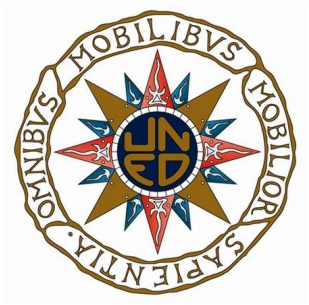


ETSI Informatica - UNED Departamento de Inteligencia Artificial

Máster Universitario En I.A. Avanzada: Fundamentos, Métodos Y Aplicaciones

---

VARIABILIDAD EN GRS1915+105. Aplicación de Minería de Datos.

José Luis Camps Palou de Comasema

Director:
Luis Manuel Sarro Barro

Madrid 15 de septiembre del 2013

A mi familia y amigos

Agradecimientos

Quiero agradecer el apoyo, confianza y tiempo de Luis, director del trabajo, y al que solo pude corresponder llevando su paciencia a límites cercanos al sistema del que versa el trabajo.

A Marion, Celia y Peggy, mencionadas en el trabajo de manera abstracta como "expertos de la ESAC", sin cuyo entusiasmo por la materia esto no tendría sentido.

Y por último a todas las personas que poco a poco van colaborando en aumentar el conocimiento científico, hoy tan poco valorado, pues el trabajo se nutre de trabajos anteriores, sin los cuales no habría sido posible su realización, y donde muchas personas han dedicado grandes esfuerzos.

Resumen

El Universo está repleto de objetos lejanos que comenzamos a descubrir y a entender. GRS1915+105 es uno de estos objetos, un sistema binario compuesto por una estrella y un agujero negro. Parece tener un mecanismo de autorregulación de la tasa de crecimiento, que lo hace de especial interés.

En la variabilidad del sistema podemos encontrar la clave para su compresión, y la búsqueda de patrones en esta variabilidad el camino que nos lleva a encontrarla.

La minería de datos engloba algoritmos, técnicas y métodos que nos permiten la búsqueda de estos patrones, y por tanto es una herramienta fundamental para comprender el sistema GRS1915+105.

Siendo el objetivo del trabajo la búsqueda y selección algoritmos, técnicas y métodos dentro de la minería de datos, y que se ajuste a los requisitos impuestos por las características del sistema GRS1915+105. Para ello se ha seguido una metodología clásica de minería de datos.

Índice general

Índice de figuras

Índice de cuadros

Parte I

Introducción

CAPÍTULO 1

Introducción

1.1. Contexto del problema tratado

El 15 de agosto de 1992 la sonda soviética Granat descubrió un sistema binario compuesto de una estrella y un agujero negro [Finley, 1994]. Recibió el nombre de GRS1915+105, GRanat Source, ascensión 19 horas 15 minutos y declinación 10,5. Situado en la constelación de Aquila a unos 40000 años luz de distancia, el agujero negro que compone el sistema es el más pesado conocido de la Vía Láctea, con una masa de entre 10 a 18 veces la del Sol. El agujero negro es además un microquasar con una rotación que podría estar en 1150 veces por segundo. Fue conocido por ser la primera fuente en nuestra galaxia en expulsar material superlumínica, velocidades superiores a la de la luz, aunque ya ha quedado demostrado que se debe a un efecto conocido como aberración relativista, y que realmente la velocidad es del 90 % de la velocidad luz [ESO, 2001].

El sistema cuenta con una gran variabilidad, o variaciones en la intensidad de señal. Se cree que el sistema puede contar con un mecanismo de autorregulación de la tasa de crecimiento. El chorro de materiales expulsados es ahogado por un viento caliente que sopla proveniente del disco de acreción, privando al chorro de los materiales necesarios para sostenerlo. Cuando el viento amaina, el chorro de materiales regresa [Greiner, 2001].

Estas características han hecho que sea objeto de observación durante una década, recogiendo el material expulsado en forma de rayos-X. Los datos del observatorio Chandra de rayos-X han sido los proporcionados por la ESAC para la realización del trabajo. También se han realizado estudios intentando identificar patrones en su comportamiento [Belloni et al., 2000]. Sin embargo, este comportamiento aún no termina de estar claro, y se cree que existen más patrones que escapan a un análisis normal. Es por ello que se ha acudido a la minería de

datos.

1.2. Objetivos del trabajo

Es en este punto cuando los expertos de la ESAC acudieron a la minería de datos, buscaban esos patrones en la variabilidad que se escapa a los análisis realizados hasta el momento.

Objetivo General. Siendo el objetivo de los astrónomos encontrar nuevos patrones, el **objetivo del trabajo es proveer de los mecanismos necesarios para realizar dicha búsqueda de patrones**, buscando y seleccionando algoritmos, técnicas y métodos dentro del ámbito de la minería de datos que se ajusten más al problema planteado.

Objetivos Específicos. Los objetivos generales pueden desglosarse en objetivos más específicos.

- **Transformar los datos** para ser utilizados con técnicas de minería de datos.
- **Seleccionar los métodos** que se adapten a los requisitos del problema.
- **Analizar los resultados** de los modelos generados.
- **Evaluación los resultados**.
- **Obtener futuras lineas de trabajo**.

Hipótesis. La hipótesis en la que se fundamenta el trabajo es que la variabilidad de GRS1915+105 no es aleatoria y se pueden encontrar patrones de comportamiento.

Existen estudios previos que avalan dicha hipótesis [Belloni et al., 2000], pero además se cree que estos pueden ser más complejos que los hallados hasta el momento, y que con la ayuda de la minería de datos se pueden llegar a descubrir.

Limitaciones. La principal limitación a la hora de analizar los resultados es que el resultado final, los grupos con patrones similares, tiene un significado solo comprensible por expertos.

1.3. Organización y metodología del trabajo

Se trata pues de un trabajo de investigación practico descriptivo. Donde los puntos del trabajo vienen marcados por una metodología de minería de datos como es CRISP-DM [Chapman et al., 2000].

- **Comprensión del problema**
  Tratado en actual apartado, *introducción*. Donde se ha evaluado la situación, se describe que objetivos busca el trabajo y se genera un plan de proyecto.

- **Comprensión de datos**
  Capitulo 3 Donde los datos proporcionados serán descritos, explorados y verificada su calidad.

- **Preparación de datos**
  Capitulo 3. Los datos proporcionados no son utilizables con las técnicas y métodos de minería de datos, es necesario realizar un tratamiento previo de los datos. Es por tanto necesario seleccionar los datos útiles, limpiar aquellos que no cumplan un criterio de calidad, estructurar los datos, integrar los datos y formatear los datos.

- **Modelo de Agrupamiento**
  Capitulo 4 Con los datos obtenidos en la fase anterior buscaremos patrones. En esto consiste el agrupamiento, donde se crearán grupos con muestras que sean similares. Para ello seleccionaremos técnicas de agrupamiento, diseñaremos la evaluación, ejecutaremos el agrupamiento y evaluaremos los resultados.

- **Modelo de Clasificación**
  Capitulo 5 Generaremos un modelo capaz de predecir los patrones obtenidos a partir de una enterada dada. Para ello seleccionaremos la técnica de modelado, diseño de la evaluación, construiremos un modelo y evaluaremos los resultados.

- **Evaluación**
  Capitulo 6 Donde, tanto para el modelo de agrupamiento como para el modelo de clasificación, se realizara una evaluación de los resultados, revisaremos el proceso y se establecerán los siguientes pasos a seguir.

Frente a la metodología CRISP-DM, la fase de modelado genera dos modelos diferentes como son de agrupamiento y clasificación, esta decisión se explicar más adelante. Por otra parte queda fuera del trabajo el despliegue.

Parte II

Revisión de la cuestión

CAPÍTULO 2

Revisión de la cuestión

Los objetivos de este capítulo son dar una introducción al estado de las investigaciones de GRS1915+105 y al estado actual de la minería de datos, plantear el problema y requisitos del mismo, y centrar el trabajo en aquellas tareas y métodos dentro de la minería de datos que se ajustan al problema.

2.1. Contexto de GRS1915+105

La astronomía siempre ha estado ligada a la humanidad. En el último siglo los avances han superado con creces al conjunto de los siglos anteriores, telescopios cada vez más potentes nos acercan a mundos cada vez más lejanos. Mediante el estudio de la longitud de onda de la radiación electromagnética se han realizado grandes avances, buscando patrones en estas señales han permitido la clasificación de gran cantidad de objetos varios como tipos de estrellas, planetas, meteoritos o sistemas. Estas radiaciones se recogen en una amplia longitud de onda, incluidos los rayos gamma, los rayos-X, los ultravioletas, los infrarrojos, etcétera.

El problema que tiene esta radiación electromagnética es que es absorbida por la atmósfera, por lo que los instrumentos de detección deben situarse a gran altitud, en la actualidad dichos instrumentos se instalan en satélites. La fuente de los rayos-X son remanentes estelares, estrellas de neutrones o agujeros negros. No fue un satélite, sino la sonda Granat la que detecto el GRS1915+105, y desde entonces ha sido objeto de estudio por sus especiales características.

El observatorio de rayos-X Chandra ha recogido muestras de lo que puede ser un mecanismo de autorregulación en la tasa de crecimiento del sistema. El sistema muestra gran variabilidad, o variación en la longitud de onda, donde pasaría por un estado en el que se expulsa materiales, y un estado en el que un viento

caliente que sopla desde el disco de acreción ahogaría dicho chorro. El viento priva de materiales necesarios para sostener el chorro. Cuando el viento amaina, se vuelve al estado inicial [Greiner, 2001].

2.2. Variabilidad de GRS1915+105.

La investigación se centra en el estudio de la variabilidad de GRS1915+105, buscando patrones en el comportamiento de la señal que ayuden a entender el sistema y su mecanismo de autorregulación . Actualmente en la definición de los estados de GRS1915+105 se tienen detectado doce estados distintos [Belloni et al., 2000], sin embargo esto no termina de satisfacer las necesidades de los astrónomos, ya que con estos patrones no dan respuestas a todas sus preguntas.

La información recogida se corresponden con tres canales de rayos-X, sobre el comportamiento de la señal y la relación entre los canales se centra los trabajos realizados [Belloni et al., 2000]. Esta información viene representada en forma de curvas de luz. Las curvas de luz es una manera de representar la intensidad en función del tiempo. Las curvas de luz son una herramienta fundamental en astronomía, que nos permite aplicar técnicas de espectrometría, o estudiar la variabilidad de los objetos celestes[Ibanogammalu, 1998] [Debosscher et al., 2011]. Por ejemplo, centrándonos en la variabilidad, se puede estimar la rotación de un objeto asumiendo que los picos de luz se corresponden con regiones más luminosas del objeto, y que la franja entre dos picos iguales se corresponde con una rotación completa del objeto [Blomme et al., 2007].

Los datos proporcionados vienen en ficheros de curvas de luz, un formato de fichero que almacena curvas de luz. Además de guardar la curva de luz contiene información sobre la calidad de la señal recogida [NASA, 1997].

Busqueda de frecuencias significativas. Para describir el comportamiento de la señal se utiliza la búsqueda de frecuencias significativas [Birney et al., 2006]. Existen distintas técnicas en la búsqueda de estas, aunque no todas resultan válidas.

- Los métodos de **cadena** (*string methods*).
- Método **de fase de organización de la dispersión** (*phase dispersion minimization*).
- **Análisis de Fourier.**
- **Transformada rápida de Fourier.**
- **Periodograma Lomb-Scargle.**

Lo normal es que las muestras tengan un nivel de ruido comparable a la amplitud de la variabilidad, que existan múltiples periodos de frecuencias significativas, o que se exista una baja densidad de muestreo. Cuando analicemos las muestras veremos que esto se corresponde con la situación presentada en el trabajo.

El problema de métodos de cadena y el método de la fase de minimización de la dispersión es que solo son válidos en caso de tener un único periodo significativo, por lo que son directamente descartados al no ajustarse a los requisitos impuestos por las muestras. Esto deja el análisis de Fourier, transformada rápida de Fourier y periodograma Lomb-Scargle como únicas alternativas válidas.

Lomb-Scargle tiene ciertas ventajas que hacen que sea la opción elegida frente a las otras alternativas, fue diseñada para datos espaciados irregularmente, tiene una gran tolerancia al ruido y a cortes y perdidas en la señal [Ibanogammalu, 1998], siendo una opción muy utilizada en astronomía [Birney et al., 2006]. La explicación completa del método la podemos encontrar en la sección del trabajo 3.2.1.

Hardness Ratio. Otra información relevante, y sobre la que versan los estudios previos, es la relación existente entre bandas [NASA, 2008], la forma de representar esta relación es usando los *Hardness Ratio* (HR). Si nuestra señal viene representada por tres colores distintos $TotalLightCurve = ColorA + ColorB + ColorC$, Hardness Ratios sería la relación entre color B y color A, y la relación entre color C y color A, siendo $HR1 = ColorB/ColorA$ y $HR2 = ColorC/ColorA$.

Hardness Ratio no contiene periodos, la forma de extraer información útil que describa la variable HR1 y HR2 es la utilización de la estadística, buscando describir la intensidad, la varianza o la distribución que toma los valores de ambos Hardness Ratios [Dodge and Rousson, 1997]. Esto no solo resulta útil para los Hardness Ratio, sino que la estadística resulta de utilidad para cualquier variable, es por ello que también resulta aplicable en la curva de luz total. La explicación completa a esta cuestión la encontramos en la sección del trabajo 3.2.1.

2.3. Minerías de datos

La minería de datos es un término relativamente moderno, que está teniendo una gran expansión, se encuentra dentro de las ciencias de la computación, refiriéndose al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. A pesar de la popular la minería de datos no es más que una etapa dentro de lo que se ha venido a llamar extracción de conocimiento a parir de datos.

El abanico de técnicas y métodos es muy amplio, siendo en ocasiones ambigua la frontera de cuales pertenecen la minería de datos propiamente y cuales pertenecen a otros ámbitos, y no existiendo uno claramente superior al resto. Todo depende de las necesidades, y de la base de datos con la que se cuente.

Aplicación en Astronomía. La minería de datos tiene aplicación en un amplio número de campos. Lo que hace a la minería de datos útil para conocer en que estante debe ir un producto en un supermercado, también la hace útil en la astronomía, y esto es la búsqueda de patrones. Por supuesto los requisitos de cada problema son distintos, y esto hace que no sean válidas las mismas técnicas. En astronomía es frecuente que las técnicas y métodos deben poder trabajar con atributos continuos y grandes volúmenes de información, lo que también se convertirá en requisitos de nuestro sistema.

En la astronomía, la minería de datos, es capaz de encontrar patrones entre los millones de imágenes y datos que componen las bases de datos astronómicas. Estos patrones ayudan a identificar objetos o estados de objetos en el Espacio. En nuestro caso no buscamos patrones que nos ayude a identificar objetos, sino patrones del estado de un único sistema, el GRS1915+105. De esta forma se pretende forma grupos con muestras de características similares.

2.4. Modelos y tareas de minería de datos

Entre los modelos y tareas existentes en la minería de datos debemos elegir cuales se ajustan a nuestros problemas.

Los modelos empleados en la minería de datos para encontrar relaciones, patrones o reglas inferidas pueden ser de dos tipos [Hernández Orallo et al., 2004].

Modelo descriptivo. Identifican patrones que describen los datos, siendo útiles para explorar determinadas propiedades de los datos. Siguen un **aprendizaje no supervisado** donde no se requiere conocimiento externo que indique el comportamiento deseado [S. Sumathi and Sivanandam, 2006]. Se encuentran las siguientes tareas.

- **Agrupamiento.** Se evalúan similitudes entre datos para construir modelos descriptivos, analizar correlaciones entre las variables o representar un conjunto de datos en un pequeño número de regiones.
- **Reglas de asociación**. Se identifican afinidades entre la colección de registros examinados, buscando relaciones o asociaciones entre ellos. Esto son reglas del tipo $SiAentoncesB$.
- **Correlaciones.** Tarea descriptiva utilizada para determinar el grado de similitud de los valores de dos variables.

Modelo predictivo. Se emplea para estimar valores futuros de variables de interés. Siguen un **aprendizaje supervisado** donde se requiere determinar la respuesta que se desea generar. Se encuentran las siguientes tareas.
[Hernández Orallo et al., 2004]

- **Clasificación.** Encuentra relación entre los atributos de entrada y los registros de salida para comprender el comportamiento de los datos.
- **Regresión.** Es el aprendizaje de una función cuyo objetivo es predecir valores de una variable continua a partir de la evolución de otra variable también continua.

Mediante un aprendizaje no supervisado buscamos la descripción de muestras sin clasificar, buscando afinidades entre muestras, y obteniendo grupos de muestras con similares características. Esto nos lleva a la generación de **un modelo descriptivo, abordando la tarea de agrupamiento**.

Ahora bien, los modelos de agrupamiento son difíciles de evaluar. Una de las técnicas consiste en la comparación con los resultados con otro modelo. La generación de un modelo de clasificación nos permitirá no solo la predicción de resultados, sino sobre todo poder evaluar los resultados del modelo de agrupamiento. Es por ello que generaremos **un modelo predictivo, abordando la tarea de clasificación**.

2.4.1. Técnicas de generales

Cada una de las tareas descritas requiere métodos, técnicas o algoritmos. Una única tarea puede ser resuelta por varios métodos distintos, y un método puede ser válida en distintas tareas [Hernández Orallo et al., 2004].

Técnicas algebraicas y estadísticas. Generalmente se basan en expresar modelos y patrones mediante fórmulas algebraicas, funciones lineales, funciones no lineales, distribuciones o valores agregados estadísticos como medias, varianzas, correlaciones, etc. Generalmente cuando obtienen un patrón es mediante el uso de un modelo predeterminado, de ahí su no nombre de técnicas paramétricas, aunque también existen técnicas no paramétricas dentro de este tipo de técnicas.

Técnicas bayesianas. Mediante el teorema de bayes, se basan en la estimación de probabilidades de pertenencia a un grupo o clase. Algunos algoritmos que pertenecen a estas técnicas son clasificador bayesiano naive, métodos basados en máxima verosimilitud y el algoritmo EM (en cierta medida). Tienen como ventaja que pueden ser representados de forma gráfica.

