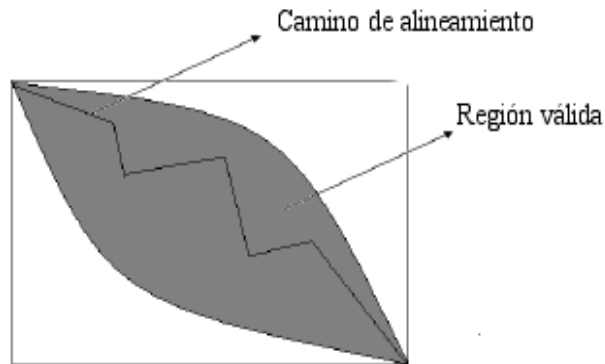


Figura C.3: Restricciones globales de DTW

En [76] se presenta un método para la obtención de una región válida óptima, definida con rectas paralelas alrededor de la diagonal principal de la matriz \mathbf{G} .

D. Discusión académica

Con el desarrollo de este trabajo de tesis se obtuvo la siguiente producción académica.

2009:

- [1]. José L. Rodríguez, Diego Peluffo, David Cuesta, Germán Castellanos. “*Non-parametric Density-Based Clustering for Cardiac Arrhythmia Analysis.*”. En: Computers in Cardiology. CINC, 09. Utah, USA.
- [2]. Diego Peluffo, José L. Rodríguez, Germán Castellanos. “*Clasificación de arritmias ventriculares empleando clustering particional y características morfológicas del QRS.*”. En: XIV Simposio de Tratamiento de Señales, Imágenes y Visión Artificial. STSIVA, 2009. Pereira, Colombia.
- [3]. José L. Rodríguez, David Cuesta, Diego Peluffo, Germán Castellanos. “*Un-supervised feature selection in cardiac arrhythmias analysis.*”. En: Engineering in Medicine and Biology Conference. EMBC, 2009. Minneapolis, USA.
- [4]. Diego Peluffo, José Luis Rodríguez, German Castellanos. “*Unsupervised classification of cardiac arrhythmias using partitional clustering.*”. Encuentro Nacional de Investigación en Posgrados. Universidad Nacional de Colombia. ENIP, 2009.
- [5]. Diego Peluffo, Carolina Ospina, Gerard Olivard Tost, José L. Rodríguez, Germán Castellanos. “*Caracterización de señales ECG empleando un modelo del pulso cardíaco y dinámica no lineal: Una prueba piloto.*”. Semana Técnica de Ingenierías Eléctrica y Electrónica. Universidad Nacional de Colombia, sede Manizales.
- [6]. Diego Peluffo, José Luis Rodríguez, Germán Castellanos. “*Estudio comparativo de métodos de selección de características de inferencia supervisada y no supervisada.*”. Revista: Tecno-Lógicas. ITM - Medellín.
- [7]. Diego Peluffo, José Luis Rodríguez, Germán Castellanos. “*Diseño de un sistema de clasificación no supervisada basado en análisis espectral.*”. Revista: Ingeniería e Investigación. Universidad Nacional de Colombia.

- [8]. Diego Peluffo, José Luis Rodríguez, Germán Castellanos. “*Clasificación de arritmias en registros Holter empleando análisis no supervisado de tipo partitional y espectral*”. En: V Seminario Internacional de Ingeniería Biomédica (SIB). Universidad de los Andes, Bogotá.

2008:

- [1]. Diego Peluffo, José L. Rodríguez, Germán Castellanos. “*Metodología para la reconstrucción y extracción de características del complejo QRS basada en el modelo paramétrico de Hermite*”. En: V Semana Técnica de Ingenierías Eléctrica y Electrónica. Universidad Nacional de Colombia. Manizales, Colombia.
- [2]. Diego Peluffo, Edilson Delgado, José L. Rodríguez, Germán Castellanos. “*Optimización del algoritmo DTW basada en la Naturaleza Estadística de las Disimilitudes*”. En: XIII Simposio de Tratamiento de Señales, Imágenes y Visión Artificial. STSIVA, 2008. Bucaramanga, Colombia.
- [3]. Diego Peluffo, José L. Rodríguez, Germán Castellanos. “*Detección de Arritmias de tipo Bloqueo de Rama mediante Análisis no Supervisado y Morfología del QRS*”. En: III Congreso de Bioingeniería e Ingeniería Biomédica. Pereira, Colombia.

