ESCUELA TÉCNICA SUPERIOR DE INGENIEROS
INDUSTRIALES Y DE TELECOMUNICACIÓN

UNIVERSIDAD DE CANTABRIA

Proyecto Fin de Carrera

TÉCNICAS ÓPTICAS APLICADAS A LA MONITORIZACIÓN HEMODINÁMICA CEREBRAL EN BEBÉS

(Optical techniques applied to monitoring cerebral hemodynamics in infants)

Para acceder al Título de

INGENIERO DE TELECOMUNICACIÓN

Autor: Víctor Pérez Maza

Octubre - 2012

AGRADECIMIENTOS

A mis padres y mis hermanos, que siempre han estado apoyándome cuando lo he necesitado y a los que debo mis virtudes y casi ninguno de mis defectos.

A mi abuela, con la que he convivido durante mis años de carrera. Gracias a ella solo me he tenido que preocupar de mis estudios.

A todos los compañeros que han estado a mi lado durante todo este camino y con los que he compartido experiencias difícilmente olvidables.

Gracias a todos.

Tabla de contenido

1 Introducción **1**

1.1 Motivación 1

1.2 Objetivos 1

1.3 Organización de la Memoria 1

2 Estado del Arte 3

2.1 Introducción 3

2.2 Monitorización cerebral en neonatos 3

2.2.1Tecnologías aplicadas no invasivas 4

2.3 Espectroscopía en el infrarrojo cercano 7

2.3.1 Contexto de la NIRS 7

2.3.2 Aplicaciones 8

2.3.3 NIRS para la monitorización hemodinámica (Oximetría cerebral) 9

2.3.4 Espectroscopía óptica difusa 11

2.3.5 Espectroscopía de correlación difusa 12

2.3.6 Estudios experimentales de la NIRS aplicados a la monitorización cerebral 13

2.4 Minería de Datos 14

2.4.1 Introducción 14

2.4.2 Aplicación en la medicina 15

2.4.3 Series temporales 16

3 Técnicas generales para el análisis de series temporales **18**

3.1 Introducción 18

3.2 Suavizado o filtrado 18

3.3 Wavelets 22

3.4 Reducción de la dimensionalidad 25

3.4.1 Análisis de Componentes Principales 25

3.4.2 PAA – PieceWise Aggregate Approximation 26

3.5 Clasificación de patrones 27

4 Diseño de la técnica del estudio **32**

4.1 Suavizado o Smoothing 32

4.2 Normalización 33

4.3 Reducción de la dimensionalidad 35

4.3.1 Symbolic Aggregate Approximation (SAX) 35
4.4 Principio MDL para el descubrimiento del patrón 37
4.5 Modelo de agrupamiento de patrones 39

5 Resultados de los test evaluados 44
5.1 Serie Simple (Dificultad Baja) 44
5.1.1 Longitud del patrón 45
5.1.2 Conclusiones 46
5.2 Prueba Victoria-1(Dificultad Media) 47
5.2.1 Estudio de los resultados 48
5.2.2 Conclusiones-Prueba Victoria-1 52
5.3 Prueba Victoria-2 (Dificultad Media) 53
5.3.1 Estudio de los resultados 54
5.3.2 Conclusiones-Prueba Victoria-2 56
5.4 Prueba EDAT (Dificultad Media-Alta) 57
5.4.1 Estudio de los resultados 59
5.4.2 Conclusiones-EDAT 61
5.5 Prueba ECG (Prueba real) 61
5.5.1 Medida aVR 63
5.5.1.1 Estudio de los resultados de aVR 63
5.5.2 Medida I 64
5.5.2.1 Estudio de los resultados de I 65
5.5.3 Medida V2 66
5.5.3.1 Estudio de los resultados de V2 67
5.5.4 Conclusiones 67
5.6 Prueba Real (Muestra obtenida mediante NIRS) 68
5.6.1 Oxihemoglobina 69
5.6.2 Desoxihemoglobina 72
5.6.3 Hemoglobina Total 74
5.6.4 Conclusiones de los resultados de las muestras de la NIRS 76
6 Conclusiones y Líneas futuras 77

6.1 Conclusiones 77
6.2 Líneas futuras 78
7 Referencias **80**

1. INTRODUCCIÓN

1.1. Motivación

El control de los parámetros hemodinámicos cerebrales es un objetivo prioritario en el ámbito de la medicina en las últimas décadas. Numerosas investigaciones se centran en el estudio de dichos elementos con técnicas de procesamiento de datos como el reconocimiento de patrones. Sin embargo, el uso de dichas técnicas de exploración de la hemodinámica cerebral viene limitado por una serie de consideraciones generales prioritarias: inocuidad de la técnica, facilidad de aplicación y precisión de la medida.

El hecho de que mediante técnicas de minería de datos y reconocimiento de patrones podamos entender los distintos daños cerebrales que se pueden producir, ayuda a anticiparse a hechos posteriores y por lo tanto dar un tratamiento óptimo al paciente.

Este Trabajo Fin de Carrera intenta aportar una solución al procesamiento y posterior estudio de los datos hemodinámicos cerebrales mediante técnicas de reconocimiento de patrones, de forma que se pueda adquirir un conocimiento de dichas secuencias. Debido a la multitud de documentación que existe al respecto en este campo, es necesario realizar primero un estudio profundo del estado de arte y analizar las distintas propuestas al respecto para apoyar la decisión de la técnica final aplicada.

1.2. Objetivos

El objetivo de este proyecto es realizar un sistema de reconocimiento de patrones que permita conocer características repetitivas relevantes en un conjunto de datos de carácter biomédico. Tras presentar algunas de los métodos más comunes aplicados en series temporales multivariantes, el estudio se centrará en la explicación detallada de la técnica aplicada. Una vez se conozca su funcionamiento, se probará dicho sistema con series temporales en las que se incrementará la dificultad de los patrones introducidos hasta la utilización de secuencias hemodinámicas reales obtenidas mediante espectroscopia en el infrarrojo cercano.

1.3. Organización de la memoria

La memoria constará de los siguientes capítulos

- ***Capítulo 1***: Introducción: motivación, objetivos y organización de la memoria.
- ***Capítulo 2:*** Estado del Arte: repaso a trabajos previos en técnicas de monitorización hemodinámica y sus limitaciones.
- ***Capítulo 3:*** Técnicas generales para el análisis de series temporales: explicación de las diferentes técnicas para el tratamiento de señales de carácter temporal multivariante.
- ***Capítulo 4:*** Diseño de la técnica del estudio: explicación de las diferentes etapas que contiene la solución propuesta que han sido implementadas en el estudio.
- ***Capítulo 5:*** Resultados de los test evaluados: exposición de los resultados obtenidos de las diferentes secuencias de entrada al sistema.
- ***Capítulo 6:*** Conclusiones y líneas futuras: a partir de las pruebas y resultados realizados, se exponen las conclusiones de las mismas y posibles líneas futuras de trabajo.

2. ESTADO DEL ARTE

2.1. Introducción

La Biología moderna ha tenido una importante evolución en las últimas décadas debido a las múltiples opciones que se tienen para poder obtener información de las diferentes patologías. Esto provoca que se tengan que analizar un conjunto de datos muy amplios, de ahí la necesidad de técnicas computacionales que permitan el estudio de dichos elementos de una manera rápida y eficaz. De esta manera nace la Bioinformática como la aplicación de las Matemáticas y la Informática en el tratamiento de datos biológicos. Este campo se encarga de la recogida, mantenimiento, distribución, análisis y uso de las inmensas cantidades de información biológica disponible.

Las investigaciones biológicas basadas en la luz, especialmente en la luz láser, están en pleno auge. Con la ayuda de los conocimientos biológicos, matemáticos y fotónicos se pueden evaluar órganos y células con un detalle antes inalcanzable. Este hecho ayuda en el diagnóstico, seguimiento y terapia en campos médicos como la oncología o la neurología.

La espectroscopía en el infrarrojo cercano proporciona un método no invasivo para medir de forma continua el grado de oxigenación de los tejidos, el metabolismo y el flujo de sangre en el cerebro. Estos datos permiten estudiar el desarrollo del cerebro en bebés con sensores ópticos en forma de casco. Se trata de tener una alternativa a la resonancia magnética funcional, que permite visualizar las áreas del cerebro que se activan con cada función realizada y que no se pueden aplicar a los niños pequeños. Algo que tiene especial importancia como en el caso del autismo. Las aplicaciones terapéuticas, todavía en estado experimental, especialmente e bebés prematuros en cuidados intensivos, y otras no terapéuticas, como las relacionadas con la oxigenación muscular en atletas de élite, son el fruto de los grandes avances recientes en instrumentación y métodos de análisis.

Focalizando en la lesión cerebral en niños prematuros, este hecho representa un importante problema debido a la gran cantidad de bebés que nacen cada año y presentan secuelas neurológicas. De ahí que la neonatología haya sufrido un notable desarrollo en las últimas décadas al verse aumentada su demanda, hecho que en principio puede parecer paradójico, pues coincide con un descenso de la natalidad. La explicación a este fenómeno radica en varias razones: control de la natalidad, con mayor exigencia en el bienestar materno-neonatal, incremento del número de recién nacidos (RN) con un bajo peso que requieren cuidados y vigilancia intensiva, incremento del número de embarazos múltiples… , lo cual ha empujado a la creación de Unidades de Cuidados Intensivos Neonatales (UCIN) en muchos de los hospitales.

Este capítulo se encargará de aportar una información general de la técnicas existentes no invasivas para la monitorización de bebés con problemas al nacer, haciendo más hincapié en la espectroscopía en el infrarrojo cercano ya que es el sistema por el que se han recogido los datos que se analizarán finalmente en el estudio.

2.2. Monitorización cerebral en neonatos

La monitorización del Sistema Nervioso Central ha experimentado un importante avance en los últimos 40 años. Todo ello conlleva a un mejor conocimiento en la fisiopatología

del daño cerebral, que a su vez permite un diagnóstico y tratamiento más acorde en cada caso. Los diferentes métodos de monitorización cerebral tienen como objetivo común la detección de la isquemia cerebral. La monitorización del flujo sanguíneo y metabolismo cerebral permite instaurar un tratamiento adecuado.

La regulación del flujo sanguíneo cerebral es un factor determinante entre los mecanismos implicados en el desarrollo de la lesión cerebral neonatal, siendo por ello objetivo prioritario de muchos investigadores en las últimas décadas, con el fin de encontrar técnicas fiables de medición de los parámetros hemodinámicos cerebrales. Sin embargo, el uso de diferentes técnicas de exploración de la hemodinámica cerebral vienen limitado por una serie de consideraciones generales prioritarias: la inocuidad de la técnica, su facilidad de aplicación y la fiabilidad de la medida.

Las técnicas de monitorización se pueden organizar en dos grupos: las invasivas y las no invasivas. La diferencia entre ambas radica en que las "no invasivas" no requieren de daño físico alguno para su aplicación.

2.2.1. Tecnologías aplicadas no invasivas

El desarrollo de la Neurología Neonatal se ha visto favorecido por la conjunción de la aparición de innovaciones tecnológicas (ecografía doppler-color, resonancia magnética nuclear con espectroscopía, espectroscopía en el infrarrojo cercano, análisis espectral con EEG digital…), introducción de marcadores bioquímicos de lesión cerebral y excelentes aportaciones de autores que han permitido una mejor evaluación neurológica de los recién nacidos de alto riesgo y su valoración pronóstica. A continuación se dará una breve explicación de los métodos más comunes.

ÍNDICE BIESPECTRAL [1]

El índice biespectral (BIS) es un número que evalúa el grado de hipnosis al estimar el nivel de actividad eléctrica cerebral mediante el análisis de las frecuencias de las ondas del electroencefalograma (EEG). Desarrollado fundamentalmente para controlar la hipnosis durante la cirugía, ha empezado a utilizarse en los pacientes críticos, aunque todavía existe poca experiencia en niños.

El BIS estima el grado de actividad eléctrica cerebral, y por lo tanto, el de sedación del paciente mediante el análisis de las frecuencias de las ondas del EEG (% frecuencias rápidas / % frecuencias lentas). La información del EEG se obtiene a través de un sensor que se coloca en la frente del paciente. Su valor puede oscilar entre 0 y 100; 0 en el caso de supresión completa del EEG y 100 cuando el paciente está completamente despierto.

Figura 1: Aplicación del sistema BIS a un paciente pediátrico crónico. (DM)

ULTRASONOGRAFÍA CEREBRAL [2]

En las últimas décadas el incremento de la supervivencia de los neonatos de bajo peso, así como de los niños con asfixia prenatal, ha suscitado un interés creciente por el desarrollo del conocimiento de la neurología neonatal. La ultrasonografía cerebral (USC) es la técnica que probablemente más ha contribuido a ese desarrollo, tanto desde el punto de vista clínico-asistencial como investigador.

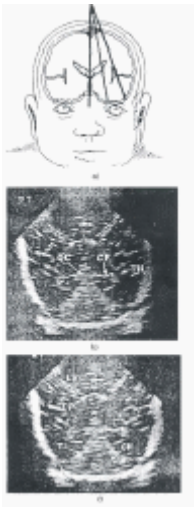


Figura 2: Análisis cerebral mediante ultrasonografía

DOPPLER TRANSCRANEAL [3] [4]

Esta técnica no invasiva permite la medición del flujo sanguíneo cerebral gracias a la emisión de ondas sonoras de baja frecuencia que atraviesan la barrera ósea craneal.

Christian Doppler introdujo en 1842 lo que hoy conocemos como el "Principio de Doppler", según el cual las señales emitidas por una fuente de ultrasonido chocan y se reflejan en un objeto en movimiento, siendo las frecuencias de las señales reflejadas directamente proporcionales a la velocidad de dicho objeto. Si los objetos en movimiento son hematíes circulantes en el interior de vasos sanguíneos, la determinación de su velocidad puede proporcionar una estimación indirecta del valor del flujo sanguíneo.

Inicialmente la sonografía de Doppler se utilizó para el estudio de vasos periféricos y, con posterioridad, para la valoración de las arterias carótidas, utilizando ultrasonido de frecuencias de entre 2-20 MHz. Estas frecuencias resultaron idóneas para el cálculo de velocidades de flujo de las grandes arterias extra-craneales. No obstante, en su aplicación craneal el rango utilizado no era válido. Para ello se introdujo un sistema que emitía pulsos a menores frecuencias (entre 1 y 2 MHz) lo que permitía la lectura de velocidades en las grandes arterias intracraneales.

La mayor desventaja que tiene el Doppler Transcraneal es que su precisión diagnóstica depende sensiblemente de la posición del instrumento emisor/detector de las ondas sonoras.

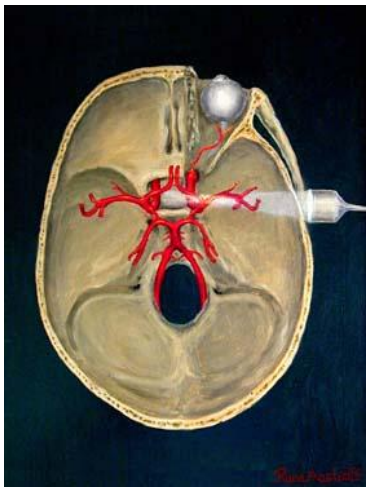


Figura 3: Doppler Transcraneal

TOMOGRAFÍA POR EMISIÓN DE POSITRONES (PET) [5]

La tomografía por emisión de positrones (PET) es una herramienta que evalúa la función cerebral a través de la administración de radiofármacos. La PET es anterior a la resonancia y proporciona información sobre el metabolismo de la glucosa y flujo sanguíneo cerebral en varias enfermedades neurológicas. Además, la PET también puede ser utilizada para la evaluación de la distribución regional de receptores y neurotransmisores aunque su uso se produce generalmente en investigación científica. En la práctica neurológica, la PET es utilizada predominantemente en el diagnóstico y manejo de pacientes pediátricos con epilepsia y enfermedades oncológicas.

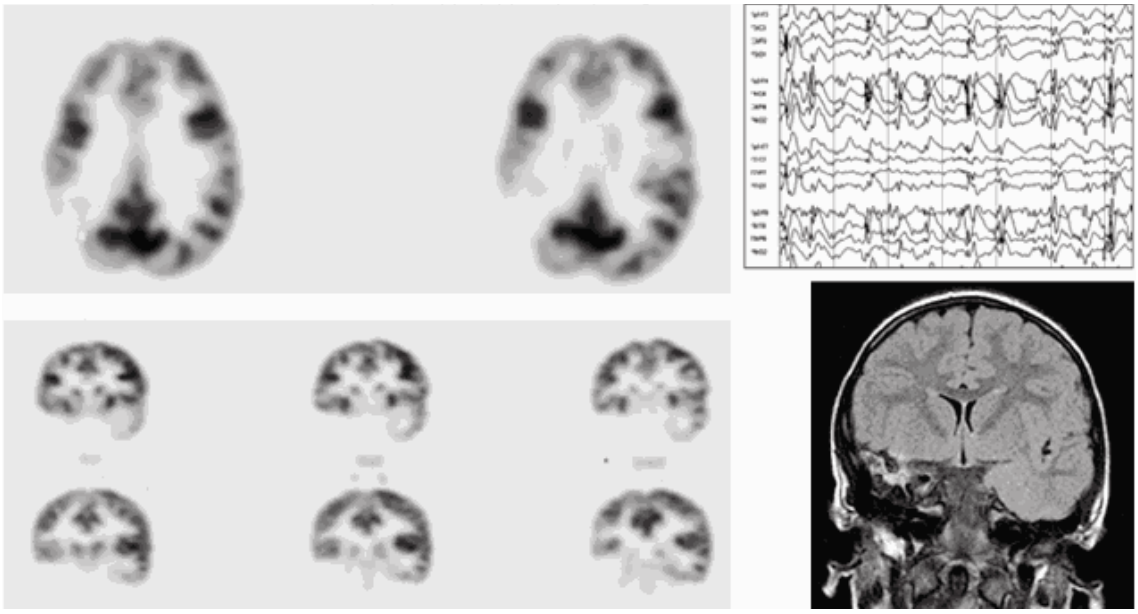


Figura 4: Niña de 6 años de edad con encefalomalacia en la región temporal derecha secundaria a encefalitis herpética (observe lesión temporal derecha en la resonancia magnética de cráneo). FDG-PET muestra hipometabolismo en el lóbulo temporal derecho y región inferior del lóbulo frontal y parietal a la derecha. El EEG muestra estado de mal eléctrico durante el sueño.

SPET CEREBRAL [6]

El SPET cerebral es una técnica que utiliza un radio-trazador para obtener información referente al flujo y volumen sanguíneo cerebral. Varios estudios han objetivado una excelente correlación entre la reactividad cerebrovascular determinada por SPET y por Doppler transcraneal. Sin embargo, a pesar de su uso extendido en la reserva hemodinámica en pacientes con estenosis carotidea, el SPET es una técnica radioactiva, y requiere de dos exploraciones (basal y acetazolamida) espaciadas en el tiempo para el estudio de la reactividad cerebrovascular.

ESPECTROSCOPÍA EN EL INFRARROJO CERCANO

La espectroscopía cercana al infrarrojo (NIRS) es una técnica no invasiva que ha sido utilizada en diferentes protocolos experimentales para el estudio de los cambios hemodinámicos y metabolismo del oxígeno a nivel cerebral y sistémico. Como esta técnica es la que se ha utilizado en el presente estudio, se describirá más en profundidad a continuación.

2.3. Espectroscopia en el infrarrojo cercano

La espectroscopía en el infrarrojo cercano (Near Infrarred Spectroscopy, NIRS) ha alcanzado un gran desarrollo a nivel mundial por su precisión y exactitud. Se trata de una técnica que no destruye ni contamina y tiene, además, potencialidades para ser automatizada. Su práctica permite analizar, de forma rápida y relativamente fácil, un gran número de muestras. Numerosos estudios demuestran la eficacia de esta técnica y sus ventajas frente a otras en la detección precoz del daño cerebral sin intervención quirúrgica. Investigadores de la Universidad de Pensilvania probaron la técnica en cerdos a los que habían provocado daños cerebrales de distinta consideración. Demostraron la capacidad de la técnica para el análisis y la discriminación de los daños cerebrales en los cerdos.

En el presente apartado se pretende dar a conocer las ventajas de este sistema y sus potencialidades para la evaluación del estado neurológico en el que se encuentra un paciente. Primeramente se dará una idea del uso del NIRS en la medicina y posteriormente se explicarán detalles más técnicos acerca de su instrumentación.

2.3.1. Contexto de la NIRS

Este proyecto nace de la idea de buscar relaciones concluyentes entre el flujo sanguíneo y la presión sanguínea cerebral en bebés prematuros. Encontrar correlaciones entre estos dos parámetros puede ser un paso importante para evaluar el estado de un bebé prematuro ya que se podría determinar daños potenciales sabiendo el comportamiento de estas dos variables bajo distintas patologías. Esto permitiría tener una rápida reacción frente a determinados daños que puedan sufrir pacientes neonatos.

La NIRS aparece como una técnica capaz de satisfacer una serie de requerimientos importantes: es inocua, se puede aplicar en la cabeza del paciente, y permite la monitorización continua de la saturación de la sangre cerebral, así como de la utilización del oxígeno por parte de las células. La técnica permite supervisar en tiempo real la concentración de

oxihemoglobina (HbO), desoxihemoglibina (Hb), hemoglobina total (THb) y el citocromo aa3 (cytaa3), siendo esta última la enzima que cataliza más del 95% del total del oxígeno utilizado en la célula [7].

Siguiendo las directrices del principio de Fick, y utilizando el oxígeno como trazador, puede obtenerse por NIRS cuantificaciones absolutas del flujo sanguíneo cerebral (FCS). A partir de ese parámetro se puede calcular la cesión de oxígeno a los tejidos.

La medición del volumen sanguíneo cerebral (VSC) puede realizarse en términos absolutos, introduciendo un cambio en la saturación del oxígeno (lo que obliga a la medición intermitente del mismo), o de forma continua, por medio de la monitorización de los cambios en la THb a lo largo del tiempo, obteniéndose entonces una cuantificación absoluta de las variaciones del volumen a partir de un punto inicial de medida.

Además de los parámetros señalados, por medio de la NIRS se puede calcular la saturación cerebral venosa (CSvO); a partir de la cual permite obtener la extracción cerebral de oxígeno, así como su la tasa metabólica cerebral (conociendo el FCS). Todos estos parámetros posibilitan la evaluación del acoplamiento entre el flujo sanguíneo cerebral y las necesidades metabólicas.

En bebés prematuros con estado de circulación presión-pasiva, esta técnica podría ser de gran ayuda en la prevención de la lesión cerebral, dada la gran cantidad de elementos hemodinámicos y metabólicos que permite monitorizar. Además, por medio de la NIRS se podrá ayudar a obtener criterios de intervención en el recién nacido prematuro ya que está basada en medidas directas de la oxigenación cerebral.

2.3.2. Aplicaciones

La espectroscopía infrarroja es una técnica que permite explorar las características de los tejidos aplicando luz en el infrarrojo. Entre las técnicas utilizadas cabe destacar la espectroscopía de absorción, la cual permite evaluar la composición de un material.

Dentro de las tres posibilidades de estudio en el infrarrojo, el documento se va a centrar en el infrarrojo cercano (NIRS), que se extiende aproximadamente entre las longitudes de onda 780 – 3000 µm. El uso de este rango de longitudes se utiliza para el análisis cuantitativo de compuestos que contengan agrupaciones funcionales con hidrógenos unidos a carbonos, nitrógenos y oxígenos. Estos compuestos se pueden determinar a menudo con exactitudes y precisiones bastante altas.

PRINCIPALES APLICACIONES DE LA NIRS [8]

REGIONES ESPECTRALES	TIPO DE MEDIDA	TIPO DE ANÁLISIS	TIPO DE MUESTRAS
INFRARROJO CERCANO	Reflectancia difusa Absorción	Cuantitativa Cuantitativa	Materiales comerciales sólidos o líquidos Mezclas gaseosas
INFRARROJO MEDIO	Absorción Reflectancia Emisión	Cualitativa Cuantitativa Cromatográfico Cualitativa Cuantitativa	Compuestos sólidos, líquidos o gaseosos puros Mezclas complejas de gases, líquidos y sólidos Mezclas complejas de gases, líquidos o sólidos Compuestos sólidos, líquidos o gaseosos puros Muestras atmosféricas
INFRARROJO LEJANO	Absorción	Cualitativa	Especies inorgánicas puras organometálicas

2.3.3. NIRS para la monitorización hemodinámica (oximetría cerebral)

La oximetría cerebral es una técnica no invasiva que permite la monitorización de los cambios en el metabolismo cerebral del oxígeno.

El método se basa en la emisión de fotones cercanos al infrarrojo (NIR) en la piel de la frente del paciente. Después del sufrir el fenómeno de dispersión por el tejido capilar, cráneo y cerebro, parte de estos fotones se ven “rebotados” por el efecto de la reflectancia. Al medir la cantidad de fotones que regresan se puede inferir la absorción espectral del tejido subyacente y extraer conclusiones sobre su oxigenación media [10].

Al dispersarse por los tejidos, los fotones que no son reflejados son absorbidos, produciendo unas señales características en el espectro de luz emergente. El cromóforo con mayor absorción en el tejido corporal es la hemoglobina, cuyo espectro de absorción varía según su estado de oxigenación.

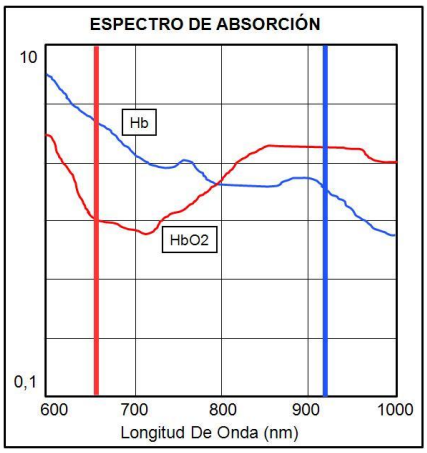


Figura 5: Espectro de absorción de la *HbO2* y la *HB* para las longitudes de onda de 660 y 920 nm.[9]

La medición selectiva del tejido cerebral puede hacerse por el principio de resolución espacial. La profundidad a la que penetran los fotones emitidos desde la piel depende de la distancia a la que se encuentre el detector. Se utilizan dos detectores situados a distancias diferentes del punto emisor: el más cercano recibe la luz del haz superficial, correspondiente a la piel, tejido celular, y el cráneo; el más lejano recibe la señal de estos tejidos más la del tejido celular subyacente. La resta de las dos señales permite obtener la correspondiente a la corteza cerebral situada debajo de los sensores.

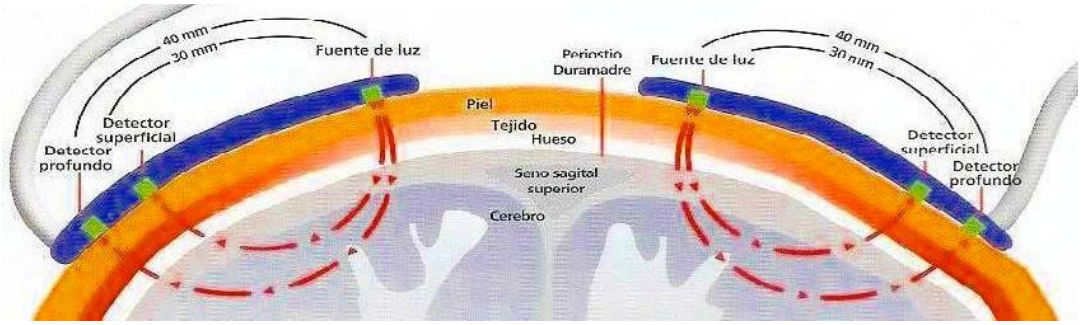


Figura 6: Esquema de funcionamiento de la oximetría cerebral

En la imagen de la figura se puede observar cómo se utilizan dos sensores a ambos lados de la línea media. En cada sensor hay un punto emisor de luz y dos puntos de detección de señal, situados a 3 y 4 cm aproximadamente del punto emisor.

La fuente de luz emite dos haces en el rango próximo al infrarrojo (730-810 nm) e ilumina el tejido. La intensidad de la luz que recibe cada detector se convierte en una señal eléctrica que se procesa y digitaliza.

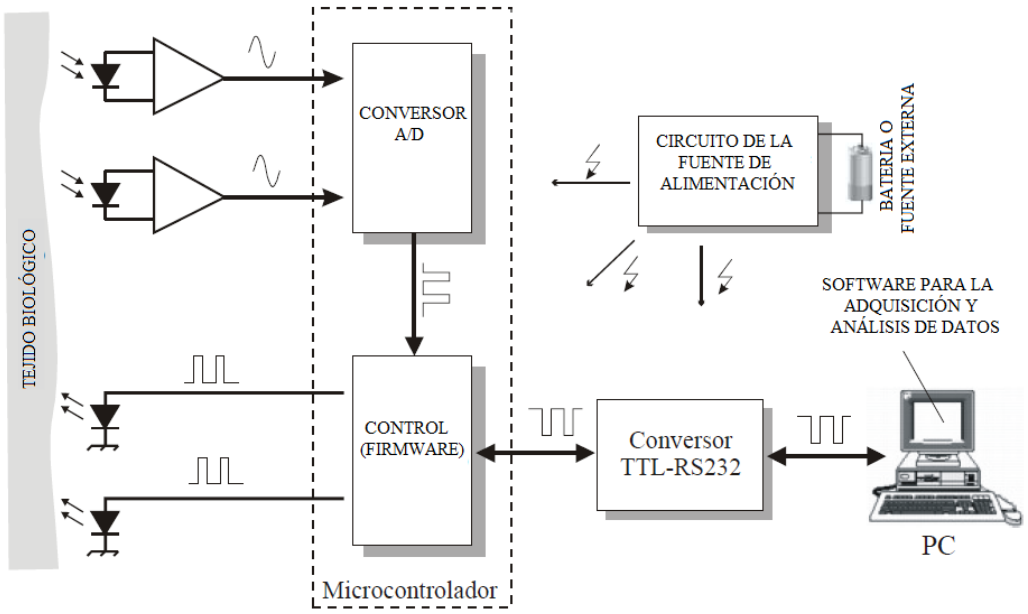


Figura 7: Esquema básico del procesamiento de señales recogidas de la espectroscopía

2.3.4. Espectroscopía óptica difusa

Al igual que la NIRS, la espectroscopia óptica difusa (*Diffuse Optical Spectroscopy - DOS*) [10] permite la exploración en profundidad de los tejidos (~ mm-cm) usando una ventana espectral de baja absorción como es la del infrarrojo cercano (700-900 nm). Sin embargo, a diferencia de la NIRS, el DOS se apoya en la teoría de difusión del fotón para extraer información más rigurosa sobre la absorción y esparcimiento o dispersión óptica en el interior de los tejidos basándose en la atenuación de la luz del tejido superficial. La DOS mide variaciones leves de la dispersión/esparcimiento y absorción cerebral y es sensible a la concentración absoluta del cromóforo, y por lo tanto, es capaz de medir valores absolutos de saturación de oxígeno cerebral (SbO2) y volumen sanguíneo cerebral (CBV).

Una configuración típica de DOS es la que se emplea en el dominio de la frecuencia. Se utiliza una fuente láser sinusoidal en el rango de los 100MHz modulada en amplitud por un oscilador de frecuencias RF. Esta señal modulada se envía a través de fibra óptica experimentando durante el trayecto una atenuación de amplitud y un desplazamiento de fase. La detección óptica y la electrónica permiten cuantificar la atenuación y el desviamiento de la fase con el fin de reconstruir la señal original.