Técnicas basadas en conteos de frecuencias y tablas de contingencia. Basadas en contar la frecuencia en la que dos o más sucesos se presentan conjuntamente. Cuando el conjunto es grande existen algoritmos que van comenzando

con pares de sucesos e incrementando el conjunto solo en aquellos casos en el que las frecuencias conjuntas superen un umbral. Este el caso de los algoritmos "Apriori" y similares.

Técnicas basadas en arboles de decisión y sistemas de aprendizaje de reglas. Algoritmos cuyo modelo se puede representar en forma de reglas. Existen dos tipos: "divide y vencerás", como ID3/C4.5; y "separa y vencerás" como el CN2.

Técnicas relacionales, declarativas y estructurales. Los modelos de estas técnicas se representan por medio de lenguajes declarativos, lenguajes lógicos, funcionales o lógico-funcionales. La más conocida es ILP (programación lógica inductiva).

Técnicas basadas en redes neuronales artificiales. Técnicas donde se aprende un modelo mediante el entrenamiento de pesos que conecten un conjunto de nodos o neuronas. La topología y el peso de las conexiones determinan el patrón aprendido. Existen gran cantidad de variantes como: perceptrón simple, redes multicapa, redes de base radial, redes de Kohonen, etc, que a su vez tienen gran cantidad de algoritmos para su organización.

Técnicas basadas en núcleo y máquinas de soporte vectorial. En ella el algoritmo intenta maximizar el margen entre grupos o clases mediante transformaciones que pueden aumentar la dimensional, llamadas *kernels*. Existen muchísimas variantes dependiendo del núcleo utilizado y la manera de trabajar con el margen.

Técnicas estocásticas y difusas. Junto a las redes neuronales, incluye la mayoría de las técnicas de computación flexible. Técnicas que o bien los componentes aleatorios son fundamentales, como el *simulated annealing*, los métodos evolutivos y genéticos, o bien utilizan funciones de pertenencia difusa (*fuzzy*)

Técnicas basadas en casos, en densidad o distancia. Métodos basados en la distancia de los elementos, ya sea de forma directamente, como los vecinos más próximos (los casos más similares) o de manera más complicada mediante la estimación de funciones de densidad. Además de los vecinos más próximos algunos algoritmos muy conocidos son los jerárquicos, como Two-step o COBWEB, y los no jerárquicos, como K-medias.

También existen técnicas y métodos híbridos. Resulta difícil realizar una taxonomía óptima e indiscutible de todas las técnicas, por ejemplo EM, mencionado en esta taxonomía dentro técnicas bayesianas porque se nutre de técnicas bayesianas, también puede estar clasificado dentro de las técnicas basadas en densidad, ya que calcula las funciones de densidad para cada uno de los grupos.

2.4.2. Técnicas de agrupamiento

¿Qúe técnicas pueden ser válidas para el agrupamiento? En la tarea del agrupamiento dividimos una población heterogénea, en nuestro caso las muestras, en subgrupos homogéneos de acuerdo a las similitudes de sus registros.

Existen unas cuantas técnicas utilizadas en el agrupamiento, pero no todas son válidas para nuestro sistema. En trabajos anteriores se clasifica la variabilidad del sistema en distintas clases, sin embargo se busca **no predefinir el número de grupos**, quedando como requisitos.

- No requerir el número de grupos buscados.
- Manejar grandes de datos.
- Manejar con atributos continuos.
- Tolerancia al ruido y datos perdidos.

En función de en qué se basan las aproximaciones se cuenta con aproximaciones: [Bishop., 2006] [Hand and Kamber, 2006].

Basadas en particionados. Donde se dividen el conjunto de entrenamiento de **k** particiones o grupos. El óptimo local vendría dado por **K Medias** o **K Medianas**.

Algunos algoritmos basados en particiones solo trabajan con conjuntos de entrenamiento pequeños, lo que hace no cumplan con los requisitos. Pero entre las limitaciones generales de estos métodos está que son poco robustos ante el ruido, y es necesario introducir k.

Basado en redes neuronales. Como Mapas auto-organizativos de Kohonen cuya capa final contiene tantos nodos como grupos buscados, y que por tanto queda descartadas, por la misma razón que la aproximación anterior.

Jerárquicos. En el agrupamiento jerárquico se va construyendo un árbol donde se parte de todos los individuos en la raíz, y se considera cada subnodo como un subconjunto del nodo anterior, hasta llegar a las hojas o grupos finales. El criterio para realizar cada separación de cada nodo se realiza disminuyendo la distancia posible entre individuos. Este tipo de árbol se puede construir comenzando desde las hojas hasta llegar a la raíz, *aglomerativos*, o comenzando en la raíz hasta llegar a las hojas, *divisivos* [Larose, 2005].

Las técnicas basadas agrupamiento jerárquico son una posible opción, no siendo necesario introducir como parámetro inicial el número de grupos buscados cumplen con el primer requisito.

COBWEB, al estar dentro del agrupamiento jerárquico, por lo que cumple con el primer requisito, y además maneja un gran número de muestras gracias a que incorpora de manera incrementa las muestras al dendograma.

Siendo aparentemente una alternativa viable COBWEB tiene una serie de limitaciones que finalmente han hecho que sea descartado [Bishop., 2006] [Hand and Kamber, 2006].

- Solo trabaja con atributos discretos. Habría que transformar los atributos continuos.
- Asume atributos independientes. No podemos garantizar este hecho sin ningún género de duda.
- No hay garantía de mínimo local.
- Es sensible al orden en que se muestran las muestras.

Basados en densidades o probabilísticos **EM (Expectation Maximization)**, donde se busca la función de densidad probabilística desconocida a la que pertenece un conjunto complejo de datos. En los métodos basados en densidades un conjunto no pertenece a un grupo sino que puede pertenecer a distintos grupos con distintas probabilidades [Bishop., 2006] [Hand and Kamber, 2006].

Cumple con el primer requisito, estimando k optimo, maneja bien grandes volúmenes de datos, trabaja con variables continuas, y además resulta tolerante al ruido y perdida de datos [Garre et al., 2005] [Garre et al., 2007]. Además trabajar probabilidades es una ventaja añadida, puesto que las muestras no siempre pertenecen a una clase, sino que pueden contener características que lo hagan pertenecer a varias clases a la vez.

Es por ello que **EM es el mejor candidato para realizar el agrupamiento**.

2.4.3. Técnicas de clasificación

Dentro de las tareas del modelo predictivo nos centramos en la clasificación. En ella tendremos como datos de entrada los atributos seleccionados, ya en el modelo descriptivo, y las salidas, o grupos obtenidos en el modelo descriptivo. Al terminar el aprendizaje obtendremos un modelo capaz de clasificar cada nueva entrada en un grupo distinto. Esto se engloba dentro del aprendizaje supervisado.

El principal problema es que el aprendizaje adquirido es poco representativo y no nos proporciona un conocimiento detallado, siendo como una caja negra de entradas y salidas. Existen algoritmos que si solucionan este tema como son las clasificaciones bayesinas, árboles de decisión, etc [Hernández Orallo et al., 2004].

Tipos de datos. Las técnicas antes mencionadas son utilizables en una o varias tareas de minería de datos. Pero también es necesario tener en cuenta que tipo de datos estamos tratando. Pudiendo ser datos complejos y/o heterogéneos, pero también puede darse el caso de tener que aplicarlo a bases de datos espaciales, temporales, secuenciales, con contenido multimedia, o documentos.

Clasificación de datos secuenciales. En el problema tratado en el trabajo nos encontramos ante una secuencia de eventos, ver aparto 4.4.6, donde datos continuos muestran algún tipo de relación. Un ejemplo típico de datos secuenciales está en el reconocimiento de caracteres, donde hay que tener en cuenta una serie de normas a la hora de interpretarlos, así en castellano la mayoría de las veces después de dos consonantes nos encontraremos una vocal [Hernández Orallo et al., 2004]. Algo parecido nos ocurre con los estados donde la probabilidad de encontrarse un estado determinado dependerá del anterior.

Existen tres técnicas muy similares entre ellas para tratar este tipo de datos:

- **Predicción de secuencia de un objeto.** Consiste en dada una secuencia $y_i = \langle y_1, y_2, y_3, ..., y_T \rangle$. predecir y_{T+1}.
- **Aprendizaje supervisado secuencial.** El problema de la clasificación secuencial puede ser formulado de la siguiente manera [Dietterich, 2002]. Sea x_i, y_i, $1 \leqslant i \leqslant N$ en un conjunto de N ejemplos. Cada ejemplo es un par de secuencias. (x_i, y_i), y donde $x_i = \langle x_{i,1}, x_{i,2}, x_{i,3}, ..., x_{i,Ti} \rangle$ y $y_i = \langle y_{i,1}, y_{i,2}, y_{i,3}, ..., y_{i,Ti} \rangle$. El objeto de aprendizaje es obtener un modelo h que pueda predecir una secuencia "clase" $y = h(x)$ dado una secuencias de entrada x. Es decir la clase no es un atributo sino una secuencia de atributos.
- **Clasificación de secuencias.** Dado una secuencia $x_i = \langle x_1, x_2, x_3, ..., x_T \rangle$ el objetivo es predecir un clase x_T, ahora si, como un atributo aislado y no una secuencia de atributos.

Las tres técnicas están muy relacionadas, y un problema puede ser abordado por varias de ellas al mismo tiempo [Hernández Orallo et al., 2004].

Algunas de las técnicas utilizadas para el aprendizaje supervisado secuencial son:

- **Ventana deslizante**, donde el problema se convierte en un problema clásico de aprendizaje supervisado. [Fawcett, 1997]
- **Ventana recurrente**, similar a la primera [Bakiri and Dietterich, 2002].
- **N-gramas**, que es una particularizan de las ventanas aplicadas a las ventanas de texto.

- **Modelos ocultos de Markov o HMM)** (*Hidden Márkov Model*) que representan modelos probabilísticos de cómo se generan las secuencias x_i e y_i [Rabiner and Juang, 1986]

En el trabajo buscamos la validación del modelo de agrupamiento, para ello realizaremos una clasificación de frecuencias. Dentro de las técnicas bayesianas hemos optado por el **modelo oculto de Márkov** o **HMM** (*Hidden Markov Model*), que ha demostrado su eficacia en gran número de utilidades como el reconocimiento de la voz, reconocimiento del gesto, criptoanálisis, etc [Huang et al., 2001], demostrando buenos resultados tanto en problemas de aprendizaje supervisado secuencial como en problemas de clasificación secuencial [Dietterich, 2002], siendo esta última la aplicada en el trabajo. Una explicación detallada del algoritmo la encontramos en el apartado 5.2.

Por otra parte el modelo oculto de Márkov tiene una ventaja importante, y es que al mostrarnos las hipótesis con las probabilidades nos permite medir la verisimilitud.

2.5. Técnicas de selección de atributos

Además de las técnicas aplicadas a los modelos y las tareas, es necesario buscar resolver ciertas cuestiones no pertenecientes estrictamente al modelado.

Tras la extracción de la información de la señal y los colores que componen esta, que abordaremos en el apartado 3.2, tendremos el problema de validar cuán de útil resulta cada uno de los atributos extraídos. No todos los atributos serán igualmente relevantes, y lo que resulta de utilidad para casos aparentemente similares, no tiene porqué ser de utilidad en nuestro problema. Por ejemplo, en estrellas variables resulta de utilidad atributos tales como la amplitud y la fase, y no ser útiles en nuestro sistema.

Realizando un análisis individual de los atributos, se deben eliminar atributos identificadores de la muestras, o eliminar atributos recomendados por expertos. Después recurrimos a técnicas que permitan medir la importancia de cada atributo con respecto al resto. Existen dos grandes grupos
[Hernández Orallo et al., 2004].

Métodos de selección, que proporcionan un conjunto básico de atributos significativos. Existen dos tipos generales de métodos para la seleccionar características.

- **Métodos de filtro** o métodos previos, donde los atributos irrelevantes se filtran antes de cualquier proceso de minería de datos. Son técnicas fundamentalmente estadísticas, donde el conjunto optimo se basa en medidas de calidad previa que se calculan a partir de los datos mismos [Liu and Dash, 1997].

- **Métodos basados en modelo** o métodos envolventes (*wrapper*), donde los atributos se evalúan con respecto a la calidad de un modelo de minería de datos o estadístico extraído a partir de los datos. Estos métodos resultan más costosos de evaluar, además los métodos seleccionados para generar el modelo y para la evaluación no necesariamente tiene que ser los mismos.

Métodos de evaluación, que proporcionan un ranking de atributos de más a menos significativos. Completando de esta forma la información dada por el primer conjunto de métodos.

2.6. Estrategias de evaluación de modelos

Una vez terminado el modelo es necesario marcar una estrategia de evaluación. Cada tipo de modelo tiene unas características que hacen que la estrategia tenga que ser diferente. A continuación vamos a resumir las principales estrategias de los modelos abordados en el trabajo, que nos permitirán establecer medidas de calidad paras las hipótesis encontradas.

2.6.1. Evaluación de modelos de agrupamiento

El problema de los modelos de agrupamientos es que no existe una clase o valor numérico donde medir las veces que el modelo aprendido predice correctamente, esto hace que sean modelos difíciles de evaluar.

Existen varias estrategias básicas para evaluar este tipo de modelo.

- **Basados en la verosimilitud.** La verosimilitud o *likelihood* significa quedarse con la hipótesis con la cual los datos sean más verosímiles, es decir donde $P(D|h)$ se maximice, siendo D los datos y h la hipótesis. Para ello, se mide $p(x)$ que es la probabilidad estima de observar un punto en la posición x utilizando el modelo a evaluar, p es una función de densidad. Si el modelo es bueno $p(x)$ en los puntos observados será elevada [Hernández Orallo et al., 2004].

  De tal forma que el objetivo consistirá en maximizar la función:

$$L = \prod_{i=n}^{n} p(x(i)) \tag{2.1}$$

  Utilizando un logaritmo de probabilidades:

$$\log L = \sum_{i=n}^{n} \log p(x(i)) \tag{2.2}$$

En caso de utilizar funciones negativas, el objetivo es minimizar la función:

$$S_L = -\log L = -\sum_{i=n}^{n} \log p(x(i)) \tag{2.3}$$

Siendo S_L una medida de error, o un tipo de entropía que mide cómo el modelo describe los datos de entrenamiento.

En caso de no utilizar modelos que no proporcionen la probabilidad, se puede utilizar una medida que estudie cuán compactos son los grupos, como por ejemplo el error cuadrático de cada grupo.

$$S_{EC}(m) = \sum_{i \in cluster_k} |A|\, x(i) - c_k \left\|B\right\|_2 \tag{2.4}$$

- **Basados en la distancia.** Donde se mide la distancia entre los grupos como medidas de calidad del modelo de agrupación. A mayor distancia entre grupos, mejor será la separación, entre ellos.
- **Basadas en modelos.** Consiste en la generación de varios modelos, donde utilizaríamos distintas algoritmos. Si los resultados son similares, podemos pensar que la agrupación es acertada. Una variante sería la utilización de un modelo de agrupación / test de clasificación y regresión, y donde compararemos los resultados de ambos.
- **Basado en la complejidad de la hipótesis.** En el que se evalúa el modelo en función de la complejidad de la hipótesis. La hipótesis más simple, en igualdad de condiciones, es la correcta, siguiendo lo que se conoce como navaja de Occan. Existen métodos que nos permiten medir la complejidad de las hipótesis como el principio *MDL (Minimum Dresciption Length)*. Donde se mide el tamaño de bits (unidades de información) $K(h)$ de la descripción de una hipótesis más la descripción de los ejemplos que no son cubiertos (excepciones) $K(D|h)$. Es decir el objetivo es minimizar la función:

$$K(h) + K(D|h) \tag{2.5}$$

La concreción de que métodos abordaremos para evaluar el agrupamiento será abordada en la sección del trabajo 4.5.

2.6.2. Evaluación de modelos de clasificación

Resulta más sencillo de medir que para modelos de agrupamiento, ya que se cuenta con un conjunto de entrada cuya salida es conocida. Existen distintas aproximaciones para la evaluación de la hipótesis.

- **Basadas en la precisión.** Consistiría en el cálculo del error de muestreo 2.6, que calcula el error a partir de los acierto, $\delta(verdadero) = 1$ y de los

fallos $\delta(falso) = 0$, siendo h la hipótesis y f la función objeto dentro de un número de componentes S.

$$error_s(h) = \frac{1}{n} \sum_{x \in S} \delta(f(x) \neq h(x)) \tag{2.6}$$

y error verdadero 2.7, que indica la probabilidad de que h prediga un resultado diferente a f

$$error_t(h) = Pr_{x \in D} \left[(f(x) \neq h(x)) \right] \tag{2.7}$$

donde $Pr_{x \in D}$ denota que x se toma siguiendo una distribución de probabilidad D.

El error verdadero es lo que desearíamos conocer, sin embargo solo podemos conocer el error de muestreo.

Existen distintos mecanismos para el cálculo de este error, utilizando el conjunto de entrenamiento como conjunto de test corremos el riesgo de favorecer un sobreajuste, esto se da porque el modelo está ajustado al mismo conjunto de test, y no funciona bien con otros ejemplos. Para evitar el sobreajuste existen técnicas como **validaciones cruzadas**, que consiste en dividir el conjunto de evidencias en k subconjunto, utilizando $k - 1$ para el entrenamiento y el resto para el test, se repite esto por cada uno de los conjuntos, el resultado será la medía [Dietterich, 1998]. El **bootstrap** que resulta útil en el caso de contar con pocos ejemplos, es similar a las validaciones cruzadas aunque se procede de forma distinta, puesto eligiendo el conjunto de muestras inicial de manera aleatoria.

- **Basada en coste.** No todos los errores cometidos tienen el mismo valor, ejemplo no es lo mismo enviar a un paciente enfermo a casa que ingresar a un paciente sano. Para solucionar esto existen este tipo de técnicas, donde se utiliza una matriz de confusión. Si se dispone de una matriz de confusión $Coste = \sum_{1 \leq i \leq n, 1 \leq j \leq n} C_{i,j} M_{i,j}$, donde $C_{i,j}$ expresa la posición i, j de la matriz de coste, y $M_{i,j}$ en la matriz de confusión. De esta forma es posible favorecer modelos que aun cometiendo más errores de muestreo, estos sean menos costosos. En caso de no disponer de matriz de confusión existe alternativas como **análisis ROC** (*Receiver Operating Characteristic*), y que permite seleccionar un conjunto de operadores óptimo en general.