Bibliografía

- [1] E. Braunwald. *Tratado de Cardiología*, volume 1. 4 edition, 1993.
- [2] J. L. Rodríguez. Revisión de técnicas para agrupamiento no supervisado de latidos ECG. Universidad Nacional de Colombia, sede Manizales.
- [3] P. Hansen and N. Mladenovic. J-means: a new local search heuristic for minimum sum of squares clustering. *Pattern Recognition*, pages 405 – 413, 2001.
- [4] A. Guyton and J. E. Hall. *Tratado de fisiología médica*. 10 edition.
- [5] G. B. Moody and R. G. Mark. The MIT-BIH arrhythmia database on CD-ROM and software for use with it. *Computers in Cardiology*, 1999.
- [6] D. Cuesta. *Estudio de métodos para procesamiento y agrupación de señales electrocardiográficas*. PhD thesis, Universidad Politécnica de Valencia, 2001.
- [7] J. L. Rodríguez. Estudio comparativo de técnicas para la extracción de parámetros de señales electrocardiográficas. Master’s thesis. Universidad Nacional de Colombia, Sede Manizales.
- [8] M. L. Ahlstrom and J. Tompkins. Digital filters for real-time ECG signal processing using microprocessors. *IEEE Transactions on Biomedical Engineering*, 32(9):708–713, 1985.
- [9] M. Okada. A digital filter for the QRS complex detection. *IEEE Transactions on Biomedical Engineering*, 26(12):700–703, 1979.
- [10] A. V. Oppenheim and R.W. Schaffer. *Discrete-Time Signal Processing*. Prentice- Hall International, 1989.
- [11] N. V. Thakor, J.G. Webster, and W.J Tompkins. Estimation of QRS complex power spectra for design of a QRS filter. *IEEE Transactions on Biomedical Engineering*, 31(11):702–706, 1984.
- [12] P. Laguna, R. Jané, O. Meste, P. W. Poon, P. Caminal, H. Rix, and N. V. Thakor. Adaptive filter for event-related bioelectric signals using an impulse correlated reference input: Comparison with signal averaging techniques. *IEEE trans. biomed. Eng.*, 39(10):1032–1044, October 1992.

-
- [13] O. Sayadi and M. Shamsollahi. ECG denoising with adaptive bionic wavelet transform. In *28th IEEE EMBS Annual International Conference*, New York city, New York, USA, August-September 2006. IEEE, The Printing House, Inc.
- [14] De-noising by soft-thresholding. Technical report, 1996. Dept. of Statistics, Stanford University.
- [15] D. Donoho. De-noising by soft-thresholding. *IEEE Trans. Information Theory*, 41(3):612–627, 1995.
- [16] J. Madeiro, P. Cortez, F. Oliveira, and R. Siqueira. A new approach to QRS segmentation based on wavelet bases and adaptive threshold technique. *Medical Engineering and Physics*, 2006.
- [17] J. S. Sahambi, S. N. Tandon, and R. Bhatt. Using wavelet transform for ECG characterization. *IEEE Transactions engineering in medicine and biology*, pages 77–88, january 1997.
- [18] S. Kadambe, R. Murray, and G. F. Boudreaux-Bartels. Wavelet transform-based QRS complex detector. *IEEE Trans. on Biomed. Eng.*, 46(7):838–848, JULY 1999.
- [19] R. Almeida, S. Gouveia, P. Rocha, P. Lago, and F. De Freitas. Wavelet transform based matlab system for the detection and delineation of QRS complexes in ambulatory ECG recordings. in *BioEng 2001 - 6th Portuguese Conference on Biomedical Engineering*, 2001.
- [20] J. Pan and W.J. Tompkins. A real-time detector algorithm. *IEEE Trans. on Biom. Eng.*, 32(3):230–236, March 1985.
- [21] Natalia Arzeno, Chi-Sang Poon, and Zhi-De Deng. Quantitative analysis of QRS detection algorithms based on the first derivative of the ECG. In *28th IEEE EMBS Annual International Conference*, pages 1788–1791, New York city, New York, USA, August-September 2006. IEEE, The Printing House, Inc.
- [22] N. Hughes, S. Roberts, and L. Tarassenko. Semi-supervised learning of probabilistic models for ECG segmentation, 2005.
- [23] F. Jager. Feature extraction and shape representation of ambulatory electrocardiogram using the Karhunen-Löve transform. *Electrotechnical Review*, 69(2):83–89, march 2002.
- [24] T. Stamkopoulos, K. Diamantaras, N. Maglaveras, and Michael Strintzis. ECG analysis using nonlinear PCA neural networks for ischemia detection. *IEEE Transactions On Signal Processing*, 46(11):3058–3066, Noviembre 1989.
- [25] H. Haraldsson, L. Edenbrandt, and M. Ohlsson. Detecting acute myocardial infarction in the 12-lead ECG using hermite expansions and neural networks. Technical report, Department of theoretical Physics, Lund University, Sweden, 2002.
- [26] C. Li, C. Zheng, and C. Tai. Detection of ECG characteristic points using wavelet transforms. *IEEE Trans.*, 42(1):21–28, January 1995.