La medida de la atenuación en amplitud es un parámetro relativamente fácil de realizar. Por lo tanto, la dificultad radica en detectar el desviamiento de fase sufrido. Debido a que dicho desplazamiento es altamente dependiente de la frecuencia a la que se module la fuente, se debe tener especial cuidado en la elección de la misma. Tanto la separación de la fuente con el detector, como la profundidad de penetración de luz en el tejido se ven afectadas por la elección de la frecuencia.

2.3.5. Espectroscopia de correlación difusa

La espectroscopia de correlación difusa (*Diffuse Correlation Spectroscopy-DCS*) [11] se encarga de medir los cambios en el flujo sanguíneo cerebral mediante la monitorización de la función de autocorrelación en el tiempo de la intensidad de la luz procedente del tejido superficial. La DCS es ideal para la población de neonatos con problemas post-parto, ya que permite una evaluación continua, no invasiva y con la posibilidad de colocar junto a la cama un monitor portátil en el que se pueda monitorizar el flujo sanguíneo cerebral en la región cortical. Además, esta técnica combinada con las anteriormente comentadas como el Doppler Transcraneal o la NIRS puede dar una información completa del estado de bebé. Algunos estudios apuntan a la combinación tanto de la DOS como de la DCS, lo cual proporcionaría medidas de oxigenación y flujo sanguíneo. Estos parámetros asociados permiten la creación de índices que permitan medir el metabolismo cerebral del paciente.

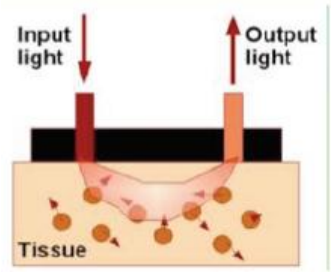


Figura 8: Representación del funcionamiento de la DCS

En la DCS, la luz coherente producida por el láser se introduce en el medio altamente dispersivo (en este caso la cabeza de paciente) en el que se introduce en profundidad produciéndose multitud de fenómenos de esparcimiento o dispersión (*scattering*) en su camino. Parte de la luz que se transmite por estos tejidos vuelve a la superficie donde es detectada por un fotodiodo. El campo eléctrico en un único punto de la superficie está constituido por la superposición de fotones provenientes de los distintos caminos que sigue la luz a lo largo de su recorrido por los tejidos. Estos campos se interfieren de forma constructiva o destructiva (dependiendo de su fase) y forman un patrón característico. Este cambio de fase se produce al entrar en contacto la luz con los distintos elementos en movimiento como pueden ser los glóbulos rojos. Por lo tanto, el patrón de fotones formado en el detector cambia velozmente en el tiempo. De esta manera la DCS es capaz de establecer el flujo sanguíneo cerebral.

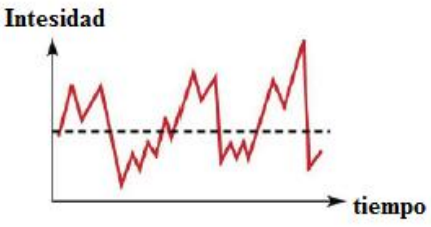


Figura 9: Ejemplo de señal temporal

2.3.6. Estudios experimentales del NIRS aplicado a la monitorización cerebral

La introducción de la espectroscopia en el infrarrojo cercano aplicado a la monitorización hemodinámica ha supuesto el desarrollo de variedad de estudios que pretenden demostrar la eficiencia de esta técnica y las ventajas con respecto a los sistemas médicos ya existentes es este área. Por este hecho se ha considerado relevante comentar algunas de las investigaciones que han servido para hacerse una idea de la importancia que podría tener la NIRS en el presente y en el futuro de la medicina para la evaluación del estado neurológico de pacientes con daños cerebrales de distinta consideración.

Monitorización de los cambios hemodinámicos en cerdos con daño cerebral severo [11]

Investigadores de la Universidad de Pensilvania llevaron a cabo pruebas con cerdos jóvenes a los que habían causado previamente daños cerebrales con el objetivo de evaluar la correlación de las medidas de la NIRS con la toma de medidas en la arteria femoral. Se estudió la presión y el flujo sanguíneo de manera continua y durante un período de tiempo significativo. Los factores que se tuvieron en cuenta en el experimento fueron la concentración de oxígeno en sangre, la concentración de hemoglobina total, la saturación de oxígeno en sangre y el flujo sanguíneo. Además se les provocó episodios críticos de apnea o paros cardíacos para la comprobación de los cambios que se producían en las medidas.

La NIRS mostró cambios transitorios en la oxigenación cerebral. La hipoxia produjo un severo compromiso hemodinámico que se correlacionó con el descenso de la oxigenación y el flujo sanguíneo cerebral. Además, las medidas obtenidas por medio del catéter insertado en la femoral y las obtenidas mediante la espectroscopía tuvieron una estrecha correlación.

Monitorización de cabecera del flujo sanguíneo cerebral para víctimas de apoplejía [12]

Los mismos investigadores de la Universidad de Pensilvania (Penn, Filadelfia, EUA) han desarrollado la espectroscopía de correlación difusa (DCS), una tecnología para la medición transcraneal no invasiva del flujo sanguíneo cerebral (FCS) que se puede hibridizar con la espectroscopía en el infrarrojo cercano. Como parte del desarrollo, la investigación examinó la utilidad de la DCS y la NIRS para medir los efectos en el flujo sanguíneo con distintas inclinaciones de la cabeza en la cama (HOB). Los ángulos que propone el estudio son 30, 15, 0 y -5 grados y se lo aplicaron a pacientes con apoplejía sistémica aguda que afectaba la corteza frontal y los controles. Se encontró que el posicionamiento HOB alteraba considerablemente el FSC, las concentraciones de oxihemoglobina (HbO2) y de hemoglobina total (HbT) [12]. Más aún, los investigadores encontraron que la presencia de un infarto ipsilateral tenía un efecto significativo sobre todos los parámetros. Se demostró que los resultados eran consistentes con la noción de la autorregulación del flujo sanguíneo en el hemisferio infartado.

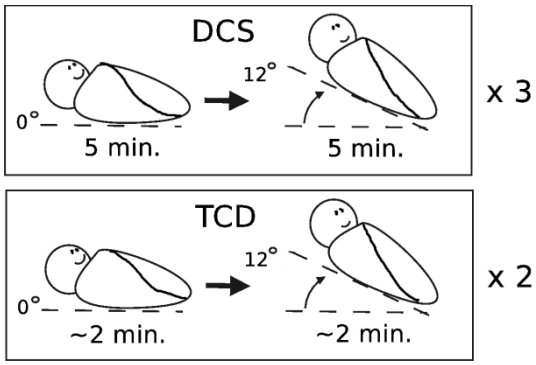


Figura 10: Protocolo para la medición del DCS

A modo de explicación breve, el sistema no invasivo emplea sondas posicionadas sobre vasos sanguíneos principales en cada hemisferio cerebral. Las sondas usan luz difusa para detectar cambios fisiológicos como el flujo sanguíneo, la saturación de oxígeno en sangre (SpO2) y la concentración de hemoglobina con la finalidad de informar a los médicos el tratamiento adecuado en cada caso. El sistema usa láseres, detectores basados en contadores de fotones, electrónica de radiofrecuencia, procesado de datos y un monitor para mostrar las imágenes que serán utilizadas por el profesional que deba evaluar el estado del paciente.

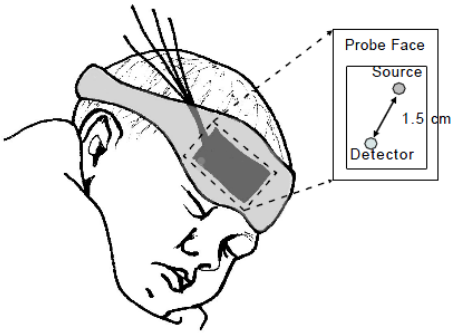


Figura 11: Instrumentación para la obtención de los datos

2.4. MINERÍA DE DATOS

2.4.1. Introducción

La minería de datos se define como [13]:

El conjunto de técnicas para la representación, análisis, manejo y descubrimiento de conocimiento a partir de diversas fuentes de datos (bases de datos, Web, archivos, sensores…). Incluye aspectos de estadística, manejo de conocimiento, computación de alto rendimiento, algoritmos genéticos, redes neuronales, sistemas de soporte a la toma de decisiones, sistemas de información, sistemas distribuidos y bases de datos. El conocimiento extraído se emplea en la toma de decisiones.

En la sociedad de la información de la que se forma parte, donde cada día se multiplican la cantidad de datos almacenados casi de forma exponencial, la Minería de Datos es una herramienta fundamental para el análisis y explotación de dichos datos de una manera eficaz.

La Minería de Datos hace uso de todas aquellas técnicas que puedan aportar información útil, desde la forma más sencilla como es el análisis gráfico, hasta aquéllos que requieren pasar por métodos estadísticos más o menos complejos. Además, estas técnicas complementadas con métodos y algoritmos del campo de la Inteligencia Artificial y el aprendizaje automático son capaces de resolver problemas más típicos de agrupamiento automático, clasificación, predicción de valores, detección de patrones, asociación de atributos, etc. Es, por lo tanto, un campo multidisciplinar que cubre numerosas áreas y se aborda desde múltiples puntos de vista, como la estadística, la informática o la ingeniería.

2.4.2. Aplicación en la medicina

Durante las últimas décadas, el campo del descubrimiento del conocimiento en bases de datos (DCDB o "Knowledge Discovery in Database", KDD en inglés) ha conseguido atraer un considerable interés por la capacidad de extraer conocimiento de bases de datos biomédicas. En ambientes de negocio tradicionales, donde la Minería de Datos es utilizada rutinariamente por analistas y administradores, es necesario un almacén de datos adecuado (en inglés *datawarehouse*) para asegurar la fiabilidad del proceso.

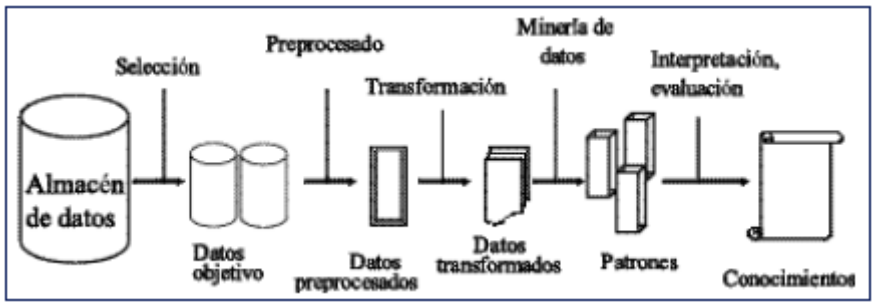


Figura 12: Modelo clásico del proceso DCDB

La mayoría de los sistemas médicos de Minería de Datos todavía no son utilizados de manera rutinaria en la práctica médica, con algunas pocas excepciones. De hecho, se puede establecer una sorprendente similitud entre sistemas médicos expertos y la Minería de Datos. De ahí que sea interesante preguntarse si la Minería de Datos está proporcionando el tipo de información y conocimiento que los profesionales esperan, ya sea para progresar en la ciencia biomédica como en las consultas médicas. Estos profesionales basan generalmente sus argumentos en mucho más que una simple base de datos o un conjunto de signos externos y síntomas. La evaluación exhaustiva de la apariencia del paciente, sus circunstancias personales, o los informes y rasgos psicológicos, además de los datos aportados por las bases de datos, son fundamentales para el diagnóstico médico y la gestión integral del paciente. Además, los médicos recopilan datos con apreciaciones subjetivas de los enfermos y usan diferentes teorías.

En las últimas décadas se han producido cambios revolucionarios en el ámbito de la investigación biomédica y la biotecnología. Las nuevas tecnologías emergentes permiten evaluar, medir y cuantificar de forma simultánea una gran diversidad de situaciones, lo que ha provocado un incremento de la cantidad de datos biológicos disponibles. Áreas que van desde la terapia genética, investigación del cáncer, detección de marcadores regulares, desarrollo de fármacos, etc., han experimentado cambios drásticos en la cantidad de información generada. Esto ha dado lugar que metodologías de análisis que venían siendo desarrolladas y aplicadas preferentemente en otras disciplinas hayan desembarcado en el campo de las ciencias de la vida.

La generación y administración de los datos experimentales no es un proceso trivial y mucho menos cuando la cantidad de datos generados empieza a crecer. Esto hace necesario tanto la generación de estándares de laboratorio como la generación de estándares de procesamiento. Además, esta generación de estándares da lugar a las consiguientes ventajas administrativas como puede ser la capacidad de intercambio de información entre distintos centros.

2.4.3. Series temporales

Los tipos de datos se pueden definir de varias maneras. Los atributos que están contenidos en las bases de datos pueden ser cualitativos o cuantitativos, además de que algunos conjuntos de datos pueden tener características especiales como son series de tiempo u otros objetos con relaciones explícitas entre ellos. Está claro que el tipo de datos determina las herramientas y técnicas que pueden ser utilizadas para analizar los mismos.

Las series de tiempo son un tipo especial de datos secuenciales donde cada registro es una serie de tiempo. Una serie de tiempo es una secuencia de datos, medidos en tiempos sucesivos espaciados por intervalos uniformes de tiempo. Por ejemplo, un conjunto de datos financieros debe contener objetos que son series de tiempo de los precios diarios de varias acciones. Por lo tanto, trabajar con series temporales implica que es importante considerar autocorrelaciones temporales. Como ejemplo se dice que si dos mediciones son cercanas en el tiempo, entonces los valores de estas mediciones a menudo son similares [14].

El análisis de series de tiempo tiene como objetivo extraer estadísticas y otras características significativas de los datos.

El tiempo es un hecho importante en la explicación de algunos fenómenos. Algunas bases de datos incluyen esta propiedad, por lo que los trabajos relacionados con la Minería de Datos cambian su interpretación, así se desprende una nueva rama dentro de la Minería de Datos denominada Minería de Datos sobre Series de Tiempo. Los principales retos son:

a) Búsqueda de eficientes representaciones.
b) Detección de puntos de transición.
c) Clasificación.
d) Agrupación.

Recientemente se ha incrementado el interés en la Minería de Datos de series temporales. Como en la mayoría de los problemas de ciencia de la computación, la representación de los datos es la clave para soluciones eficientes y efectivas. Reducir el tamaño de los datos hace que el almacenamiento, transición y cómputo de los mismos sea más rápido y eficaz. En esta reducción de tamaño es importante no perder información valiosa de las series de tiempo.

En la Minería de Datos existen generalmente dos etapas claramente diferenciables, como son el pre-procesamiento y el post-procesamiento de los datos.

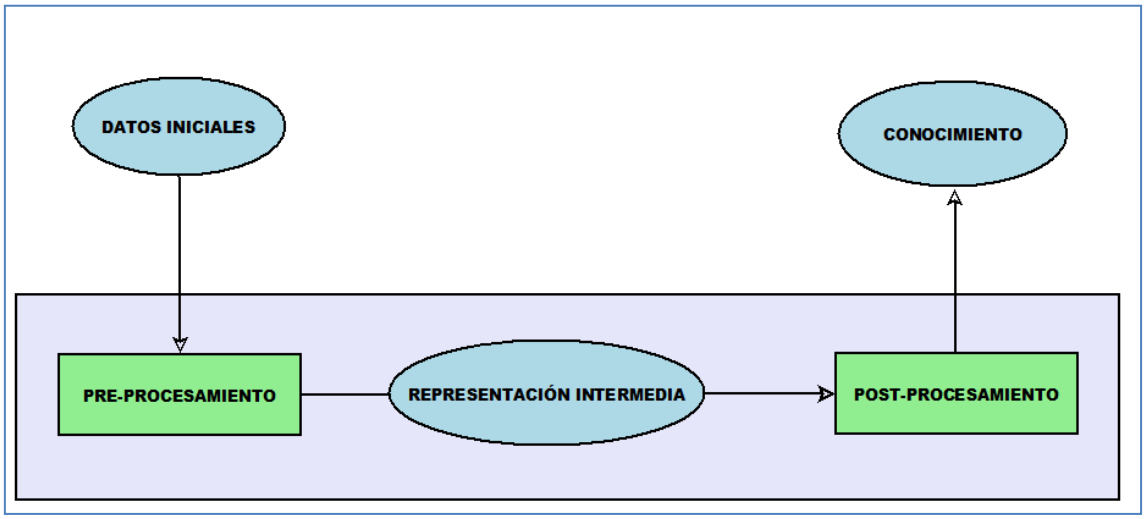


Figura 13: Diagrama conceptual sobre la etapas en la aplicación de la Minería de Datos

3. TÉCNICAS GENERALES PARA EL ANÁLISIS DE SERIES TEMPORALES

3.1. Introducción

Una serie temporal es un conjunto de datos (medidas) numéricos que se obtienen en períodos de tiempo regulares. La unidad de tiempo se establece dependiendo de la naturaleza de los datos a evaluar.

El objetivo del estudio de las series temporales es la identificación y aislamiento de los factores de influencia con el fin de hacer predicciones (pronósticos), y prevenir hechos potencialmente peligrosos.

Con el análisis de series temporales se pretende extraer el patrón de comportamiento sistemático contenido en una sucesión de observaciones que se recoge de forma regular y homogénea a lo largo del tiempo. Con este patrón es posible:

a) Caracterizar el comportamiento del fenómeno estudiado.
b) Predecir su evolución futura
c) Extraer componentes no observables (señales) que reflejan más fielmente la evolución subyacente de la variable de interés.

Para obtener el conocimiento 'oculto' en las series temporales, se debe someter a las mismas a un proceso de tratamiento que permita eliminar aquellas componentes que puedan provocar que las conclusiones del estudio no sean acertadas. El hecho, por ejemplo, de someter a una de estas series temporales a una etapa de suavizado o filtrado que elimine las componentes ruidosas de las señal original, puede hacer entender una manera más fidedigna la naturaleza y el comportamiento de la señal a examinar.

En este apartado se mostrarán algunas de las técnicas más comunes en el tratamiento de series temporales multivariantes.

3.2. Suavizado o filtrado

Como se ha comentado en apartados anteriores, una etapa importante en el tratamiento de series temporales es la eliminación del ruido y la tendencia. A continuación se mostrarán algunas de las técnicas más comentadas en el filtrado (suavizado) de series temporales.

MEDIAS MÓVILES [15]

Una media móvil se calcula, para cada punto, como un promedio del mismo número de valores a cada lado de ese punto. Por lo tanto, una media móvil genérica se calcula como:

$$X_t = \frac{Z_{t-k}+Z_{t-k+1}+\cdots+Z_{t-1}}{k} \qquad (1)$$

Donde X_t es el valor de la predicción para el instante t, Z_t es el valor real de la serie en un instante genérico t y k es el número de valores considerados en el cálculo de la media móvil.

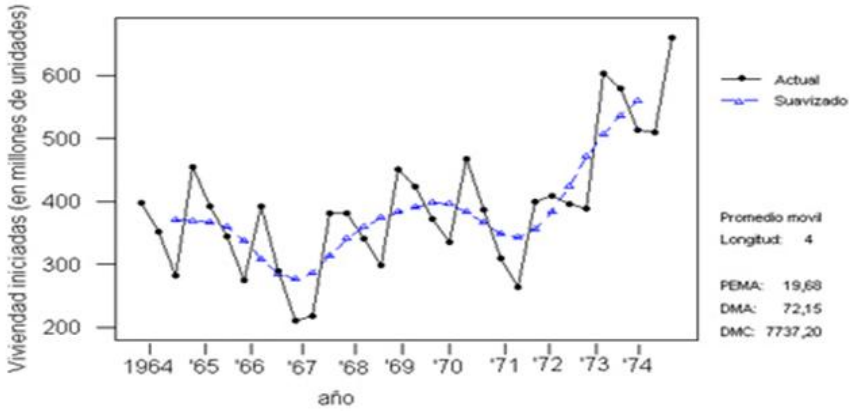


Figura 14: Ejemplo de la aplicación de la técnica de Medias Móviles en una serie temporal

Las previsiones obtenidas por este método pueden ser útiles en situaciones en que la serie es muy volátil y con una media estable. Se recomienda un k alto si la serie es muy heterogénea. Normalmente se trabaja con valores entre 3 y 11.

Este método de suavizado es bastante simple de aplicar, pero a su vez tiene algunos inconvenientes.

- "Costosa" de calcular: cuando se utilizan pesos, el cálculo hay que hacerlo desde cero para cada valor.
- Problemáticas en los extremos de la serie de datos. Dada la anchura de la ventana, no se pueden extender hasta el final de la serie, hecho que suele ser interesante.
- No se puede definir fuera de la serie temporal, por lo que no se pueden utilizar para realizar predicciones.

SUAVIZADO EXPONENCIAL

El suavizado exponencial se basa en idea muy simple, la de que es posible calcular un promedio nuevo a partir de uno anterior y también de la demanda más recientemente observada.

Dentro de las variantes que este método ofrece, se van a comentar brevemente las tres técnicas más usadas del suavizado exponencial [16].

- *Suavizado exponencial simple:* para series sin tendencia ni estacionalidad.
- *Suavizado exponencial doble:* para series con tendencia pero sin estacionalidad.
- *Suavizado exponencial triple:* para series con tendencia y estacionalidad.

Suavizado exponencial simple

Esta técnica supera al método de medias móviles en el sentido de que no elimina valores. Además, le da mayor ponderación a los últimos valores de la serie y menos peso a los primeros. Se recomienda cuando la serie no tiene tendencia ni estacionalidad, es decir, se estima el nivel de la serie. El pronóstico debe ser a muy corto plazo.

$$S_1 = y_1$$
$$S_t = \alpha y_t + (1-\alpha)S_{t-1} \qquad 0 < \alpha < 1\,,\ t = 2,3,\dots,n \qquad (2)$$

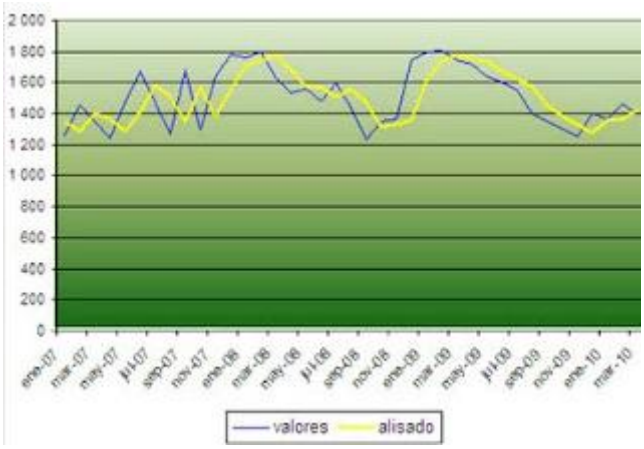


Figura 15: Ejemplo de suavizado exponencial simple

Suavizado exponencial doble

Este procedimiento de suavizado debe su nombre a que se realizan dos operaciones de "*suavizado consecutivas*". Más concretamente, en primer lugar se realiza un suavizado sobre la serie original y posteriormente se realiza un suavizado sobre la serie resultante de la primera etapa.

$$s_i = \alpha x_i + (1-\alpha)(s_{i-1} + t_{i-1})$$
$$t_i = \beta(s_i - s_{i-1}) + (1-\beta)t_{i-1} \qquad (3)$$

El suavizado exponencial doble retiene información acerca de la tendencia: la señal suavizada s_i y la tendencia t_i. El parámetro β se utiliza para realizar un suavizado exponencial sobre la tendencia.

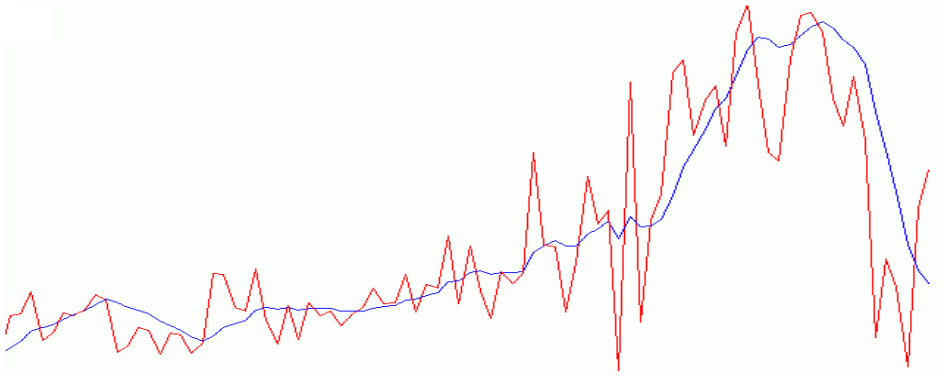


Figura 16: Ejemplo de suavizado exponencial doble (rojo->señal original; azul->señal suavizada)

Suavizado exponencial triple (Método de Holt-Winters)

El método de Holt-Winters es un método de alisado exponencial que tiene en cuenta el componente de tendencia (método de Holt, extensión del método de alisado exponencial simple) y la componente estacional (extensión por Winters del método de Holt). El proceso de

predicción de este método está formado por tres componentes: el nivel, la tendencia y el componente estacional. Estos tres componentes están relacionados por ecuaciones iterativas, que contienen tres parámetros (α, β, γ).

Si $\beta = \gamma = 0$, las ecuaciones se reducen a una, y el resultado de la predicción corresponde al proceso de suavizado exponencial simple.

Si $\gamma = 0$, las ecuaciones se reducen a dos, y el resultado es el proceso de Holt, que no tiene en cuenta el componente estacional.

El suavizado triple puede tener dos variantes de estacionalidad según interese: una aditiva y otra multiplicativa:

Estacionalidad aditiva	Estacionalidad multiplicativa
$s_i = \alpha x_i + (1-\alpha)(s_{i-1} + t_{i-1})$	$s_i = \alpha \frac{x_i}{p_{i-k}} + (1-\alpha)(s_{i-1} + t_{i-1})$
$t_i = \beta(s_i - s_{i-1}) + (1-\beta)t_{i-1}$	$t_i = \beta(s_i - s_{i-1}) + (1-\beta)t_{i-1}$
$p_i = \gamma(x_i - s_i) + (1-\gamma)p_{i-k}$	$p_i = \gamma \frac{x_i}{s_i} + (1-\gamma)p_{i-k}$
$x_{i+h} = s_i + ht_i + p_{i-k+h}$	$x_{i+h} = (s_i + ht_i)p_{i-k+h}$ (5)

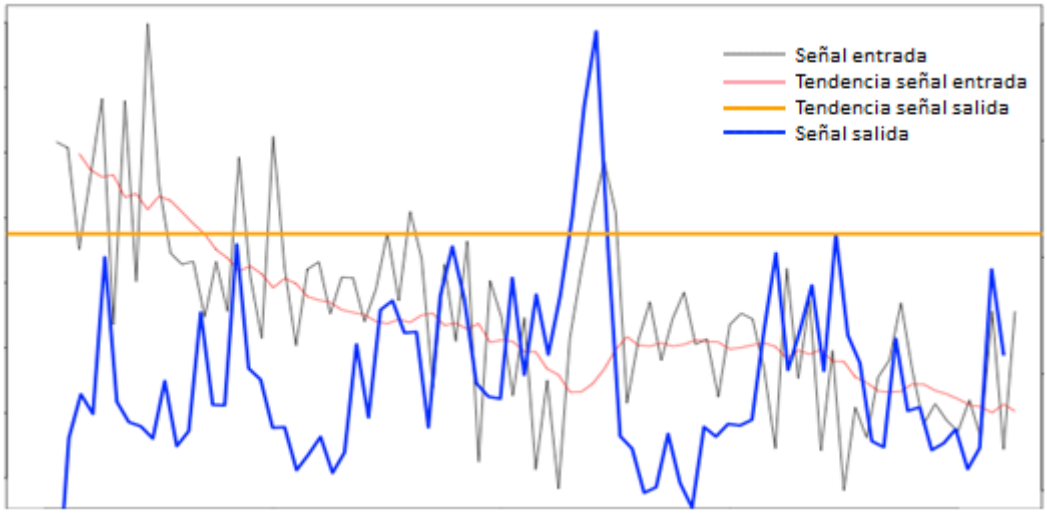


Figura 17: Ejemplo de suavizado exponencial triple, en el que se observa cómo se corrigen tanto la tendencia como la estacionalidad.

FILTRO DE PRESCOTT-HODRICK (FPH)

El filtro de Prescott-Hodrick es un método que permite extraer la componente tendencial a fin del ciclo. Descompone la serie en una componente tendencial y otra cíclica. Mediante el factor multiplicativo λ se ajusta la sensibilidad de la tendencia a las fluctuaciones a corto plazo.

La medida de las componentes cíclicas viene dada por:

$$c_t = y_t - \tau_t \quad (5)$$

siendo y_t la serie temporal para t=1,2,…, N

El componente tendencial se puede resolver con la siguiente ecuación:

$$min \sum_{t=1}^{N}(y_t - \tau_t)^2 + \lambda \sum_{t=2}^{N-1}[(\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1})]^2 \quad (6)$$

Según Hodrick y Prescott el componente tendencial de la serie es el que minimiza tal ecuación.

$$\sum_{t=1}^{N}(y_t - \tau_t) = 0 \quad (7)$$

Este tipo de filtros es ampliamente utilizado en el campo financiero-económico y es poco usual que aparezcan en estudios con señales fisiológicas.

3.3. Wavelets

La transformada Wavelet extiende cada vez más su uso en procesado de señal con resultados muy positivos en áreas como la eliminación de ruido, estudio de señales transitorias, compresión de señales e imágenes, etc., llegando a superar notablemente a otras técnicas tradicionales que siguen empleándose en la actualidad. Así mismo, nuevos resultados prometedores la hacen cada vez más empleada en otros campos como la electromedicina [17].

La posibilidad de superponer funciones seno y coseno para aproximar la representación de otras funciones fue descubierta ya hace dos siglos por Joseph Fourier, y ha sido tradicionalmente empleada con resultados satisfactorios. Pero este análisis tiene el gran inconveniente de que al transformar una señal al dominio frecuencial se pierde la información temporal. Este inconveniente cobra importancia cuando la señal a analizar tiene características transitorias o son de duración finita, situación que se da en un número muy elevado de casos en vida real. En estos casos el contenido frecuencial cambia con el tiempo.

Una primera solución para tratar de paliar este inconveniente es la aplicación de la transformada de Fourier en cortas ventanas de tiempo (STFT), obteniendo como resultado una función bidimensional tiempo-frecuencia. Pero aparece el problema de que la ventana temporal a elegir es la misma para todas las frecuencias, lo cual provoca pérdida de información frecuencial a resoluciones de bajas frecuencias y temporal a resoluciones de altas.

La transformada Wavelet (WT) salva estos inconvenientes al analizar la señal a diferentes frecuencias con diferentes resoluciones, utilizando regiones con enventanados de distinto tamaño, aplicando ventanas temporales de mayor tamaño donde se requiera extraer información de baja frecuencia más precisa, y recíprocamente, ventanas temporales más cortas o resolutivas en el tiempo donde se desee extraer información de alta frecuencia. Como consecuencia se consigue buena resolución temporal y baja resolución frecuencial en las altas frecuencias, y viceversa en las bajas frecuencias (buena resolución frecuencial y baja resolución temporal).

El análisis Wavelet utiliza formas de onda de duración finita, oscilantes, de media cero y que tienden a ser irregulares y asimétricas.