Existen otras medidas de evaluación. La estrategia concreta utilizada para la evaluación de la clasificación se realizara en la sección 5.6

2.7. Metodologías de minería de datos

Resulta peligroso aprender cosas que no son útiles, tomando decisiones importantes en información y conocimiento incorrecto. Esto se da cuando los patrones obtenidos no representan reglas, el modelo desarrollado no representa la

población relevante o los datos pueden estar a un nivel equivocado de detalles.

También es posible llegar a datos no útiles. Esto se da cuando: aprendemos cosas ya conocidas o cosas no utilizables. Llegar a conclusiones conocidas no siempre es inútil, en nuestro caso si los grupos encontrados coinciden con la clasificación existente tendríamos una demostración de que la clasificación es razonablemente precisa.

Seguir una metodología procura evitar resultados indeseables y se basa en las mejores prácticas. En la minería de datos es recomendable seguir una metodología. Entre ellas esta: CRISP-DM de *Cross Industry Standard Process for Data Mining* que es la principal métodologia; SEMMA de *Sample Explore Modify Model Assess* que es la segunda metodología más utilizada; Y Berry y Linoff.

Nos hemos decantado en el trabajo por seguir CRIPS-DM por las mismas razones que la hacen la más popular. Es bien conocida y está bien especificada, aunque muchas de estas metodologías comparten tareas o tienen tareas similares y podrían ser igualmente válidas.

2.7.1. CRISP-DM

La metodología se fundamenta en las siguientes fases [Chapman et al., 2000].

1. **Entendimiento del Negocio.** *Objetivos y requerimientos desde una perspectiva no técnica.*
   Comprender los objetivos y requerimientos del proyecto, definir el problema de minería de datos y trazar un plan preliminar son los objetivos a alcanzar en esta fase.

2. **Entendimiento de los datos** *Familiarizarse con los datos teniendo presente los objetivos del negocio.*
   Recopilar, explorar y medir la calidad de los datos.

3. **Preparación de los datos** *Obtener una vista minable o dataset*
   Seleccionar datos, limpiar los que no cumplan con los criterios de calidad, estructurar los datos, integrar los datos y formatear los datos.

4. **Modelado** *Aplicar técnicas de minería de datos a los dataset.*
   Seleccionar la técnica de modelado, diseñar la evaluación, construir el modelo y evaluar el modelo.

5. **Evaluación** *Determinar que modelos de las fases anteriores son útiles.*
   Evaluar resultados, revisar el proceso y establecer siguientes pasos.

6. **Despliegue** *Explorar utilidad del modelo e integración en las tareas de tomas de decisión*
   Planificar el despliegue, planificar monitorización y mantenimiento, generación del informe final y revisión del proyecto.

2.8. Herramientas de minería de datos

Actualmente existen gran número de herramientas que se pueden utilizar en la minería de datos, como SPSS Clementine, WEKA, Kepler, ODMS Oracle Data Mining Suite (Darwin) DBMiner, o librerías.

El trabajo se centra dos herramientas gratuitas bajo la licencia GNU y GLP respectivamente:

Weka. Para la selección de atributos. Herramienta hecha en Java por la universidad de Waikato (Nueva Zelanda),

Librerias R. Para la generación de los modelos un conjunto de librerías en R que engloba desde el tratamiento de los datos, lectura de ficheros Fits, búsqueda de frecuencias, estadística hasta los modelos. El generar los modelos con R y ser este un lenguaje de programación nos permite la definición de nuevas funciones si fuera necesaria. Además tiene la potencia suficiente como para la ejecución de grandes volúmenes de datos, siendo utilizado como sustituto de Matlab.

Parte III

Modelado y Evaluación

CAPÍTULO 3

Selección de Información

3.1. Comprensión de los datos

El objetivo de esta sección del trabajo es familiarizarse con los datos con los que cuanta el trabajo.

3.1.1. Recopilación de los datos.

Todos los datos pertenecen a un único sistema celeste, el GRS1915+105, también llamado V1487 Aquilae. Es un sistema binario compuesto por un agujero negro y una estalla [Finley, 1994]. Entre sus características cuanta con un sistema de autorregulación del crecimiento [Greiner, 2001]. Se pretende, por medio del análisis de la variabilidad, encontrar patrones que ayuden a la compresión de dicho sistema y como se autorregula.

Observatorios de rayos-X, entre ellos Chandra, han recogido muestras durante 10 años. Son estos los datos proporcionados por la ESAC y que serán objetos de estudio del trabajo.

3.1.2. Descripción de los datos.

Las muestras vienen en formato de curvas de luz. Las curvas de luz es una forma gráfica de representar la intensidad de la luz en función del tiempo. Por cada muestra se cuenta con cuatro ficheros que recogen tres colores y la señal completa o sumatorio de los tres restantes colores.

- **ColorA** Resta el nivel constante en el fondo de la señal de 10 cts/seg.
- **ColorB** Resta el nivel constante en el fondo de la señal de 20 cts/seg.

- **ColorC** Resta el nivel constante en el fondo de la señal de 100 cts/seg.
- **Curva de Luz Total**, R ,Recoge la señal completa o sumatorio de los tres colores anteriores. $R = ColorA + ColorB + ColorC$

Las muestras de curva de luz vienen en fichero Light Curve FITs, los cuales pueden tener distintos formatos. En nuestro caso contienen cuatro columnas [NASA, 2011].

- **TIME** Fase.
- **RATEn** Ctns/seg para la serie n.
- **ERRORn** Error en Rate para la serie n.
- **FRACEXP** Exposición fraccional.

Para el trabajo solo resulta de interés las dos primeras columnas, que contienen la información de la curva de luz, las restantes columnas recogen información correspondiente a la calidad de los valores recogidos.

En un alto porcentaje de muestras además se cuenta con una clasificación previa, basada en estudios anteriores no relacionados con la minería de datos [Greiner, 2001] [Belloni et al., 2000].

3.1.3. Exploración de los datos.

Se cuenta con 1912 muestras, y un total de 7490545 segundos o 86,7 días de señal en total.

Clasificación. En cuanto a la clasificación, existe un alto porcentaje de muestras clasificadas, 1066 muestras, aunque no todas las muestras clasificadas pertenecen a una clase claramente, y en muchos casos se tiene dudas sobre a que clase podría pertenecer, 267 tiene una clasificación dudosa pudiendo pertenecer a varias clases.

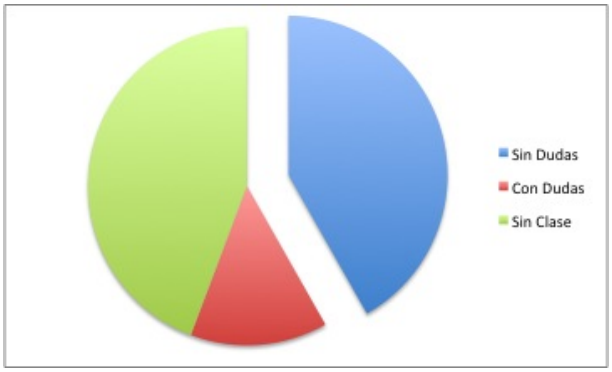


Figura 3.1: Resumen de los datos

Distribución de las clases. La distribución de las muestras tampoco es homogénea, existiendo clases representadas por muchas muestras, como por ejemplo la clase *Chi* representada por 490 muestras, y clases representadas por pocas muestras, como *Mu*.

clase	Total	sin dudas	con dudas	\%sin dudas	\%con dudas
phi	126	70	56	6,59%	8,76%
chi	490	447	43	25,63%	55,94%
gamma	15	7	8	0,78%	0,88%
mu	65	37	28	3,40%	4,63%
delta	76	39	37	3,97%	4,88%
theta	53	40	13	2,77%	5,01%
lambda	3	3	0	0,16%	0,38%
kappa	19	9	10	0,99%	1,13%
rho	154	124	30	8,05%	15,52%
upsilon	0	0	0	0,00%	0,00%
alpha	50	15	35	2,62%	1,88%
beta	15	8	7	0,78%	1,00%
sin clase	846			44,25%	
RESUMEN	1912	799	267		

Cuadro 3.1: Distribución de las muestras en clases

En el gráfico 3.2 se incluye todas las muestras clasificadas, incluidas las

dudosas, y se comparando con las muestras no clasificadas.

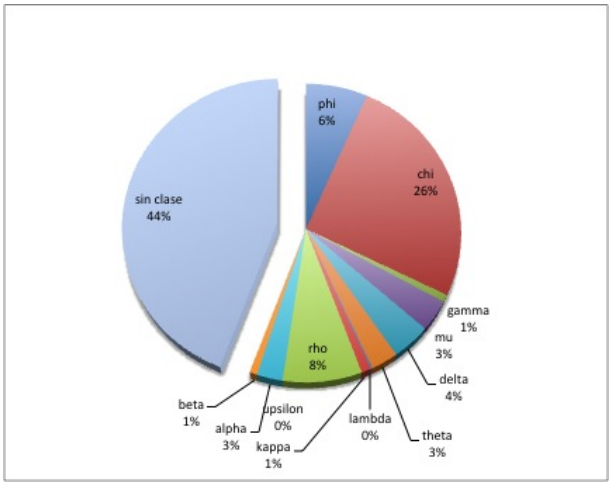


Figura 3.2: Distribución de las muestras en clases

En caso de eliminar muestras no clasificadas y muestras cuya clasificación es dudosa, la distribución de las muestras en clases cambia, quedando como se muestra en el siguiente gráfico.

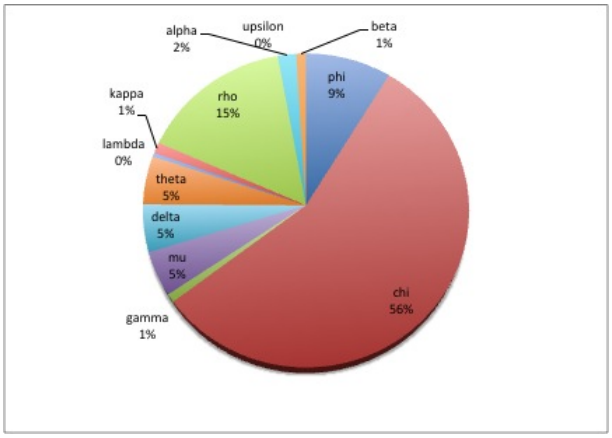


Figura 3.3: Distribución de las muestras clasificadas no dudosas

Contenido de las muestras. Al explorar el contenido de las muestras en estas curvas de luz puede existir **ruido, uno o varias frecuencias**. Además estas pueden ser **frecuencias altas**, **frecuencias bajas** o ambas en una sola

muestras, con distintas amplitudes. Dado que se trata de un sistema variable es la situación que cabría esperar.

Encontramos algunos problemas que describiremos y mediremos en el próximo apartado del trabajo.

3.1.4. Verificación de calidad de los datos.

Al realizar una exploración de las muestras nos encontramos con una serie de situaciones que merman la calidad de los datos.

Muestras incompletas. No se cuenta con todos los ficheros en todos los datos. En algunos casos faltan muestras correspondientes a alguno de los colores, si en total existen 1912, muestras con todos los colores son 1629,

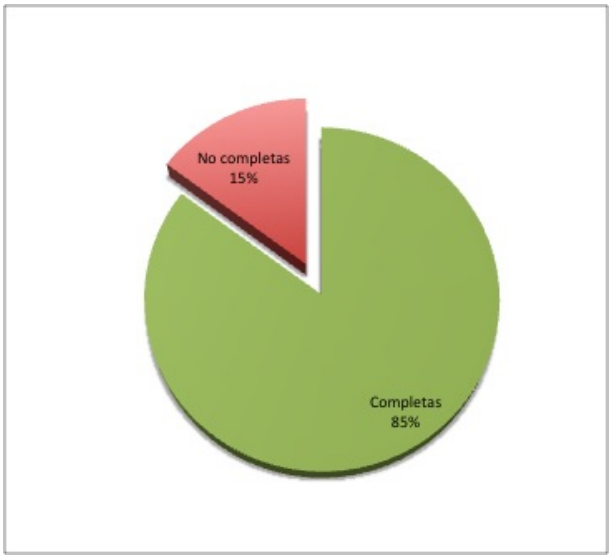


Figura 3.4: Muestras no completas

Cortes en la señal. Resulta más difícil determinar los cortes en la señal, analizando los puntos no existen punto en la señal sin definir, o nulos, significativos. Sin embargo explorando las muestras en muchos casos contienen cortes o huecos en la señal. Esto es debido a que no siempre los puntos son equidistantes en el tiempo, existen distintas densidades de muestreo. El gráfico que se muestra a continuación es un histograma del porcentaje de puntos nulos por señal, donde se puede apreciar el bajo porcentaje de ellos.

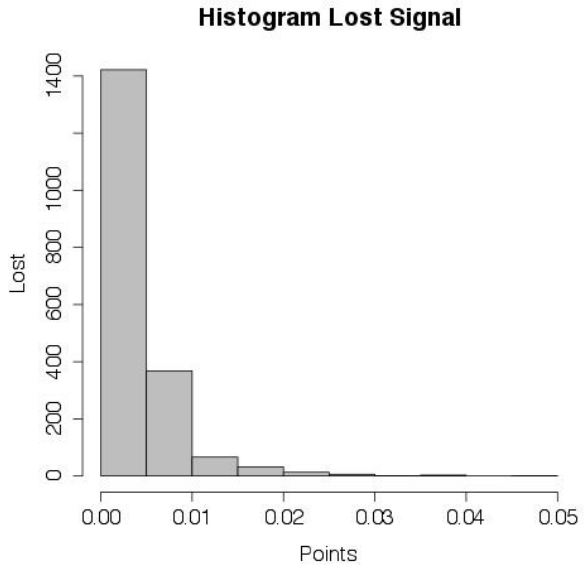


Figura 3.5: Histograma de porcentaje de puntos nulos por muestra

Variabilidad en la longitud. Dentro del conjunto de muestras totales se tiene muy diferentes longitudes de tiempo. En el siguiente gráfico se nos muestra el histograma de la distribución de la longitud de las muestras. En él se puede apreciar como la gran mayoría se encuentran entre 0 y 5000 segundos, pero existen gran cantidad de muestras con longitudes de tiempo más largos.

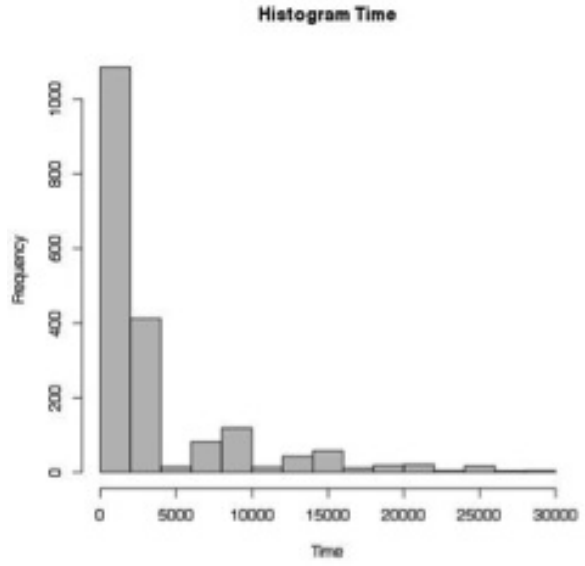


Figura 3.6: Histogramas de las longitudes de tiempo de las muestras

Esto puede resultar un problema que hay que solventar por dos motivos

- El agrupamiento toma en cuenta la longitud de las muestras en la búsqueda de patrones. Lo que puede generar agrupaciones incorrectas.
- En las muestras con una mayor longitud pueden existir uno o varios estados intermedios de la señal que no se analiza. Es decir, nos podemos perder estados en las muestras de mayor longitud.

3.2. Preparación de los datos

El objetivo de está parte del trabajo es la extracción de datos que pueda ser utilizables en la minería de datos y óptimos en la generación de un modelo.

3.2.1. Construcción de los datos

Las curvas de luz contienen información de la intensidad de la luz con respecto al tiempo [NASA, 1997]. Esto viene representado en las columnas *TIME* y *RATEn* de nuestros las muestras [NASA, 2011].

En las estrellas variables se producen periódicamente repeticiones de picos en la señal. Una forma de saber a cuanto rota una estrella es la medición de periodos en los que se producen picos de intensidad [Ibanogammalu, 1998]

[Debosscher et al., 2011]. En nuestro sistema la variabilidad resulta más compleja, sin embargo sigue siendo de utilidad la búsqueda de periodos en la señal. Existen varios algoritmos que nos permiten la búsqueda de frecuencias significativas [Birney et al., 2006].

Por otro lado, al margen de los periodos o frecuencias significativas podemos encontrar que la señal toma distintas formas. Una forma de describir la forma de la señal es mediante el análisis estadístico.

Buscamos extraer son de dos tipos de atributos.

- **Frecuencias significativas**. Búsqueda de frecuencias más significativas y valores derivados de estas frecuencias, que describan el comportamiento de la señal.
- **Descripción de la señal**. Describe la forma de la señal y de los Hardness Ratios mediante el análisis estadístico.

Colores y Hardness Ratio. Se cuenta con el desglose de los distintos colores que componen la señal. No resulta útil la búsqueda de periodos en los colores, ya que no los hay, pero si aporta mucha información la relación existente entre los colores, y *Hardness Ratio* (HR), es una forma de representar la relación de recuentos existentes entre bandas o colores [NASA, 2008].

$$TotalLightCurve = ColorA + ColorB + ColorC \tag{3.1}$$

$$HR_1 = ColorB/ColorA \tag{3.2}$$

$$HR_2 = ColorC/ColorA \tag{3.3}$$

Sobre HR no tiene sentido búsqueda de frecuencias significativas, si tiene sentido describir la forma que toma. Para describir la forma podemos utilizar el análisis estadístico.