-
- [27] A. Rosado, G. Camps, J. Guerrero, JV Francés, J. Muñoz, and AJ Serrano. Enhancing feature extraction for VF detection using data mining techniques. Technical report, Grup Processament Digital de Senyals (GPDS), Universitat de Valencia, Spain, 2002.
- [28] Recommended practice for testing and reporting performance results of ventricular arrhythmia detection algorithms. Technical report, AAMI (Association for the Advancement of Medical Instrumentation), 1998.
- [29] U. Kunzmann, G. Von Wagner, J. Schochlin, and A. Bolz. Parameter extraction of ECG signals in real-time. *Biomedizinische Technik*, 47(1):875–878, 2002.
- [30] Vessela Krasteva and Irena Jekova. QRS template matching for recognition of ventricular ectopic beats. *Annals of Biomedical Engineering*, 35(12):2065 – 2076, 2007.
- [31] Vessela T. Krasteva, Irena I. Jekova, and Ivaylo I Christov. Automatic detection of premature atrial contractions in the electrocardiogram. *Electrotechniques and Electronics E and E*, 2006.
- [32] Ivaylo I Christov. Real time electrocardiogram QRS detection using combined adaptive threshold. *BioMedical Engineering OnLine*, 2004.
- [33] R. Olszewski. *Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data*. PhD thesis, School of Computer Science Carnegie Mellon University, Pittsburgh, PA 15213, 2001.
- [34] L. Sörnmo, P.O. Börjesson, M. Nygard, and O. Pahlm. A method for evaluation of qrs shape features using a mathematical model for the ECG. *IEEE Transactions on Biomedical Engineering*, 28(10), 1981.
- [35] P. Aguirre, J. Cardelino, and N. Loeff. Sistema de detección, clasificación e identificación en línea de complejos QRS. Technical report, Instituto de Ingeniería Eléctrica Facultad de Ingeniería Universidad de la República, Montevideo, Uruguay, 2002.
- [36] D. Peluffo, J. L. Rodríguez, and G. Castellanos. Detección de arritmias de tipo bloqueo de rama mediante análisis no supervisado y morfología del QRS. 2008.
- [37] D. Peluffo, J. L. Rodríguez, and G. Castellanos. Metodología para la reconstrucción y extracción de características del complejo QRS basada en el modelo paramétrico de hermite. *V Semana Técnica de Ingenierías Eléctrica y Electrónica*, 2008.
- [38] Kei ichiro Minami, Hiroshi Nakajima, and Takeshi Toyoshima. Real-time discrimination of ventricular tachyarrhythmia with fourier transform neural network. *IEEE Transactions on Biomedical Engineering*, pages 179 – 185, 1999.
- [39] J. García and P. Laguna. Diseño y evaluación de un detector de isquemia basado en la transformada de Karhunen - Løeve. *Grupo de tecnología de las telecomunicaciones*.
- [40] S. Olmos, J. García, R. Jané, and P. Laguna. Análisis de la señal ECG con expansiones ortogonales de reducido número de coeficientes. España, 1997. En Comunicaciones del XV Congreso Anual de la Sociedad Española de Ingeniería Biomédica.