TRANSFORMADA WAVELET CONTINUA

La señal a analizar se descompone a base de versiones desplazadas y dilatadas de la wavelet madre o wavelet que hayamos decidido emplear, y todo ello por medio de un proceso

de correlación entre la señal a descomponer y las mencionadas versiones de la wavelet madre. Matemáticamente, la transformada wavelet continua (CWT) se define como:

$$C(a,b) = \int_{-\infty}^{\infty} f(t) \cdot a^{-\frac{1}{2}} \cdot \Psi\left(\frac{t-b}{a}\right) dt \quad (8)$$

siendo el resultado $C(a,b)$ un número de coeficientes que son función de la escala a de la wavelet y de su posición o desplazamiento b. La función a analizar o descomponer es $f(t)$, y las distintas versiones dilatadas o desplazadas de la wavelet $\Psi_{a,b}(t)$. El resultado $C(a,b)$ incluye información en el tiempo y en la frecuencia, que se representa gráficamente en lo que se reconoce como escalograma, un plano tiempo-frecuencia donde la amplitud de los coeficientes $C(a,b)$ pueden ser representados en una tercera dimensión, o más comúnmente, en una determinada escala de colores y grises.

Los pasos implicados cuando calculamos la CWT con una determinada wavelet son los siguientes:

a) Calcular un coeficiente $C(a,b)$ con la ecuación anteriormente expuesta, y para un determinado par de valores a y b de escala y desplazamiento. Dicho coeficiente representará el índice de similitud o correlación entre la wavelet que estamos empleando y la sección de señal que cae bajo aquella.
b) Se incrementa b, esto es, se desplaza la wavelet hacia la derecha en el eje de tiempos.
c) Se repiten los dos pasos anteriores hasta recorrer todo el intervalo de duración de la señal.
d) Se incrementa a, con lo que se consigue dilatar en el tiempo la wavelet, y se repiten los pasos anteriores hasta llegar a la escala a deseada. A medida que aumenta la escala, la información extraída es referente a las frecuencias más bajas.

TRANSFORMADA WAVELET DISCRETA

Cuando los parámetros a y b son valores enteros, la expresión de WT se convierte en la escrita más abajo, es decir, en una transformada Wavelet discreta (DWT). En la práctica dichos parámetros están definidos según lo que se llama una escala diádica, esto es, $a = 2^j$ y $b = k \cdot a$, con $j, k \in Z$, con objeto de obtener versiones dilatadas y trasladadas de la wavelet analizante que definen lo que se denominan una base de funciones ortogonal. Dos funciones son ortogonales entre sí cuando su producto integral a lo largo de todo su dominio es cero.

$$C(j,k) = \int_{-\infty}^{\infty} f(t) \cdot 2^{-\frac{j}{2}} \cdot \Psi\left(2^{-j} \cdot n - k\right) \quad (9)$$

Análogamente, la Transformada Wavelet Inversa (IWT) recupera la señal original a partir de los coeficientes del siguiente modo:

$$f(n) = \sum_{j \in Z} \sum_{k \in Z} C(j,k) \cdot 2^{-\frac{j}{2}} \cdot \Psi\left(2^{-j} \cdot n - k\right) \quad (10)$$

Para calcular la DWT en la práctica se usan algoritmos rápidos de cálculo basado en el esquema clásico ya conocido como análisis multi-resolución, y cuyo paso básico consiste en introducir simultáneamente la señal discreta por un par de filtros de espejo en cuadratura, uno paso bajo y otro paso alto, para analizar la señal a diferentes escalas. A causa de que estas operaciones cambian la resolución de la señal, para evitar redundancia a la salida de ambos

filtros, se lleva una operación de sub-muestreo o diezmado tomando solamente una de cada dos muestras. A la salida de la rama que incluye el paso bajo se obtienen los coeficientes de aproximación al nivel 1 de descomposición cA1, mientras que en la otra salida se obtienen los coeficientes que contienen la información de los detalles de la señal al nivel 1 cD1. Este proceso se puede repetir sucesivamente sobre los coeficientes de la aproximación que se van obteniendo, hasta descomponer la señal al nivel deseado.

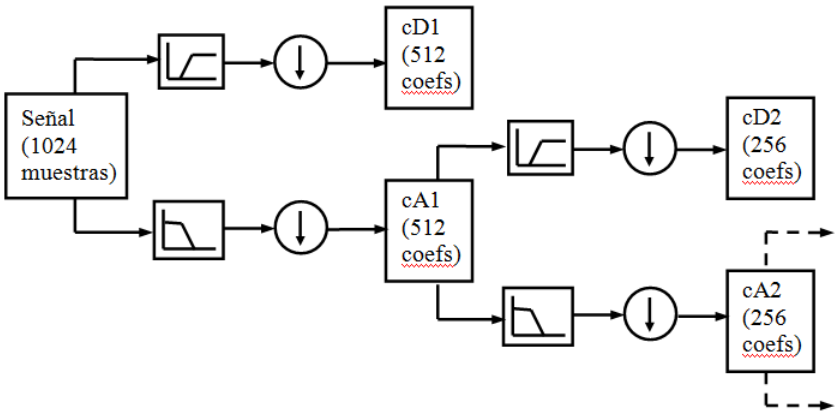


Figura 18: Proceso de la descomposición wavelet multinivel.

Para dar una información más gráfica acerca del funcionamiento de la descomposición Wavelet multinivel:

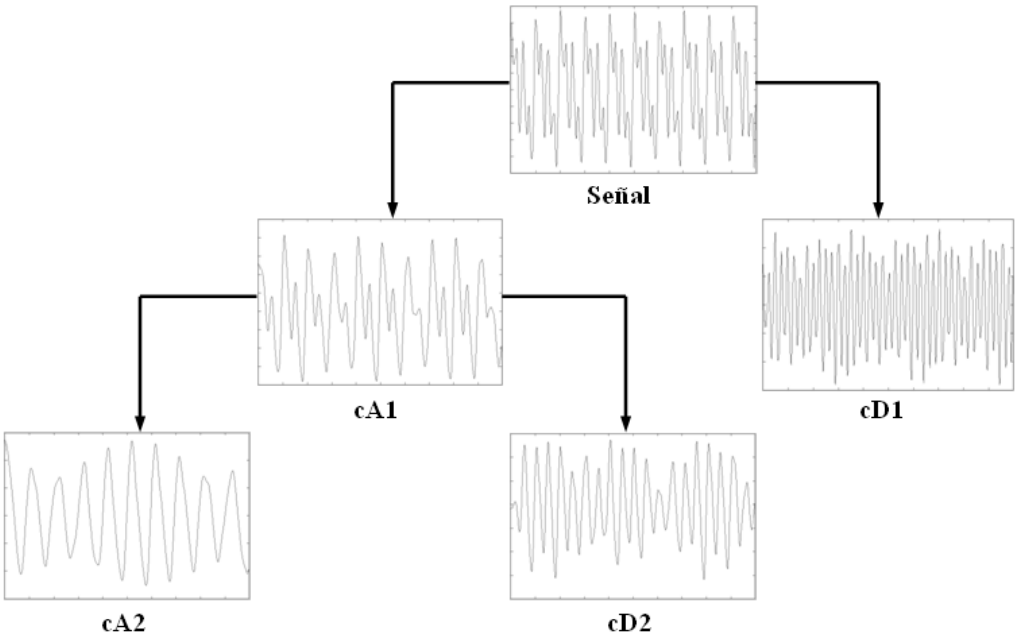


Figura 19: Señales resultantes de una descomposición Wavelet de dos niveles

Los coeficientes de detalle al primer nivel cD1 contienen la información de más alta frecuencia y, en la práctica, el ruido de más altas frecuencias. A medida que se consideran más coeficientes, ya sea de aproximación o de detalles, correspondientes a un nivel de

descomposición más elevado, la información obtenida es más suave o de baja frecuencia. Además, si se consideran las señales restauradas a partir de los coeficientes, se cumple:

$$Señal = cA_1 + cD_1 = cA_2 + cD_2 + cD_1 = cA_n + cD_n + cD_{n-1} + \cdots + cD_1 \qquad (11)$$

La síntesis, o la descomposición Wavelet a partir de los coeficientes de la descomposición, se realiza de una manera similar a la indicada en el esquema Wavelet multinivel pero en sentido inverso y haciendo interpolaciones por dos o sobre-muestreos, y sumando las señales a las salidas de los filtros. Los filtros usados para la reconstrucción son muy parecidos, aunque no idénticos, a los usados durante la descomposición. Existen algunas diferencias en dichos filtros, con objeto de eliminar el efecto de aliasing que se produjo durante la fase de descomposición por efecto del diezmado. Todo este conjunto de filtros, los usados en el análisis junto a los homólogos usados en la síntesis, se denomina sistema de filtros de espejo en cuadratura.

3.4. Reducción de la dimensionalidad

3.4.1. Análisis de Componentes Principales

El Análisis de Componentes Principales (ACP, en inglés *Principal Component Analysis - PCA*) es una técnica multivariada exploratoria para la formación de nuevas variables llamadas componentes principales (CP) las cuales son combinaciones lineales de las variables originales. El número máximo de nuevas variables que pueden ser formadas es igual al número de variables originales y las nuevas variables son autocorrelacionadas entre sí [18]. Esta técnica no sólo reduce la complejidad y dimensionalidad de los datos observados a una forma más simple, sino que también estos datos pueden ser utilizados en aplicaciones de otros métodos multivariantes. El objetivo principal del ACP es construir un indicador de ciertas características con base en las 'p' variables originales de interés para reducir la complejidad de las interrelaciones entre el número elevado de variables observadas a un número relativamente más pequeño de variables lineales consideradas como componentes principales. En gran medida, la interpretación de las componentes principales es, en general, guiada por el grado en que cada variable esté asociada con un componente particular.

Las componentes principales tienen una variedad de propiedades útiles: la primera componente principal tiene la varianza máxima entre todas las CP, la segunda CP la segunda varianza máxima, etc. Además, las componentes principales obtenidas son independientes entre sí. Para la construcción de un conjunto de datos con p variables numéricas Y_i $(i = 1,2, \dots, p)$, las q componentes principales Z_j $(j = 1,2, \dots, q)$ pueden ser obtenidas con ayuda de programas computacionales. Cada componente principal es una combinación lineal de las variables originales, con coeficientes iguales a los vectores propios de la matriz de correlación o covarianza y se expresa de la siguiente forma:

$$Z_i = a_1X_1 + a_2X_2 + \cdots + a_pX_p \qquad (12)$$

Los vectores propios de los componentes principales son ortogonales para maximizar la discriminación entre variables. Cuando se busca una componente de Z, el vector de constantes y el vector que tiene las p variables originales de interés, son determinados por los vectores propios de la matriz de covarianza o correlación donde se desea que la varianza de cada componente principal de Z tenga el valor máximo obtenido a través de los valores propios de esas matrices. El valor propio de la matriz de correlación y/o covarianza de los

componentes formados, así como el valor acumulativo de las componentes seleccionadas (el cual representa el porcentaje de confiabilidad de dicha selección) podría contribuir a la selección del número de variables de los vectores propios formados.

3.4.2. PAA – Piecewise Aggregate Approximation

En la mayoría de los casos de estudio de Minería de Datos que se dan en la actualidad se necesita un paso previo de limpieza y transformación de los datos debido a la gran cantidad de información innecesaria que poseen. En un gran número de los casos las representaciones de las secuencias, según la transformación aplicada, afecta en gran medida al problema de Minería de Datos. El objetivo es tratar de disminuir la dimensionalidad para aumentar la eficiencia del algoritmo, facilitar la comparación o sesgarlo en un determinado sentido. Para ello se suelen utilizar transformaciones junto con una estructura de indexado utilizada por el algoritmo de Minería de Datos.

Unas de las transformaciones que se están utilizando ampliamente para este tipo de problemas son las transformaciones *PieceWise*, que se basan en aproximar la secuencia temporal por un conjunto de segmentos de valor constante.

La transformación PAA [19] consiste en la división de la secuencia temporal en *N* segmentos de igual longitud. Por cada segmento de la secuencia se calcula la media que puede ser utilizada como un coordinador de un vector de indexación *N-dimensional* [20]. Esta simple reducción de la dimensionalidad presenta notables ventajas:

- La propia transformada es más rápida que la mayoría de las otras transformaciones.
- Es fácil de entender e implementar.
- El índice para una estructura de indexación se puede construir en un tiempo lineal.

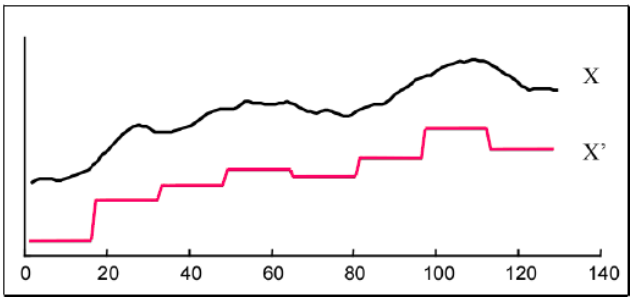


Figura 20: Representación de la Piecewise Aggregate Approximation

Una vez se tenga la transformada de la secuencia mediante PAA se debe utilizar una medida de la distancia que cumpla $D_{PAA}(x^{PAA}, y^{PAA}) \leq D(x, y)$.

$$D_{PAA}(x_{PAA}, y_{PAA}) = \sqrt{\frac{n}{N}} \sqrt{\sum_{i=1}^{N} (x^{PAA}, y^{PAA})^2} \quad (13)$$

La transformación *Adaptative Piecewise Constant Approximation (APCA)* es similar al PAA excepto que los segmentos no necesitan ser de igual longitud. Así, las regiones con más fluctuaciones podrían ser representadas con segmentos cortos mientras que las regiones con

fluctuaciones más razonables podrían representarse con menos segmentos (segmentos más largos). Esto hace que la compresión sea más efectiva que en el caso anterior.

3.5. Clasificación de patrones

Las técnicas de suavizado y de reducción de dimensionalidad son una parte fundamental en el procesamiento de los datos dentro de las redes neuronales. Pero dichas técnicas no tendrían sentido si no se dispusiese de un sistema de clasificación de patrones con el que obtener información precisa del comportamiento de la secuencia temporal. Existen variedad de clasificadores, de los cuales se expondrá el funcionamiento de los siguientes: el clasificador Bayesiano, y los basados en redes neuronales, lógica difusa y *neurofuzzy*.

CLASIFICADOR BAYESIANO

Es el método estadístico clásico, donde se estima la probabilidad a posteriori de la pertenencia de una muestra de prueba a una de las clases dadas. Esta probabilidad es evaluada en función de las probabilidades a priori de cada clase y la probabilidad condicional, resultante de la distribución de las muestras de entrenamiento y evaluada según el teorema de Bayes [20].

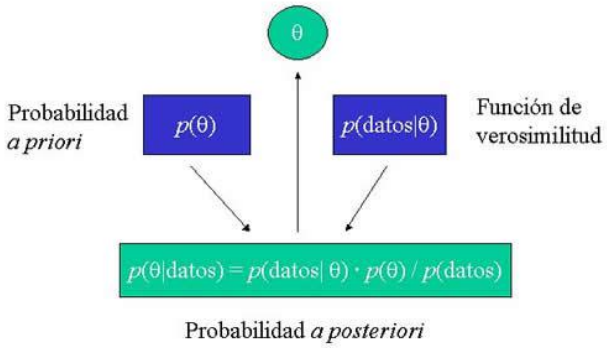


Figura 21: Esquemática del enfoque bayesiano.

ANÁLISIS DISCRIMINANTE LINEAL

El Análisis Discriminante Lineal (LDA) [20] es un método estadístico que discrimina la separabilidad de clases en forma lineal, lo cual no se ajusta de manera eficiente a los patrones de dinámica estocástica. Sin embargo, ha sido utilizada en trabajos de investigación para efectos comparativos con otras técnicas como las referencias [17], [18] y [19].

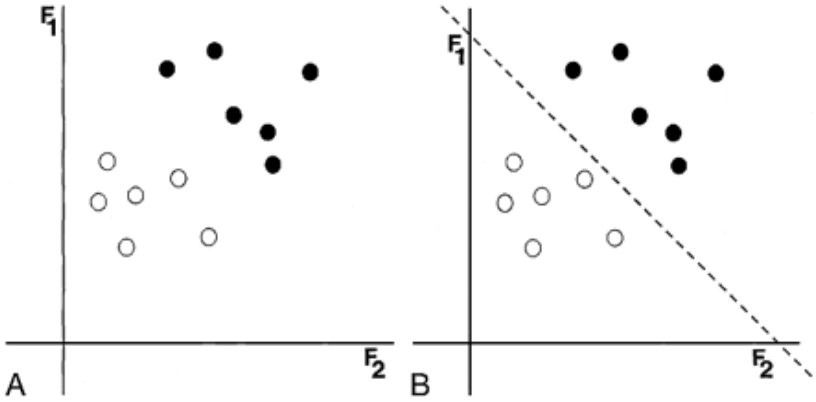


Figura 22: Ejemplo de la aplicación del LDA bidimensional

REDES NEURONALES ARTIFICIALES

Se puede definir a una Red Neuronal Artificial (RNA) como un modelo matemático inspirado en sistemas biológicos [21], adaptado y simulado en computadores convencionales. Las RNAs están inspiradas en el sistema biológico natural. Como es conocido, en este sistema la neurona es la unidad de procesamiento, y aunque las RNAs sean mucho menos complejas que una red neuronal biológica, también realizan cálculos complejos para procesar información.

Por lo tanto, en términos de computación, una RNA es una estructura compuesta de un número de unidades interconectadas (neuronas artificiales). Cada unidad posee una característica de entrada/salida e implementa una computación local o función. La salida de cualquier unidad está determinada por su característica de entrada/salida, su interconexión con otras unidades, y, posiblemente, de sus entradas externas. Sin embargo, es posible un "trabajo a mano", ya que la red desarrolla usualmente una funcionalidad general a través de una o más formas de entrenamiento.

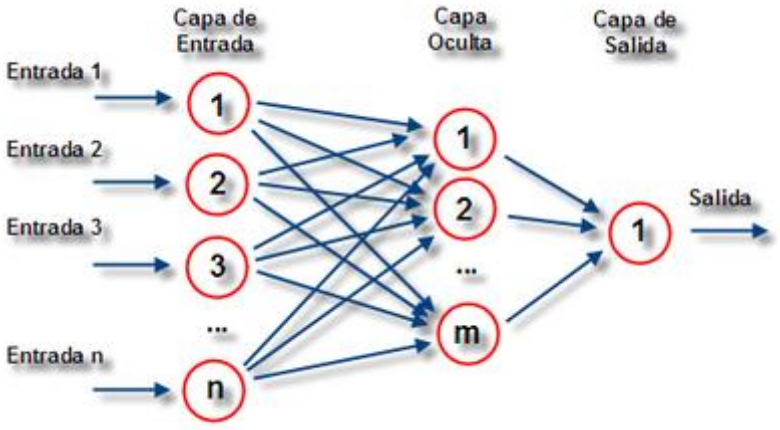


Figura 23: Red neuronal artificial perceptrón simple con *n* neuronas de entrada, *m* neuronas en su capa oculta y una neurona de salida.

Por lo que respecta a la topología de la red, las redes pueden clasificarse de acuerdo con el número de capas o niveles de neuronas, el número de neuronas por capa y el grado y

tipo de conectividad entre las mismas. La primera distinción a establecer es entre las redes *Monocapa* y las *Multicapa.*

Las redes Monocapa sólo cuentan con una capa de neuronas, que intercambian señales con el exterior y que constituyen al mismo tiempo la entrada y salida del sistema. En las redes Monocapa (red de Hopfield o red Brain-State-in-Box, máquina de Boltzman, máquina de Cauchy) se establecen conexiones laterales entre neuronas, pudiendo existir también conexiones autorrecurrentes (la salida de la neurona se conecta con su propia entrada), como en el caso del modelo Brain-State-in-Box.

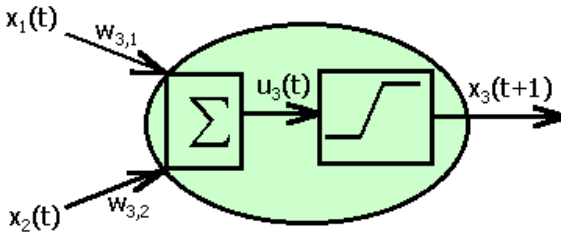


Figura 24: Modelo BSB (Brain-State-in-Box) [22]

Las redes Multicapa disponen de conjuntos de neuronas jerarquizadas en distintos niveles o capas, con al menos una capa de entrada y otra de salida, y eventualmente, una o varias capas intermedias (capas ocultas).

Una segunda clasificación que se suele hacer es en función del tipo de aprendizaje de que es capaz (si necesita o no un conjunto de entrenamiento supervisado). Para cada tipo de aprendizaje encontramos varios modelos propuestos por diferentes autores.

- *Aprendizaje supervisado:* necesitan un conjunto de datos de entrada previamente clasificado o cuya respuesta objetivo se conoce.
- *Aprendizaje no supervisado o auto-organizado*: no necesita de tal conjunto previo.
- *Redes híbridas*: son un enfoque mixto en el que se utiliza una función de mejora para facilitar la convergencia.
- *Aprendizaje reforzado*: se sitúa a medio camino entre el supervisado y el auto-organizado.

LOGICA DIFUSA

Los modelos basados en la lógica difusa [23], junto con los modelos basados en redes neuronales artificiales, constituyen un conjunto de herramientas de representación y modelización que pertenecen al campo conocido como "*soft computing*". La lógica difusa es una técnica para la incorporación del conocimiento estructurado humano en algoritmos eficientes [24].

Muchos de los procesos intelectuales humanos están basados en un razonamiento inductivo, cuyo mejor exponente son los razonamientos que lingüísticamente responden a estructuras lógicas del tipo "SI… ENTONCES…". De esta manera se puede argumentar que el conocimiento humano se estructura en reglas del tipo "SI… ENTONCES…", y que esa combinación de reglas lleva a acciones, toma de decisiones, etc.

En la Inteligencia Artificial, la *lógica difusa* o *lógica borrosa* se utiliza para la resolución de variedad de problemas, principalmente los relacionados con control de procesos industriales complejos y sistemas de decisión. Los sistemas de lógica difusa están también extendidos en el ámbito de la tecnología cotidiana, por ejemplo en cámaras digitales, en aire acondicionado, lavadoras, etc. Los sistemas basados en lógica difusa imitan la forma de toma de decisiones humana, con la ventaja de ser mucho más rápidos. Estos sistemas son generalmente robustos y tolerantes a imprecisiones y ruidos en los datos de entrada.

Como principal ventaja cabe destacar los excelentes resultados obtenidos por sistemas de control basados en lógica difusa ya que ofrecen salidas de una forma veloz y precisa, disminuyendo así las transiciones de estados fundamentales en el entorno físico que controlen.

SISTEMAS NEURO-FUZZY (NEURO-DIFUSOS)

Los sistemas neuro-difusos son una combinación entre las redes neuronales y la lógica difusa que permite una relación simbiótica en la cual se aprovecha el conocimiento de un experto y la capacidad de aprendizaje y eficiencia computacional de la red neuronal, logrando sistemas de decisiones más inteligentes.

La diferencia primaria entre los sistemas difusos y los neuro-difusos radica en que los primeros son especialistas en construcción y los segundos entrenadores de datos.

Los sistemas difusos combinan la capacidad de aprendizaje de la RNA con el poder de interpretación lingüística de la lógica difusa, obteniéndose los siguientes resultados [25]:

- Aplicabilidad de los algoritmos de aprendizaje desarrollados para redes neuronales.
- Posibilidad de promover la integración de conocimiento (implícito que pueda ser adquirido a través del aprendizaje y explícito que pueda ser explicado y entendido).
- La posibilidad de extraer conocimiento para una base de reglas a partir de un conjunto de datos.

Existen sistemas que han conseguido unir la Lógica Difusa con las Redes Neuronales:

- ANFIS (*Adaptative Neuro-Fuzzy Interference System*) es un método que permite sintonizar o crear la base de reglas de un sistema difuso, utilizando el algoritmo de entrenamiento de retropropagación a partir de la recopilación de datos de un proceso.
- FSOM (Fuzzy Self-Organizing Maps) consiste en un sistema difuso optimizado a partir de la técnica de los mapas auto-organizados de Kohonen.
- NEFCLASS está basado en la estructura de perceptrón multicapa cuyos pesos son modelados por conjuntos difusos. Así, se preserva la estructura de una red neuronal, pero se permite la interpretación del sistema difuso asociado.
- FUZZYTECH es un software que propone un método de desarrollo de sistemas Neuro-Difusos similar a ANFIS.

MEMORIA JERÁRQUICA TEMPORAL

La Memoria Temporal Jerárquica (HTM – Hierarchical Temporal Memory) es una tecnología de aprendizaje maquina que pretende capturar las propiedades estructurales y algorítmicas del neocórtex [26].

El neocórtex es la base del pensamiento inteligente en el cerebro de los mamíferos. Permite tener conciencia y control de las emociones, a la vez que desarrolla las capacidades cognitivas: memorización, concentración, autor reflexión, resolución de problemas… Teniendo en cuenta la diversidad de funciones que lleva a cabo el neocórtex, se podría esperar que también implementara el mismo número de algoritmos neuronales. Al contrario que esta deducción, el neocórtex implementa un grupo común de algortimos para llevar a cabo diferentes funciones relacionadas con la inteligencia.

La HTM es una colección de nodos interconectados, organizados con forma jerárquica de árbol [27]. Está formada por varias capas o niveles de nodos con un solo nodo en la parte superior. Este sistema funciona en dos fases: la fase de aprendizaje y la fase de inferencia. Durante la etapa de aprendizaje, la red se expone a patrones de entrenamiento y se construye in modelo que estructura los patrones en categorías. Durante la fase de inferencia la red generará la distribución en estas categorías para nuevos patrones. Todos los nodos (excepto el nodo inicial) procesan la información aproximadamente de la misma forma y consta de dos módulos: temporal y espacial. Comprender un nodo de la HTM se reduce a comprender el funcionamiento de estos módulos durante la dase de aprendizaje y entrenamiento.

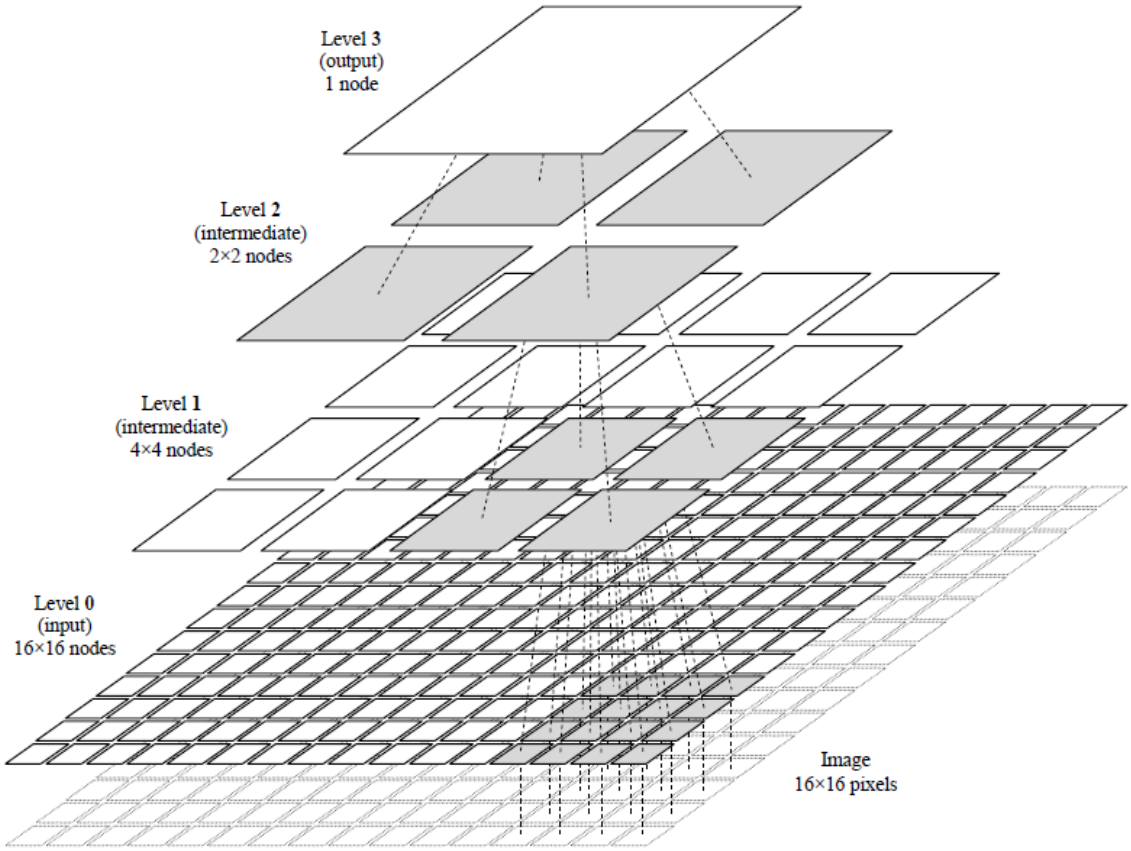


Figura 25: Diseño basado en HTM de cuatro niveles para trabajar con imágenes de 16x16 píxeles [28]. El nivel 0 tiene 16x16 nodos de entrada, cada uno asociado a un único pixel. Cada nodo del nivel 1 tiene 16 nodos del nivel inferior (regiones de 4x4) y un campo receptivo de 16 píxeles. Cada nodo del nivel 2 tiene 4 nodos del nivel 1 (2x2

regiones) y un campo receptivo de 64 píxeles. Finalmente, el único nodo de salida del nivel 3 tiene 4 nodos del nivel inferior, y el campo receptivo es de 256 píxeles.

4. DISEÑO DE LA TÉCNICA DEL ESTUDIO

En el capítulo anterior se han abordado los sistemas más relevantes de análisis de series temporales y de clasificación de patrones de manera que se tenga una visión general de los procedimientos necesarios para obtener conocimiento acerca de su comportamiento.

En este capítulo se expone de manera más profunda los procedimientos que se han seguido y las técnicas utilizadas para la resolución del problema del presente estudio. La herramienta que se usa en la implementación de cada bloque es Matlab.

A continuación se detallan las diferentes alternativas que se han implementado y validado en cada una de las etapas que componen el sistema: pre-procesamiento del sistema (suavizado), reducción de la dimensionalidad, estimador de longitud óptima de patrón y búsqueda y clasificación del patrón.