Lomb-Scargle. Lomb-Scargle es el método utilizado para la búsqueda de frecuencias significativas, tiene como ventajas frente a métodos similares como Fourier, fue diseñada para datos especialmente irregulares, tiene tolerancia al ruido y a cortes o perdidas de la señal [Graymer, 1998] [Birney et al., 2006] [Ibanogammalu, 1998]. Es comúnmente usado en la búsqueda de frecuencias significativas de señales en el Espacio, como las procedentes de estrellas variables. Datos irregulares, ruido y cortes o perdidas en la señal son problemas de calidad comunes en nuestras muestras, como vimos al explorar los datos, lo que hace que sus ventajas sean un requisito en nuestro sistema, y sea por tanto la opción más adecuada.

Lomb-Scargle realiza un muestreo de frecuencias, donde se evalúan su significancia con la siguiente ecuación [Graymer, 1998]:

$$P_x(\omega) = \left(\frac{\left[\sum x_j \cos\omega\left(t_j - \tau\right)\right]^2}{\sum \cos^2 \omega\left(t_j \tau\right)} \right) + \left(\frac{\left[\sum x_j \sin\omega\left(t_j - \tau\right)\right]^2}{\sum \sin^2 \omega\left(t_j \tau\right)} \right) \tag{3.4}$$

Siendo:

- ω: frecuencia angular (una variable continua). Siendo $TestFrequency$ un valor que ira desde el mínimo hasta el máximo de las posibles frecuencias, se calcula como $\omega = 2\pi TestFrequency$.
- x_j: Datos con su promedio restado
- t_j: Tiempos sin restricciones de ser equidistantes.
- j: Cantidad de datos. $j = 0..N-1$
- τ: Definido por

$$\tau = \tan\left(2\omega t\right) = \sum\left(cos\left(2\omega t_j\right)\right) \tag{3.5}$$

  En nuestro caso utilizaremos la formula

$$\tau = \frac{atan2\left(\sum \sin\left(2\omega t_j\right), \sum \cos\left(2\omega t_j\right)\right)}{\sum \cos\left(2\omega t_j\right)} \tag{3.6}$$

- $TestFrequency$. Es el conjunto de frecuencias de muestreo sobre la que se calcula ω.

$$TestFrequencies[j] = MinFrequency + (MaxFrequency - MinFrequency)\left(\frac{(1:M-1)}{M-1}\right) \tag{3.7}$$

- M. Aumenta el número de frecuencias que se seleccionan entre $MaxFrequency$ y $MinFrequency$. En nuestro caso y dado que se trata de muestras con bastante longitud $M = N$.
- $MinFrequency$. Dado que $MinFrecuency = \frac{1}{MaxPeriod}$. $MaxPeriod$ es el máximo periodo que podemos esperar en un muestra y este nunca será mayor que la propia muestra.

$$MaxPeriod_j = \text{máx}(t_j) - \text{mín}(t_j) \tag{3.8}$$

- $MaxFrequency$. Dado que $MaxFrecuency = \frac{1}{MinPeriod}$. $MinPeriod$ es el mínimo periodo que podemos esperar en nuestra muestra. Siendo

$$MinPeriod_j = Nyquist = \frac{1}{\left(2\left(\frac{\text{máx}(t) - \text{mín}(t)}{length(t)}\right)\right)} \tag{3.9}$$

Se selecciona $P_x(\omega)$ con el valor máximo y la frecuencia a la que corresponde como la más significativa.

La **probabilidad de que el resultado, $P_x(\omega)$, sea una falsa alarma** viene calculado por.

$$P_r = 1 - (1 - \exp(P_x(\omega)))^{Nindependent} \tag{3.10}$$

Siendo:

- $NIndependent$ es el número de frecuencias independientes.

$$NIndependient = \left(-6{,}362 + 1{,}193N + \left(0{,}00098N^2\right)\right) \tag{3.11}$$

Como datos derivados se obtienen amplitud y fase. Cuya ventaja estriba en que son invariables a transacciones en el tiempo. En término de amplitud A y fase ϕ se define como:

$$x_j = A\cos(\omega t_j + \phi) \tag{3.12}$$

Con que:

$$a = \sum\left(\frac{(x_j \cos\omega(t_j\tau))}{\sum(\cos^2\omega(t_j - \tau))}\right) \tag{3.13}$$

$$a = \sum\left(\frac{(x_j \sin\omega(t_j\tau))}{\sum\left(\sin^2\omega(t_j - \tau)\right)}\right) \tag{3.14}$$

$$A = \sqrt{a^2 + b^2} \tag{3.15}$$

$$\phi = -\omega\tau - arctan\left(\frac{b}{a}\right) \tag{3.16}$$

Para reproducir la señal generada por la frecuencia encontrada en Lomb-Scargle podemos utilizar la siguiente formula.

$$y = A\cos\left(\frac{t - PeakMaximum}{PeakPeriod}\right) + mean(x) \tag{3.17}$$

Como ejemplo del resultado obtenido al ejecutar Lomb-Scargle mostramos una representación gráfica donde se ve de arriba a abajo y de izquierda a derecha.

- **Gráfico de la señal**. La señal original de viene en color azul.
- **Gráfico de la medida de la irregularidad de los intervalos**. En este caso no aplica al ser intervalos regulares.

- **Gráfico de Lomb-Scargle**. Indicando *PeakPeriod*, o periodo de la frecuencia más significativa, y resultado del muestreo para cada una de los periodos en color rojo.
- **Gráfico de probabilidad de falsa alarma.** Indican p para el *PeakPeriod* y el resultado de calcular la probabilidad de falsa alarma para cada una de las muestras en color rojo.

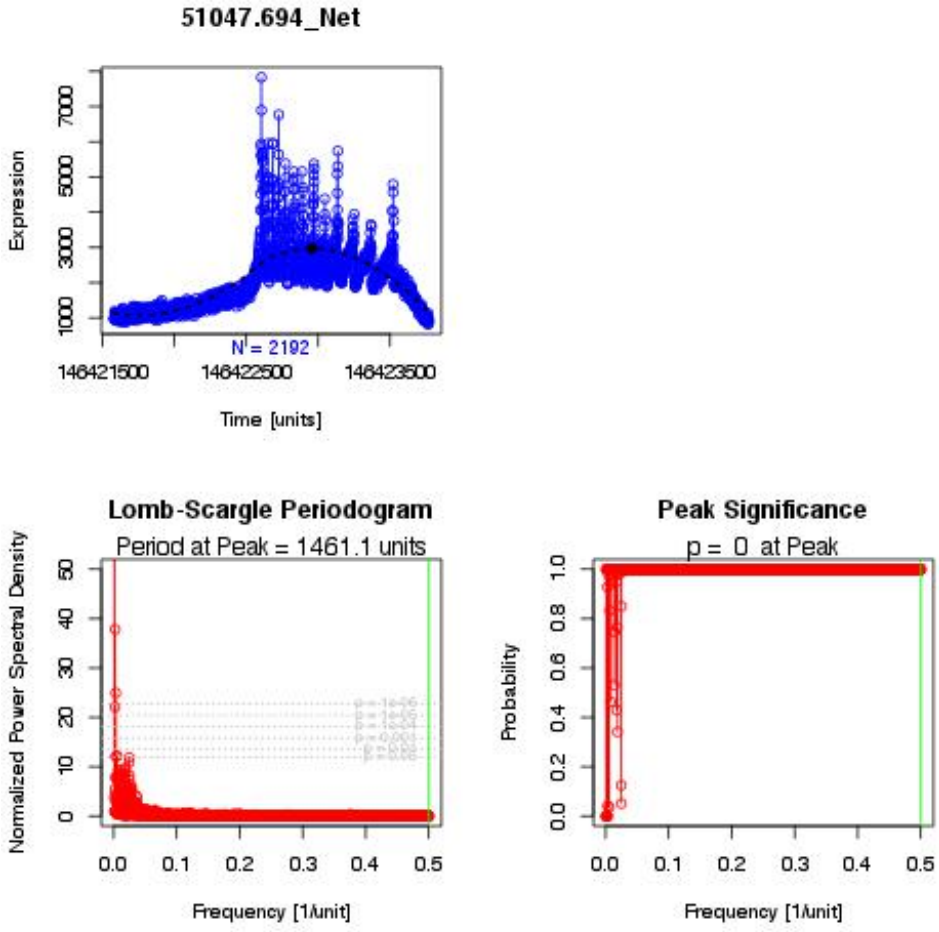


Figura 3.7: Ejemplo resultado Lomb-Scargle

Aplicación de filtros. Para la separación de frecuencias bajas y frecuencias altas se ha utilizado dos tipos de filtros [Cartwright et al., 2012].

- **Filtro Paso Bajo**. Permite el paso de frecuencias bajas, atenuando el paso de frecuencias más altas.
- **Filtro Paso Alto**. Permite el paso de frecuencias altas, atenuando el paso de frecuencias más bajas.

Tras la aplicación de los filtros se vuelve a aplicar los algoritmos de búsquedas de frecuencias. Esto nos permite centrarnos en frecuencias altas y bajas, comprobando que no existan frecuencias distintas a las detectadas.

Esto nos permite centrarnos en frecuencias altas y bajas según sea nuestra necesidad. Por ejemplo, en el caso de encontrar frecuencias bajas, podemos usar un Filtro Paso Alto para contrastar la existencia de frecuencias altas.

Como ejemplo de los resultados obtenidos tras aplicar filtros mostramos los resultados de Lomb-Scargle de forma gráfica.

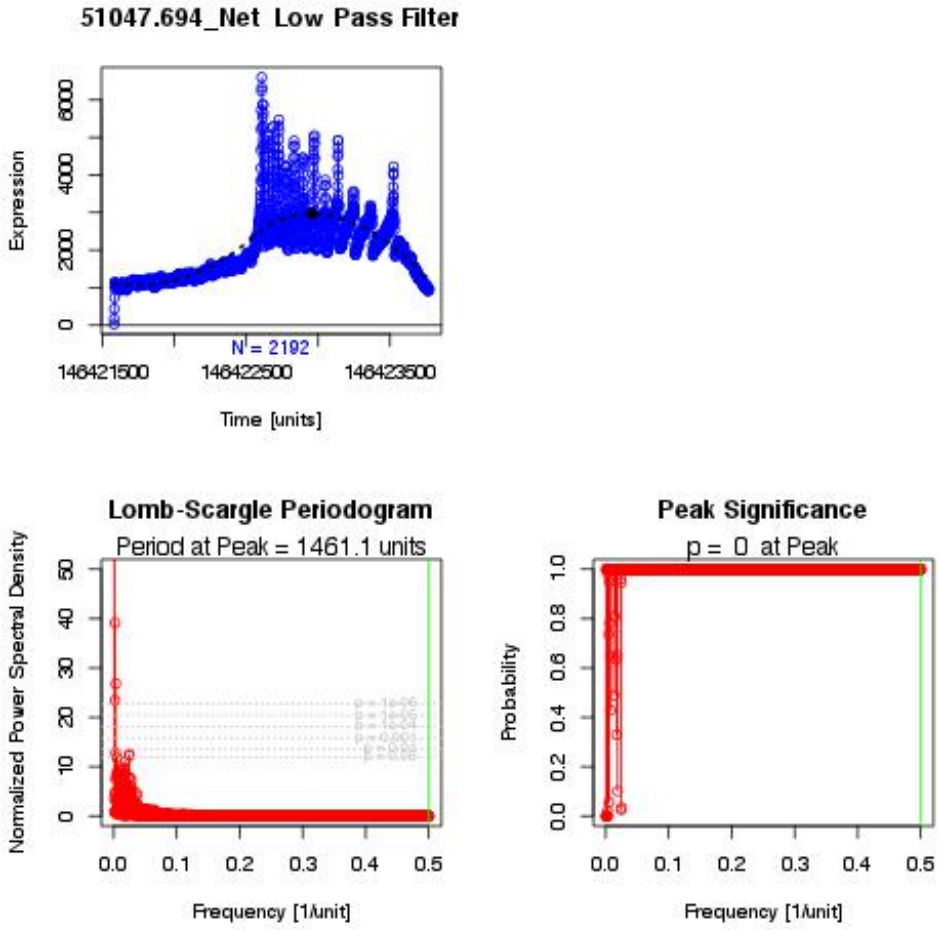


Figura 3.8: Ejemplo Lomb-Scargle tras filtro paso bajo

En el ejemplo que muestra la figura 3.8 se observa como tras aplicar un filtro paso bajo el resultado es similar al original, ya que el filtro lo que hace es dejar pasar las frecuencias bajas, y estas las detectadas en la ejecución original.

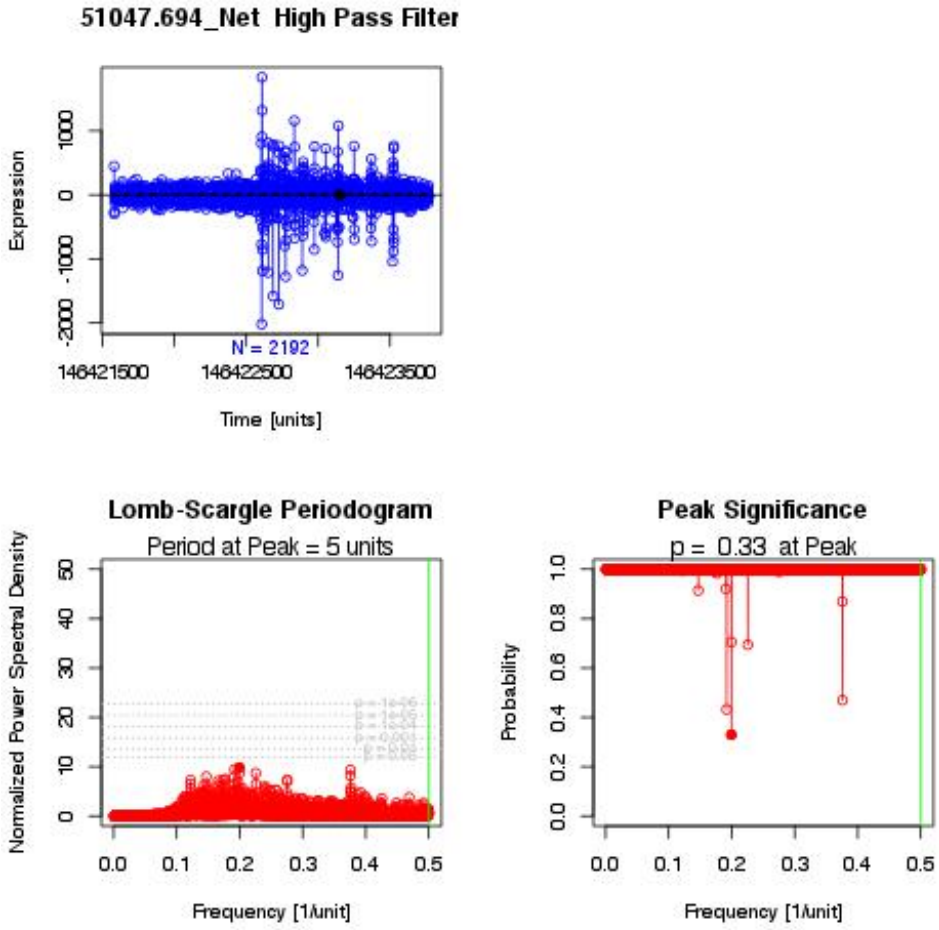


Figura 3.9: Ejemplo Lomb-Scargle tras filtro paso alto

En caso de aplicar un filtro paso alto en la misma muestra, desaparecen todas las frecuencias significativas, la figura 3.9 muestra esta situación.

Como se observa en el ejemplo el resultado es similar a Lomb-Scargle sin aplicación de filtro. Explorando el resto de las muestras nos encontramos con similares situaciones. Esto confirma en cierto modo que los resultados iniciales son correctos.

Resumen estadístico. Con ello pretendemos realizar una descripción de la forma de la señal y los colores, y de cómo se distribuyen [Dodge and Rousson, 1997]. Esta descripción se aplica a Total Light Curve, HR1 y HR2.

Se extraen los siguientes datos estadísticos:

- **Media.**
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.18}$$

- **Desviación media.**
$$D_m = \frac{1}{N}\sum_{i=1}^{n} |x_i - \overline{x}| \tag{3.19}$$

- **Varianza.**
$$S^2 = \frac{1}{N}\sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{3.20}$$

- **Desviación típica.**
$$S = \sqrt{S^2} \tag{3.21}$$

- **Mediana.**
$$M_e = \left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right)^2 \tag{3.22}$$

- **Desviación Absoluta Mediana** (MAD en inglés)
$$MAD = M_i \left(|x_i - M_j(x_j)|\right) \tag{3.23}$$

- **Cuantiles.** Que dividen a la distribución en cuatro partes (corresponden a los cuantiles 0'25, 0'50 y 0'75. Siendo 0'50 igual a la mediana.

Por otro lado se buscan medidas que describan la forma de la señal:

- **Curtosis (k).** Es una medida de la forma [Dodge, 2003]. Cuanto más elevado sea, más picuda es la distribución. De tal forma que puede ser.
  - **Leptocúrtica** $k > 3$. Más apuntada y con colas más anchas que la normal.
  - **Mesocúrtica** $k = 3$. La distribución normal.
  - **Platicúrtica** $k < 3$. Menos apuntada y con colas menos anchas que la normal.

- **Skewness ($Skew$).** Es una medida de simetría [Kendall and Stuart, 1969].
  - **Simétrica** $Skew = 0$. La distribución tiene la cola más larga a la derecha.
  - **Asimétrica positiva** $Skew > 0$. La distribución tiene la cola más larga a la derecha.

- **Asimétrica negativa** $Skew < 0$. La distribución tiene la cola más larga a la izquierda.

- **Jarque-Bera.** Es una prueba de normalidad que conjuga Curtosis y Skewness. A mayor valor más alejado nos encontramos de la normalidad [Jarque and Bera, 1980].

$$JB = \frac{n}{6}\left[Skew^2 + \frac{1}{4}(k-3)^2\right] \tag{3.24}$$

  Siendo $Skew = Skewness$ y $K = Curtosis$.

Mostramos a modo gráfico un ejemplo de la extracción de los atributos estadísticos donde se muestra:

- **Gráfico de caja**. Donde se muestra los cuantiles. La caja representa el 50 % de las muestras, la línea central la mediana, cada uno de los brazos representa el 25 % y los puntos fuera de los brazos representan los valores atípicos.

- **Histograma de la señal**. Representa la distribución de los valores que toma la señal.