-
- [41] Lior Wolf and Amnon Shashua. Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach. *Journal of machine learning*, 6:1855 – 1887, 2005.
- [42] R. Andreato, B. Dorizzi, and J. Boudy. ECG signal analysis through hidden Markov models. *IEEE Transactions on Biomedical Engineering*, 53(8):1541–1549, 2006.
- [43] C. Torrence and G. Compo. A practical guide to wavelet analysis. *Bull. Am. Meteor. Soc.*, 79(1):61–78, 1998.
- [44] Dean Cvetkovic, Elif Derya Ubeyli, and Irena Cosic. Wavelet transform feature extraction from human PPG, ECG, and EEG signal responses to elf pemf exposures: A pilot study. *ELSEVIER*, 2007.
- [45] Chia-Hung Lin, Yi-Chun Du, and Tainsong Chen. Adaptive wavelet network for multiple cardiac arrhythmias recognition. *Expert systems with applications. ELSEVIER*, 2008.
- [46] Szi-Wen Chen. A wavelet-based heart rate variability analysis for the study of non-sustained ventricular tachycardia. *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING*, 49(7), 2002.
- [47] M. H. Song and K. J. Lee. Real-time classification of heartbeats using least square acceleration filter for ambulatory monitoring. *Computers in cardiology*, 2006.
- [48] Chia-Hung Lin. Frequency-domain features for ECG beat discrimination using grey relational analysis-based classifier. *ELSEVIER*, 2007.
- [49] Seedahmed S. Mahmoud, Qiang Fang, Dragomir M. Davidovic, and Irena Cosic. A time-frequency approach for the analysis of normal and arrhythmia cardiac signals.
- [50] Xusheng Zhang and Yisheng Zhu. Detection of irregularity index series for QRS complex of ECG. institute of biomedical engineering, department of instrumentation. *Proceedings of ICSP*, 1996.
- [51] Stephane Mallat and Wen Liang Hwang. Singularity detection and processing with wavelets. *IEEE transaction on information theory*, page 617, 1992.
- [52] M. Lagerholm, C. Peterson, G. Braccini, L. Edenbrandt, and L. Sörnmo. Clustering ECG complexes using hermite functions and self-organising maps. *IEEE trans. on. Biomed.*, 48:838–847, 2000.
- [53] Y. Özbay, R. Ceylan, and B. Karlik. A fuzzy clustering neural network architecture for classification of ECG arrhythmias. *Computers in Biology and Medicine*, 36:376–388, 2006.
- [54] S. Osowski, L. Hoai, and T. Markiewicz. Support vector machine based expert system for reliable heart beat recognition. *IEEE trans. on. Biomed.*, 61:1–7, 2002.
- [55] Y. Hu, S. Palreddy, and W. Tompkins. A patient adaptable ECG beat classifier using a mixture of experts approach. *IEEE trans. on. Biomed.*, 44:891–900, 1997.

-
- [56] P. Chazal, M. O'Dwyer, and R. Reilly. Automatic classification of hearbeats using ECG morphology and heartbeat interval features. *IEEE trans. on. Biomed.*, 51:1196–1206, 2004.
- [57] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):1855 – 1887, 1999.
- [58] Greg Hamerly and Charles Elkan. Alternatives to the k-means algorithm that find better clusterings. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pages 600–607, New York, NY, USA, 2002. ACM.
- [59] D. Cuesta, J. Pérez-Cortés, and G. Andreau-García. Clustering of ECG signals in computer-aided holter analysis. *Computer Methods and Programs in Biomedicine*, 72:179–196, 2003.
- [60] B. Köhler, C. Henning, and R. Orglmeister. The principles of software QRS detection: reviewing and comparing algorithms for detecting this important ECG waveform. *IEEE tran. on Biomed.*, 21:42–57, 2002.
- [61] D. Cuesta, M. Biagetti, R. Quintero P. Micó-Tormos, and M. Aboy. Unsupervised detection of ventricular extrasystoles using bounded clustering algorithms and morphology matching. *IEEE tran. on Biomed.*, 2006.
- [62] M. Paoletti and C. Marchesi. Discovering dangerous patterns in long-term ambulatory ECG recordings using a fast QRS detection algorithm and explorative data analysis. *Computer Methods and programs in biomedicine*, 82:20–30, 2006.
- [63] B. Zhang, M. Hsu, and U. Dayal. K-harmonic-means, a data clustering algorithm. Technical report, Hewlett Packard Labs, 1999.
- [64] K. Chang, R. Lee, C. Wen, and M. Yeh. Comparison of similarity measures for clustering ECG complexes. In *32 annual international conference on computers in cardiology*, Lyon, France, September 2005. INSERM, UCBL, INSA, HCL.
- [65] P. K. Chan, M. D. F. Schlag, and J. T. Zien. K-way ratio cut partitioning and clustering. *IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems*, 13:1088 – 1096, 1994.
- [66] Bruce Hendrickson and Robert Leland. An improved spectral graph partitioning algorithm for mapping parallel computations. *SIAM Journal on Scientific Computing*, 16(2):452–469, 1995.
- [67] H. Ishikawa and D. Geiger. Segmentation by grouping junctions. In *CVPR '98: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, page 125, Washington, DC, USA, 1998. IEEE Computer Society.
- [68] Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, Aug 2000.