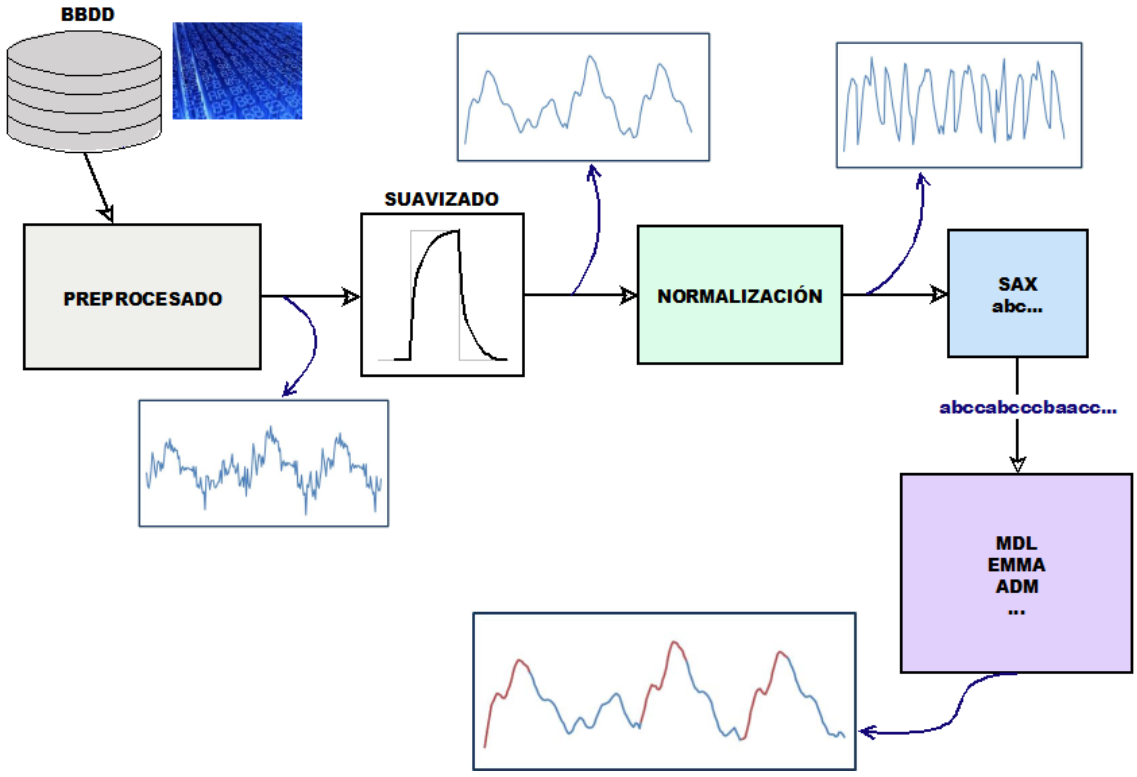


Figura 26: Esquema del diseño propuesto en el TFC

4.1. Suavizado o Smoothing

El primer problema que se plantea a la hora de obtener la mejor representación del flujo sanguíneo y la presión sanguínea es la eliminación del ruido que aparece en las medidas con las que se trabaja. La alta sensibilidad de los aparatos de medida provoca que junto con la señal deseada aparezcan variaciones externas que nada tienen que ver con la señal a estudio. Por este motivo el primer obstáculo que se debe superar es la eliminación de señales espurias mediante la aplicación de técnicas de suavizado.

La solución que se plantea a la problemática es la aplicación de un suavizado o "smoothing" que se basará en el método *loess*, técnica demostrada en el tratamiento de series temporales ruidosas.

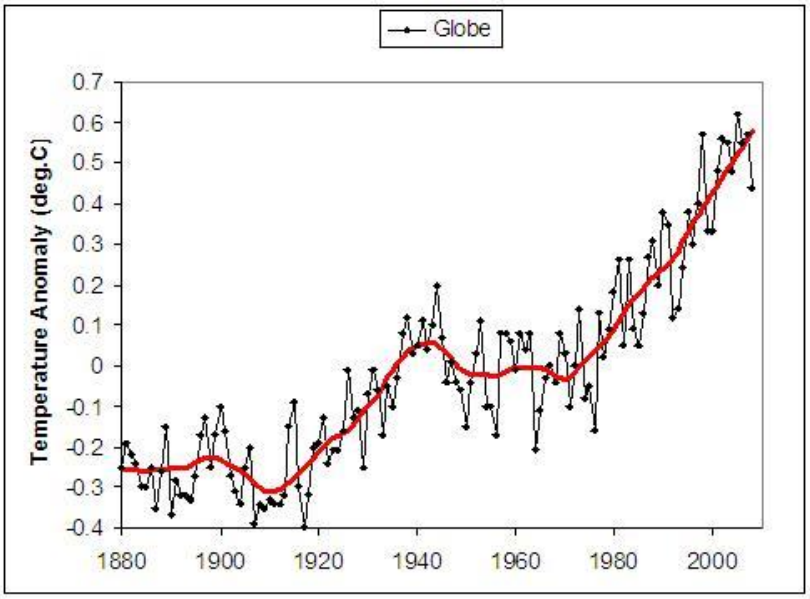


Figura 27: Ejemplo de la aplicación de la técnica de suavizado sobre datos temporales

Con este método se eliminarán los cambios pronunciados de la serie base sin alteraciones significativas en la señal original. La regresión lineal ponderada (*loess*) consiste en que si x_0 es un punto donde se desea hallar el suavizado, lo primero que se calcula es una "vecindad" usando los k vecinos más cercanos (k generalmente se expresa como un porcentaje del total de datos y es llamado el parámetro "smoother pan"). Posteriormente se calcula una regresión ponderada en dicha vecindad del valor ajustado de y en x_0 que será el valor del suavizador. Matlab ofrece varios métodos de suavizado, de los cuales se ha decidido utilizar el "rloess" que consiste en un método de regresión lineal polinómico ponderado que asigna pesos locales a los datos con un ajuste por mínimos cuadrados lineal utilizando un polinomio de segundo grado y es resistente a los valores extremos.

En el estudio se ha configurado el parámetro k a 0.04 debido a que medidas experimentales previas han demostrado que este valor ofrece un suavizado óptimo en series temporales de carácter biomédico.

4.2. Normalización

En muchos campos de la ciencia como son la medicina, astronomía, bioinformática, etc., se manejan cantidades muy grandes de información descritas en su mayoría en valores continuos. A partir de la adquisición de los datos reales, se identifica la problemática de requerir grandes espacios de almacenamiento para contener toda la información. La necesidad de tal cantidad de información es debida a la enorme cantidad de datos requeridos para el estudio y explicación de un hecho. Para trabajar con esta información se busca reducir su dimensión, con la problemática de no desechar datos útiles en el proceso. Además, la mayoría de los problemas de Minería de Datos sobre series temporales requieren de una información

discretizada, por lo que empuja a la necesidad de transformar los datos continuos originales en datos discretizados.

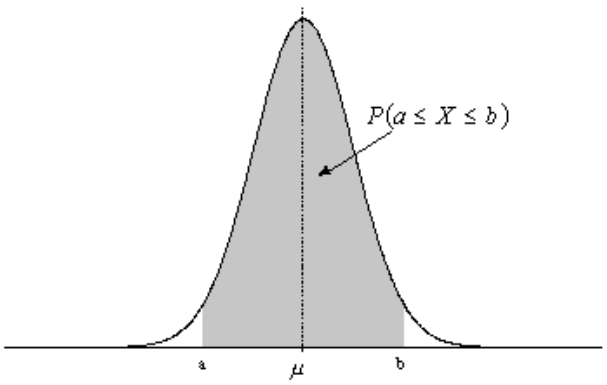
Uno de los factores más relevante en la discretización de la señal es la dimensionalidad de los datos, donde se pueden distinguir dos propuestas. La primera se centra en la reducción de los datos con un impacto mínimo sobre la información contenida en ellos. Con respecto a esta primera propuesta, únicamente se requiere de un porcentaje de reducción, mientras que para la segunda se requieren métodos que definan las partes más relevantes del conjunto de datos.

La siguiente etapa básica en el tratamiento de los datos es la normalización. Esta fase es básica para la obtención de información significativa en la información de alta dimensionalidad del flujo sanguíneo y presión sanguínea. Debido a ciertos factores no controlados en el estudio y a las variaciones en las condiciones experimentales, se produce una variabilidad significativa en los datos proporcionados, por lo que es necesario un paso de normalización previo al análisis cuantitativo. Esta normalización permite aumentar la sensibilidad en la detección de patrones.

Como se ha podido constatar en investigaciones previas, las medidas de flujo y presión sanguínea siguen una distribución gaussiana, por lo que la normalización se constituirá respecto a la distribución de Gauss norma [29].

La distribución de una variable normal está completamente caracterizada por dos parámetros, su media y su desviación estándar, denotadas generalmente por μ y σ. Con esta notación, la función densidad de probabilidad viene dada por la siguiente ecuación:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\};\ -\infty < x < \infty \qquad (14)$$



Así, se dice que una característica X sigue una distribución normal de media μ y varianza σ^2, y se denota como $X \approx N(\mu, \sigma)$.

Las propiedades de la distribución normal son:

a) Tiene una única moda, que coincide con su media.
b) La curva normal es asintótica al eje de abscisas. Por ello, cualquier valor entre $-\infty$ y ∞ es teóricamente posible. El área total bajo la curva es, por tanto, igual a 1.

c) Es simétrica con respecto a su media μ . Según esto, para este tipo de variables existe una probabilidad de un 50% de observar un dato mayor que la media y otro 50% menor que la media.
d) La distancia entre la línea trazada en la media y el punto de inflexión de la curva es igual a la desviación típica (σ). Cuanto mayor sea σ, más aplanada será la curva densidad.
e) El área bajo la curva comprendida entre los valores situados aproximadamente a dos desviaciones estándar de la media es igual a 0.95. En concreto, existe un 95% de probabilidades de observar un valor comprendido en el intervalo $(\mu - 1.96\sigma,\ \mu + 1.96\sigma)$.

Por lo tanto, debido a las características de los datos a analizar, la distribución que se usará es la distribución normal estándar, que corresponde a la siguiente ecuación:

$$f(x) = \frac{1}{\sqrt{2\pi}} exp\left\{\frac{-x^2}{2}\right\};\ -\infty < x < \infty \quad (15)$$

4.3. Reducción de la dimensionalidad

Anteriormente se habló sobre la discretización de datos y las propuestas más relevantes en esa área. Sin embargo, el presente enfoque se centra en la discretización de series de tiempo. Es muy diferente el proceso que se lleva a cabo en cada caso. Respecto a la discretización de datos, se asume que se tienen atributos independientes y en la mayoría de los casos se discretiza por atributo. Para la discretización de series temporales, el tiempo es un factor determinante y no se pueden tratar los atributos de forma independiente. La mayoría de las series temporales tienen longitud grande por lo que se buscan representaciones más cortas sin repercutir de manera notable en la serie original.

La técnica de discretización que se ha probado en el estudio permite la reducción de una serie temporal arbitraria de longitud n a una cadena de caracteres de longitud w ($w < n$ y normalmente $w \ll n$). Esta técnica es formalmente denominada PAA [30].

Esta reducción en la dimensión se entiende mejor si se define que una serie de tiempo $\bar{C}$ de longitud n que puede ser representada en el espacio $w - dimensional$ por un vector $\bar{C} = \overline{c_1} \dots \overline{c_w}$. El $i - ésimo$ elemento $\bar{C}$ es calculado mediante la siguiente ecuación:

$$c_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} c_j \quad (16)$$

El resultado de la aplicación del método es una reducción de la dimensionalidad de la serie original en ventanas de tamaño w, lo cual generará a la salida del sistema una señal cuadrada. En este procesamiento se perderá el detalle de la señal, por lo que la buena elección del tamaño de ventana es fundamental para que la información contenida en la serie no se vea seriamente perjudicada. Este tipo de técnica es intuitiva y simple, pero se ha demostrado en estudios previos que puede competir con técnicas como la transformada de Fourier o la Wavelets [31]. Además posee un tiempo de computación bajo y soporta diferentes funciones de distancias.

4.3.1. Symbolic Aggregate Aproximation (SAX)

Una vez se tenga la serie temporal en transformada a PAA, se debe aplicar la técnica de SAX. La transformación a valores discretos está dada mediante una función de densidad de probabilidad. Las series de tiempo que se analizan en el estudio tienen una distribución gaussiana, por lo que a partir de la curva gaussiana se determinan los puntos de corte con un número de a áreas del mismo tamaño. Se va tomando cada valor de la representación PAA de la serie y se le asigna un símbolo iniciando por la letra "a" para el primer intervalo, "b" para el segundo, y así sucesivamente. A esta secuencia de símbolos se la llamará *palabra* [32].

La determinación de los umbrales de decisión ($\beta = \beta_1 \dots \beta_{a-1}$) de los símbolos se realizará de manera equi-probable y teniendo en cuenta el área encerrada por la curva gaussiana de media 0 y varianza 1. Se ilustra en la siguiente figura:

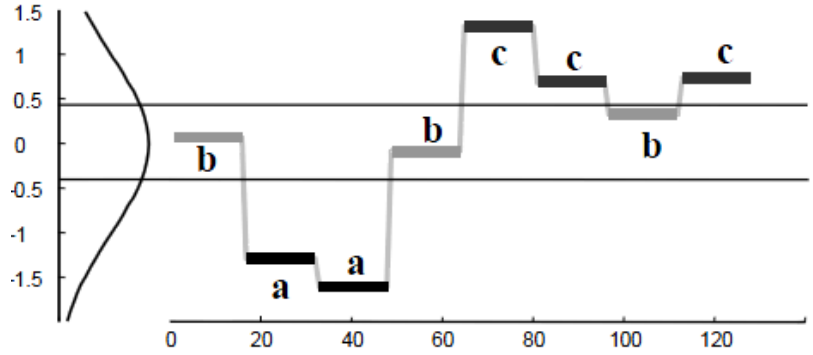


Figura 28: Ejemplo de la aplicación del algoritmo SAX.

Previo a este paso se debe escoger un tamaño de ventana deslizante (T_{min}) al que se le aplicará la normalización a la distribución gaussiana. El parámetro T_{min} se debe seleccionar como un tamaño mínimo de la longitud del motivo. Se debe escoger dicho valor con sumo cuidado, ya que un tamaño demasiado grande provocaría que el método no encontrase patrones de una longitud menor, por lo que perdería precisión.

Se muestra a continuación una demostración visual del proceso que sufre la señal temporal hasta llegar a la cadena de símbolos deseada.

$$\tilde{C} = \tilde{c}_1, \dots, \tilde{c}_{n_a} \quad con\ n_a = n - T_{min} + 1$$

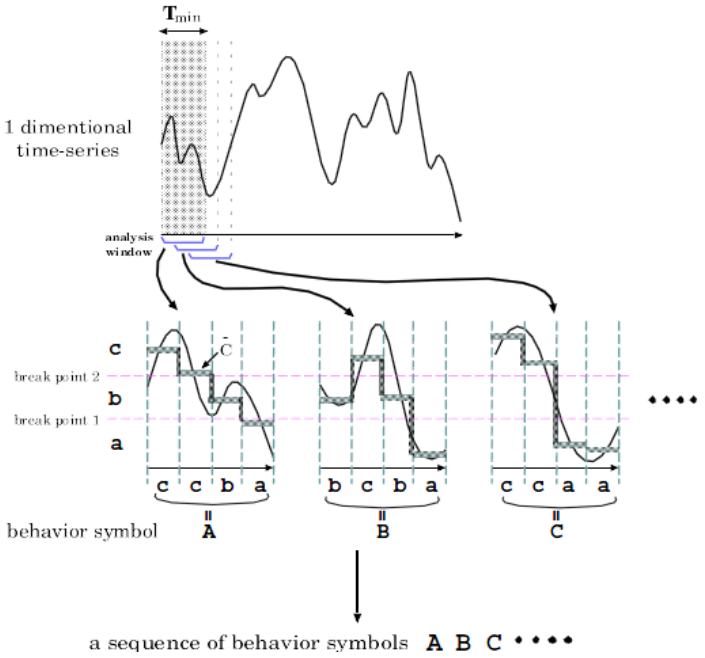


Figura 29: Ejemplo de funcionamiento de SAX

En el caso particular de este TFC se han producido cambios respecto al modelo teórico debido a la longitud de las muestras de entrenamiento y a la variabilidad de las mismas. Como consecuencia del carácter finito del alfabeto se originaron problemas cuando las sustituciones eran numerosas. La solución que se propone es la sustitución de las secuencias simbólicas por números. Se va a ilustrar la solución de una manera gráfica, utilizando la misma imagen que en el apartado anterior:

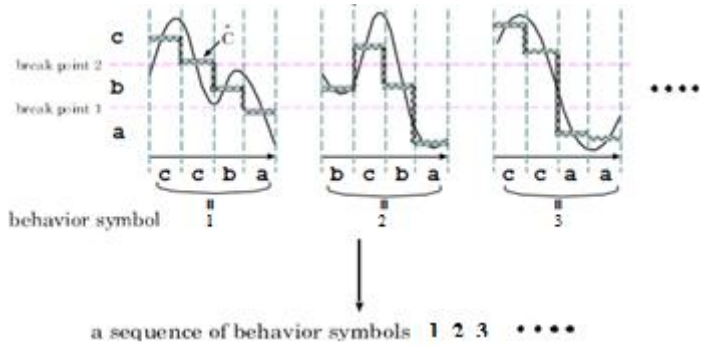


Figura 30: Adaptación de SAX a la solución propuesta

De esta manera no aparecen limitaciones en la transformación de la secuencias de símbolos, y por lo tanto se solventa el problema surgido en la propuesta teórica.

Una vez conseguida la representación "simbólica" se pueden aplicar los métodos propuestos de minería de datos con el objetivo de extraer comportamientos repetitivos de las series temporales.

4.4. Principio MDL para el descubrimiento del patrón

En los problemas de minería de datos, un objetivo que suele ser común es la búsqueda de secuencias con apariciones regulares. De ahí el hecho de crear una función que permita determinar el rango de los patrones.

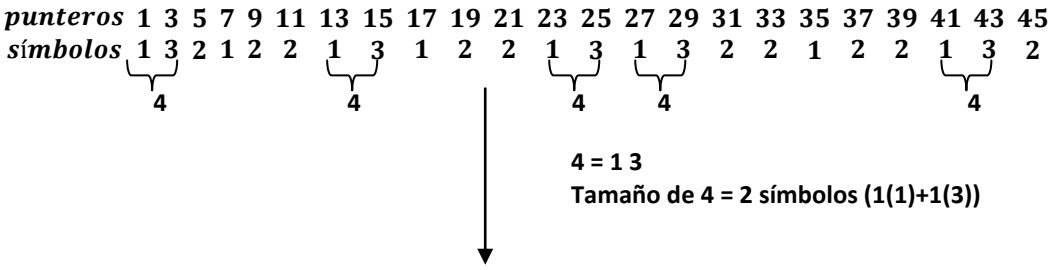
La técnica de búsqueda de patrones tiene una base en el principio de Longitud de Descripción Mínima (MDL – *Minimum Description Length*) [30] y que se puede resumir, según Grünwald como "… cualquier regularidad en un determinado conjunto de datos pueden ser utilizados para *comprimir* los datos, es decir, para describirlo con menos símbolos de los necesarios para describir los datos literalmente".

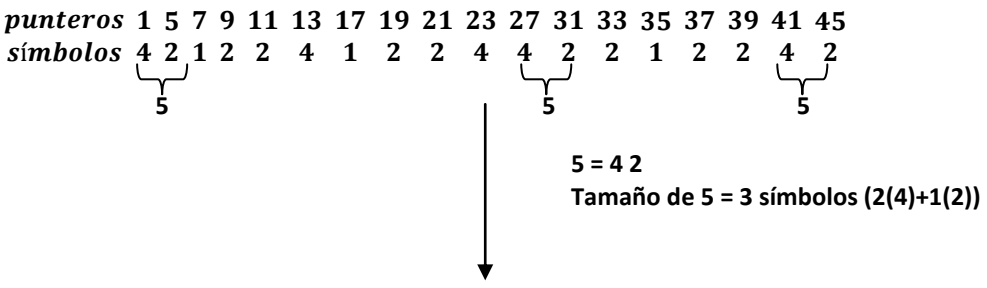
La presente técnica consiste en la búsqueda de la secuencia que más apariciones tiene en la muestra de una manera iterativa. En cada iteración, esta secuencia se sustituirá por un nuevo valor ("símbolo") de manera que la muestra inicial irá reduciéndose paulatinamente y la longitud de los nuevos símbolos irá incrementándose. Sin embargo, la complejidad no reside en este punto. La parte que más dificultad entraña es obtener el patrón con mayor longitud y llevar una correcta referencia a los punteros que lo señalan. Para solucionar esta problemática se propone lo siguiente. Se diseña una matriz de tres columnas en la que se introducirán las siguientes variantes:

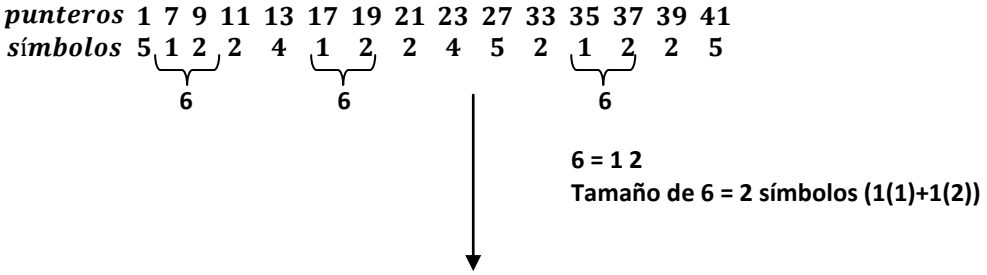
- Los punteros de la muestra principal, con lo que se tendrá total información acerca de el comportamiento de cada segmento inicial.
- El número de veces que se asocia el elemento para formar un símbolo mayor, lo que informará de la longitud del patrón establecido.
- El valor de la primera iteración en la que se utilizó el segmento para formar un símbolo de mayor longitud. Esto es de utilidad ya que la técnica trabaja agrupando de más general a particular. Con esta afirmación se puede asegurar que los símbolos iniciales en el patrón encontrado pertenecerán a la misma iteración inicial.

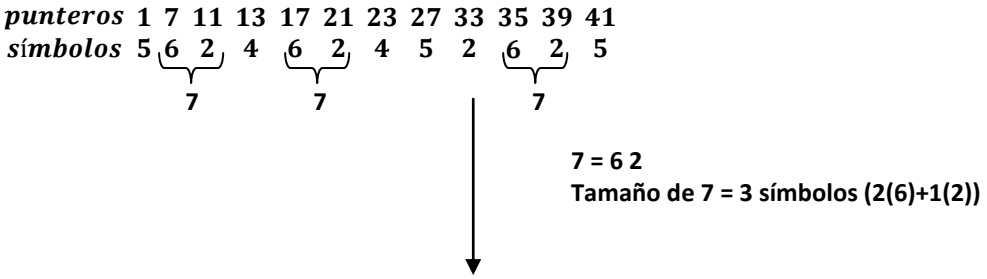
El procedimiento de actuar del método se puede explicar de la siguiente manera:

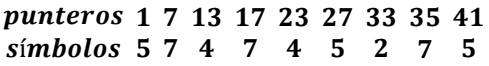
Muestra original











1	3	1
3	0	0
5	0	0
7	3	3
9	0	0
11	0	0
13	2	1
15	0	0
17	3	3
19	0	0
21	0	0
23	2	1
25	0	0
27	3	1
29	0	0
31	0	0
33	0	0
35	3	3
37	0	0
39	0	0
41	3	1
43	0	0
45	0	0
47	0	0

EXPLICACIÓN DEL EJEMPLO

En el presente ejemplo se expone una serie de prueba en la que se va a aplicar el método implementado con las características particulares de que la codificación se establece a 3 símbolos (1, 2 y 3). Para entenderlo de una manera más clara se hará un estudio de los cambios que se producen en la serie por cada iteración.

- **Iteración 1**: La agrupación de símbolos más repetitiva es "1 3" por lo que se sustituye está secuencia por un nuevo símbolo, en este caso "4". Es importante tener actualizada la lista de punteros en cada iteración, de esta manera se puede llevar referenciado cualquier símbolo de la serie original. Los símbolos de la secuencia original tienen una longitud igual a "1", esto quiere decir que no forman parte de una agrupación. Sin embargo, el nuevo valor tiene una longitud de 2, ya que es la suma de los dos símbolos primitivos (1+1 [longitud de símbolo del 1 + longitud de símbolo de 3]).

- **Iteración 2:** En este caso la secuencia más repetitiva es "4 2". Como el nuevo símbolo que se produce es incremental al anterior, su valor será "5". En este caso "5" tendrá una longitud de 3, ya que como se ha comentado anteriormente "4" tiene una longitud de 2 y el símbolo "2" tiene una longitud de 1.

- **Iteración 3:** La agrupación que más veces aparece en la serie es "1 2" y el nuevo símbolo insertado para sustituirla será "6". Su longitud será 2 ya que es la suma de dos símbolos iniciales.

- **Iteración 4:** De la misma manera que en las anteriores iteraciones, se introducirá el símbolo "7" en lugar de "6 2" y su longitud será 3 (2 de la longitud de "6" y 1 de la longitud de "2").

4.5. Modelo de agrupamiento de patrones

Cuando, mediante la técnica anterior, se alcanzan las características del patrón (longitud y secuencia de caracteres), el último paso que queda es buscar en la secuencia de entrenamiento otros segmentos de la misma longitud que tengan una tendencia parecida.

La base del funcionamiento del método implementado es el algoritmo EMMA [31]. El presente método está estrechamente ligado al anterior procedimiento de búsqueda de la longitud óptima del patrón. La técnica mixta que se utiliza recibe la longitud del motivo como en el algoritmo EMMA, pero además se apoya en la transformación PAA que se ha realizado anteriormente, con sus respectivas características de ventana de trabajo, desplazamiento de la misma y codificación simbólica. Es interesante esta visión ya que en vez de ver dos entidades independientes, se tienen dos funciones con segmentos reutilizables. Esto favorece que el trabajo de computación sea inferior.

La forma de trabajar del modelo mixto se fundamenta en una técnica de descubrimiento del motivo que por otro lado se basa en la subrutina ADM. La primera hace una primera criba de los segmentos candidatos a patrón. Para ello recorre la muestra (transformada en una representación de símbolos) y agrupa aquellos segmentos que tengan similitud con el patrón obtenido con el trabajo de la función MDL. Esta 'similitud' está determinada por un radio que se fijará y que se apoya en la función MINDIST y en la distribución de Gauss. Se ha considerado interesante el cálculo del radio de esta manera ya que mediante estos parámetros contienen las características básicas de la muestra en

representación simbólica y, por lo tanto, el radio estará estrechamente ligado a la configuración del modelo de trabajo utilizado. Una vez que se obtiene un conjunto de secuencias con un comportamiento similar, se empleará la función ADM que formará definitivamente el grupo de segmentos que formarán el vecindario. Una vez introducido el funcionamiento del enfoque del algoritmo EMMA, se procederá a una explicación más en profundidad de los métodos utilizados en el mismo.

Partiendo de la cadena de símbolos, se agruparán aquellos conjuntos de símbolos con un comportamiento parecido en lo que se va a llamar 'vecindario'. Para ello se empleará una ecuación que examinará el grado de similitud entre 2 cadenas. Esta función devolverá un valor que se comparará con un umbral establecido y proporcionará información acerca del grado de concordancia entre ambas cadenas.

El pseudocódigo implementado para encontrar el 'vecindario' y su centro es el siguiente:

Algoritmo motivo (T, motif, pointers_motifs, R, a, gauss_matrix, RR)

```
i=1;
while (i<=T.length-motif.length)
 if   MINDIST( motif , T(i) ) < R
    neighborhood = append (neighborhood, T(i) );
    i = i + motif.length ;
 else
    i = i + 1 ;
 end
end
[ central_motif, final_ neighborhood] = adm (neighborhood, R,
a, neighborhood.pointers, gauss_matrix )
return  final_ neighborhood;
return  central_motif;
return  final_neighborhood.pointers;
```

Las variables utilizadas en el pseudocódigo motivo son:

- *T*: es la cadena de símbolos que representa a la muestra a examinar.
- *motif*: es el patrón que se obtuvo con la función MDL.
- *pointers_motif*: son los punteros a los patrones.

- *R*: es el radio del vecindario. Todos los patrones que se encuentren en un radio inferior a este valor serán parte del vecindario.
- *a*: es el número de símbolos que se ha utilizado para codificar la muestra original.
- *gauss_matrix*: son los valores de la matriz de Gauss.
- *RR*: es la relación de compresión entre la señal original y su representación simbólica. Se utiliza para relacionar los punteros entre las dos señales.

Entre los mecanismos que se pueden elegir para la agrupación de patrones, se ha decidido utilizar la técnica ADM (Average Distance Matrix). Este método tiene 2 finalidades:

- La búsqueda del centro del 'vecindario', o lo que es lo mismo, encuentra el segmento que más se parece al conjunto del total.
- Agrupa los segmentos que se encuentren dentro del umbral (radio) establecido, teniendo como referencia el centro del 'vecindario'.

Esta función se basa en el cálculo de las distancias entre los distintos candidatos a patrón. En una primera exploración se genera una matriz que contiene la distancia de cada candidato a los demás, para posteriormente hacer un promedio de las distancias de cada posible patrón. Una vez obtenida dicha matriz con los promedios, se busca el patrón que tenga más patrones dentro del radio del vecindario, estableciendo dicho motivo como el centro del vecindario.

El funcionamiento de la función se puede apreciar en el pseudocódigo que se muestra a continuación:

Algorithm ADM (neighborhood, R, a, neighborhood.pointers, gauss_matrix)

```
N = neighborhood.size ;
for i=1 to N
 for j=1 to (i-1)
    a = neighborhood(i);
    b = neighborhood(j)
    adm_m(i,j) = adm(j,i) = MINDIST(a , b);
 end
end
for  k=1 to N
 for  l=1 to N
    adm_N(k , l) = adm_N(l , k) = adm_N (k , l) + adm_m (k, l) ;
 end
end
adm_N = adm_N / (N-1)^2
for  i = 1  to  N
 for  j = 1  to N
    if  adm_N (i, j) < R
       count (1, i) = count (1, i) + 1;
    end
 end
end
central_motif.index = max(count).index;
central_motif = neighborhood(central_motif.index);
for i=1  to N
 if adm_N(central_motif.index, i) < R
    final_neighborhood = append (final_neighborhood, neighborhood(i))
 end
end
return central_motif;
return final_neighborhood;
```

Tanto el pseudocódigo que se encarga de encontrar el motivo como el de la función ADM se basan en un estudio de la Universidad de California [31].

La técnica anteriormente descrita funciona haciendo un promedio de las distancias de cada candidato a patrón con todos los demás. Una vez se tenga una matriz con los promedios de cada candidato, se busca el que menor valor tenga. Este será el centro del 'vecindario'. Una vez encontrado el centro, tendremos que fijar un umbral de decisión que servirá para concretar qué candidatos a patrón están dentro del 'vecindario' y cuales fuera. Tanto las distancias promedio como este umbral se obtendrán mediante una función que se denomina MINDIST.

Algorithm MINDIST (b, c, w, a, gauss_matrix)

```
distancias = calculaDistancia (a, gauss_matrix);
for  i=1 to w
     dist  += ||distancias(b(i))| − |distancias(c(i))||;
end
return  √(w/a) · dist²
```

La función MINDIST se apoya en los umbrales de decisión de la matriz de gauss para calcular las distancias entre los símbolos de las secuencias. Además, MINDIST penaliza en mayor medida cuando las diferencias entre símbolos son mayores. Esto quiere decir que, por ejemplo, dos veces la distancia entre dos símbolos consecutivos 'a' y 'b' no es igual que la distancia entre los símbolos 'a' y 'c'. Se penaliza en mayor grado los símbolos que estén más alejados. De esta manera se tiene un ajuste más fino de los patrones y sus tendencias similares.

5. RESULTADOS DE LOS TEST EVALUADOS

5.1. Serie Simple (Dificultad Baja)

Para la primera prueba se va a utilizar una serie construida manualmente con el objetivo de probar la fiabilidad del programa de detección de patrones implementado. Existen dos pasos significativos en la búsqueda de los mismos:

1. Averiguar la longitud óptima del patrón.
2. Agrupar los segmentos de comportamiento similar, teniendo en cuenta la longitud obtenida anteriormente.