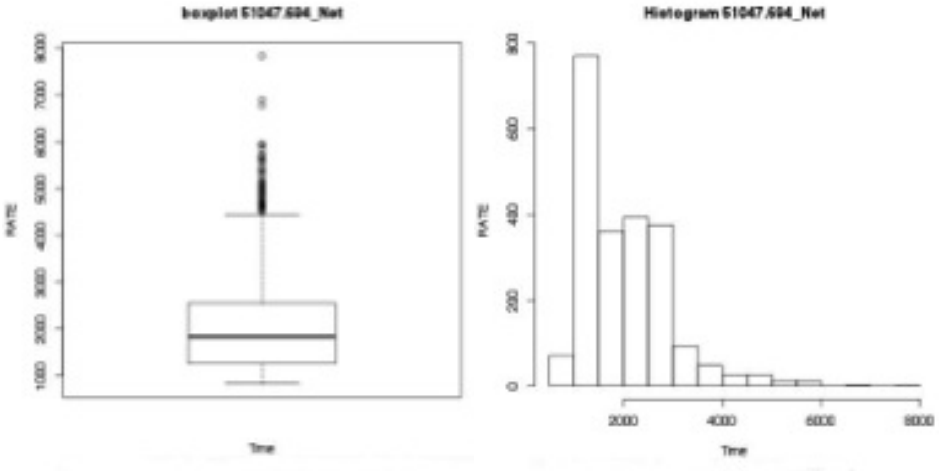


Figura 3.10: Ejemplo resultados estadísticos

En los gráficos se puede apreciar como los valores de la señal contiene más valores bajos. Sigue una distribución leptocúrtica $k = 6{,}26$ con un skewness positivo $skew = 1,433$ que indica una cola larga a la derecha.

3.2.2. Limpieza de datos

Clasificación dudosa. En realidad no es un problema de calidad de los datos, sino que existen muestras cuya clasificación resulta ambigua al poseer las muestras características que la sitúan en varias clases, de tal forma que $P(D|h)$ es baja, es decir la probabilidad que la muestra D pertenezca a la clase y se cumpla la hipótesis h es baja.

Esto no afectara a la agrupación ya que la clase a la que pertenece la muestra no se tiene en cuenta. Si puede afectar a la selección de atributos, y al análisis de resultados de la agrupación. El modelo predictivo se genera a partir de los resultados obtenidos en la agrupación, por lo que tampoco se verá afectado.

Para solucionar esto creamos tres conjuntos de entrenamiento, cada uno será usado según necesidades.

- **Conjunto 1**. Muestras sin dudas en su clasificación. Cuenta con 808 muestras.
- **Conjunto 2**. Muestras con dudas y sin dudas en su clasificación. Cuenta con 1078 muestras.
- **Conjunto 3**. Todas las muestras, con dudas en su clasificación, sin dudas en su clasificación y sin clasificar. Cuanta con 1912 muestras.

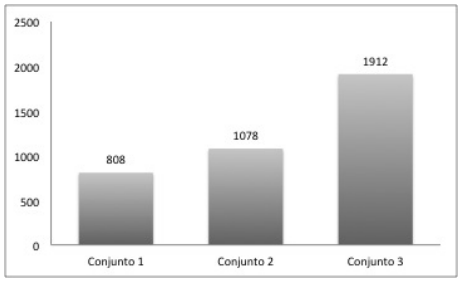


Figura 3.11: Comparativo conjuntos iniciales

Muestras incompletas. Se ha detectado muestras para las que no se cuenta con todos los ficheros, en algunos casos faltan alguno de los colores que componen la señal. En estos casos se ha optado por la eliminación del modelo de dichas muestras. El resultado para cada uno de los conjuntos anteriores es:

- **Conjunto 1**. 682 muestras.
- **Conjunto 2**. 871 muestras.
- **Conjunto 3**. 1629 muestras.

El siguiente gráfico muestra la comparativa de los conjuntos tras eliminar las muestras no completas.

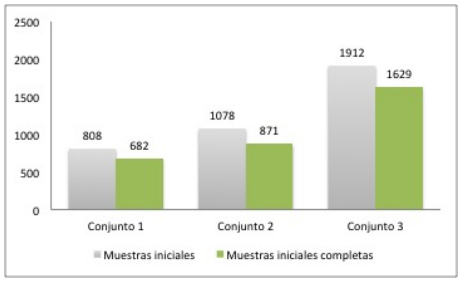


Figura 3.12: Comparativo conjuntos iniciales completos

Cortes en la señal. Para solucionar las muestras con cortes no realizamos ninguna acción concreta. Los métodos utilizados para la extracción de los datos trabajan bien con cortes en la señal [Birney et al., 2006].

Variabilidad en la longitud. Por último, teníamos el problema de la variabilidad de la longitud de los tiempos en la duración de cada muestra. Para solventar esto estandarizaremos la longitud de las muestras, siendo necesario definir dos parámetros.

- **Longitud**. Que longitud o ventana de tiempo que puede tener la muestra, siendo lo suficiente pequeña como para no perder estados intermedios, y lo suficiente grande como para no perder frecuencias significativas.
- **Desplazamiento**. Cada cuanto se selecciona el punto de corte en la muestra original.

Para abordar esto es necesario el análisis de la información que contienen las muestras originales. Al no ser un problema de la calidad de los datos, sino más bien un problema de presentación, lo abordaremos más adelante en el formateo de los datos.

3.2.3. Selección de los datos

En este punto nos encontramos tenemos un conjunto grande de datos, es necesario realizar una selección de cuáles de ellos resultan relevantes y cuáles no.

En la selección manual de atributos. Existen atributos para los que es necesario la colaboración de un experto en la materia si se desean eliminar mediante el análisis, estos atributos no se han eliminado. Otros en cambio no necesitan la colaboración de dicho experto, los cuales si han sido eliminados [Hernández Orallo et al., 2004] [Liu and Dash, 1997].

- Identificadores de la muestra.
- Atributos relacionados con la longitud de la muestra. No nos interesa que la longitud se tenga en cuenta en el agrupamiento.
- Atributos generados para el muestreo de señal, atributos que podemos variar de forma arbitraria y que solo darán errores en el agrupamiento. Ejemplo: Frecuencia máxima y frecuencia mínima.

Tras la selección manual de atributos recurrimos a métodos de selección y evaluación. Son métodos que permiten medir lo significativo que resultan los atributos para la clasificación de manera automática. Para poder ejecutarlo necesitamos que sean muestras sin atributos nulos, en nuestro caso se da cuando contienen todos los ficheros, y que sean muestras clasificadas.

Se han utilizado los dos conjuntos con las muestras clasificadas.

- **Conjunto 1**. Muestras sin dudas en su clasificación.
- **Conjunto 2**. Muestras con dudas y sin dudas en su clasificación.

Métodos de selección. Nos proporcionan un conjunto básico de atributos significativos, dentro de las opciones existentes se han seleccionado lo siguiente métodos, por comportarse bien con atributos continuos.

- **Selección de Atributos por Correlación** o CFS (*Correlation Feature Selection*), en la tabla *CfsSubsetEval*. Genera un conjunto de atributos óptimos basándose en la capacidad individual de cada uno de ellos y de su grado de redundancia. El método de búsqueda utilizado es *BestFirst*, que busca el espacio de subconjuntos de atributos por hillclimming codicioso.

$$CFS = \max_{s_k} \left[\frac{r_{cf1} + r_{cf2} + ... + r_{cfk}}{\sqrt{k + 2\left(r_{f1f2} + ... + r_{fifj} + ... + r_{fkf1}\right)}} \right] \tag{3.25}$$

  Siendo $\overline{r_c f}$ la media de todas las correlaciones de los atributos-clasificación, y $\overline{r_f f}$ la media de las correlaciones atributo-atributo.

  Esto nos proporcionara que atributos tiene mayor capacidad de clasificar por sí mismo, calculando la correlación de la clase con cada atributo, y eliminando tanto atributos que tengan una correlación muy alta como atributos redundantes.

- **Métodos basados en modelo** En la tabla *WrapperSubsetEval*. Evalúa conjuntos de atributos usando un esquema de aprendizaje. Esto nos proporcionara un conjunto óptimo de atributos evaluándolo en base a la ejecución de métodos de aprendizaje. Lógicamente esto resulta más costoso. En nuestro caso se han seleccionado dos, cuyo comportamiento con variables continuas es correcto.

- **Árbol C4.5**. En la tabla *J48* [Quinlan, 1993], sin poda, es decir no controlamos el sobreajuste. En C4.5 en cada nodo se elige un atributo de los datos que más eficientemente divida el conjunto de muestras. El criterio para esta división es la ganancia de información (diferencia de entropía) que resulta de la elección de un atributo para dividir los datos.
- **Redes bayesianas**. En la tabla *Bayes Net*. Basado en el teorema de Bayes.

$$P(h|O) = \frac{P(O|h)P(h)}{P(O)} \tag{3.26}$$

  Donde $P(h)$ es la probabilidad de que se la hipótesis, $P(O)$ la probabilidad de las observaciones, y $P(h|O)$ y $P(O|h)$ las probabilidades condicionales.

  En las redes bayesianas se crea una red, donde cada uno de los nodos tiene una probabilidad condicional a los padres de que se cumpla. Para generar esta red se ha usado un algoritmo K2, el cual utiliza un esquema voraz en su búsqueda de soluciones candidatas cada vez mejores, y parte de que las variables de entrada estén ordenadas.

  K2, siendo X los nodos o variables ordenados y D los datos [Hernández Orallo et al., 2004].

```
# Fase de inicialización
    for i:= to n
       set all X[i]=0.0
# Fase Iterativa
for i:= to n
    ok:=true
    do while ok
        sea X[j] el nodo tal que j<I y X[j] no pertenezca a
        Pa(X[i]) que maximiza f[i](X[i]|Pa(X[i] U X[j] D)

        if f[j](X[i]|Pa(X[i]) U X[j]D)>f[i](Xi|Pa(Xi):D)
            Pa(X[i]=Pa(X[i])U X[j]
       else
            ok=false
```

Métodos de evaluación. Nos proporcionan un ranking de atributos de más a menos significativos, complementando la información de los métodos anteriores.

- **ReliefF**. En la tabla *ReliefFAttribute*. Evalúa el valor de un atributo por muestreo repetitivo de una instancia y teniendo en cuenta el valor del atributo dado por el más cercano ejemplo de la misma clase y el más cercano de diferente clase [Kononenko, 1994].

Es una variación de *Relief* [Kira K., 1992]

```
set all weigths W[A]:=0.0

for i:=1 to m do
    begin
        randomly select an instance R;
        find nearest hit H and nearest miss M;
        for A:=1 to all_attributes do
            W[A]:=W[A]- diff(A,R,H)/m + diff(A,R,M)/m
    end
```

En el caso de ReliefF

$$W[A] = W[A] - \frac{diff(A,R,H)}{m} + \sum_{c \neq class(R)} \frac{[P(C)diff(A,R,M(C))]}{m} \tag{3.27}$$

- **Ganancia de Información** o IG (Information Gain). En la tabla *InfoGainAttributeEval*. Evalúa el valor de un atributo mediante la medición de ganancia de información con respecto a la clase [Irani and Fayyad, 1993]. Dado un conjunto de entrenamiento S_x y un vector con i variables x_i. El porcentaje de ejemplos que representa una variable i es $|S_{x_{i=v}}|/|S_x|$ dado el valor v.

$$IG(S_x, x_i) = H(S_x) - \sum_{v=values(x_i)}^{|S_{x_i=v}|} H(S_{x_i=v}) \tag{3.28}$$

con entropía

$$H(S) = -p_+(S)\log_2 p_+(s) - p_-(S)\log_2 p_-(S) \tag{3.29}$$

$p \pm (S)$ es la probabilidad de que el ejemplo de entrenamiento S sea positivo/negativo. Las variables continuas se discretizan.

- **Ratio de Ganancia** o GR (Gain Ratio). En la tabla *GainRatioAttributeEval*. Evalúa el valor de un atributo mediante la relación de ganancia con respecto a la clase. Elimina atributos redundantes. Basándonos en IG la evaluación de cada atributo se realiza con la formula.

$$GR(S_x, x_i) = \frac{H(S_x) - H(S_x|x_i)}{H(x_i)} \tag{3.30}$$

En todos los métodos de evaluación se ha utilizado *Rank* como método de búsqueda, el cual evalúa los atributos de manera individual.

Selección de atributos. Las siguientes tablas muestran el resumen de los resultados obtenidos con los métodos de selección y evaluación teniendo como entrada las tablas 3.2, 3.3 y 3.4.

ID	Atributos	Descripción
1	PeakSPD	Valor de Lomb-Scargle de la frecuencía seleccionada.
2	PeakPeriod	Periodo de la frecuencia seleccionada.
3	PeakPvalue	Probabilidad de falsa alarma de la frecuencia seleccionada.
4	SignFreq2	Número de frecuencias con probabilidad de falsa alarma menor de 0,2.
5	SignFreq1	Número de frecuencias con probabilidad de falsa alarma menor de 0,1.
6	Mean	Media de la señal.
7	MeanDeviation	Desviación media de la señal.
8	Variance	Varianza de la señal.
9	Deviation	Desviación tipica de la señal.
10	Median	Median de la señal.
11	Mad	Desviación Absoluta Mediana de la señal.
12	Quantile000	Cuantiles 0,00 de la señal.
13	Quantile025	Cuantiles 0,25 de la señal.
14	Quantile050	Cuantiles 0,50 de la señal.
15	Quantile075	Cuantiles 0,75 de la señal.
16	Quantile100	Cuantiles 1 de la señal.
17	highCut5	Corte cinco veces por encima de MAD.
18	lowCut5	Corte cinco veces por debajo de MAD.
19	highCut2	Corte dos veces por encima de MAD.
20	lowCut2	Corte dos veces por debajo de MAD.
21	Num5MAD	Puntos de la muestra que salen del corte 17 y 18.
22	Num2MAD	Puntos de la muestra que salen del corte 19 y 20.
23	kurtosis	Curtosis de la señal.
24	skewness	Skewness de la señal.
25	jarque.statistic	Jarque-Bera de la señal.
26	jarque.p.value	Jarque-Bera valor probabilistico de la señal.
27	PeakAmplitude	Amplitud de la frecuencia seleccionada por Lomb-Scargle.
28	PeakPhaseShift	Fase de la frecuencia seleccionada por Lomb-Scargle.

Cuadro 3.2: Tabla de atributos evaluados de la señal total

ID	Atributos	Descripción
29	HR1.Mean	Media de HR1.
30	HR1.MeanDeviation	Desviación media de HR1.
31	HR1.Variance	Varianza de HR1.
32	HR1.Deviation	Desviación tipica de HR1.
33	HR1.Median	Median de HR1.
34	HR1.Mad	Desviación Absoluta Mediana de HR1.
35	HR1.Quantile000	Cuantiles 0,00 de HR1.
36	HR1.Quantile025	Cuantiles 0,25 de HR1.
37	HR1.Quantile050	Cuantiles 0,50 de HR1.
38	HR1.Quantile075	Cuantiles 0,75 de HR1.
39	HR1.Quantile100	Cuantiles 1 de HR1.
40	HR1.kurtosis	Curtosis de HR1.
41	HR1.skewness	Skewness de HR1.
42	HR1.jarque.statistic	Jarque-Bera de HR1.
43	HR1.jarque.p.value	Jarque-Bera valor probabilístico de HR1.
44	HR2.Mean	Media de HR1.
45	HR2.MeanDeviation	Desviación media de HR2.
46	HR2.Variance	Varianza de HR2.
47	HR2.Deviation	Desviación tipica de HR2.
48	HR2.Median	Median de HR2.
49	HR2.Mad	Desviación Absoluta Mediana de HR2.
50	HR2.Quantile000	Cuantiles 0,00 de HR2.
51	HR2.Quantile025	Cuantiles 0,25 de HR2.
52	HR2.Quantile050	Cuantiles 0,50 de HR2.
53	HR2.Quantile075	Cuantiles 0,75 de HR2.
54	HR2.Quantile100	Cuantiles 1 de HR2.
55	HR2.kurtosis	Curtosis de HR2.
56	HR2.skewness	Skewness de HR2.
57	HR2.jarque.statistic	Jarque-Bera de HR2.
58	HR2.jarque.p.value	Jarque-Bera valor probabilistico de HR2.

Cuadro 3.3: Tabla de atributos evaluados de Hardness Ratio

ID	Atributos	Descripción
59	PeakPeriod.low	Periodo de la frecuencia seleccionada tras aplicar un filtro paso bajo.
60	SignFreq2.low	Número de frecuencias con probabilidad de falsa alarma menor de 0,2 tras aplicar un filtro paso bajo.
61	SignFreq1.low	Número de frecuencias con probabilidad de falsa alarma menor de 0,1 tras aplicar un filtro paso bajo.
62	SignFreq2.low.boolean	Indica si hay frecuencias con probabilidad de falsa alarma menor de 0,2 tras aplicar un filtro paso bajo.
63	SignFreq1.low.boolean	Indica si hay frecuencias con probabilidad de falsa alarma menor de 0,1 tras aplicar un filtro paso bajo.
64	PeakPeriod.high	Periodo de la frecuencia seleccionada tras aplicar un filtro paso alto.
65	PeakPvalue.high	Probabilidad de falsa alarma de la frecuencia seleccionada tras aplicar un filtro paso alto.
66	SignFreq2.high	Número de frecuencias con probabilidad de falsa alarma menor de 0,2 tras aplicar un filtro paso alto.
67	SignFreq1.high	Número de frecuencias con probabilidad de falsa alarma menor de 0,1 tras aplicar un filtro paso alto.
68	SignFreq2.high.boolean	Indica si hay frecuencias con probabilidad de falsa alarma menor de 0,2 tras aplicar un filtro paso alto.
69	SignFreq1.high.boolean	Indica si hay frecuencias con probabilidad de falsa alarma menor de 0,1 tras aplicar un filtro paso alto.
70	CLASS	Clasificación de la muestra.

Cuadro 3.4: Tabla de atributos evaluados tras aplicación de filtros

Conjunto 1. Muestras clasificadas no dudosas.