-
- [69] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, 2001.
- [70] Yu Stella X. and Shi Jianbo. Multiclass spectral clustering. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 313, Washington, DC, USA, 2003. IEEE Computer Society.
- [71] Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means, spectral clustering and normalized cuts. 2004.
- [72] Desmond Higham and Milla Kibble. A unified view of spectral clustering. In *Mathematics Research Report 2*, University of Strathclyde, 2004.
- [73] Lihi Zelnik-manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17*, pages 1601–1608. MIT Press, 2004.
- [74] Bin Zhang. Generalized k-harmonic means - boosting in unsupervised learning. *Software technology laboratory HP*, 2000.
- [75] Robert Breunig. Nonparametric density estimation for stratified samples. *The Australian National University working papers in economics and econometrics*, (459), 2005. Center for Economic Policy Research, Economics Program, Research School of Social Sciences, The Australian National University, Canberra ACT 0200, AUSTRALIA.
- [76] D. Peluffo, E. Delgado, J. L. Rodríguez, and G. Castellanos. Optimización del algoritmo DTW basada en la naturaleza estadística de las disimilitudes. 2008.
- [77] Bui The Anh and D. D. X. Thanh. A perron-frobenius theorem for positive quasipolynomial matrices associated with homogeneous difference equations. *Journal of applied mathematics*, 2007.
- [78] Pak K. Chan, Martine Schlag, and Jason Y. Zien. Spectral k-way ratio-cut partitioning and clustering. In *Proceedings of the 1993 symposium on Research on integrated systems*, pages 123–142, Cambridge, MA, USA, 1993. MIT Press.
- [79] Marina Maila and Jianbo Shi. Learning segmentation with random walk. In *Neural Information Processing Systems (NIPS)*, 2001.
- [80] Sangkeun Lee and M.H. Hayes. Properties of the singular value decomposition for efficient data clustering. *Signal Processing Letters, IEEE*, 11(11):862–866, Nov. 2004.
- [81] J. Rodríguez-Sotelo, D. Cuesta-Frau, and G. Castellanos-Domínguez. Unsupervised classification of atrial heartbeats using a prematurity index and wave morphology features. *Medical and Biological Engineering in biomedicine*, 2008.
- [82] M. Lagerholm, C. Peterson, G. Braccini, L. Edenbrandt, and L. Sörnmo. Clustering ECG complexes using hermite functions and self-organising maps. *IEEE trans. on. Biomed*, 48:838–847, 2000.

-
- [83] Jonathon Shlens. A tutorial on principal component analysis. april 2009.
- [84] Hui Y. Wang and Xiao J. Wu. Weighted PCA space and its application in face recognition. In *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, volume 7, 2005.
- [85] H. Henry Yue and Masayuki Tomoyasu. Weighted principal component analysis and its applications to improve fdc performance. In *Conference on decision and control*, 2004.
- [86] Pierre Dutilleul. Multi-frequential periodogram analysis and the detection of periodic components in time series. *Communications in statistics*, 2001.
- [87] F. Casacuberta and E. Vidal. *Reconocimiento automático del habla*. 1987.
- [88] M. Brown and L. Rabiner. Dynamic time warping for isolated word recognition based on ordered graph searching techniques. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82.*, volume 7, pages 1255–1258, May 1982.
- [89] David Cuesta. Aplicación de la normalización temporal no lineal al alineamiento de señales electrocardiográficas. 1999. Trabajo de investigación de 6 créditos.