Por lo tanto, la serie va a contener el siguiente patrón:

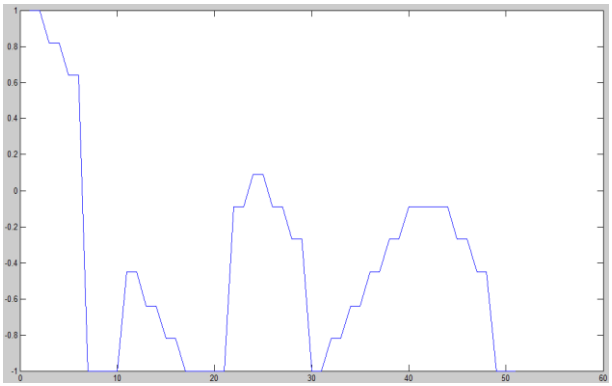


Figura 31: Patrón introducido

La longitud de dicho patrón es de 51, por lo que en el caso que el código implementado opere de la mejor manera se obtendría un patrón de longitud óptima de 51.

Este 'motivo' se va a introducir de una manera modulada (multiplicado por un factor) para que su detección no sea obvia. La cadena o señal final que se va a tratar es la siguiente:

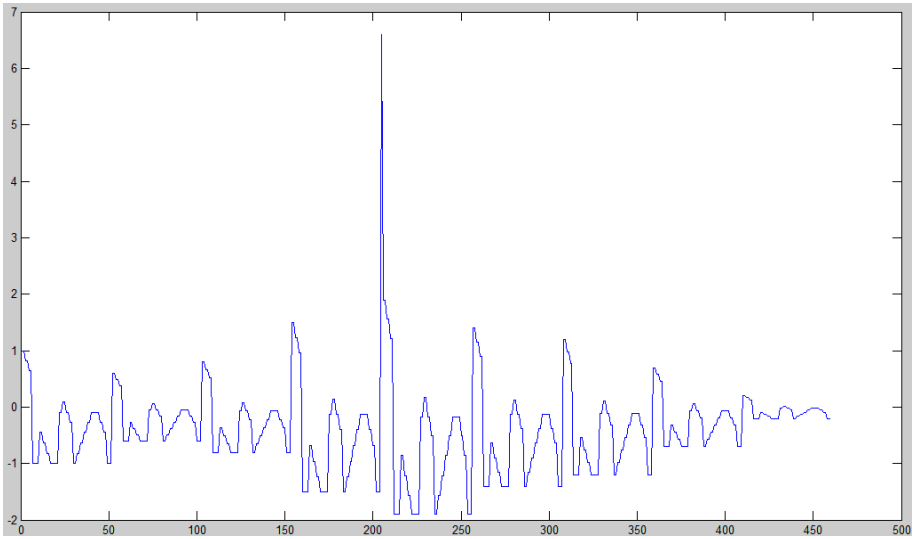


Figura 32: Serie artificial a evaluar

5.1.1. Longitud del patrón

La prueba que se va a realizar consistirá en un 'barrido' de las variables necesarias para solucionar el problema con la intención de poder conocer la combinación que mejor resultado produce. Además, de esta manera, se podrá tener información del tipo de patrones que se obtienen con una determinada configuración. Los factores que se van a tener en cuenta en el barrido van a ser:

- Ventana utilizada (T_{min})
- Número de valores numéricos que formarán un símbolo (N_s).
- Desplazamiento de la ventana (D).
- Número de símbolos utilizados para la codificación (S).

Resultados obtenidos

Valores numéricos/símbolo = 2

	DESPLAZAMIENTO	
	Tmin/3	**Tmin/5**
Tmin=6	46	52
Tmin=8	43	47
Tmin=10	43	43
Tmin=12	8	41
Tmin=14	-	43
Tmin=16	-	40

Valores numéricos/símbolo = 3

	DESPLAZAMIENTO	
	Tmin/3	**Tmin/5**
Tmin=9	52	41
Tmin=12	8	41
Tmin=15	-	40
Tmin=18	-	-
Tmin=14	-	-
Tmin=16	-	-

Tabla 1: Longitudes de patrón obtenidas para distintas combinaciones de ventanas y desplazamientos de ventana

Influencia de los resultados dependiendo del número de símbolos de la codificación (Desplazamiento=Tmin/5)

	CODIFICACIÓN	
	3 símbolos	**7 símbolos**
Tmin=6	52	52
Tmin=8	47	45
Tmin=10	43	43
Tmin=12	41	41
Tmin=14	43	40
Tmin=16	40	38

Tabla 2: Longitudes de patrón que se obtienen variando la ventana y el nº de símbolos

Representación de los resultados más interesantes

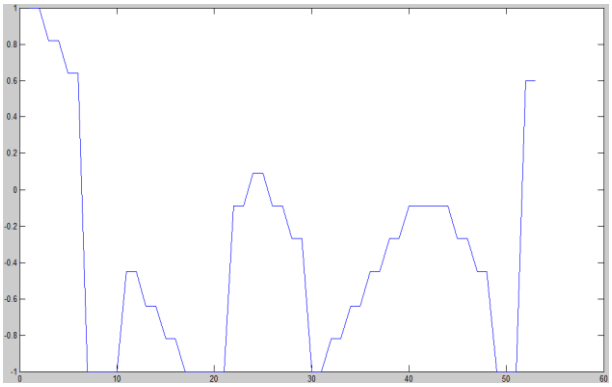


Figura 33: Patrón con una configuración de Tmin = 6, desplazamiento = Tmin/5 y 3 símbolos utilizados para la codificación

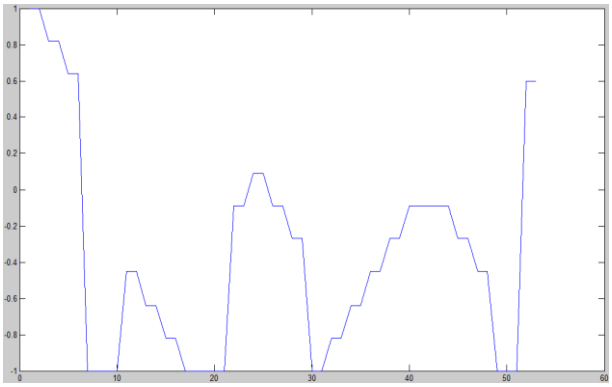


Figura 34: Patrón con una configuración de Tmin = 9, desplazamiento = Tmin/3 y 3 símbolos utilizados para la codificación

5.1.2. Conclusiones – Prueba Simple

Hay varios aspectos que se pueden destacar a la vista de los resultados obtenidos.

1. Influencia del número de símbolos por ventana (T_{min}/N_s) : la elección de un tamaño adecuado de ventana pueden aumentar la eficiencia de la solución considerablemente. Cuanto menor es la ventana, el resultado es más preciso debido a que los fragmentos que se agrupan son más pequeños, con lo que el ajuste es mejor. Es más fácil encontrar segmentos de 3 símbolos que de 5, por ejemplo. El punto en contra es que el tiempo de cálculo tiene una penalización con esta configuración.

2. Influencia del desplazamiento: Es obvio que el desplazamiento juega un papel fundamental en los resultados que se requieren. Pero, como ocurre en el punto 1, el tiempo de computación se ve influenciado de una manera importante. Es crucial

buscar un compromiso entre la longitud de las medidas a analizar y el desplazamiento de la ventana elegida.

3. Dependencias con el número de símbolos de la codificación: La consecuencia directa de aumentar el número de símbolos de la codificación es el aumento de la precisión en la secuencia. Esto provoca que sea más difícil encontrar patrones ya que se complica la búsqueda de secuencias con los mismos códigos. Esta es la razón por la que en la presente comparación entre una codificación con 3 símbolos y una con 5, en la última la longitud del patrón sea menor. Cuando se aumente la complejidad de la serie temporal a analizar se podrá observar más claramente este factor.

5.2. Prueba Victoria-1 (Dificultad Media)

En este caso se ha aumentado la dificultad con el objetivo de verificar que la solución resuelve el problema de búsqueda de patrones. Además, se comprobará que las conclusiones sacadas en la prueba simple son correctas.

En esta ocasión se va a utilizar un patrón más complejo y se combinará con ruido blanco para dificultar aún más la detección del patrón. El motivo a introducir tiene la siguiente forma:

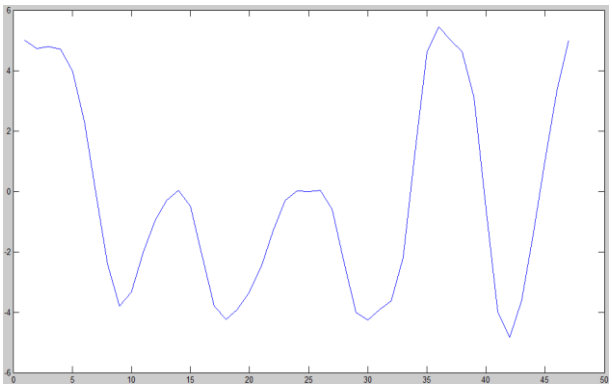


Figura 35: Patrón introducido en la serie Victoria 1

Este patrón tiene una longitud de 46 y se ha insertado en la secuencia un número de 33 veces. Por lo tanto, la mejor combinación debería dar como resultado:

- **Longitud del Patrón = 46**
- **Nº de patrones = 33**

Para dar más verosimilitud a la muestra a procesar se introduce el patrón y el ruido multiplicados por determinados factores que harán que la detección sea menos trivial que si se introdujeran en su forma natural. Por lo tanto, la secuencia temporal a analizar tiene la siguiente constitución:

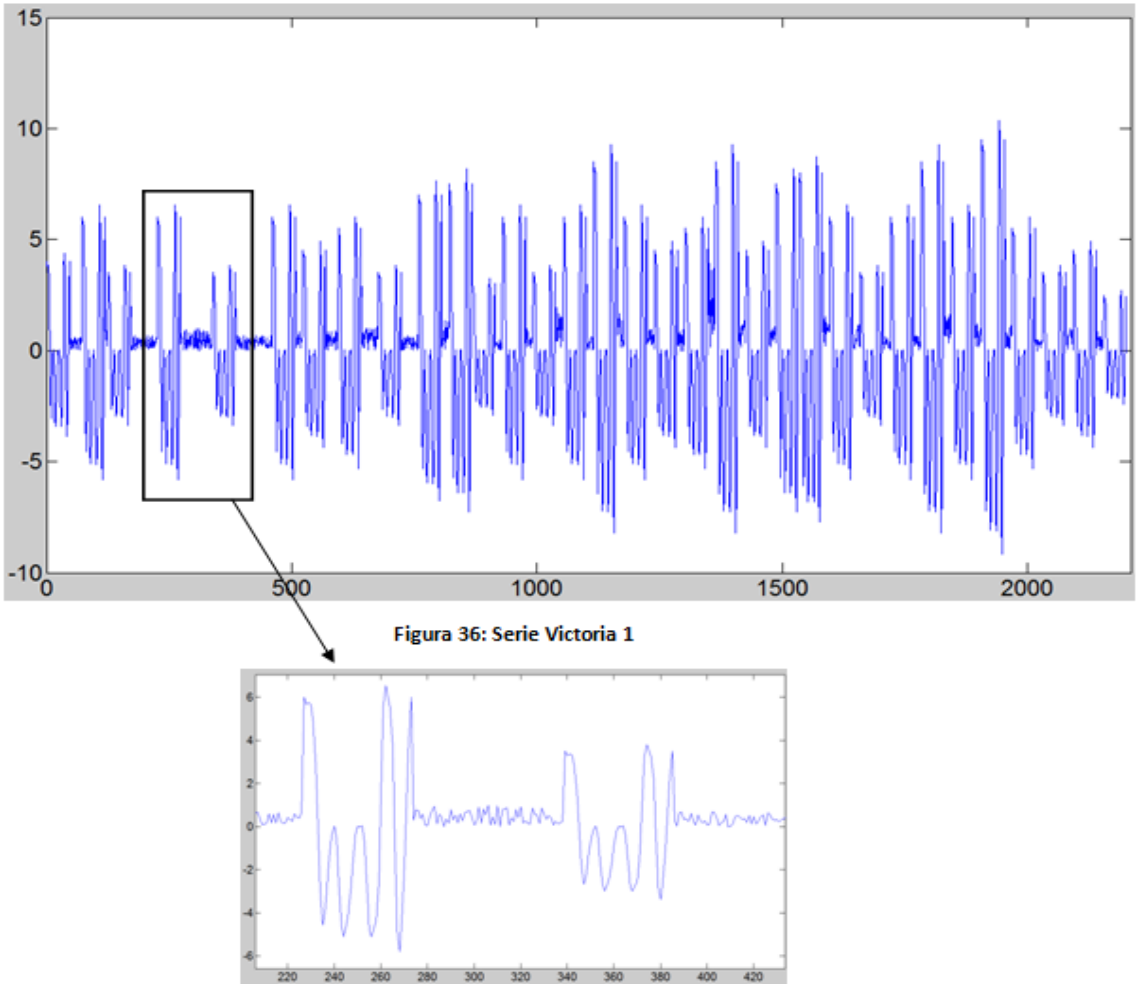


Figura 36: Serie Victoria 1

Además de mostrar, como en la ocasión anterior, los patrones detectados, se realizará un estudio de la tendencia de la longitud del patrón óptima comparado con las variables del sistema y se mostrará de forma gráfica.

5.2.1. Estudio de los resultados

FASE I

	DESPLAZAMIENTO						
	Tmin/2	**Tmin/3**	**Tmin/4**	**Tmin/5**	**Tmin/6**	**Tmin/7**	**Tmin/8**
Tmin=6	48 / 3	46 / 7	46 / 5	45 / 3	47 / 5	46 / 3	47 / 5
Tmin=8	44 / 11	45 / 5	44 / 12	44 / 8	44 / 5	44 / 3	44 / 5
Tmin=10	45 / 4	42 / 11	42 / 11	44 / 6	44 / 6	43 / 6	43 / 6
Tmin=12	42 / 8	44 / 4	42 / 3	42 / 4	42 / 9	42 / 4	42 / 9
Tmin=14	42 / 3	40 / 5	40 / 4	36 / 11	40 / 4	40 / 3	40 / 4
Tmin=16	40 / 3	35 / 5	36 / 9	36 / 10	39 / 3	38 / 5	38 / 7
Tmin=18	36 / 6	36 / 6	35 / 5	40 / 3	33 / 5	33 / 8	34 / 11

Tmin=20	30 / 5	35 / 5	30 / 7	32 / 9	33 / 8	36 / 3	33 / 8
Tmin=22	33 / 4	35 / 4	30 / 8	32 / 7	32 / 5	33 / 3	30 / 11
Tmin=24	36 / 4	24 / 7	36 / 3	30 / 4	32 / 5	30 / 4	30 / 5
Tmin=26	26 / 5	36 / 6	28 / 6	25 / 7	32 / 3	32 / 5	27 / 6
Tmin=28	28 / 4	27 / 6	28 / 5	24 / 9	25 / 7	28 / 3	24 / 11
Tmin=30	24 / 4	30 / 3	24 / 7	24 / 7	25 / 4	28 / 4	24 / 8
Tmin=32	-	22 / 4	24 / 4	24 / 5	25 / 3	20 / 7	24 / 4
Tmin=34	-	22 / 4	27 / 3	21 / 4	24 / 3	20 / 7	20 / 7
Tmin=36	-	24 / 4	18 / 6	28 / 3	24 / 3	20 / 6	20 / 3
Tmin=38	-	26 / 4	20 / 5	16 / 4	18 / 7	20 / 3	15 / 7
Tmin=40	-	-	-	16 / 3	14 / 5	18 / 3	15 / 6

Tabla 3: ***Longitudes de patrón/nº de patrones*** **obtenidos en la fase I**

TENDENCIA DE LA LONGITUD ÓPTIMA OBTENIDA

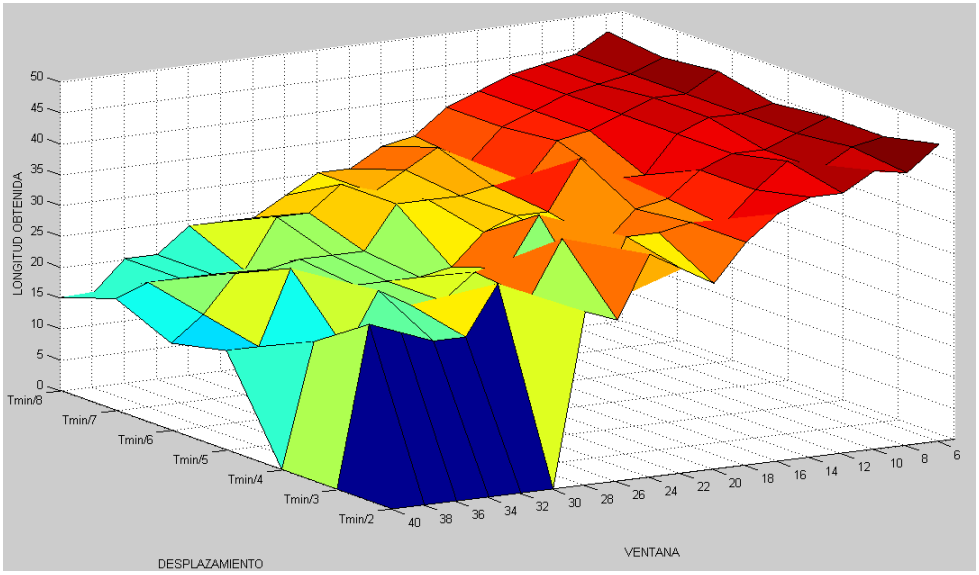


Figura 36: Longitud del patrón en función de la ventana y el desplazamiento

PATRONES ENCONTRADOS EN FUNCIÓN DE LA VENTANA Y EL DESPLAZAMIENTO

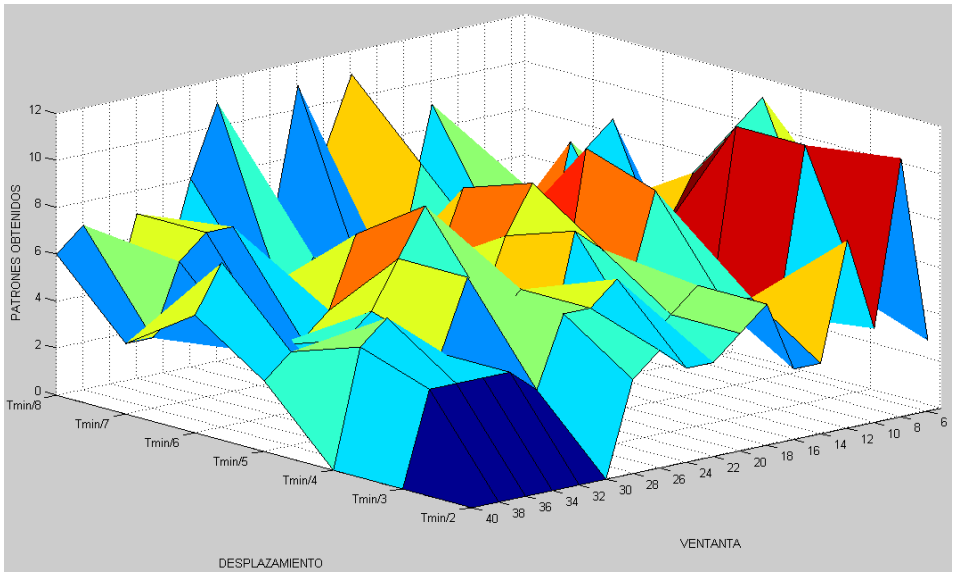


Figura 37: Número de patrones que se obtienen en función de la ventana y el desplazamiento

FASE II

Una vez conseguida la longitud óptima que va a tener el patrón, se somete la secuencia a estudio con el objetivo de encontrar todas aquellas sub-cadenas que tengan un comportamiento afín al patrón que se ha obtenido anteriormente. Los resultados se muestran de la misma forma que en la fase anterior (Longitud patrón/nº patrones encontrados):

	DESPLAZAMIENTO						
	Tmin / 2	**Tmin / 3**	**Tmin / 4**	**Tmin / 5**	**Tmin / 6**	**Tmin / 7**	**Tmin / 8**
Tmin=6	48 / 18	46 / 30	46 / 31	45 / 32	47 / 31	46 / 31	47 / 31
Tmin=8	44 / 13	45 / 20	44 / 31	44 / 27	44 / 31	44 / 31	44 / 31
Tmin=10	45 / 8	42 / 11	42 / 11	44 / 21	44 / 20	43 / 32	43 / 31
Tmin=12	42 / 19	44 / 27	42 / 32	42 / 25	42 / 25	42 / 25	42 / 24
Tmin=14	42 / 10	40 / 14	40 / 25	36 / 32	40 / 29	40 / 29	40 / 29
Tmin=16	40 / 18	35 / 36	36 / 28	36 / 32	39 / 32	38 / 21	38 / 21
Tmin=18	36 / 19	36 / 22	35 / 25	40 / 23	33 / 32	33 / 32	34 / 21
Tmin=20	30 / 30	35 / 24	30 / 26	32 / 28	33 / 31	36 / 31	33 / 31
Tmin=22	33 / 25	35 / 14	30 / 23	32 / 32	32 / 32	33 / 32	30 / 32
Tmin=24	36 / 25	24 / 25	36 / 28	30 / 32	32 / 30	30 / 32	30 / 33
Tmin=26	26 / 20	36 / 31	28 / 25	25 / 33	32 / 26	32 / 25	27 / 32
Tmin=28	28 / 26	27 / 33	28 / 30	24 / 33	25 / 33	28 / 33	24 / 33
Tmin=30	24/27	30 / 30	24 / 33	24 / 34	25 / 33	28 / 32	24 / 36
Tmin=32	-	22 / 31	24 / 27	24 / 35	25 / 32	20 / 34	24 / 36
Tmin=34	-	22 / 32	27 / 32	21 / 33	24 / 34	20 / 35	20 / 37
Tmin=36	-	24 / 27	18 / 34	28 / 31	24 / 32	20 / 32	20 / 34
Tmin=38	-	26 / 27	20 / 31	16 / 34	18 / 47	20 / 33	15 / 57
Tmin=40	-	-	-	16 / 52	14 / 41	18 / 37	15 / 34

Tabla 4: *Longitudes de patrón/nº patrones del vecindario* obtenidos en la fase II

PATRONES ENCONTRADOS EN FUNCIÓN DE LA VENTANA Y EL DESPLAZAMIENTO

Como se ha comentado anteriormente, el objetivo de esta fase es el encontrar el mayor número de patrones con las características especificadas en la etapa anterior, por lo tanto el estudio se centrará en el resultado del número de patrones obtenidos (lo que se denomina vecindario). Para tener una idea más clara de la mejora se presentarán los resultados de una manera porcentual entre el número de patrones que ha devuelto el código y el número de patrones que contenía la muestra de entrenamiento, que como ya se ha comentado, son 33.

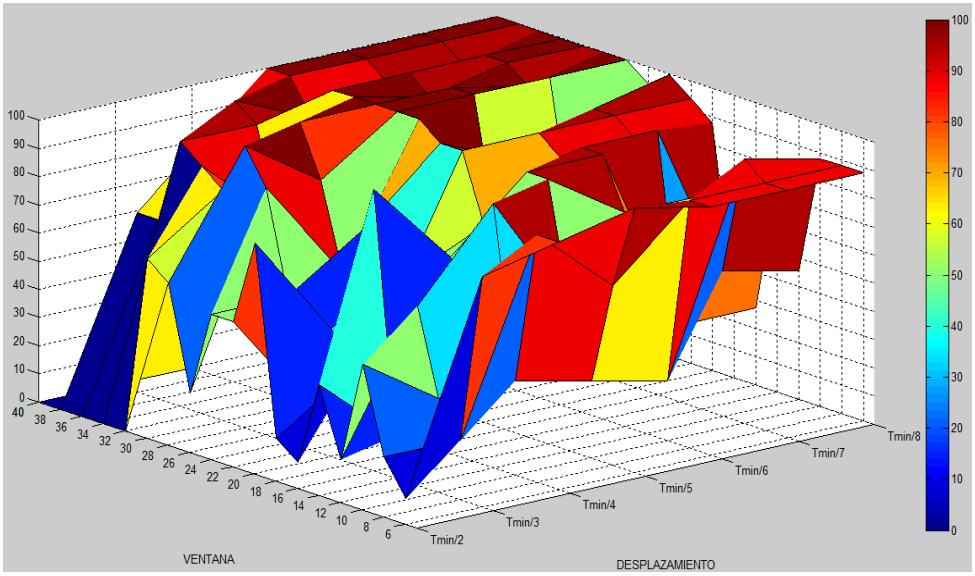


Figura 38: Relación entre patrones obtenidos con el método y los patrones insertados manualmente en función de la ventana y el desplazamiento

Este gráfico muestra la influencia tanto de la ventana, como del desplazamiento, a la hora de encontrar el número óptimo de patrones. Para un desplazamiento mayor de Tmin/5 y una ventana mayor de 30 se obtienen los mejores resultados. Este resultado es lógico que ya el patrón que necesitan encontrar es de una longitud reducida. También añadir que en la mayoría de los casos se obtienen resultados superiores al 80% de acierto, lo que da una idea de que el funcionamiento del código es bastante aceptable.

A la vista de los resultados obtenidos, se elige la solución que mejor se acerca a la ideal para su estudio. Se ha optado por la combinación de una ventana de valor 6 y un desplazamiento de Tmin/6. El estudio identifica lo siguiente:

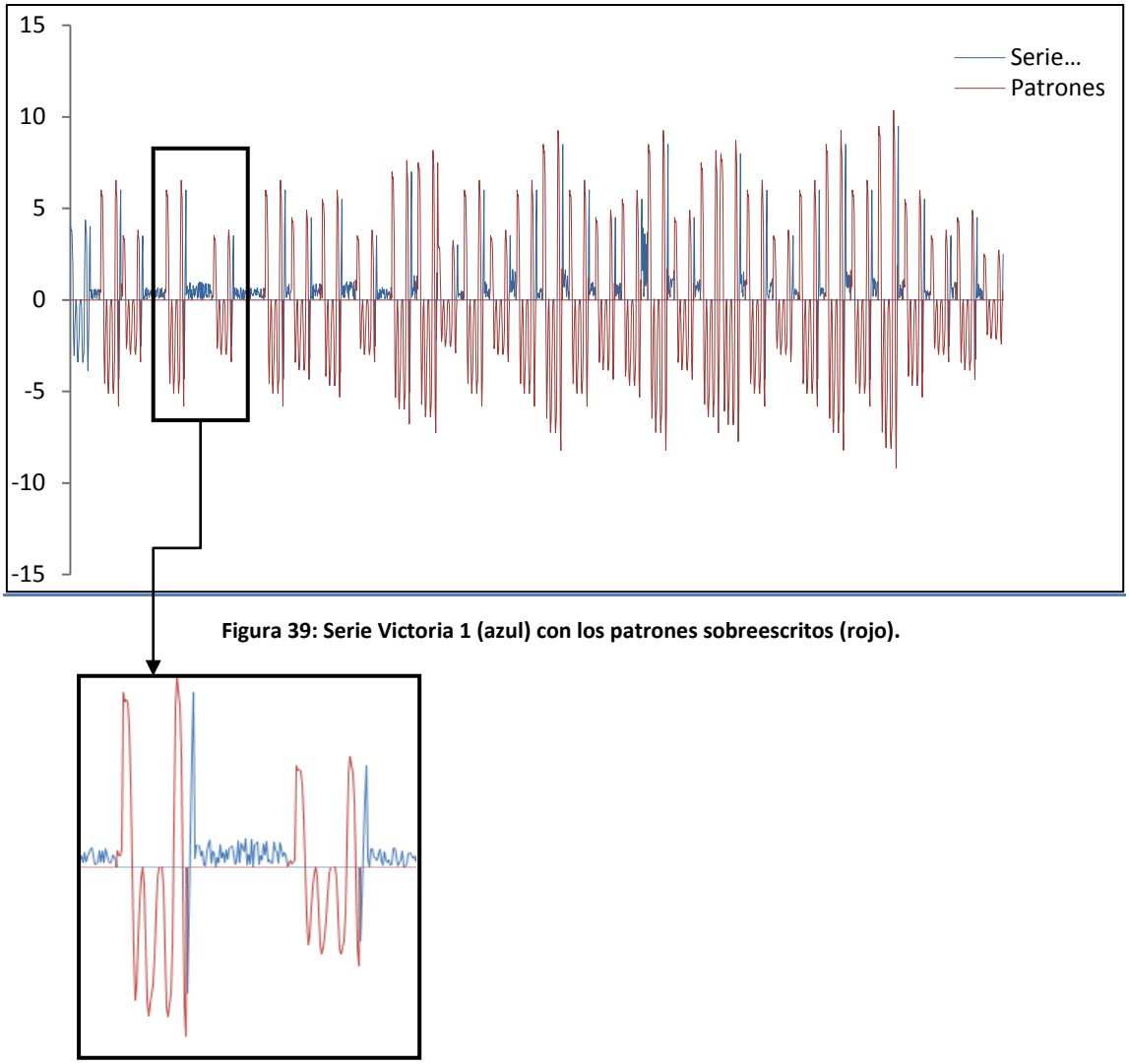


Figura 39: Serie Victoria 1 (azul) con los patrones sobreescritos (rojo).

5.2.2. Conclusiones – Prueba Victoria-1

Con los datos mostrados anteriormente, se pueden obtener algunas conclusiones interesantes.

- Teniendo en cuenta las especificaciones del patrón, longitud de 46 y está inmerso en la secuencia un total de 33 veces, se puede afirmar que los mejores resultados se dan con una ventana de análisis pequeña.
- Se observa que el desplazamiento ayuda a una detección más en profundidad. Este hecho es trivial, ya que con un desplazamiento de ventana menor se aumenta la redundancia.
- Como contrapunto, hay que destacar que la fase I tiene dificultades a la hora de encontrar un número alto de patrones, debido posiblemente a que los patrones encontrados tienen que ser idénticos. Este número aumenta en la fase II ya que a estos patrones idénticos se le suman segmentos que tienen un comportamiento muy similar a ellos, de acuerdo a la fórmula mostrada en el apartado.