Métodos Selección	Filtro/Modelo	Opciones	Estrategia	Atributos Seleccionados.
CfsSubSetEval	Filtro	Valores Por defecto	BestFirst	2,6,8,27,29,30,31,35,36,39,44,45,46 ,48 : 14
		J48	BestFirst	8,27,31,36 : 4
WrapperSubSetEval	Modelo	Bayes Net	BestFirst	7,10,12,14,15,16,19,20,27,32,36,41 : 12
Métodos Evaluación	Filtro/Modelo		Estrategia	Orden de los atributos
ReliefFAttributeEval	Filtro		Ranker	9,16,7,47,12,13,6,45,14,10,15,19,27 ,17,11,20,32,8,30,18,46,38,36,33,37 ,41,24,43,31,56,69,23,40,68,22,59,1 ,65,2,26,58,49,50,44,48,53,54,51,52 ,35,29,34,39,55,21,42,25,60,61,3,4, 5,57,66,67,64,62,63,28 : 69
InfoGainAttribute	Filtro		Ranker	51,50,54,52,53,35,29,48,39,44,9,8,7 ,30,18,20,6,46,16,45,15,14,10,47,13 ,31,32,36,41,11,37,33,38,42,43,27,1 7,19,25,24,26,12,58,57,21,23,56,2,5 9,1,5,3,4,64,63,65,49,60,55,62,61,6 9,34,22,28,40,66,68,67 : 69
GainRatioAttributeEval	Filtro		Ranker	32,31,19,17,30,27,43,42,11,38,16,4 6,45,36,47,8,9,6,37,33,18,7,13,15,1 0,14,25,2,59,12,24,41,26,20,29,53,4 8,51,52,50,39,35,44,54,56,21,23,57, 58,67,64,66,65,3,1,69,68,4,5,40,55, 49,28,22,62,63,34,61,60 : 69

Cuadro 3.5: Resultado Selección y Evaluación Conjunto 1

Conjunto 2. Todas las muestras clasificadas.

Métodos Selección	Filtro/Modelo	Opciones	Estrategia de b	Atributos Seleccionados.
CfsSubSetEval	Filtro	Valores Por defecto	BestFirst	1,2,7,8,11,12,16,18,19,20,23,24,27, 29,30,31,32,36,38,40,41,45,47,56,5 9,66 : 26
		J48	BestFirst	3,7,11,36,38,47,58 : 7
WrapperSubSetEval	Modelo	Bayes Net	BestFirst	2,16,20,24,32,36,41,45,69 : 9
Métodos Evaluación	Filtro/Modelo		Estrategia de b	Orden de los atributos
ReliefFAttributeEval	Filtro		Ranker	16,9,7,69,15,45,27,19,6,68,14,10,17 ,47,13,12,36,33,37,11,3,38,8,41,18, 20,32,24,30,56,46,65,23,43,40,1,22, 2,31,59,55,26,21,28,25,58,42,4,5,60 ,61,57,66,67,64,51,49,50,54,52,53,3 5,29,34,48,39,44,63,62 : 69
InfoGainAttribute	Filtro		Ranker	51,50,54,52,53,35,29,48,39,44,8,9,1 1,7,16,17,19,18,38,15,31,32,27,36,6 ,37,33,30,14,10,41,24,47,46,45,20,1 ,13,2,56,5,4,59,12,23,3,40,25,21,42, 55,57,26,43,67,65,66,22,58,69,68,6 4,61,62,63,49,34,28,60 : 69
GainRatioAttributeEval	Filtro		Ranker	16,7,3,11,8,9,27,18,19,20,15,17,30, 5,4,10,14,6,1,56,31,32,45,55,66,24, 41,13,2,43,12,37,33,38,47,46,59,42, 36,40,21,26,50,44,48,53,54,51,52,2 9,39,35,23,58,25,65,57,67,22,69,68, 64,61,28,49,34,62,60,63 : 69

Cuadro 3.6: Resultado Selección y Evaluación Conjunto 2

Como se puede apreciar en el **conjunto 1** los atributos estadísticos resultan más relevantes. Sin embargo cuanto tenemos en cuenta todas las muestras clasificadas, en el **conjunto 2**, los atributos más relevantes son los derivados las frecuencias significativas.

Algunos atributos no aparecen en ninguna selección y aparecen en los últimos puestos de la evaluación. En el caso de la fase (*PeakPhaseShift*) resulta no relevante, sin embargo lo dejamos por ser de utilidad en las estrellas variables.

En el caso de atributos extraídos tras la aplicación de filtros paso alto y bajo no son relevantes, para no eliminar totalmente toda la información nos limitaremos a indicar la existencia o no de frecuencias altas y frecuencias bajas.

Los atributos estadísticos son relevantes, pero resulta difícil saber cuál es más relevante. Según el método utilizado es más relevante uno u otro. Por ello optamos por un conjunto que describa distintos elementos de la señal.

- **Intensidad**. Para medir la intensidad de la señal seleccionamos **mediana**.
- **Variabilidad**. Utilizamos una medida robusta de la varianza como es la **desviación mediana absoluta** o **MAD**.
- **Distribución**. Para describir la distribución utilizamos la **curtosis** y **skewness**.

ID	Atributos	Descripción
2	PeakPeriod	Periodo de la frecuencia seleccionada.
3	PeakPvalue	Probabilidad de falsa alarma de la frecuencia seleccionada.
5	SignFreq1	Número de frecuencias con probabilidad de falsa alarma menor de 0,1.
10	Median	Median de la señal.
11	Mad	Desviación Absoluta Mediana de la señal.
23	kurtosis	Curtosis de la señal.
24	skewness	Skewness de la señal.
25	jarque.statistic	Jarque-Bera de la señal.
27	PeakAmplitude	Amplitud de la frecuencia seleccionada por Lomb-Scargle.
28	PeakPhaseShift	Fase de la frecuencia seleccionada por Lomb-Scargle.
33	HR1.Median	Mediana de HR1.
34	HR1.Mad	Desviación Absoluta Mediana de HR1.
40	HR1.kurtosis	Curtosis de HR1.
41	HR1.skewness	Skewness de HR1.
42	HR1.jarque.statistic	Jarque-Bera de HR1.
48	HR2.Median	Mediana de HR2.
49	HR2.Mad	Desviación Absoluta Mediana de HR2.
55	HR2.kurtosis	Curtosis de HR2.
56	HR2.skewness	Skewness de HR2.
57	HR2.jarque.statistic	Jarque-Bera de HR2.
63	SignFreq1.low.boolean	Indica si hay frecuencias con probabilidad de falsa alarma menor de 0,1 tras aplicar un filtro paso bajo.
69	SignFreq1.high.boolean	Indica si hay frecuencias con probabilidad de falsa alarma menor de 0,1 tras aplicar un filtro paso alto.
70	CLASS	Clasificación de la muestra.

Cuadro 3.7: Tabla de atributos seleccionados

En los próximos capítulos veremos como este conjunto de atributos iniciales sufrirán modificaciones.

Además de esta selección principal, se han realizado ejecuciones alternativas con otras selecciones de atributos, aunque sin profundizar en el análisis. En total tenemos las siguientes selecciones de atributos:

1. **Selección de atributos 1.** *PeakPeriod, PeakPvalue, SignFreq1, Median, Mad, kurtosis, skewness, PeakAmplitude, PeakPhaseShift, HR1.Median, HR1.Mad, HR1.kurtosis, HR1.skewness, HR2.Median, HR2.Mad, HR2.kurtosis y HR2.skewness.* Eliminando atributos derivados de aplicar filtros y Jarque-Bera.

2. **Selección de atributos 2.** *PeakPeriod, PeakPvalue, SignFreq1, Median, Mad, kurtosis, skewness, PeakAmplitude, HR1.Median, HR1.Mad, HR1.kurtosis, HR1.skewness, HR2.Median, HR2.Mad, HR2.kurtosis y HR2.skewness.* Eliminando fase.

3. **Selección de atributos 3.** *PeakPeriod, PeakPvalue, SignFreq1, Median, Mad, kurtosis, skewness y PeakAmplitude.* Sobre la selección 2, eliminando atributos de Hardness Ratio

4. **Selección de atributos 4.** textitHR1.Median, HR1.Mad, HR1.kurtosis, HR1.skewness, HR2.Median, HR2.Mad, HR2.kurtosis y HR2.skewness. Sobre la selección 2, eliminando atributos de la señal.

3.2.4. Formateo de los datos

Los datos originales, curvas de luz, no son utilizables directamente. Los atributos derivados, tanto extraídos a partir de la búsqueda de frecuencias significativas como los estadísticos, son atributos continuos. Hemos eliminado las muestras que contenían algún atributo nulo, y nos quedamos con un subconjunto de los atributos, los atributos más significativos. Pero no es necesario ningún formateo de los datos.

Es por ello que el apartado se centra exclusivamente en el formateo de la longitud de la señal. Siendo el objetivo del apartado decidir que **longitud** tomamos como estándar, y con qué **desplazamiento** tomamos las muestras segmentadas.

Análisis de frecuencias. Las frecuencias detectadas se pueden ver afectadas si segmentamos las muestras. Es necesario una longitud mínima para la detección de un periodo, si segmentamos por debajo de esta longitud mínima no será posible detectar el periodo.

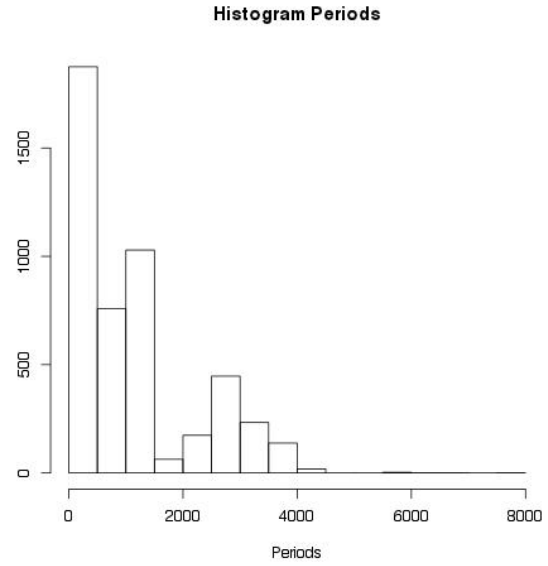


Figura 3.13: Histogramas de los periodos de las muestras significativas

Como se observa en el histograma de periodos, la mayor parte de las frecuencias están por debajo de periodos de 5000 segundos.

- **Frecuencias alta**. La frecuencia más alta se corresponde con la muestra *53227,947*. Tiene una $longitud = 279 segundos$ y un $PeakPeriod = 139,5 segundos$, siendo $PeakPeriod$ el periodo de la frecuencia más significativa encontrada. Este tipo de frecuencias no se ven afectadas al segmentar.

- **Frecuencias bajas**. La mayoría de las frecuencias no superan el periodo de 5000 segundos, por ejemplo la muestra *52364,609* tiene una $longitud = 16300 segundos$, y un $PeakPeriod = 5220,888$. La gran longitud con respecto al periodo puede indicarnos que la muestra puede ser segmentada en ventanas de tiempo más pequeñas sin perder frecuencias significativas.

- **Frecuencias espurias**. También podemos encontrarnos frecuencias que aparentemente son significativas, pero que si las analizamos son claramente espurias, estos son frecuencias donde se cumple $longitud < PeakPeriod$ aunque $P_r < 0,1$.

Por otro lado las frecuencias significativas parecen estar aglutinadas en el mismo espacio.

Análisis de segmentación. A la hora de segmentar las muestras en muestras más pequeñas se pierden frecuencias bajas significativas.

Una cosa que vemos en la ejecución es que la perdida de frecuencias bajas no implica necesariamente una pérdida de información equivalente. Que no se detecte una frecuencia baja, no implica que no se detecte una frecuencia baja cercana. La pérdida de información es difícil de medir sin realizar la agrupación, si se puede medir la perdida de las frecuencias más significativas.

En este apartado el objetivo es llegar a un equilibrio entre segmentos lo más pequeños posibles, que faciliten encontrar estados intermedios, y lo suficientemente grandes como para no perder frecuencias bajas, con periodos muy largos. Esto se hace variando la **longitud** de las muestras. Cambiando el **desplazamiento** no afecta a la pérdida de frecuencias, si afecta a la precisión con la que se buscan estados intermedios, como inconveniente se tiene que a mayor precisión mayor tiempo de cálculo es necesario.

Para ello se realizamos el análisis de frecuencias con distintas parametrizaciones en la segmentación. Variando la longitud hasta 200 segundos. En el desplazamiento hemos mantenido el criterio de que sea la mitad de la longitud, de tal forma que la segunda mitad de una muestra segmentada corresponda con la primera mitad de la siguiente muestra. Por ejemplo dividiendo la muestra inicial en segmentos de 4000 segundos, la primera muestra segmentada ira de 1 a 4000 segundos, la segunda muestra segmentada de 2000 a 6000 segundos.

		N.ALL		PeakPeriod			
Segmentos	Num Muestras	Max	Min	Max	Min	> 4000	> 1000
NA	1910	16300	279	19088.92	2.00	134	506
8000s	2100	8000	279	12779.74	2.00	180	649
4000s	2781	4000	279	10603.79	2.00	99	1032
2000s	4753	2000	279	1232448.00	2.00	11	1392
1000s	9756	1000	279	828125.00	2.00	3	523
400s	25671	400	186	415625.00	2.00	4	284
200s	53231	200	81	405848.00	1.99	10	229

Cuadro 3.8: Resumen pérdidas de frecuencia en la segmentación

La tabla muestra la siguiente información.

- **Segmentos**. Longitud con la que se ha realizado la división de las muestras.
- **Num Muestras**. Numeró de muestras que se analizan tras la segmentación.
- **N.ALL Max** y **N.ALL MIN**. Longitud de tiempo de la muestra más larga y más corta respectivamente.

- **PeakPeriod Max** y **PeakPeriod Min**. Periodo de la frecuencia más significativa más baja, y periodo de la frecuencia más significativa más alta respectivamente.
- $>$ **4000**. Número de frecuencias significativas con un periodo superior a 4000 segundos.
- $>$ **1000**. Número de frecuencias significativas con un periodo superior a 1000 segundos.

Se confirma que las frecuencias altas no se ven afectadas manteniéndose estable. Las frecuencias bajas si se van perdiendo a medida que disminuimos la longitud de tiempo. También existen frecuencias claramente espurias al ser mayor el periodo de la frecuencia que la propia muestra.

Al segmentar las muestras originales en muestras más pequeñas el número aumenta, al igual que el número de frecuencias significativas de un periodo. Esto ocurre hasta que los segmentos son tan pequeños que no es posible detectar el periodo, en cuyo caso comienza a disminuir. Esto ocurre en el caso de periodos mayores de 4000 cuando el segmento es inferior a 8000 segundos. Estas frecuencias significativas son un porcentaje pequeño.

En el caso de las frecuencias con periodos mayores de 1000 segundos, comienza a disminuir al ser inferior la longitud de la muestra de 2000 segundos. El porcentaje de muestras con periodos superiores a 1000 segundos es del 25

Resultado Segmentación. En base a lo descrito, y viendo con la intención de mantener el mayor número de frecuencias significativas, inferiores a periodos de 1000 segundos, optamos por mantener **longitudes de 2000 segundos**.

Por otro lado, y al margen del análisis realizado, la clasificación de la biografía [Belloni et al., 2000], coincide que los datos relevantes se encuentran en periodos de entre pocos segundos y 2000 segundos.

En cuanto al desplazamiento del punto de corte de cada segmento no tiene repercusiones negativas en cuanto a la perdida de frecuencias significativas, pero si en la precisión de detectar estados intermedios. Tampoco nos resulta posible realizar desplazamientos muy pequeños al aumentar el tiempo de cálculo necesario para procesar el aumento en el número de muestras. Es por ello que hemos optado por un **desplazamiento de 1000 segundos**. Esto permitirá detectar las transiciones entre estados, sin aumentar mucho el coste de cálculo.

Con estos parámetros en la segmentación se vuelve a calcular los atributos seleccionados, que son los que utilizaremos a partir de ahora.

CAPÍTULO 4

Modelo de Agrupamiento

4.1. Selección del Método de Agrupamiento

El objetivo de este apartado es la selección del método de agrupamiento más adecuado a las necesidades de nuestro problema.

Requisitos. El problema tratado tiene una serie de requisitos que debe cumplir cualquier método de agrupamiento que vayamos a utilizar.

- **Manejar grandes volúmenes de dados**. Se cuenta con un conjunto de muestras elevado correspondiente a diez años de observaciones.
- **Manejar atributos continuos**. Los atributos seleccionados son continuos.
- **No requerir el número de grupos buscados**. No predeterminamos la búsqueda a un número concreto de grupos se obtiene más libertad a la hora de buscar patrones.

Métodos descartados. El tercer requisitos no se cumple por los siguientes métodos de agrupamiento [Bishop., 2006] [Hand and Kamber, 2006].

- **Mapas auto-organizativos de Kohonen**. Modelo basado en las redes neuronales, cuya capa final contiene tantos nodos como grupos buscados, por lo que es necesario predefinir el número de grupos buscados [Hernández Orallo et al., 2004].
- **K-Medias**, Donde se parte de un número de prototipos alrededor de donde se irán agrupando las muestras, formando los grupos, y por tanto también es necesario predefinir el número de grupos buscados [Hernández Orallo et al., 2004].

Métodos posibles. El **agrupamiento jerárquico** está compuesto por un conjunto de métodos de agrupamiento que si cumplen con el requisito de no predefinir el número de grupos. Se construye un árbol donde se considera que cada subnodo es un subconjunto del anterior nodo, hasta llegar a las hojas. Siendo el nodo raíz el grupo formado por todos los individuos, este se va separando sucesivamente hasta llegar a las hojas. La separación se hace disminuyendo la distancia que puede haber entre individuos. Se pueden construir de dos tipos: **Aglomerativos**, donde se comienza por las hojas hasta llegar a la raíz; o **divisivos** donde se comienza desde la raíz hasta llegar a las hojas.

La alternativa dentro del agrupamiento jerárquico que nos interesa es **COBWEB**, que aporta de lo ya dicho para el agrupamiento jerárquico, añadiendo que se comporta bien con grandes volúmenes de datos, al incorporar de manera incremental los ejemplos al dendograma [Hernández Orallo et al., 2004]. Ahora bien, se ha descartado finalmente porque:

- Trabaja con atributos discretos, y tenemos atributos continuos.
- Los atributos debe ser independientes, y no podemos asegurar que lo sean.
- No hay garantía de encontrar un mínimo local.
- Es sensible al orden en que se presentan las muestras.