5.3. Prueba Victoria-2 (Dificultad Media)

De manera similar a lo realizado en el apartado anterior, se va a estudiar la siguiente serie temporal con la introducción del patrón mostrado:

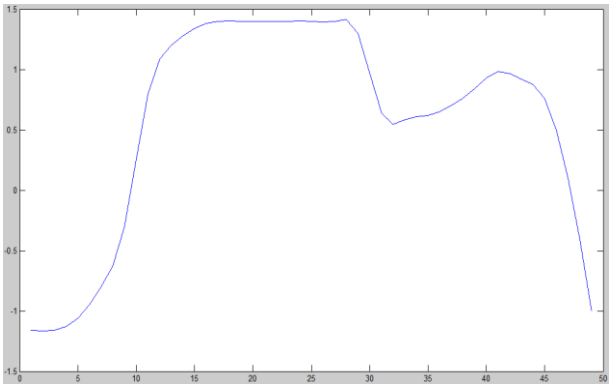


Figura 40: Patrón a introducir

Este patrón tiene una longitud de 49 y se ha insertado en la secuencia un número de 33 veces. Por lo tanto, la mejor combinación debería dar como resultado:

- **Longitud del Patrón = 49**
- **Nº de patrones = 33**

La serie temporal que vamos a analizar se muestra a continuación:

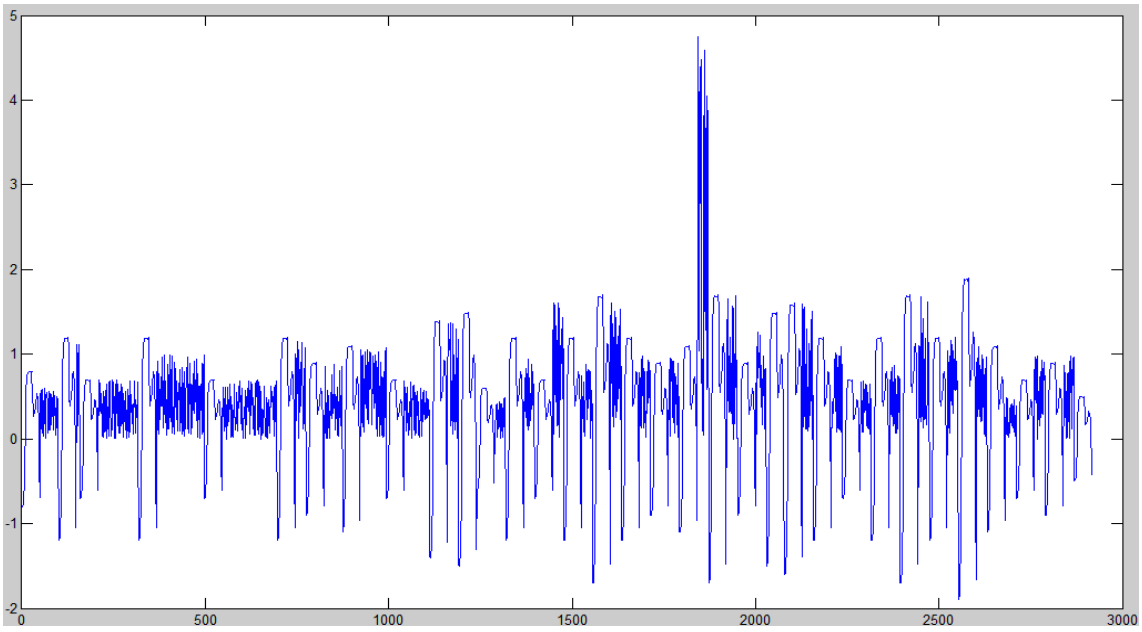


Figura 41: Muestra Victoria 2

5.3.1. Estudio de los resultados

FASE I

	DESPLAZAMIENTO						
	Tmin/2	**Tmin/3**	**Tmin/4**	**Tmin/5**	**Tmin/6**	**Tmin/7**	**Tmin/8**
Tmin=6	54/3	50/3	54/4	50/6	51/3	51/3	51/3
Tmin=8	48/6	48/4	48/5	50/3	47/4	47/4	47/4
Tmin=10	45/4	45/8	45/5	46/3	46/3	42/14	42/14
Tmin=12	42/6	44/6	42/4	44/5	44/5	44/5	44/5
Tmin=14	42/5	45/5	44/4	42/5	44/3	44/3	44/3
Tmin=16	40/6	40/4	40/3	39/4	39/4	40/3	40/3
Tmin=18	36/6	36/6	36/6	40/4	36/8	36/8	38/5
Tmin=20	40/4	35/5	35/7	40/3	36/3	36/3	36/3
Tmin=22	33/5	35/4	35/5	36/3	36/3	33/4	33/4
Tmin=24	36/6	32/5	30/6	30/7	32/4	30/7	30/7
Tmin=26	26/3	27/6	30/6	30/5	28/4	28/4	27/7
Tmin=28	28/4	27/6	28/5	30/3	25/7	28/4	28/4
Tmin=30	-	30/3	28/5	24/6	25/7	24/7	24/7
Tmin=32	32/4	22/5	24/7	24/4	25/3	25/3	28/4
Tmin=34	-	22/5	24/5	21/5	24/3	25/3	20/7
Tmin=36	-	-	18/6	21/4	18/6	20/4	20/4
Tmin=38	-	-	20/4	16/7	18/5	20/3	20/3
Tmin=40	-	-	-	16/6	14/5	18/3	15/6

Tabla 5: Longitudes de patrón/nº de patrones obtenidos en la fase I

TENDENCIA DE LA LONGITUD ÓPTIMA OBTENIDA

El procesamiento de los datos en la fase I ha dado una longitud con las características que tiene la gráfica que se muestra a continuación.

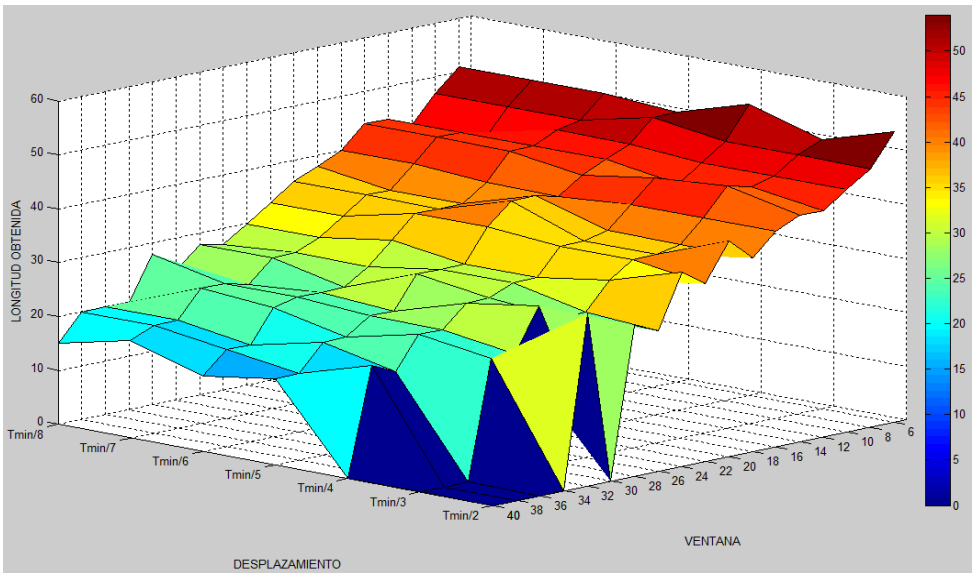


Figura 42: Longitud del patrón en función de la ventana y el desplazamiento

Como ocurría en la prueba 'Victoria 1', los mejores resultados se agrupan en la zona en la que se tienen valores de ventana pequeños. Como ya se podía intuir, este hecho se da a causa de que es más fácil agrupar pequeños segmentos; su contrapunto es que la sobrecarga aumenta.

PATRONES ENCONTRADOS EN FUNCIÓN DE LA VENTANA Y EL DESPLAZAMIENTO

La otra variante que interesa es la evolución del número de patrones con respecto al desplazamiento y la ventana.

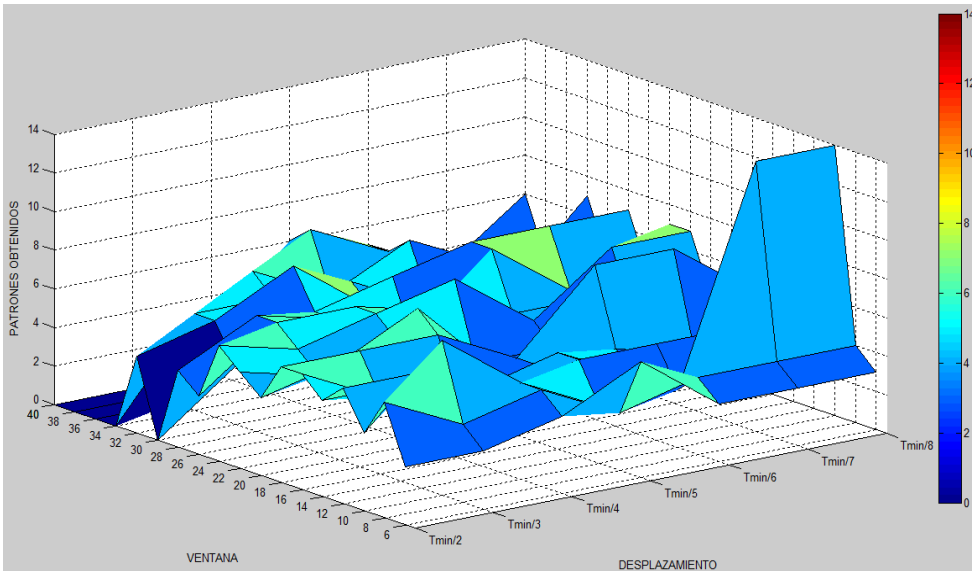


Figura 43: Número de patrones que se obtienen en la fase I en función de la ventana y el desplazamiento

Como se puede apreciar en la imagen, el número de patrones que detecta la fase I es bajo. Esta es una de las características que tiene mayor margen de mejora.

FASE II

	DESPLAZAMIENTO						
	Tmin/2	**Tmin/3**	**Tmin/4**	**Tmin/5**	**Tmin/6**	**Tmin/7**	**Tmin/8**
Tmin=6	54/26	50/31	54/29	50/31	51/31	51/31	51/31
Tmin=8	48/32	48/32	48/32	50/31	47/32	47/32	47/32
Tmin=10	45/19	45/31	45/32	46/32	46/32	42/32	42/32
Tmin=12	42/25	44/26	42/32	44/32	44/32	44/32	44/32
Tmin=14	42/19	45/26	44/32	42/31	44/32	44/32	44/32
Tmin=16	40/26	40/32	40/32	39/32	39/32	40/32	40/32
Tmin=18	36/20	36/32	36/33	40/32	36/33	36/33	38/32
Tmin=20	40/19	35/32	35/32	40/32	36/32	36/32	36/32
Tmin=22	33/22	35/30	35/33	36/32	36/32	33/32	33/32
Tmin=24	36/20	32/27	30/33	30/33	32/32	30/32	30/32
Tmin=26	26/19	27/27	30/33	30/32	28/32	28/32	27/34
Tmin=28	28/21	27/23	28/31	30/32	25/34	28/33	28/33
Tmin=30	-	30/28	28/32	24/33	25/33	24/38	24/38
Tmin=32	32/23	22/26	24/32	24/32	25/32	25/32	28/32
Tmin=34	-	22/24	24/32	21/34	24/32	25/32	20/46
Tmin=36	-	-	18/33	21/33	18/35	20/34	20/34
Tmin=38	-	-	20/30	16/38	18/35	20/33	20/33
Tmin=40	-	-	-	16/36	14/35	18/35	15/39

Tabla 6: Longitudes de patrón/nº patrones del vecindario obtenidos en la fase II

PATRONES ENCONTRADOS EN FUNCIÓN DE LA VENTANA Y EL DESPLAZAMIENTO

Al igual que en la prueba 'Victoria 1' se han introducido 33 patrones, por lo tanto el resultado ideal del código de búsqueda de patrones será 33. De acuerdo a esto, el grado de acierto que tiene la evaluación es:

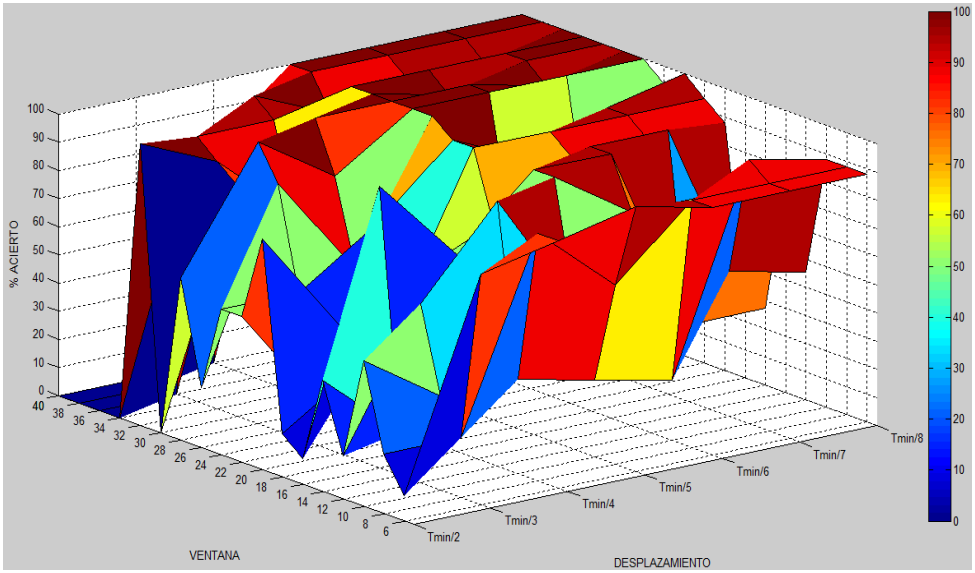


Figura 44: Porcentaje de acierto entre el número de patrones obtenidos en la fase II y el número de patrones introducidos.

5.3.2. Conclusiones – Prueba Victoria 2

Una vez más se comprueba que los datos que se obtienen tienen una estrecha relación con los ya vistos anteriormente. A partir del desplazamiento Tmin/5 se puede observar que los resultados tienen un grado de acierto alto. Pero lo que más interesa es que estos resultados optimistas coincidan con una longitud de patrón que sea correcta. Por lo tanto, a pesar de que con ventanas de valor alto se obtiene un resultado mejor, el estudio se centrará en aquellas ventanas que tengan un valor cercano a los 49 ideales con un número de patrones encontrados lo más próximo a 33. Esta vez se ha elegido la combinación de ventana de valor 8 y desplazamiento Tmin/5.

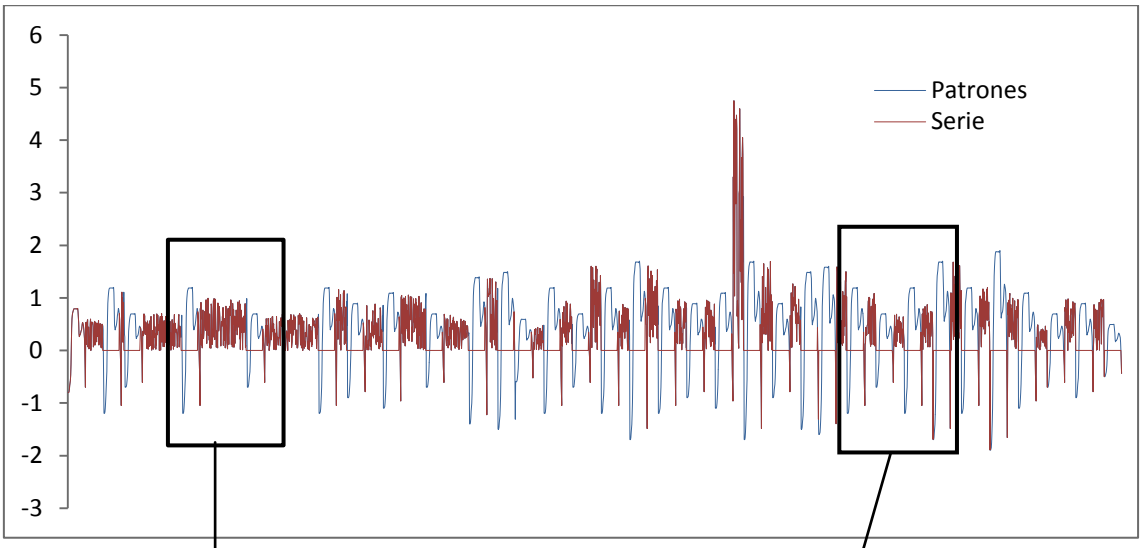
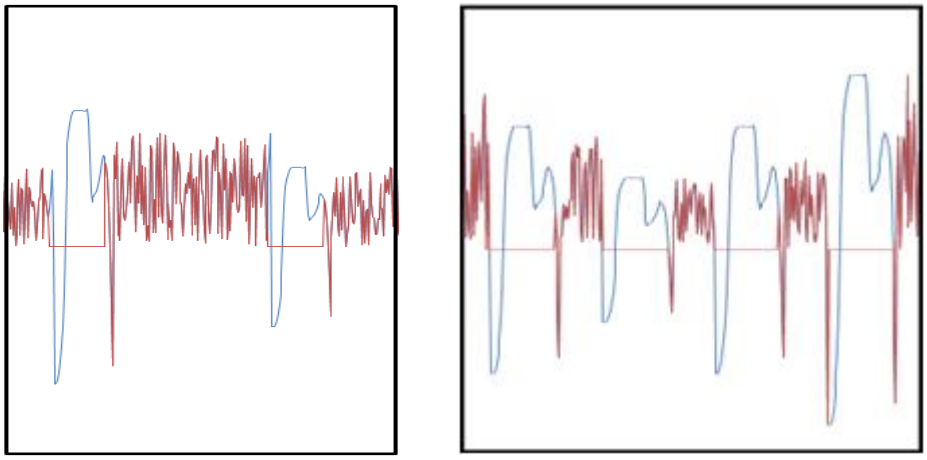


Figura 45: Muestra Victoria 2 (Rojo) con los patrones sobreimpresionados (azul)



5.4. Prueba EDAT (Dificultad Media-Alta)

Como se viene produciendo a lo largo de este apartado, se prosigue en el aumento de la dificultad de las secuencias de entrada con el fin de evaluar la respuesta del sistema. Esta vez la entrada será una secuencia temporal obtenida de la referencia [30].

Esta señal proviene de un conjunto de mediciones realizadas por el telescopio Whole Earth (en realidad es un conjunto de telescopios coordinados) que recoge información de la intensidad de la luz de una estrella, más concretamente de una enana blanca, durante el mes de Marzo de 1989.

Como inicialmente esta serie no tiene ningún carácter repetitivo conocido, se ha procedido a introducir un patrón un número determinado de veces. Esto da la ventaja de conocer cuál debe ser el comportamiento frente a la secuencia original.

Para complicar aún más la prueba, este patrón se introducirá antes del pre-procesamiento y se multiplicará su valor por un factor aleatorio, lo cual tiene como objetivo que la escala de ese factor varíe, como se apreciará más tarde.

De manera similar a lo realizado en el apartado anterior, vamos a estudiar la serie temporal con la introducción del patrón mostrado:

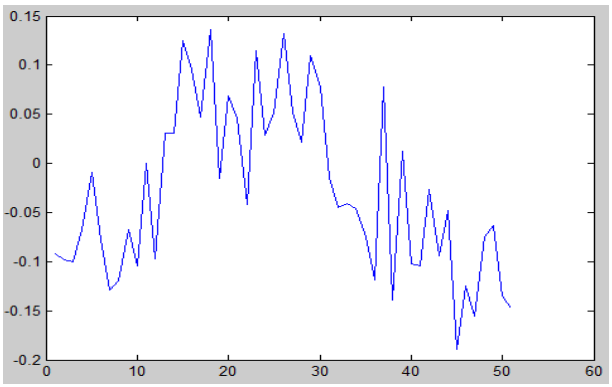


Figura 46: Patrón introducido de manera externa.

Este patrón tiene una longitud de 51 y se ha insertado en la secuencia un número de 8 veces. Por lo tanto, la mejor combinación debería dar como resultado:

- **Longitud del Patrón = 51**
- **Nº de patrones $\geq$ 8**

Una vez introducido el patrón deseado el número de veces especificado anteriormente, comprobamos el resultado de la secuencia original después de traspasar la etapa de filtrado.

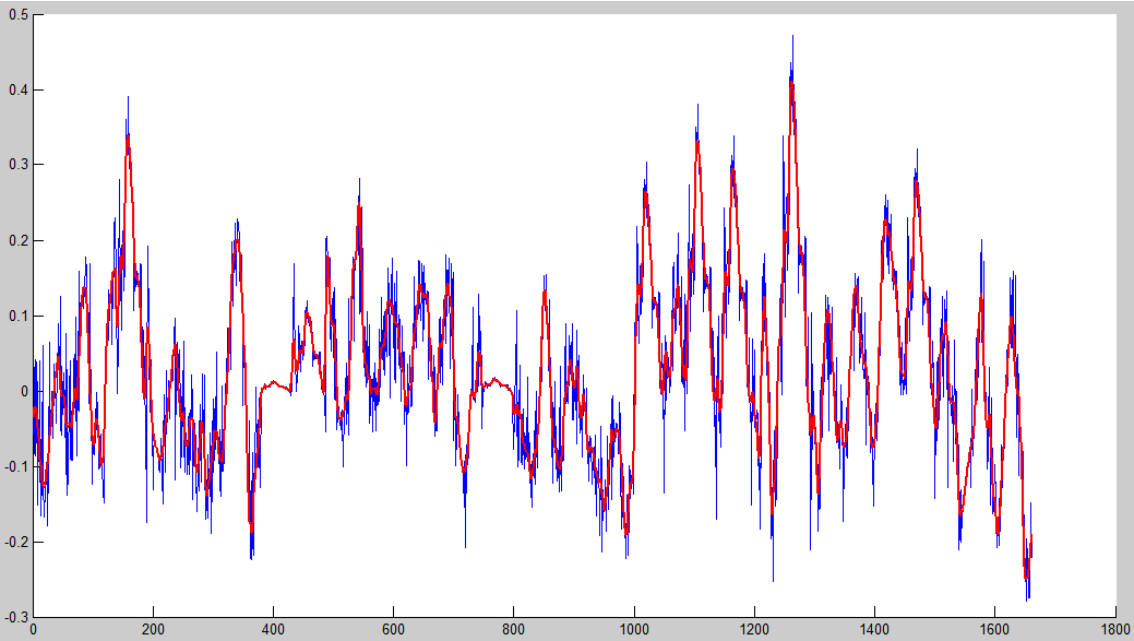


Figura 47: Señal de entrada (azul) frente a la seña suavizada (rojo).

Como se puede observar en la figura, la etapa de suavizado ha transformado las variaciones rápidas (frecuencias altas) en variaciones lentas sin influir determinantemente en la información de la señal. De esta manera la búsqueda del patrón es más sencilla.

Una vez pre-procesada la señal, se muestra el resultado con los patrones introducidos marcados de manera que se pueda observar en su totalidad la señal y sus redundancias.

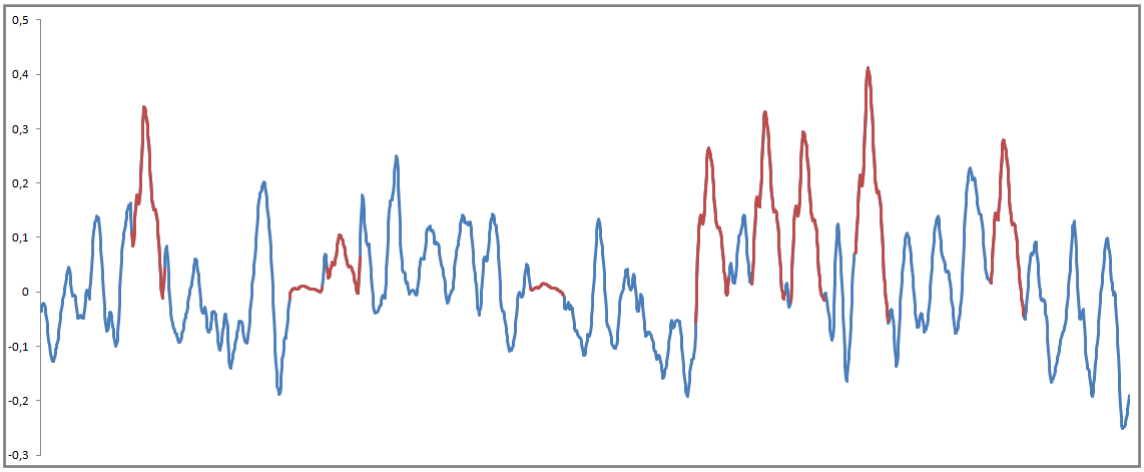


Figura 48: Muestra EDAT Filtrada

Es interesante destacar como algunos de los patrones que se introducen no son fácilmente detectables a causa del factor por el que se modulo dicho patrón. Esto aumenta aún más la dificultad de descubrir el patrón oculto en la secuencia original.

5.4.1. Estudio de los resultados obtenidos

Para evaluar los resultados se deben tener en cuenta 2 parámetros de información que contiene el patrón insertado: su longitud y el número de ocurrencias en la secuencia.

Posteriormente se realizará un barrido sobre los parámetros de configuración que posee la solución propuesta en el estudio (T_{min}, tamaño del símbolo y número de símbolos de la codificación). Con el fin de exponer una solución más clara y directa, se mostrarán sólo aquellas combinaciones que mejor resultado hayan mostrado. Cabe destacar que para evaluar dicha combinación se ha establecido un criterio que tiene en cuenta el porcentaje de acierto entre la longitud del patrón que detecta y el número de aciertos en los elementos obtenidos. De esta forma, el mejor porcentaje conjunto de estos dos valores será la mejor solución al sistema. Las combinaciones que mejor resultado han obtenido son las siguientes:

- *Tmin=12 | Valores/símbolo=2 | Símbolos=3*
- *Tmin=12 | Valores/símbolo=3 | Símbolos=3*
- *Tmin=12 | Valores/símbolo=4 | Símbolos=4*

Tmin=12 | Símbolos=3 | Valores/Símbolo=2

Los resultados hallados aplicando estos parámetros han sido una longitud óptima de 36 y un total de 16 patrones, entre los cuales se encuentran los 8 patrones introducidos externamente. Por lo tanto, se puede establecer el grado de acierto como:

$$\frac{L_{opt}}{L_{real}} = \frac{36}{50} = 72\% \qquad \frac{n_{patrones\_opt}}{L_{patrones\_real}} = \frac{7}{8} = 87{,}5\%$$

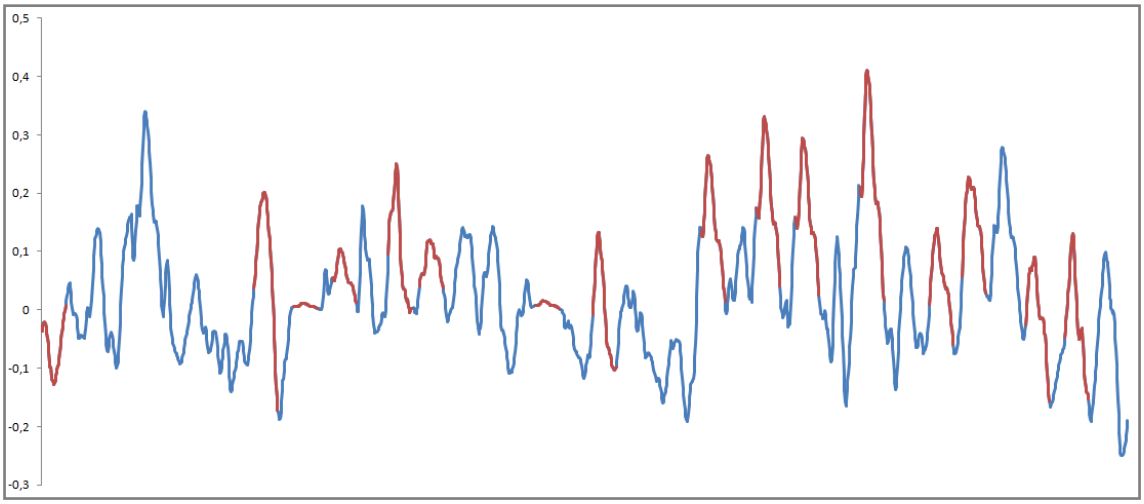


Figura 49: Señal con los patrones detectados de la primera solución

Tmin=12 | Símbolos=3 | Valores/Símbolo=3

Los resultados hallados aplicando estos parámetros han sido una longitud óptima de 36 y un total de 13 patrones, entre los cuales se encuentran los 8 patrones introducidos externamente. Por lo tanto, se puede establecer el grado de acierto como:

$$\frac{L_{opt}}{L_{real}} = \frac{36}{50} = 72\% \qquad \frac{n_{patrones\_opt}}{L_{patrones\_real}} = \frac{7}{8} = 87{,}5\%$$

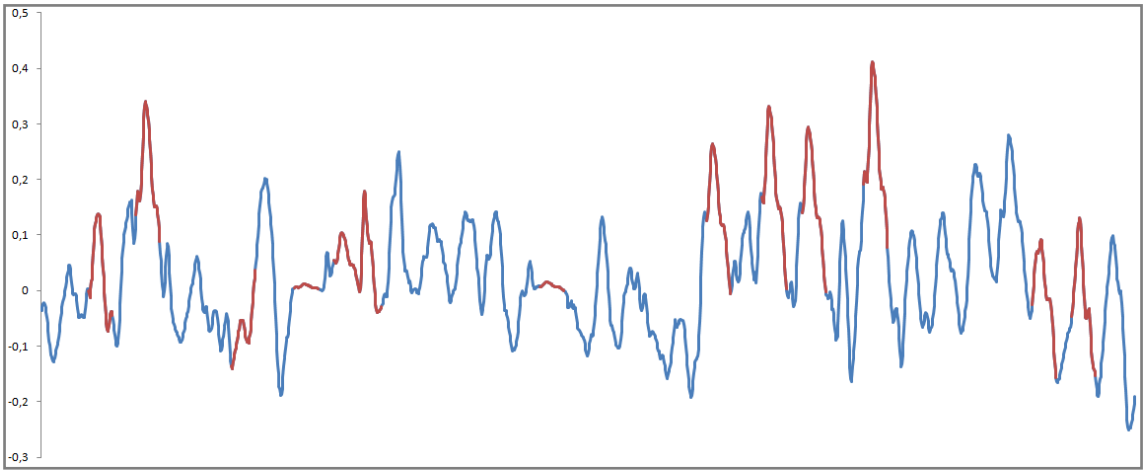


Figura 50: Señal con los patrones detectados de la segunda solución.

Tmin=12 | Símbolos=4 | Valores/Símbolo=4

Los resultados hallados aplicando estos parámetros han sido una longitud óptima de 36 y un total de 12 patrones, entre los cuales se encuentran los 6 patrones introducidos externamente. Por lo tanto, se puede establecer el grado de acierto como:

$$\frac{L_{opt}}{L_{real}} = \frac{36}{50} = 72\% \qquad \frac{n_{patrones\_opt}}{L_{patrones\_real}} = \frac{6}{8} = 75\%$$

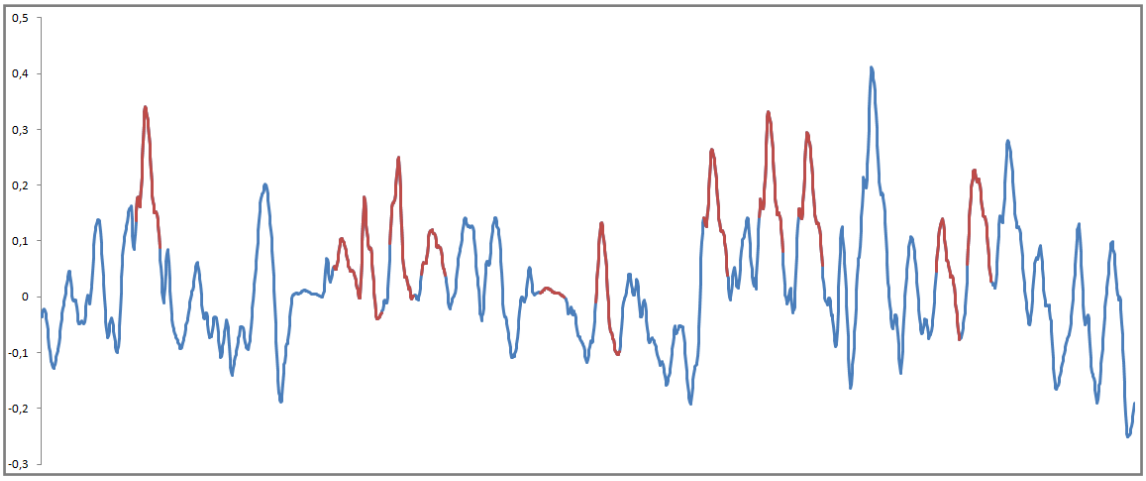


Figura 51: Señal con los patrones encontrados de la tercera solución

5.4.2. Conclusiones – EDAT

Una vez verificado el sistema de búsqueda de patrones con series de una dificultad media-baja, se precisaba de evaluar el rendimiento para secuencias reales, de manera que se pueda verificar la fiabilidad de la respuesta. Los resultados que se han obtenido de la prueba EDAT corroboran la hipótesis de las pruebas anteriores y fortalece aún más la idea de utilizar la solución propuesta en el estudio como un sistema de descubrimiento de patrones.