Otra alternativa son, dentro de los métodos probabilísticos, **EM (Expectation Maximization)**, cumple con todos los requisitos, comportándose bien con grandes volúmenes de datos, uso de variables continuas, y no requiriendo indicar el número de grupos buscados. Como añadido incorpora una ventaja, que puede no ser un requisito exactamente, frente COWEB, es permisivo a los datos perdidos debido a que trabaja con funciones de densidad.

Conclusión. Es por este buen comportamiento ante perdidas de datos, y que cumple con todos los requisitos que nos **decantamos por el uso de EM** (Expectation Maximization) frente COBWEB, que sería la alternativa cercana más viable. Existencia de estudios previos que dejan mejor situado a EM frente a COWEB [Garre et al., 2005, Garre et al., 2007].

4.2. EM (Expectation Maximization)

El método EM es un método densidad o probabilístico, en este método se busca la función de densidad probabilística (FPD) desconocida a la que pertenece un conjunto complejos de datos. Esta FPD se aproxima mediante una combinación lineal de N componentes, definida a falta de una serie de parámetros $\{\theta\} = \cup\{\theta_j \forall_j = 1..N\}$, que hay que averiguar.

$$P(x) = \sum_{j=1}^{N} \pi_j P(x;\theta_j) \tag{4.1}$$

$$\sum_{j=1}^{N} \pi_j = 1 \tag{4.2}$$

Las probabilidades a priori de cada grupo (π_j) suman 1. Forman parte de la solución buscada, $P(x)$.

La función de densidad del componente j viene indicada por $P(x;\theta_j)$, representando lo que sería el grupo o conjunto de datos con características similares.

El algoritmo pasa por dos pasos.

1. **Expectation**. Utiliza los valores de los parámetros iniciales o proporcionados por el paso de Maximization de la iteración anterior, obteniendo diferentes formas de la FDP (función de densidad) buscada.

   Se pueden estimar FDP de formas arbitrarias, utilizándose FDP normales, n-dimensionales, t-Student, Bernoulli, Poisson y log-normales. A la hora de indicar que modelo de distribución puede seguir cada uno de los grupos, e imaginando un espacio tridimensional, hemos buscado una forma elipsoidal, variando el volumen, la forma y la orientación dentro de una función normal. Esto se traduce en que todos los atributos son linealmente independientes y se da total liberta a la hora de ajustar la función de densidad.

2. **Maximization**. Obtiene nuevos valores de los parámetros a partir de los datos proporcionados por el paso anterior.

   Likehood es el valor que indica el ajuste de los datos de la distribución sobre los datos que esta representa dicha distribución. Se trataría de buscar los parámetros θ_j que maximicen el *likehood* 4.3. Normalmente este cálculo se realiza sobre el $log-likehood$ al resultar más sencillo de calcular que hacerlo de forma analítica y la solución obtenida es la misma debido a la propiedad de monotonicidad del logaritmo.

$$L(\theta,\pi) = \log \prod_{n=1}^{NI} P(x_n) \tag{4.3}$$

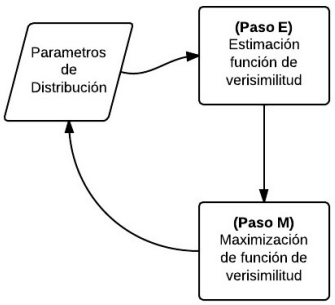


Figura 4.1: Proceso EM

Tras unas iteraciones el algoritmo tendera a obtener un máximo local de las funciones (Garre M., 2005).

Finalmente se obtienen el conjunto de grupos.

En nuestro caso la búsqueda de los parámetros óptimos de EM se realiza mediante un método de Criterio Bayesiano de Información (BIC). BIC es un criterio para la selección de un conjunto finito de modelos, el cual se basa en parte en la función de probabilidad, pero que solventa el problema del sobreajuste penalizando el número de parámetros en el modelo.

$$BIC = G - gl. \ln N \tag{4.4}$$

Siendo G el coeficiente de verosimilitud, gl los grados de libertad y N el tamaño de la muestra. Donde se escogería el modelo con BIC menor.

4.3. Ejecución del Método de Agrupamiento

Problema en la ejecución. Al ejecutar el algoritmo con la selección de atributos original, y el conjunto de datos totales, nos da un error "singular covariance", esto ocurre cuando alguno de los atributos es linealmente dependiente de otro atributo, y la matriz de covarianza no puede ser invertida. Como resultado no podríamos utilizar una forma elipsoidal, variando el volumen, la forma y la orientación para la búsqueda. En EM no es un requisito que los atributos sean linealmente independientes, pero si resulta muy recomendable [Bishop., 2006].

Se realizan varias ejecuciones añadiendo atributos hasta detectar cual es el que causa el problema. Para solucionarlo modificamos el conjunto inicial de atributos seleccionados, eliminando:

- **jarque.statistic**. Tanto los de la señal como los de los colores son combinaciones de la curtosis y skewness.
- **SignFreq1.low.boolean** y **SignFreq1.high.boolean**. Las cuales tampoco habían sido seleccionados por ningún método de selección o evaluación de atributos.

Quedando la selección de atributos tal como se describe en el conjunto de atributos seleccionados 1 del aparatado . Como añadido se han hecho ejecuciones alternativas con los conjuntos de atributos 2, 3 y 4 del mismo apartado.

Resumen de los resultados Para la ejecución principal se obtienen un total de 12 grupos distintos, analizados en el apartado 4.4.4.

Aunque el resultado coincida con el número de clases propuestas en el artículo [Belloni et al., 2000], esto parece mera coincidencia. Por ejemplo debería existir un grupo con un número elevado de muestras, donde quedase representada la clase χ. Siendo la distribución obtenida uniforme si exceptuamos el grupo 5, como se puede apreciar en la figura 4.2.

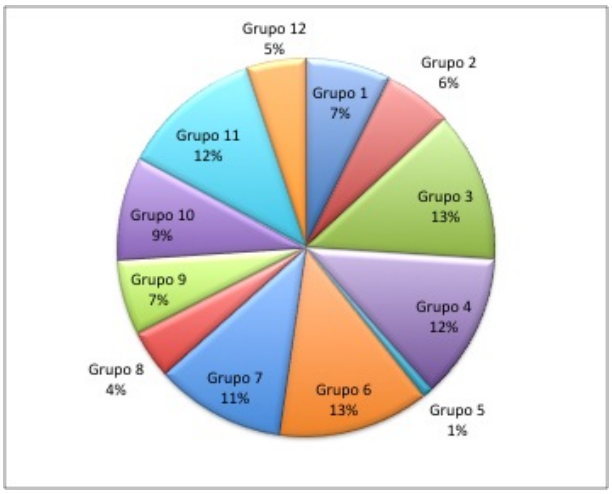


Figura 4.2: Distribución Grupos por Conjunto de Atributos 1

Las medias de cada atributo nos indica el punto medio dentro del grupo. Con la variabilidad dentro del grupo podemos ver cuánto de significativo resulta dicho valor concreto. A mayor dispersión más amplio será el rango de valores para dicho atributo.

Por otro lado consideramos que las muestras cuyos atributos contengan valores próximos a la media son los más representativos del grupo. El prototipo

del grupo se correspondería con una muestra con todos los atributos cuyo valor se encuentre en la media.

Ejecuciones alternativas. Obtenemos distintos número de grupos.

El segundo conjunto de atributos parece similar al primero, sigue existiendo un grupo con un número poco numeroso.

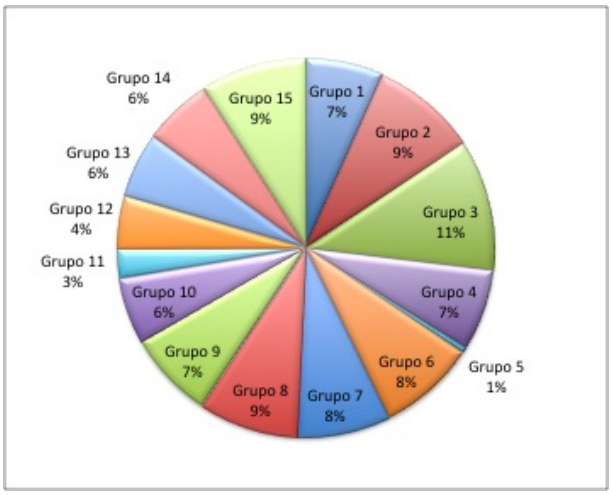


Figura 4.3: Distribución Grupos por Conjunto de Atributos 2

El tercer conjunto de atributos se tienen menos grupos, de ellos el grupo 3 parece ser muy abundante.

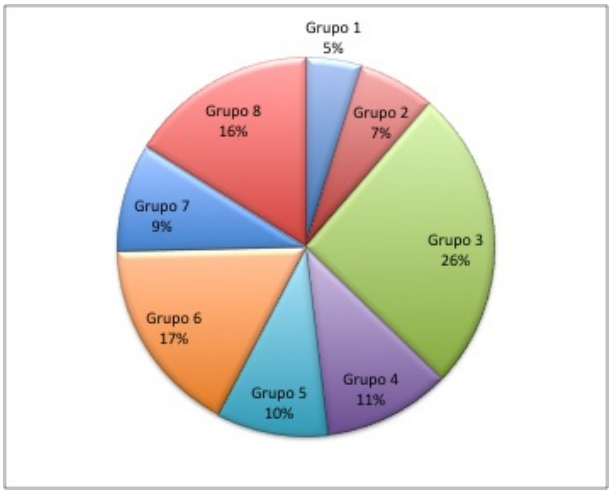


Figura 4.4: Distribución Grupos por Conjunto de Atributos 3

En el cuarto grupo se obtienen valores similares al primero, pero existe un grupo 11 con un solo elemento.

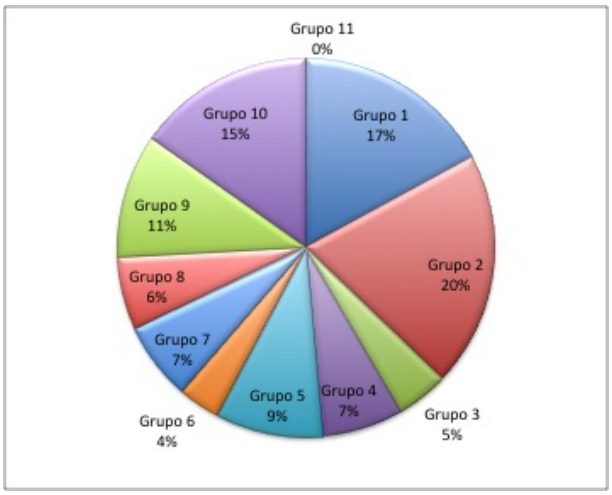


Figura 4.5: Distribución Grupos por Conjunto de Atributos 4

4.4. Análisis de Resultados del Agrupamiento

El objetivo de este apartado es un análisis inicial de los resultados y proporcionar las herramientas adecuadas a los expertos para profundizar en cada

uno de los grupos. A continuación describimos cada uno de los elementos proporcionados a la ESAC para el estudio de los grupos obtenidos, tomando como ejemplo el conjunto de atributos principal.

4.4.1. Distancia entre grupos

Una herramienta de análisis es la distancia entre grupos y entre muestras con respecto al centro del grupo. Existen distintas formas de medir dicha distancia: distancia Euclidía, distancia de Manhattan, distancia de Chebychev, distancia del coseno, o distancia de Mahalanobis.

Distancia de Mahalanobis Es una medida robusta de medir la distancia entre dos variables aleatorias multi-dimensionales, en nuestro caso cada variable se corresponde con una muestra o grupo y cada dimensión con un atributo. Su robustez radica en el uso de la covarianza S [Mahalanobis, 1936]para medir la distancia, esto permite que cada dimensión tenga un significado, valores y distribución distintos, sin embargo tener una medida de distancia fiable.

$$d(x, y) = \sqrt{(X - Y)^T S^{-1}(x - y)} \tag{4.5}$$

La distancia de Mahalanobis la hemos calculado para:

- **Grupos**. Midiendo la separación entre grupos, lo que permite comprobar la superposición de grupos o distancias destacables entre algunos de ellos.
- **Muestras con respecto el grupo**. Lo que nos permite tener ranking de muestras con respecto al grupo al que pertenecen.

Cuando hablamos de grupo tomamos como referencia el prototipo de este, el cual hemos considerado que se encuentra en la medía de todos los atributos.

Distancia de Mahalanobis entre grupos La siguiente tabla 4.1 nos indica como los grupos tienen un distancia entre 15 y 22. No existen grupos superpuestos, ni grupos alejados o cercanos entre si destacables.

	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9	Grupo 10	Grupo 11	Grupo 12
Grupo 1	0,00	21,12	21,96	21,49	21,18	21,90	11,61	21,41	20,85	21,98	21,90	15,94
Grupo 2	21,12	0,00	21,47	21,95	22,00	21,61	16,77	21,97	21,98	20,81	21,62	19,67
Grupo 3	21,96	21,47	0,00	21,75	21,51	21,99	12,90	21,69	21,25	21,87	21,99	16,92
Grupo 4	21,49	21,95	21,75	0,00	21,96	21,84	15,69	22,00	21,87	21,25	21,85	18,94
Grupo 5	21,18	22,00	21,51	21,96	0,00	21,65	16,63	21,98	21,97	20,88	21,66	19,58
Grupo 6	21,90	21,61	21,99	21,84	21,65	0,00	13,54	21,80	21,42	21,78	22,00	17,39
Grupo 7	11,61	16,77	12,90	15,69	16,63	13,54	0,00	15,96	17,38	10,60	13,57	21,42
Grupo 8	21,41	21,97	21,69	22,00	21,98	21,80	15,96	0,00	21,90	21,16	21,80	19,12
Grupo 9	20,85	21,98	21,25	21,87	21,97	21,42	17,38	21,90	0,00	20,49	21,43	20,08
Grupo 10	21,98	20,81	21,87	21,25	20,88	21,78	10,60	21,16	20,49	0,00	21,78	15,16
Grupo 11	21,90	21,62	21,99	21,85	21,66	22,00	13,57	21,80	21,43	21,78	0,00	17,41
Grupo 12	15,94	19,67	16,92	18,94	19,58	17,39	21,42	19,12	20,08	15,16	17,41	0,00

Cuadro 4.1: Distancia de Mahalanobis entre grupos

Distancia de Mahalanobis entre muestra y grupo. Como se ha mencionado tomamos el prototipo del grupo y la muestra y medimos la distancia. Se obtiene un ranking de lo representativa que resulta una muestra con respecto al grupo de mayor a menor, en la tabla 4.2 se puede ver un resumen. A menor distancia de Mahalanobis más representativo resulta del grupo, a mayor distancia menos representativo. Los ejemplos más representativos nos ayudan a analizar el grupo, pudiendo ver de manera gráfica un ejemplo próximo al prototipo del grupo [1].

	Max-Mahanalobis		Min-Mahanalobis		
	ID	Value	ID	Value	N
Grupo 1	54001.396_Net_Seg8	53,48	53280.243_Net_Seg6	3,06	307
Grupo 2	55377.241_Net	242,91	52482.298_Net_Seg5	2,99	244
Grupo 3	50217.657_Net_Seg6	47,11	,50326.326_Net_Seg7	4,43	543
Grupo 4	50371.632_Net_Seg7	51,59	51488.935_Net_Seg2	4,4	524
Grupo 5	55725.9937_Net_Seg2	36,07	50333.754_Net_Seg4	4,97	39
Grupo 6	50617.542_Net_Seg1	98,79	50700.251_Net_Seg5	4,95	543
Grupo 7	52354.233_Net_Seg2	48,18	53416.347_Net_Seg3	4,43	469
Grupo 8	54358.325_Net_Seg3	63,34	53367.419_Net_Seg1	2,43	173
Grupo 9	50945.018_Net.ls	49,44	52732.181_Net	4,01	272
Grupo 10	52099.121_Net	56,21	51331.378_Net_Seg7	4,63	370
Grupo 11	51912.571_Net_Seg2	64,19	52016.164_Net_Seg2	4,03	510
Grupo 12	51235.314_Net_Seg6	51,18	51702.995_Net	4,41	215

Cuadro 4.2: Distancia de Mahalanobis de muestras con respecto a los grupos

4.4.2. Relación entre Clases y Grupos

La relación entre clases y grupos ayuda a la comprensión de los patrones que hemos encontrado. En la clasificación manual ya vienen descritos una serie de patrones que simplifican el trabajo de describir los actuales.

Descripción de las clases . Las clases vienen descritas en otros documentos en profundidad. A modo de resumen indicar cuales son las características más destacables pasamos a describirlas.

[1] 'A la hora de mostrar los ejemplos en algunos casos se ha seleccionado muestras distintas. Bien por encontrarse las muestras óptimas segmentadas, bien por contar con trozos de señal perdidos.'

Clase ϕ. Sin amplitudes largas, menos de factor 2. Rate está en 10 kcts/seg. En el intervalo de un segundo solo veremos ruido. $HR2 = 0,1$ y $HR2 = 0,05$.

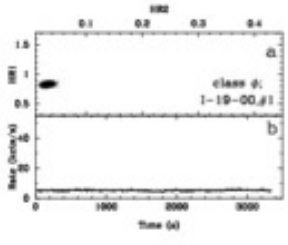


Figura 4.6: Clase ϕ

Clase χ. Rate se situado entre 3400 y 30000. $HR2 = 0,1$. El diagrama de color muestra una forma alargada.

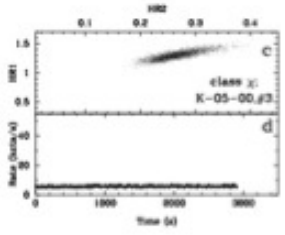


Figura 4.7: Clase χ

Clase γ. Estable en 10 kcts/seg con fuertes caídas cada par de segundos. El diagrama de color muestra una forma alargada. En los puntos correspondientes a los caídas están separadas por la distribución principal, y están situados en a lo largo de la parte izquierda de la forma alargada.