Con la presente prueba se observa como el método aplica un suavizado de la secuencia original, y como se ha podido apreciar en la figura 48, la pérdida de información es mínima. Posteriormente se evaluaron distintas combinaciones de las variables del código obteniéndose varios casos resultados destacables, como son los mostrados anteriormente.

Una vez verificada la capacidad de detección de patrones del método a estudio con una serie temporal con una dificultad considerable, se probará el comportamiento con una serie temporal multivariable procedente de un sistema de medidas NIRS.

5.5. Prueba ECG (Prueba real)

Una vez se haya verificado que el comportamiento del sistema de detección de patrones es válido tanto para señales artificiales creadas específicamente para que se detecten patrones como para una señal temporal multivariante procedente de la medición de la intensidad de la luz en una estrella, se procederá a la realización de una prueba con señales que han sido recogidas mediante un electrocardiograma [31]

El electrocardiograma (ECG) mide la actividad eléctrica del corazón en distintos puntos del cuerpo. En un ECG normal se utilizan 10 electrodos que se encargan de medir 12 derivaciones (medida del voltaje entre 2 electrodos).

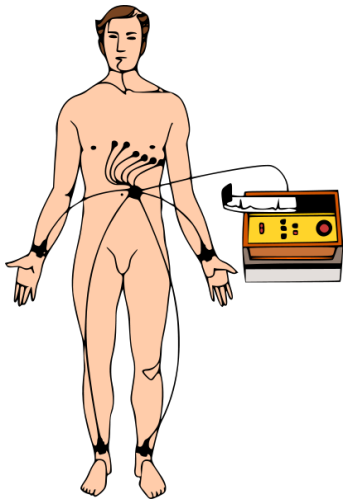


Figura 52: Ejemplo del posicionamiento de los electrodos en un paciente [35].

Por lo tanto, mediante este sistema de medida se obtienen 12 señales que indican el estado del paciente [36].

- *aVR*: medida del electrodo del brazo derecho.
- *aVL*: medida del electrodo del brazo izquierdo.
- *aVF*: medida del electrodo de la pierna izquierda.
- *I*: diferencia entre la medida del electrodo del brazo izquierdo y derecho.
- *II*: diferencia entre la medida del electrodo de la pierna izquierda y el brazo derecho.
- *III*: diferencia entre la medida del electrodo de la pierna izquierda y el brazo izquierdo.
- *V1*: medida del electrodo colocado en el cuarto espacio intercostal (entre las costillas 4 y 5) a la derecha del esternón.
- *V2*: medida del electrodo colocado en el cuarto espacio intercostal (entre las costillas 4 y 5) a la izquierda del esternón.
- *V3*: medida del electrodo colocado entre V2 y V4.
- *V4*: medida del electrodo colocado en el quinto espacio intercostal (entre las costillas 5 y 6) a en la línea medio-clavicular (línea imaginaria que baja desde el punto medio de la clavícula).
- *V5*: medida del electrodo colocado en la misma línea horizontal que V4, pero verticalmente en la línea axilar anterior (línea imaginaria que baja desde el punto medio entre el centro de la clavícula y su extremo lateral, que es el extremo más próximo al brazo)
- *V6*: medida del electrodo colocado en la misma línea horizontal que V4 y V5, pero verticalmente en la línea medio-axilar (línea imaginaria que baja desde el centro de la axila del paciente).

Para que no se haga tedioso el análisis de cada uno los parámetros que mide el ECG, solo se evaluarán las medidas aVR, I y V2.

5.5.1. Medida aVR

La señal de aVR que proporciona el dataset que se ha utilizado en esta prueba es:

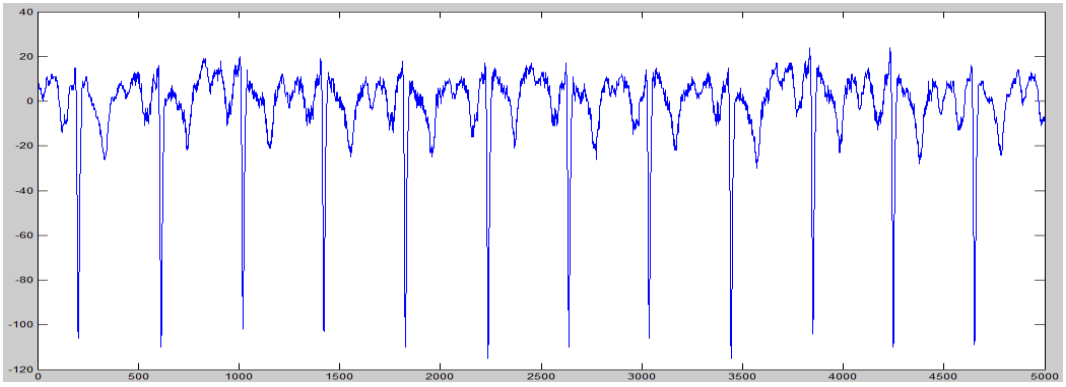


Figura 53: Señal aVR sin filtrar

Como se ha realizado en ocasiones anteriores, antes de pasar a la etapa de búsqueda de secuencias repetitivas en la señal se debe proceder a filtrarla para eliminar "ruidos" indeseados.

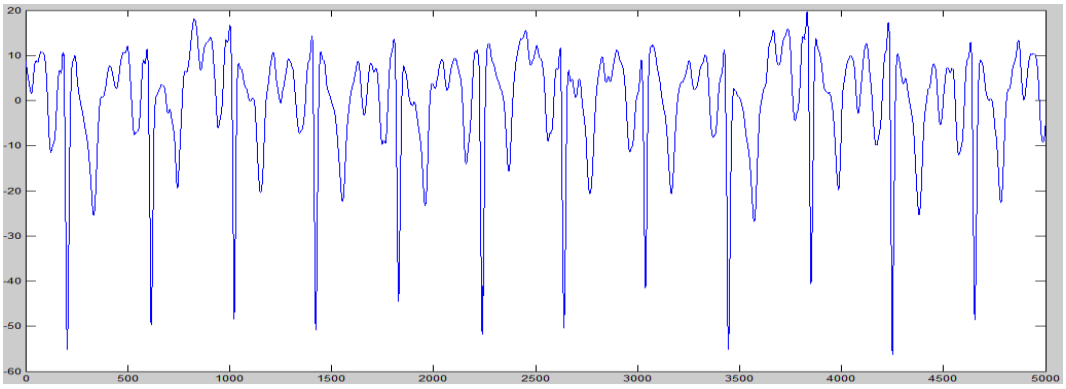


Figura 54: Señal aVR filtrada

5.5.1.1. Estudio de los resultados de aVR

Debido a que la periodicidad de los datos procedentes de la medición de la aVR, se buscará una solución que aporte datos sobre los patrones contenidos en la misma. Como en este caso se desconocen la longitud que va a tener dicho patrón y su forma (aunque de manera visual se puede intuir) la solución viene precedida de un entrenamiento previo de manera que se obtenga la combinación que ofrezca el mejor resultado. En este caso, de entre los resultados que se obtuvieron se ha elegido el siguiente:

Tmin=57 | Símbolos=3 | Valores/Símbolo=19

Los resultados han revelado un patrón que posee una longitud de 172 que se repite en la señal un total de 29 veces.

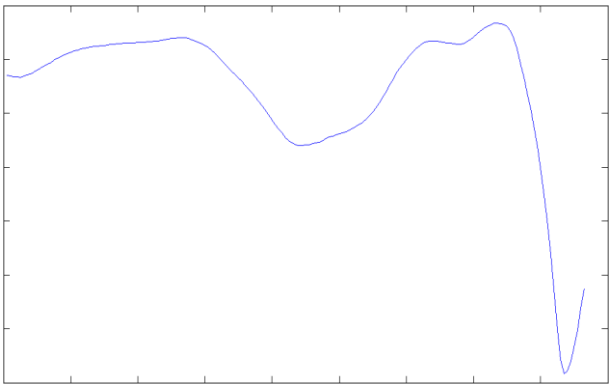


Figura 55: Patrón resultante del análisis de la señal aVR

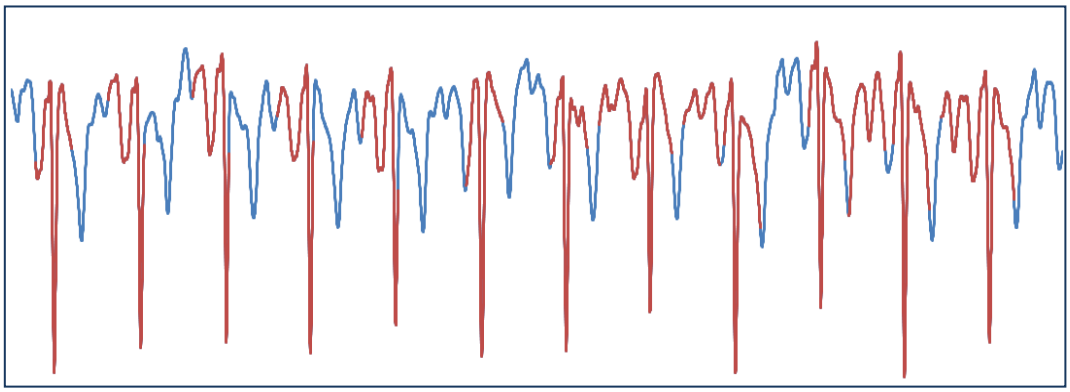


Figura 56: Señal original aVR (azul) con los patrones superpuestos (rojo)

5.5.2. Medida de I

La señal de I que proporciona el dataset que se ha utilizado en esta prueba es:

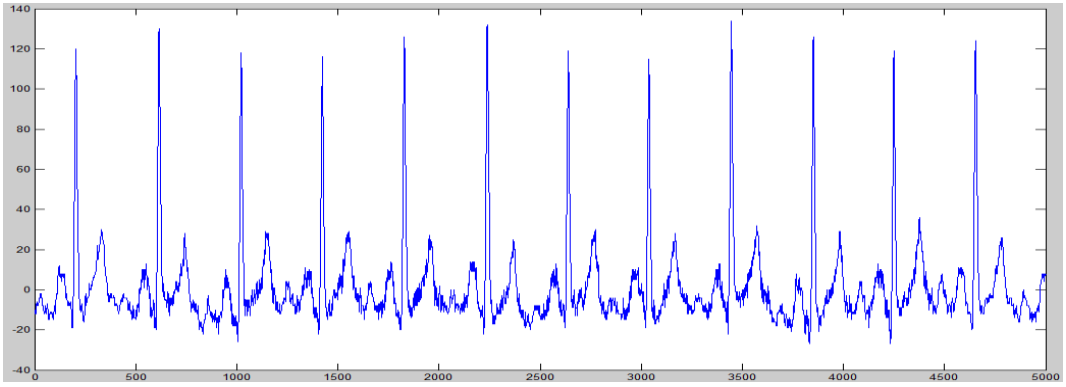


Figura 57: Señal I sin filtrar

Una vez haya sido sometida la señal a la etapa de filtrado contenida en el sistema, la señal resultante que se obtiene es:

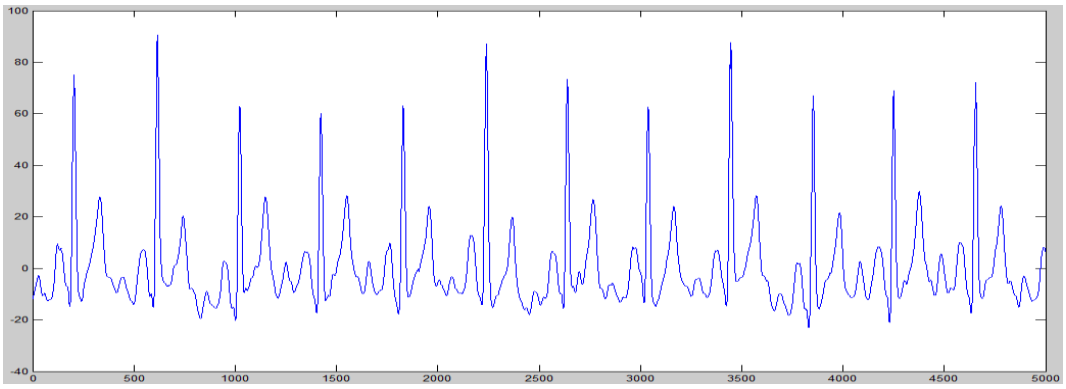


Figura 58: Señal I filtrada

5.5.2.1. Estudio de los resultados de I

Como en el caso de la señal aVR, se puede comprobar que existen determinadas secuencias que se repiten a lo largo de la señal. El hecho de que se pueda observar con facilidad dichas ocurrencias puede llevar a equívocos; la verdadera dificultad está en analizar un gran conjunto de datos de manera automática, en tiempo real y sin aportación humana.

En este caso, el mejor de los escenarios es el siguiente:

Tmin=69 | Símbolos=3 | Valores/Símbolo=23

Los resultados han revelado un patrón que posse una longitud de 208 que se repite en la señal un total de 12 veces.

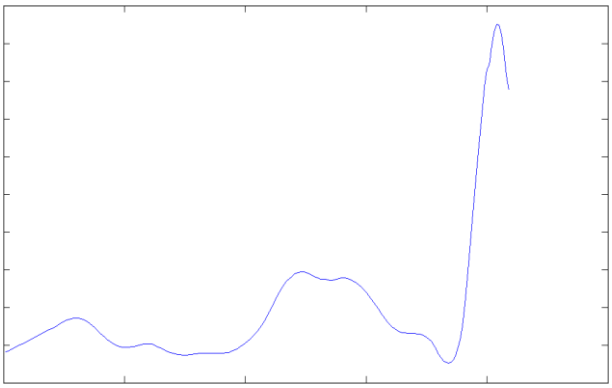


Figura 59: Patrón resultante del análisis de la señal I

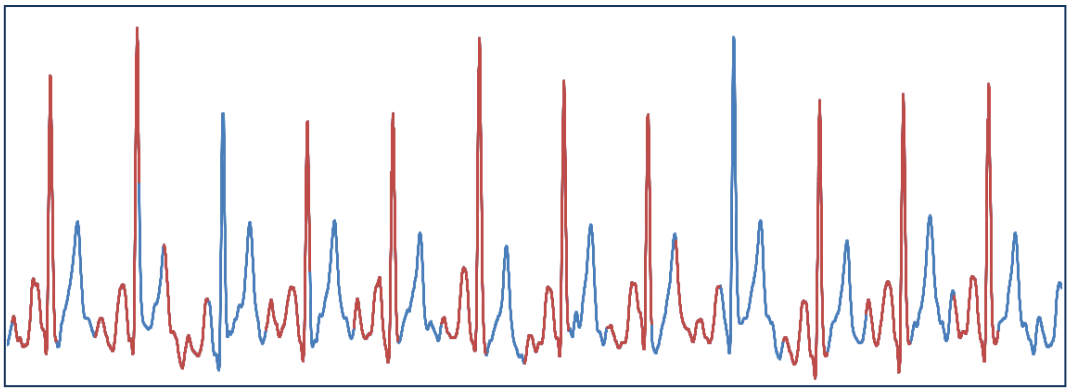


Figura 60: Señal original I (azul) con los patrones superpuestos (rojo)

5.5.3. Medida V2

La señal de V2 que proporciona el dataset que se ha utilizado en esta prueba es:

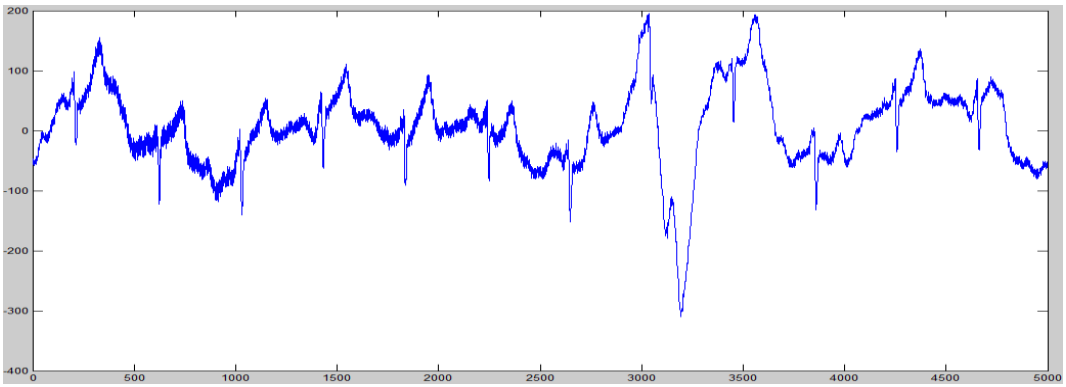


Figura 61: Señal V2 sin filtrar

Como se ha realizado en ocasiones anteriores, antes de pasar a la etapa de búsqueda de secuencias repetitivas en la señal se debe proceder a filtrarla para eliminar "ruidos" indeseados.

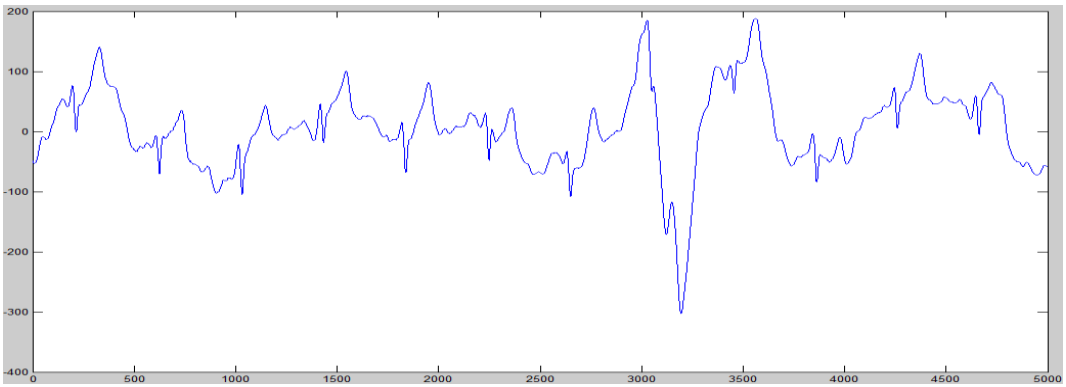


Figura 62: Señal V2 filtrada

5.5.3.1. Estudio de los resultados obtenidos

Como se puede observar en la señal de V2 filtrada, la identificación de patrones de manera visual no resulta tan obvia como en los casos anteriores. De hecho, la idea de la aplicación de técnicas de minería de datos para la búsqueda de patrones en series temporales tiene su fin en este tipo de secuencias en las que es complicado identificar el comportamiento de los valores de la señal.

En este caso, la combinación que mejor resultado que ha devuelto nuestro sistema de detección de patrones es:

Tmin=54 | Símbolos=3 | Valores/Símbolo=18

La longitud del patrón obtenido es de 163, mientras que el número de apariciones en V2 es de 20.

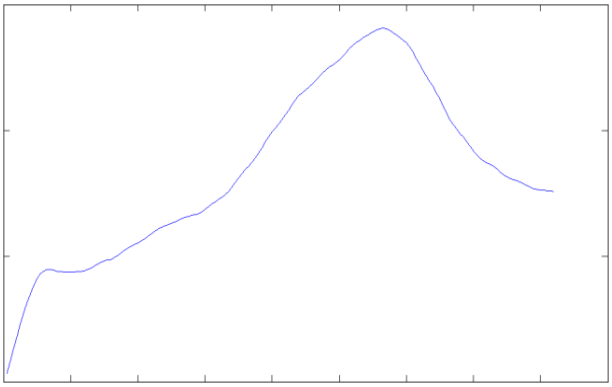


Figura 63: Patrón resultante del análisis de la señal V2

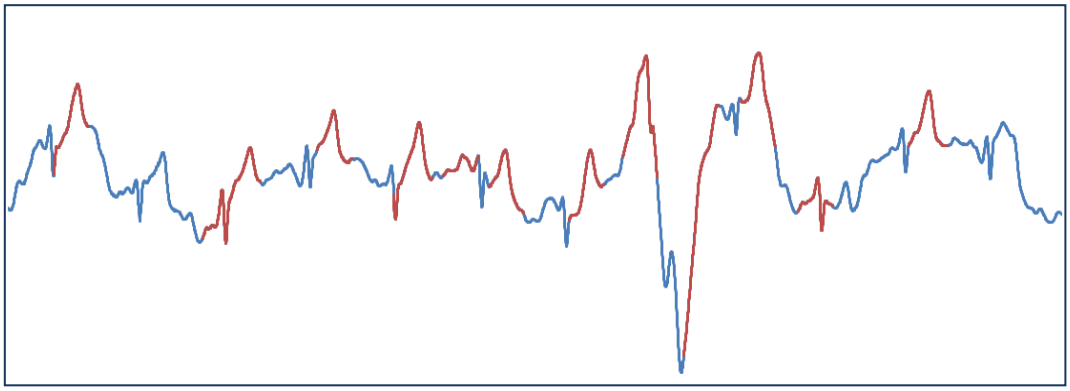


Figura 64: Señal original V2 (azul) con los patrones superpuestos (rojo)

5.5.4. Conclusiones generales de las pruebas del ECG

En base a los resultados mostrados de la aplicación del sistema que plantea el TFC a las señales medidas mediante la técnica del electrocardiograma, se puede concluir que el sistema es capaz de detectar patrones en dichas señales de manera efectiva.

Analizando las señales *aVR* y *I* por un lado, se puede observar cómo el patrón detectado se repite en multitud de ocasiones a lo largo de la señal. Analizando de forma subjetiva (visual) las señales con sus patrones superpuestos, se puede comprobar cómo, a pesar de no detectar la totalidad de los patrones en sus respectivas señales a simple vista, reconoce multitud de los mismos. En base a este examen se corrobora que el sistema funciona correctamente con señales biomédicas reales de complejidad relativa.

Por otro lado, cabe analizar por separado la *V2* debido a su mayor dificultad de comprensión. Como se ha realizado con las señales de *I* y *aVR*, el análisis visual inicial no es tan obvio como en anteriores. Por un lado, la señal de entrada al sistema tiene un nivel de ruido superior a las otras dos. Esto hace que su filtrado sea más complicado, y por lo tanto, puede que haya comportamientos en la salida del filtro que no correspondan a la naturaleza de la medida. Y por otra parte, la variabilidad del nivel de la tensión medida en *V2* es sensiblemente superior en comparación con las señales *aVR* y *I* analizadas anteriormente. Esta variabilidad dificulta la normalización de la señal. A pesar de estas dificultades, el resultado mostrado por la señal original con los patrones superpuestos es relativamente bueno, por lo que nuevamente validamos el funcionamiento del sistema propuesto.

5.6. Prueba Real (Muestra obtenida mediante NIRS)

Una vez establecidas las normas por las que se rige el código implementado y conocidos los parámetros que mejor resultado ofrecen para la extracción del conocimiento, se va a tratar de buscar patrones en series temporales capturadas mediante la espectroscopía en el infrarrojo cercano [37]. Las muestras corresponden a mediciones cerebrales de hemoglobina total (HbT), hemoglobina oxigenada (HbO2 – oxihemoglobina) y hemoglobina reducida (HbR – desoxihemoglobina).

Figura 65: Aplicación de las técnicas ópticas de monitorización en un bebé.

Esta técnica, como se ha explicado en apartados anteriores, emite luz a diferentes longitudes de onda basándose en el espectro de absorción de los parámetros que se desean medir.

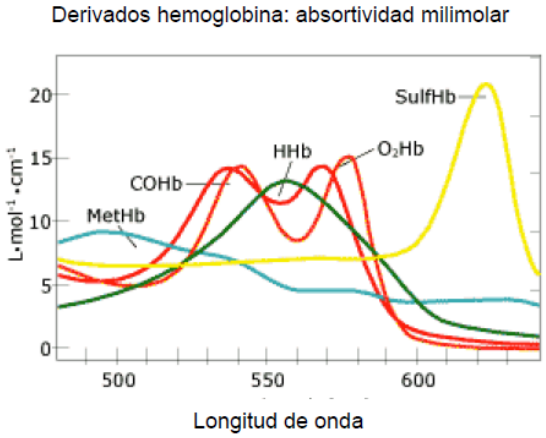


Figura 66: Espectro de absorción de cantidades equimolares de los posibles derivados de la hemoglobina [38]

Como se ha podido comprobar en las pruebas anteriores, uno de los factores más determinantes para la búsqueda de patrones en las series temporales es la ventana de trabajo. Los mejores resultados se obtienen con ventanas de longitud reducida, por los que ha configurado el procedimiento para que realice un barrido hasta una ventana de longitud 20. Además se ha configurado para que la codificación sea de 4 símbolos, y cada símbolo esté compuesto por dos valores numéricos. Para el desplazamiento de la ventana y con el objetivo de que el programa obtenga resultados satisfactorios, se ha establecido un barrido desde un símbolo de desplazamiento hasta seis.

5.6.1. Oxihemoglobina

La oxihemoglobina o hemoglobina oxigenada (HbO2) es la hemoglobina cuando está unida al oxígeno, dando el aspecto rojo intenso de la sangre arterial.

Los resultados obtenidos de la longitud del patrón y de los patrones que forman el vecindario son los siguientes:

	DESPLAZAMIENTO					
	1	**2**	**3**	**4**	**5**	**6**
Tmin=6	54	36	36	36	23	23
Tmin=8	48	40	39	34	34	32
Tmin=10	50	30	24	24	24	24
Tmin=12	48	30	28	39	26	26
Tmin=14	42	35	25	24	24	20
Tmin=16	32	24	30	20	21	21
Tmin=18	36	27	36	20	20	21
Tmin=20	40	30	34	30	20	18

Tabla 7: Datos de las longitudes óptimas de descripción obtenidas de la muestra de HbO2

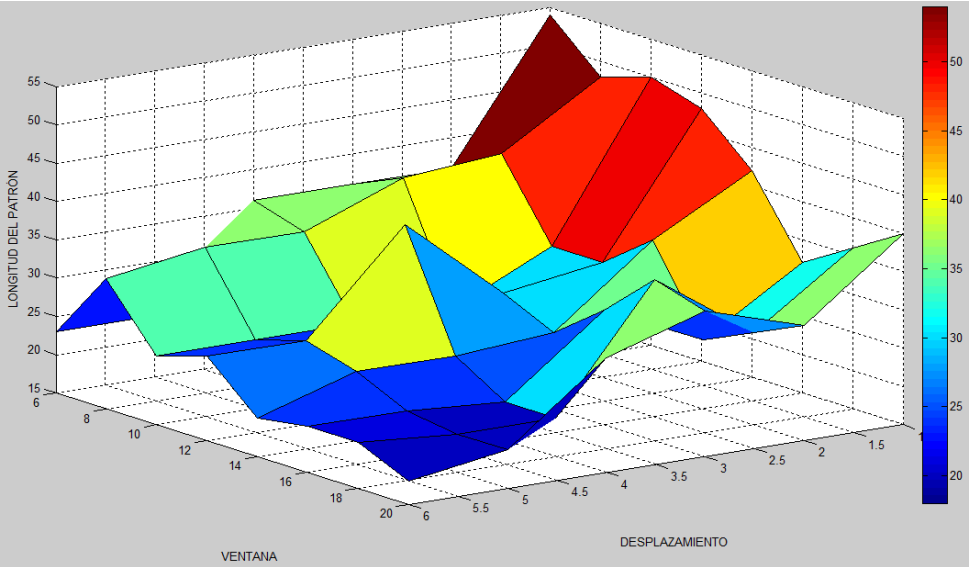


Figura 67: Gráfica de la tendencia de los valores de la longitud del patrón de la muestra HbO2

	DESPLAZAMIENTO					
	1	**2**	**3**	**4**	**5**	**6**
Tmin=6	28	31	32	50	99	99
Tmin=8	26	29	74	32	50	51
Tmin=10	17	58	46	46	82	82
Tmin=12	37	50	56	24	67	67
Tmin=14	39	35	69	74	90	103
Tmin=16	46	77	48	80	98	98
Tmin=18	47	49	56	82	99	86
Tmin=20	40	42	64	44	86	104

Tabla 8: Número de patrones que componen el vecindario de la muestra de HbO2

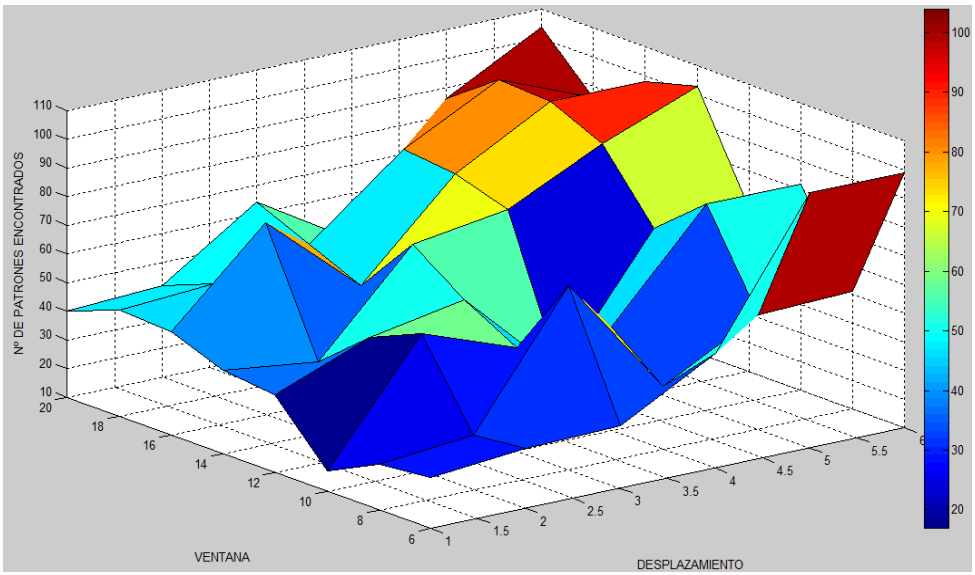


Figura 68: Gráfico con la tendencia del número de patrones que constituyen el vecindario de la muestra HbO2

Una vez mostrados los resultados generales que se han obtenido del estudio, se mostrará la solución obtenida más interesante. Este caso se corresponde con la configuración de una ventana de longitud 6 y un desplazamiento de ventana de 1 símbolo.

- **Longitud de patrón= 54**
- **Nº patrones=28**

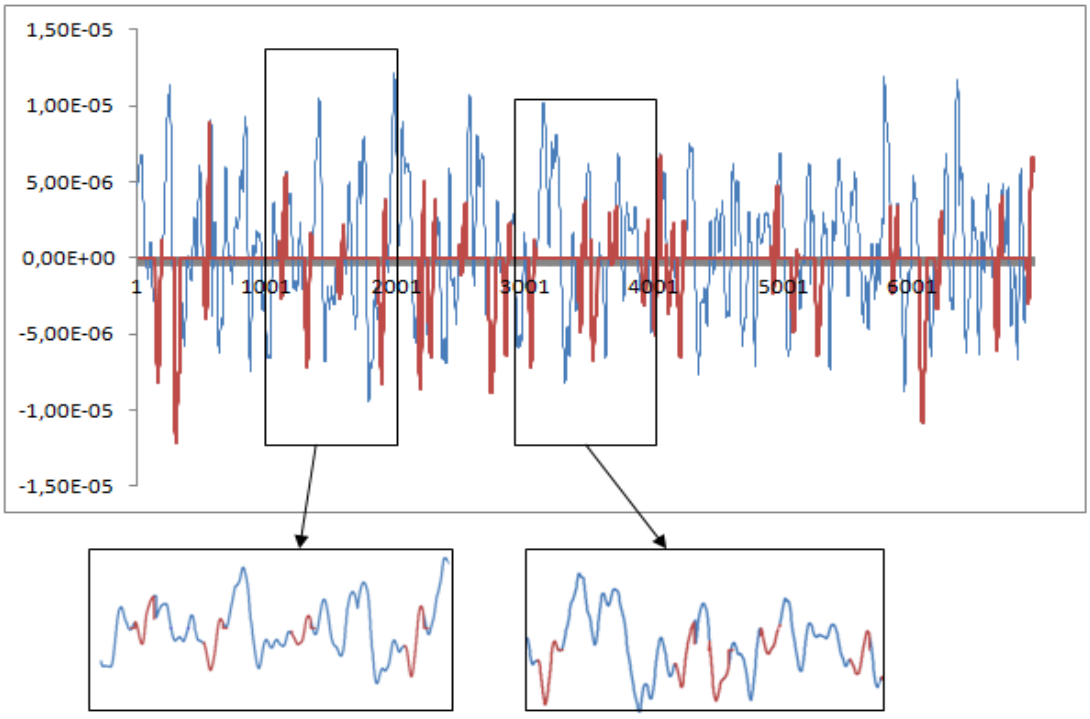


Figura 70: Datos de la oxihemoglobina (azul) con los patrones superpuestos (rojo).

5.6.2. Desoxihemoglobina

La desoxihemoglobina o hemoglobina reducida (HbR) es la molécula de hemoglobina sin oxígeno y presenta el color rojo oscuro característico de la sangre venosa.

De la misma manera que se ha realizado en la oxihemoglobina, se obtienen los siguientes resultados de las pruebas:

	DESPLAZAMIENTO					
	1	**2**	**3**	**4**	**5**	**6**
Tmin=6	48	42	34	34	24	24
Tmin=8	48	40	48	36	36	35
Tmin=10	40	30	21	21	24	24
Tmin=12	48	30	44	30	26	34
Tmin=14	42	35	30	28	24	24
Tmin=16	48	32	20	20	18	18
Tmin=18	36	27	30	30	24	24
Tmin=20	40	40	28	25	28	18

Tabla 9: Longitudes del patrón obtenidas de la muestra de HbR

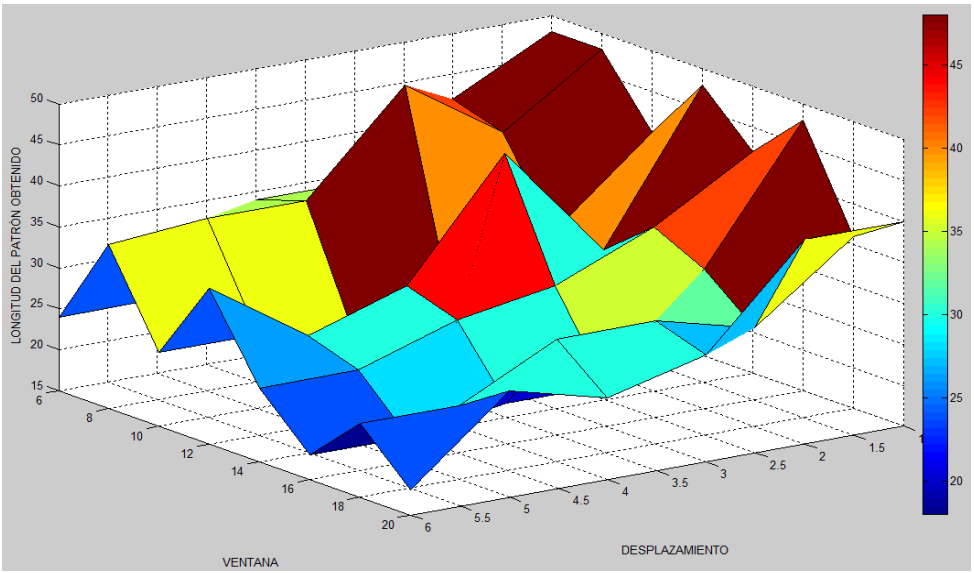


Figura 71: Gráfica que muestra la tendencia de las longitudes del patrón obtenidas de la muestra HbR

	DESPLAZAMIENTO					
	1	**2**	**3**	**4**	**5**	**6**
Tmin=6	59	35	53	53	89	89
Tmin=8	57	33	21	42	42	48
Tmin=10	28	52	96	96	72	72
Tmin=12	17	74	65	56	70	45
Tmin=14	42	42	62	30	80	79
Tmin=16	30	44	104	89	122	122

Tmin=18	51	55	50	51	81	82
Tmin=20	39	46	56	72	59	109

Tabla 10: Número de patrones que componen el vecindario de la muestra de HbR

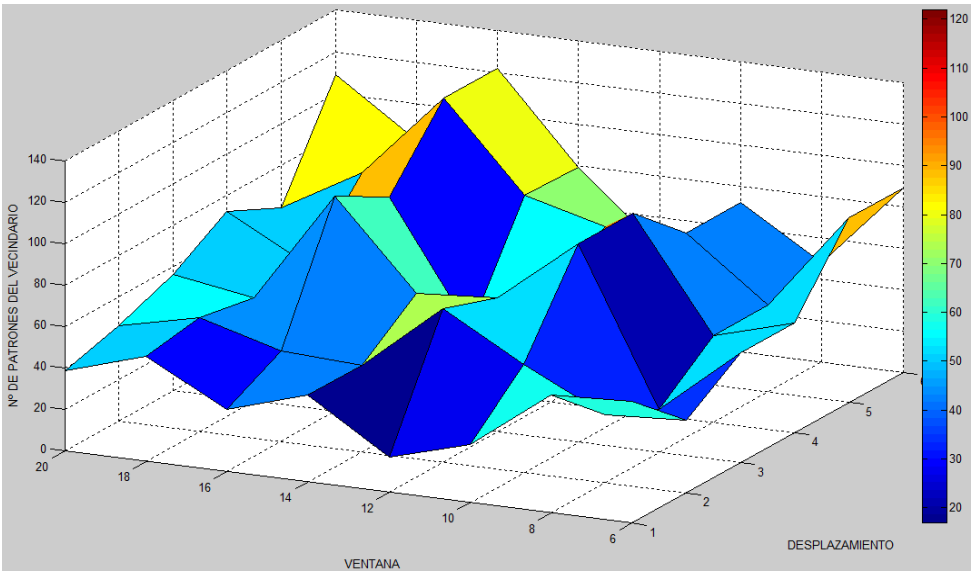


Figura 69: Gráfico con la tendencia del número de patrones del vecindario de la muestra HbR.

Analizados los datos obtenidos y sus respectivas tendencias, se evaluará la condición que más acertada. En este caso, los parámetros de configuración de dicha solución son una ventana de trabajo de longitud 8 y un desplazamiento de la ventana de 1 símbolo.

- **Longitud de patrón=48**
- **Nº patrones=57**

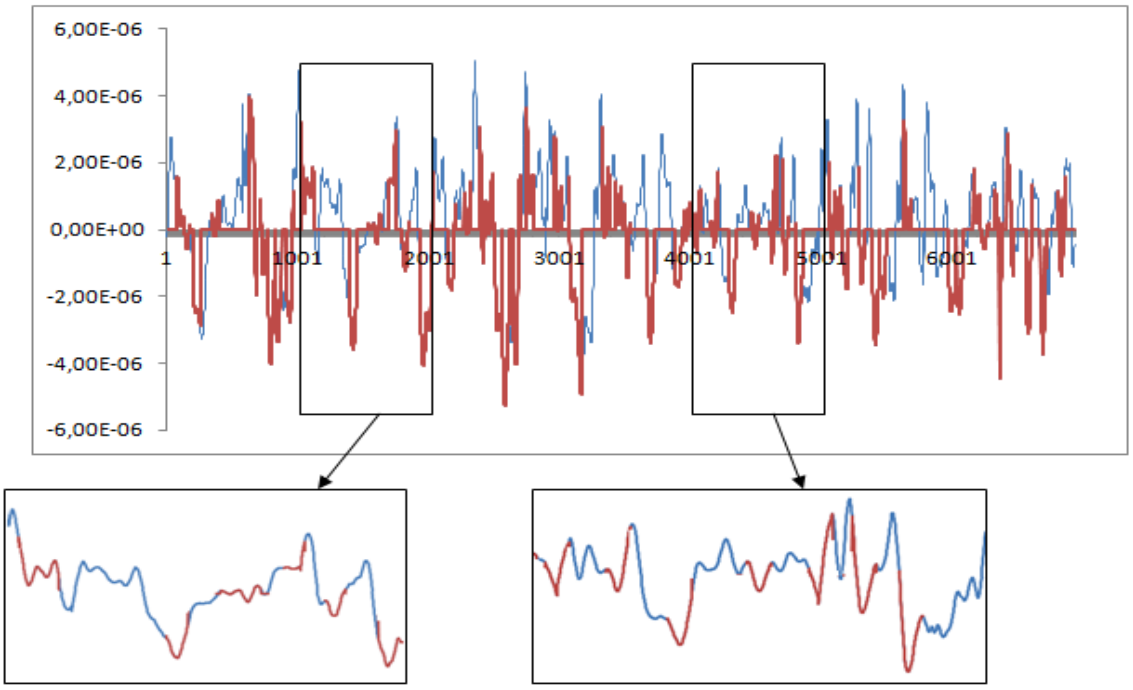


Figura 70: Muestra de HbR (azul) con los patrones superpuestos (rojo)

5.6.3. Hemoglobina Total

La hemoglobina total (HbT) es la suma de las concentraciones de todas las formas de hemoglobina. Los datos recogidos de la aplicación del programa a la muestra de HbT son:

	DESPLAZAMIENTO					
	1	**2**	**3**	**4**	**5**	**6**
Tmin=6	48	33	30	30	26	26
Tmin=8	48	48	39	46	46	46
Tmin=10	50	25	24	24	18	18
Tmin=12	48	36	28	27	30	26
Tmin=14	42	35	25	28	15	20
Tmin=16	48	24	25	24	15	15
Tmin=18	36	18	24	25	20	21
Tmin=20	-	20	21	20	24	18

Tabla 11: Longitud del patrón obtenida de la muestra de HbT.

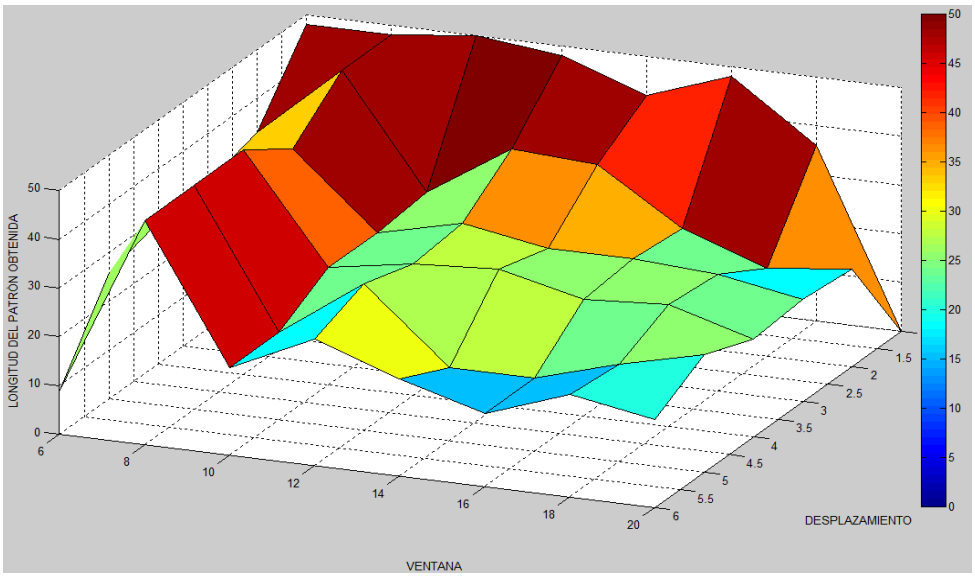


Figura 71: Tendencia de la longitud del patrón de la muestra HbT.

	DESPLAZAMIENTO					
	1	**2**	**3**	**4**	**5**	**6**
Tmin=6	50	37	48	48	83	83
Tmin=8	23	22	80	69	69	68
Tmin=10	21	80	79	79	111	111
Tmin=12	33	75	54	57	48	90
Tmin=14	36	38	72	55	117	98
Tmin=16	21	71	70	69	117	117
Tmin=18	38	101	72	58	96	81
Tmin=20	-	85	74	90	82	91

Tabla 12: Número de patrones que componen el vecindario de la muestra de HbT

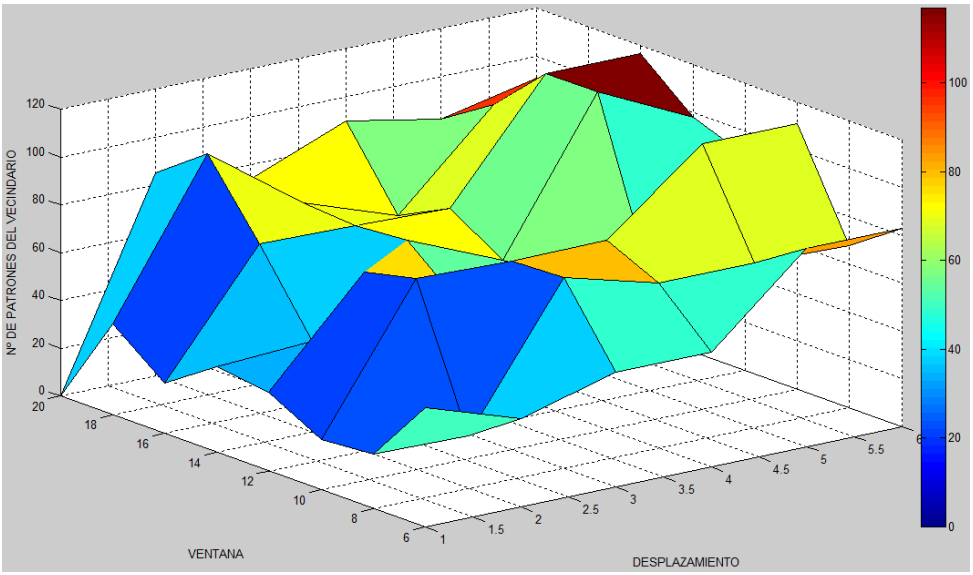


Figura 72: Tendencia del número de patrones del vecindario en la muestra HbT.

De entre todos los resultados en el estudio de la muestra de HbT, el más interesante ha sido la combinación de ventana de longitud 10 y desplazamiento de 1 símbolo:

- **Longitud de patrón=50**
- **Nº patrones->21).**

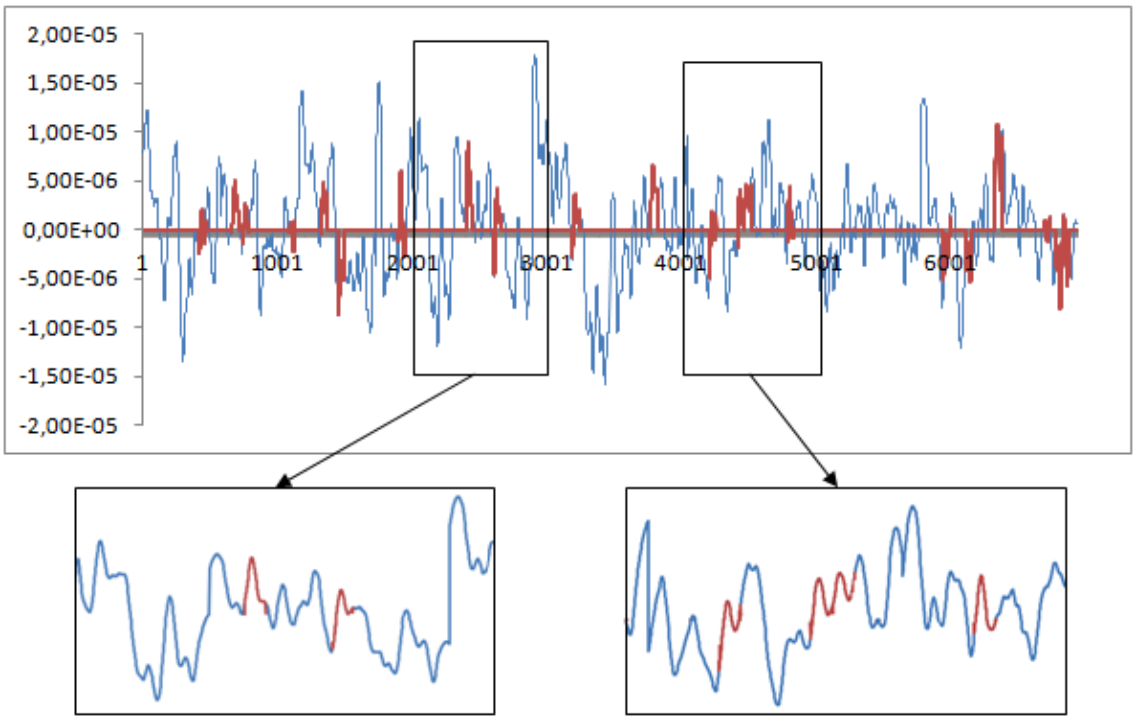


Figura 73: Secuencia HbT (azul) con los patrones superpuestos (rojo).

5.6.4. Conclusiones de los resultados de las muestras de la NIRS

A partir de los datos conseguidos en este apartado se pueden definir algunas ideas. Aunque se han probado un número considerable de combinaciones, muchas de ellas se deben desechar ya que sólo forman parte de una tendencia (por ejemplo, la pendiente de una curva).

Como se pueden comprobar en los resultados obtenidos del estudio de la oxihemoglobina, desoxihemoglobina y hemoglobina total, no pueden definirse como concluyentes porque el patrón que se obtiene es de una longitud limitada. Aún así, se puede verificar cómo la solución resultante determina una secuencia repetida y posteriormente la identifica en la muestra a estudiar.

6. CONCLUSIONES Y LÍNEAS FUTURAS

En este último capítulo se exponen las conclusiones del trabajo llevado a cabo en este trabajo fin de carrera. La mayor parte de estas conclusiones se extraen a partir de los resultados presentados en el capítulo anterior.

En el segundo apartado se exponen las posibles líneas futuras de trabajo con respecto a los técnicas que se han tratado en este trabajo, añadiendo nuevas etapas a las implementaciones, variando el funcionamiento de las mismas o introduciendo técnicas más sofisticadas y su posible desarrollo.

6.5. Conclusiones

Este proyecto se ha centrado en el estudio de los datos obtenidos a través de medidas en el cerebro de recién nacidos con técnicas ópticas avanzadas, más concretamente con espectroscopía en el infrarrojo cercano. El estudio trata el pre-procesamiento de los datos del NIRS y la aplicación de técnicas de minería de datos para la adquisición de un conocimiento acerca del comportamiento repetitivo de determinados factores hemodinámicos cerebrales.

Fundamentándose en los estudios previos realizados para la consecución del trabajo y en éste mismo, la primera conclusión que se obtiene es la complejidad que conlleva la detección de patrones en series temporales, desde la recogida de datos hasta la obtención del conocimiento. Cada sistema de medida tiene un comportamiento singular y, añadido a la naturaleza de la señal, implica un examen exhaustivo de los datos para poder aplicarles un pre-procesamiento adecuado. Ya se explicó anteriormente cómo puede afectar una mala elección en los parámetros de preprocesamiento. Una vez salvada dicha dificultad, se plantea el problema de la detección de patrones en la señal ya preparada para el estudio. Para hacerse una idea de dicho problema sólo hay que comprobar la multitud de técnicas para el descubrimiento de patrones que se tienen en la actualidad. La aplicación de la minería de datos en las series temporales y del estudio de ocurrencias en las mismas requiere de un conocimiento previo pormenorizado de la naturaleza de la señal. Por lo tanto, no todas las técnicas existentes se pueden aplicar a una determinada secuencia temporal. La mayor dificultad en esta aplicación es la efectividad, lo que se puede resumir como un resultado óptimo en el menor tiempo posible. De ahí la importancia de las secuencias de entrenamiento antes de la validación de cualquier sistema.

Para el pre-procesamiento de las muestras del NIRS se han mostrado técnicas de suavizado, normalizado, discretización y transformación de la serie temporal a estudio, con el objetivo de obtener el máximo conocimiento de la muestra con el mínimo conjunto de datos, debido a que la sobrecarga de computación es un hecho a tener en cuenta cuando el conjunto de los datos es muy grande. Estas técnicas anteriormente descritas se basan en el estudio [23].

Por otro lado, para la minería de datos se han empleado las técnicas EMMA y MDL, descritas también en el capítulo 4 del proyecto. Estos métodos proporcionan soporte para el descubrimiento de patrones en las señales cerebrales, además de agrupar aquellas secuencias que se comporten de una manera similar al patrón conocido.

Respecto a los resultados obtenidos, cabe destacar algunas características de interés:

1. Para el procesamiento de los datos, es necesario tener información acerca de la naturaleza de los mismos. Por lo tanto, se necesita de un estudio previo para la aplicación de los parámetros adecuados en el pre procesamiento.
2. Es importante elegir los parámetros adecuadamente en el pre-procesamiento. Una mala elección puede proporcionar datos adicionales que aumentaría el tiempo de cómputo de una manera crítica o, por otra parte, puede producir que se pierda información útil de la muestra a estudio.
3. Cuando se transforman y se analizan los datos, se ha comprobado que la elección del tamaño de la ventana y su desplazamiento son los factores que más afecta a la hora del descubrimiento de la longitud de descripción mínima. Cuanto más pequeña sea la ventana y menor sea el desplazamiento, los resultados que se obtienen son mejores. Estas afirmaciones se apoyan en los estudios previos realizados con una serie de dificultad baja y con las series Victoria 1 y Victoria 2.
4. El número de símbolos que se aplican en la codificación influye en menor medida que los dos parámetros expuestos en el punto 3. La longitud del patrón disminuye ligeramente con una codificación con más símbolos debido a que aumenta la precisión de la transformación mediante el PAA. Esto provoca que la serie se vuelve más compleja y es más difícil encontrar segmentos que se comporten de la misma manera.
5. Cuando se someten las medidas reales del NIRS a este procedimiento se comprueba que el rendimiento del código no es lo suficientemente alto. La razón puede ser que la aplicación del método MDL, debido a su complejidad, no tiene un funcionamiento óptimo. Hay que tener en cuenta lo complejo de buscar en series temporales multivariantes segmentos de las mismas que se comporten de la misma manera. Además, habría que tener secuencias más largas y estables, contrastadas respecto a referencias, para tomar decisiones más claras. El conjunto de datos de los que se disponía en el caso de bebés, no tenían una referencia clara que permitiera validar los resultados del análisis ya que, en todos los casos, los bebés tenían la capacidad de autorregulación cerebral en estado comprometido.
6. Respecto al método de agrupación de patrones proporcionado por el 'vecindario' se cree que su funcionamiento es válido y los resultados que se muestran en el punto 4 así lo confirman.
7. Un razonamiento a tener en cuenta es que quizá las señales hemodinámicas no contengan información suficiente como para extraer un conocimiento de ellas debido a que su longitud es limitada. Es importante tener en cuenta para estudios posteriores que las secuencias que se deseen analizar tengan la información suficiente para la extracción de pautas que puedan resultar interesantes para su estudio.

6.6. Líneas Futuras

En los últimos años las técnicas de monitorización no invasivas han tomado gran importancia en el ámbito biomédico. Se han desarrollado aparatos de medida que recogen la información interesante para conocer el estado del paciente. Por estas mismas razones es necesario mejorar en la misma medida las técnicas aplicadas a la monitorización, ya que existen muchas posibilidades de que en un futuro próximo el continuo desarrollo de los sistemas de monitorización no invasivos aporten una información fiable y clínicamente válida que permita sustituir a otros sistemas de monitorización más invasivos que en el momento actual se aplican a los pacientes neurocríticos.

6. Conclusiones y líneas futuras

Se ha resumido algunos métodos aplicables a señales procedentes de la espectroscopía en el infrarrojo cercano. A pesar de trabajo mostrado en este estudio, se necesitan más aportaciones en esta línea para que el trabajo realizado pueda ser aplicable.

Sería importante realizar investigaciones más pormenorizadas en el pre-procesamiento de los datos. Sería interesante la implementación de un método que amolde los datos de la espectroscopía a la solución sin necesidad de un conocimiento previo de los mismos. De esta manera el funcionamiento de las técnicas expuestas sería totalmente mecanizado y no requeriría de la aportación humana.

En el apartado anterior se ha aportado un resumen de las ventajas e inconvenientes de la solución propuesta al problema planteado. Por lo tanto, se estima que se debe seguir trabajando en el aspecto de encontrar la longitud de descripción del patrón, ya que este es indispensable para la extracción de las reglas de comportamiento de la serie.

Además, en esta ocasión se ha aportado una solución para el tratamiento de señales. Se necesita de una estrecha colaboración entre los campos de la Medicina y el de la Ingeniería para que la información oculta en las señales de provenientes de la monitorización sea desvelada y así poder anticiparse a males hoy en día desconocidos.

7. REFERENCIAS

1. **Mencía Bartolomé, S; López-Herce Cid, J; Lamas Ferreiro, A; Borrego Domínguez, R; Sancho Pérez, L; Carrillo Álvarez, A.** "*Aplicación del índice biespectral en la monitorización del niño enfermo crítico*". An Pediatr (Barc). 2006; 64:96-9. - vol.64 núm. 01.
2. **Hack M, Friedman H, Avroy A, Fanaroff MB**. *"Outcomes of extremely low birth weight infants"*. Pediatrics 1996; 98: 931-937.
3. **Berré, J., De Witte O, Moriane, J.J.,** *"Cerebral blood flow velocity using Doppler techniques"*, in Vincent, J.L. (ed): Update in intensive care and emergency medicine. Volume 14. Update 1991. Berlin, Springer-Verlag; 1992: 522-529.
4. **Caplan, L.R., Brass, L.M., DeWitt, L.D., et al**.: *"Transcranial Doppler ultrasound: present status"*. Neurology 1990; 40: 696-700.
5. **Augusta Montenegro, María**. *"Nueropediatría Ilustrada"*. Capítulo 16: PET cerebral en la infancia.
6. **Chamorro A., Sacco RL, Mohr JP.** *"Clinical computed tomography correlations of lacunar infarction in the Stroke Data Bank"*. Stroke 1991;22:175-181
7. **Wyatt JS, Cope M, Delpy DT, Richardson CE, Edwards AD, Wray S, Reynolds EOR**." *Quantitation of cerebral blood volume in human infants by near-infrared spectroscopy"*. J Apply Physiol 1990; 68:1086-1091.
8. **Skoog,D.A, Hiller F.J, Nieman T.A, 2001.** "*Principios de Análisis Instrumental"*, Madrid: Mc Graw Hill, 5ª Edición, 490 pgs.
9. http://portalbiomedico.com/equipamiento-biomedico/oximetro/oximetria-de-pulso-conceptos.html
10. **Eric M. Buckley**. *"Cerebral hemodinamics in high-risk neonates probed by Diffuse Optical Spectroscophy"*.
11. **C. Zhou, S.A. Eucker, T. Durduran, G. Yu, J. Ralston, S.H. Friess, R.N. Ichord, S. S. Margulies, and A. G. Yodh.** *"Diffuse Optical Monitoring of Hemodynamic*
12. **Noah Cook, Erin M Buckley, Turgut Durduran, Meeri , Chou Zhou, Susan Schultz, Chandra Sehgal, Peter Arger, Mary Putt, Daniel Licht, Hallam Hurt, and Arjun Yodh.** "*Diffuse Correlation Spectroscopy Reveals that Cerebrovascular Autoregulation is Intact in Preterm Infants Undergoing a Postural Challenge*". In The 4th International Conference on Brain Monitoring and Neuroprotection in the Newborn, February 2009 .
13. **Ms.C. Guillermo Molero Castillo,Ms.C. María Elena Meda Campaña.** *"Integración de Minería de Datos y Sistemas Multiagente:un campo de investigación y desarrollo"*. Ciencias de la Información Vol. 41, No.3, septiembre - diciembre, pp. 53 - 56, 2010
14. **Tan, Pang-Ning, Steinbach, Michae, and Kumar, Vipin ."***Introduction to Data Mining"*. Pearson Addison-Wesley.
15. **Chatfield, C.** "*The Analysis of Time Series: An Introduction***".** Ed. Chapman and Hall. London, 2003.
16. **Jiawei Han& Micheline Kamber**. *"Data Mining: Concepts and Techniques"* Morgan Kaufmann, 2006.
17. *"An Introduction to Wavelets"*, IEEE Computational Science and Engineering, 1995. Graps A.
18. **John Wiley & Sons**. *"Applied multivariate techniques"*. Inc., Canada.
19. **Yi B.-K. y Faloutsos, C.,** "*Fast time sequence indexing for arbitrary Lp norm"*, The VLDB Journal, pp. 23-27, 1998.
20. **Harold A. Romo, Esp., Judy C. Realpe, Ing., Pablo E. Jojoa, PhD.***"Análisis de Señales EMG Superficiales y su Aplicación de Control de Prótesis de Mano"*, Universidad del Cauca.

7. Referencias

21. **Wasserman, P.D. (1989)**. *"Neural computing theory and practice"*. Van Nostrand Reinhold. ISBN 0-442-20743-3.
22. **Anderson, Silverstein, Ritz, and Jones**. Psychological Review, 84, 413-451.
23. **L.A. Zadeh**, *"Fuzzy sets"*, Information and control, Vol 8, 338-353, 1965.
24. **V. Kecman**, *"Self organizing maps"*, Spinger Verlag, New York, 1995.
25. **J. Jatsen.** *"Neurofuzzy modeling"*. Tech. Report No 98-H-874. Department of Automation, Technical University of Denmark – Lyngby (1998)
26. *"HTM cortical learning algorithm"*, Numenta. Version 0.2.1. , September 12, 2011.
27. **Perea A.J., Meroño J.E., Aguilera M.J.** *"Hierarchical temporal memory for mapping vineyards using digital aerial photogra"*. African Journal of Agricultura Research Vol. 7(3), pp. 456-466, January, 2012.
28. **Davide Maltoni**. *"Pattern Recognition by Hierarchical Temporal Memory"*. Biometric System Laboratory – Universidad de Bolonia.
29. **Pértegas Díaz, S., Pita Fernández, S.,** *"La distribución normal"*, Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Universitario de A Coruña (España),CAD ATEN PRIMARIA 2001; 8: 268-274
30. **Yoshiki Tanaka and Kuniaki Uehara** ,*"Discover Motifs in Multi Dimensional Time-Series Using the Principal Component Analysis and the MDL Principle"*, Department of Computer and Systems Engineering, Kobe University, 1-1 Rokko-dai, Nada, Kobe 657-8501, Japan.
31. **Theophano Mitsa.** *"Temporal Data Mining"*. Chapman & Hall, 2010.
32. **Lin, J., Keogh, E., Lonardi, S., Patel, P.:** *"Finding motifs in time series"*. In: Proceedings of the 2nd Workshop on Temporal Data Mining, Canada, 2002.
33. http://sci2s.ugr.es/keel/datasets.php, KEEL-dataset.
34. http://www.physionet.org/cgi-bin/atm/ATM, PhysioBank database.
35. http://ecg.utah.edu/, Alan E. Lindsay ECG Learning Center, University of Utah-School of Medicine.
36. **Harrison**, "*Principios de Medicina Interna"*, 17a edición, Tomo-2.
37. http://bisp.kaist.ac.kr/, Dept.of Bio and Brain Engineering, KAIST.
38. **Moran FR**. *"Application of haemoglobin derivates in STAT analysis"*. 1999. www.bloodgas.org.