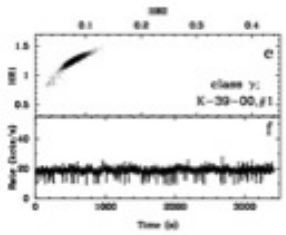


Figura 4.8: Clase γ

Clase μ. Contiene una amplitud larga, superior a factor 2. Oscilaciones entre 10 y 100 segundos. El diagrama de color muestra una forma alargada con una caída a la derecha en la parte baja.

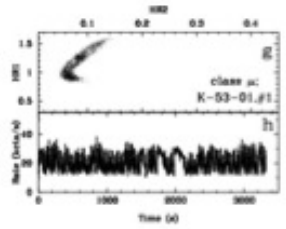


Figura 4.9: Clase μ

Clase δ. Muestra variabilidad de ruido rojo. Tiene caídas cada 10 o 20 segundos de más a menos de 10000 cts/seg. En el diagrama de color vemos un significativo suavizado de los mismos.

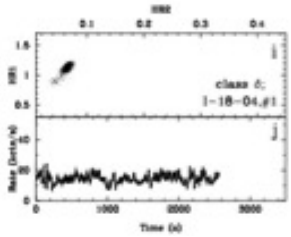


Figura 4.10: Clase δ

Clase θ. La señal sigue forma similar a una M, con intervalos de unos cientos de segundos, de 100 a 200 segundos. El Rate es bajo, inferior a 10000 cts). $HR2 = 0,1$.

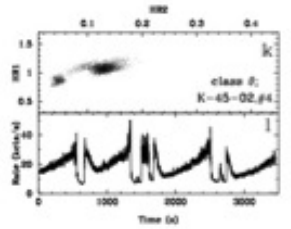


Figura 4.11: Clase θ

Clase λ. Similar a θ, forma de la señal con oscilaciones en forma de M, separados por unos cuantos centenares de segundos. Se caracteriza porque el diagrama de color muestra una forma de C.

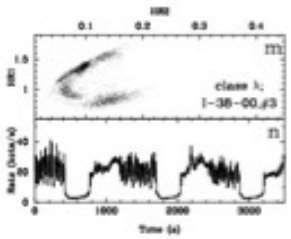


Figura 4.12: Clase λ

Clase κ. Similar a λ, pero con oscilaciones de periodos muy cortos.

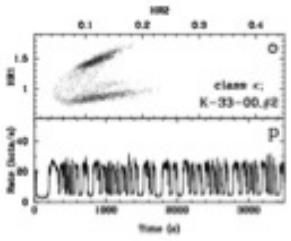


Figura 4.13: Clase κ

Clase ρ. Extremadamente regular con "bengalas" en escalas de 1 a 2 minutos. El diagrama de color sigue una forma de C.

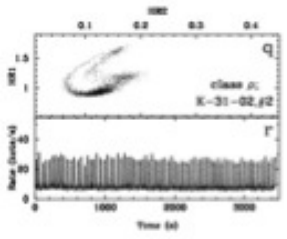


Figura 4.14: Clase ρ

Clase υ. Similar a ρ, pero más irregular, muestra intervalos tranquilos bajos. Tras el primer pico principal se muestra un segundo más notable. El diagrama de color muestra una forma con desplazamiento a la parte izquierda.

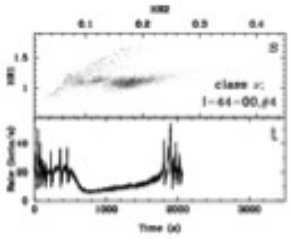


Figura 4.15: Clase υ

Clase α. Largos periodos tranquilos de 1000 segundos aproximadamente, con un Rate superior a 10000 cts/seg. Estos periodos tranquilos van seguidos de fuertes llamaradas de unos centenares de segundo con oscilaciones que van decreciendo. El diagrama de color muestra una forma alargada similar a un anillo.

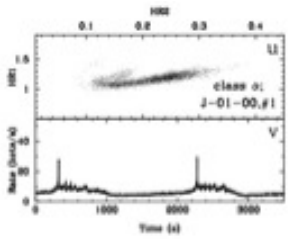


Figura 4.16: Clase α

Clase β. Es el comportamiento más complejo. La forma de identificarlo es por el diagrama de color, muestra una forma extraña alargada con un estrechamiento diagonal.

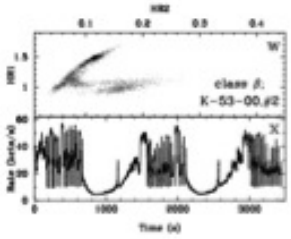


Figura 4.17: Clase β

Distribución de Clases en Grupos. Cómo ya vimos en el apartado 3.1.3 no todas las clases están representadas en igual cantidad, o tienen clasificaciones dudosas al tratarse de muestras ambiguas. Siendo χ la clase más numerosas y por tanto la que mayor representatividad tiene y μ la clase que menos representatividad tiene.

La tabla 4.3 muestra la distribución de las muestras clasificadas en los distintos grupos. Para esta comparación se ha usado el **conjunto 1** de muestras, el cual elimina muestras cuya clasificación es dudosa. Con ello nos quedamos exclusivamente con muestras cuya descripción se ajusta mejor a las clases.

	phi	chi	gamma	mu	delta	theta	lambda	kappa	rho	upsilon	alpha	beta
Grupo 1	0	0	0	0	0	0	0	0	68	7	0	0
Grupo 2	0	9	0	0	0	0	0	0	5	43	28	0
Grupo 3	19	87	11	0	54	8	0	0	0	12	0	0
Grupo 4	5	17	19	0	55	52	0	0	5	19	6	4
Grupo 5	0	4	0	0	1	1	0	1	6	0	0	0
Grupo 6	0	0	0	6	4	50	15	36	2	3	0	34
Grupo 7	36	135	0	0	0	1	0	0	0	1	1	0
Grupo 8	69	3	0	0	0	0	0	0	0	0	0	0
Grupo 9	0	146	0	0	0	0	0	0	0	0	0	0
Grupo 10	0	184	0	0	0	0	0	0	0	0	0	0
Grupo 11	0	254	0	0	0	0	0	0	0	0	1	0
Grupo 12	0	2	0	0	0	0	0	1	92	0	0	0

Cuadro 4.3: Relación de los grupos en clases

Se dan varias situaciones.

- **Grupos con varias clases**. Existen elementos comunes entre estas clases que no se tienen en cuenta en la clasificación.
- **Grupos con una clase**. El grupo se ve descrito por esta única clase.
- **Clases repartidas en un único grupo**. La clase está englobada dentro del patrón del grupo, aunque este puede contener más clases.
- **Clases repartidos en varios grupos**. Existen varios patrones dentro del conjunto de muestras que conforman la clase el cual permite dividirla.

4.4.3. Relación entre atributos

Para el estudio de la relación entre atributos se calcula la covarianza, varianza y desviación típica.

Covarianza. Nos muestra la relación entre dos atributos cualesquiera.

$$\sigma_{xy} = \frac{1}{N} \sum_{i=1}^{n} (x_i - \overline{x}) (y_i - \overline{y}) \tag{4.6}$$

Pueden darse tres situaciones distintas en la covarianza entre dos atributos.

- $\sigma > 0$ existe dependencia directa (positiva), es decir a grandes valores de x le corresponden grandes valores de y.
- $\sigma = 0$ no existe relación entre los dos atributos.
- $\sigma < 0$ existe dependencia inversa (negativa), es decir a grandes valores de x le corresponden pequeños valores de y.

Varianza. Es un caso especial de covarianza, mide la dispersión de un único atributo, por lo que nunca puede ser negativo [Dodge and Rousson, 1997], descrita en la formula 3.20.

Desviación típica. Hace más comprensible la información, al usar escalas más similares a la de los atributos. Descrita en la formula 3.21.

4.4.4. Descripción Grupos

Por cada una de los grupos obtenidos se ha realizado una exploración visual de las muestras obtenidas, siguiendo el *ranking* obtenido de la distancia de Mahalanobis 4.4.1, así como los atributos extraídos de las muestras.

Grupo 1. Representando mayoritariamente muestras pertenecientes a la clase ρ, aunque no la totalidad de esta clase. La señal comporta de forma regular a modo de *"bengalas"* en escalas de 1 a 2 minutos. El diagrama de color suele tener una forma de C.

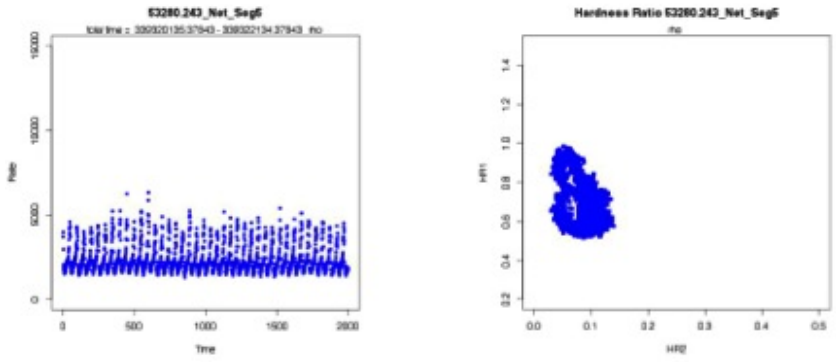


Figura 4.18: Ejemplo grupo 1

Grupo 2. Representa mayoritariamente dos clases, υ y α, que aparecen también repartidas en otros grupos en menor proporción. Estas clases tienen ciertas similitudes, comenzando con largos periodos donde la señal se muestra tranquila, seguidas de *ráfagas*.

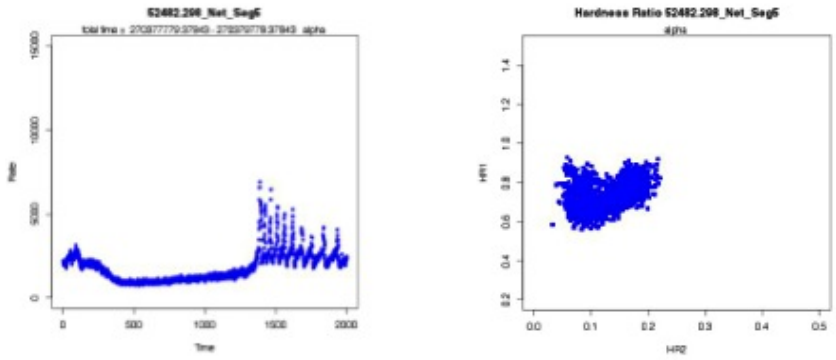


Figura 4.19: Ejemplo grupo 2

Grupo 3. Existen varias clases dominantes χ, δ y ϕ y en menor proporción otras clases. La descripción es similar a δ manteniendo una señal más elevada que la esperada para χ o ϕ, y caídas cada cada 10 o 20 segundos. El diagrama de color se nos muestra concentrado en unos pocos puntos.

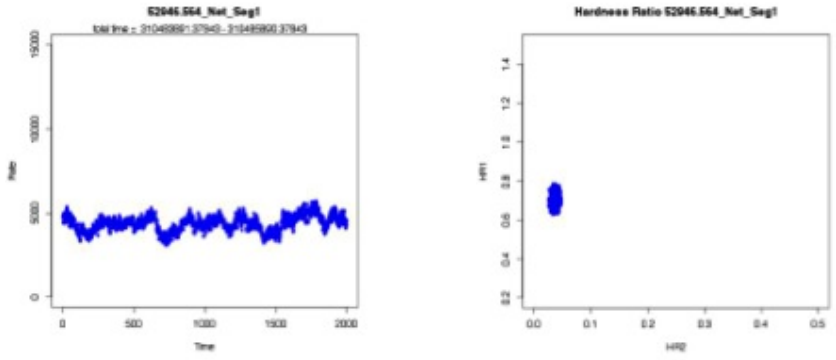


Figura 4.20: Ejemplo grupo 3

Grupo 4. Mayoritariamente formadas por δ y θ, también contiene un gran número de otras clases. Son clases que se diferencian bien visualmente y en cuanto a la descripción se refiere. Al navegar vemos que la clase θ es la que

coincide más en la descripción con este grupo, la señal muestra una forma similar a una M, con intervalos de cientos de segundos.

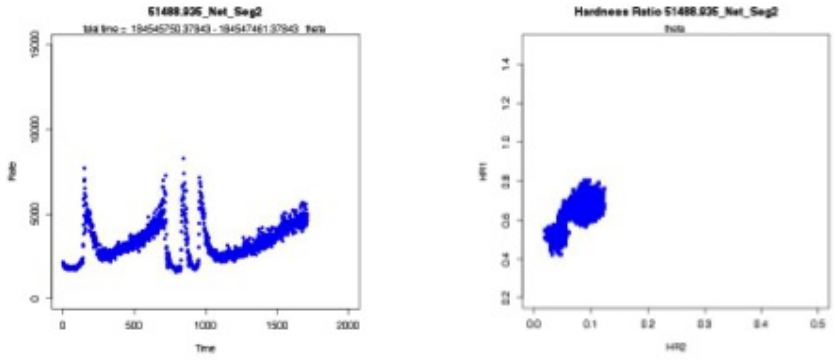


Figura 4.21: Ejemplo grupo 4

Grupo 5. Es un grupo problemático de analizar, contiene muy pocas muestras de muy diferentes clases. En un gran porcentaje de las muestras existen cortes muy prolongados, pero incluso esto no se cumple siempre, como muestra el ejemplo.

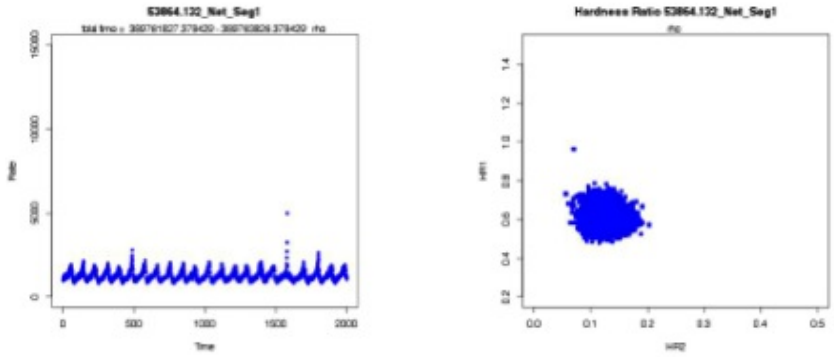


Figura 4.22: Ejemplo grupo 5

Grupo 6. Contiene mayoritariamente las clases θ, λ, γ y β, clases que son similares en algunos aspectos. Comienza con *ráfagas* que van decreciendo hasta una caída final, se mantiene estable con un leve incremento, hasta que comienza de nuevo las *ráfagas*.

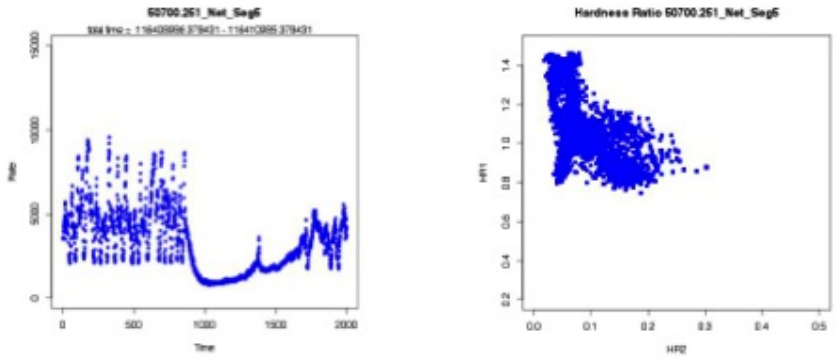


Figura 4.23: Ejemplo grupo 6

Grupo 7. Contiene en su mayoría clases χ y ϕ, siendo la señal estable con muy pocas frecuencias significativas, parece una estado intermedio entre ambas clases.

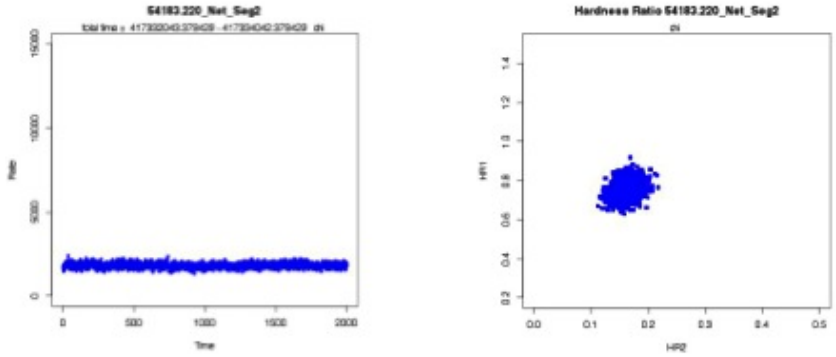


Figura 4.24: Ejemplo grupo 7

Grupo 8. Definida por *phi* mayoritariamente, contiene más frecuencias significativas, aunque no tanto como los grupos del 1 al 6. El diagrama de color se muestra circular y compacto.

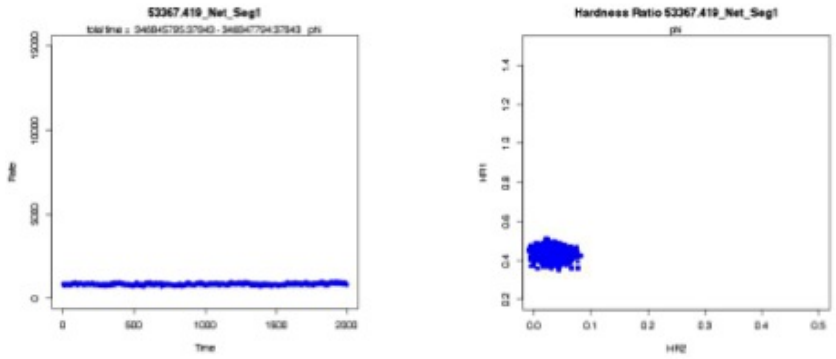


Figura 4.25: Ejemplo grupo 8

Grupo 9. Contiene muestras clasificadas como χ, sin frecuencias significativas, donde la señal se muestra estable sin grandes variaciones, y el diagrama de color tiene una forma ovaloide, estirada de derecha a izquierda y de arriba a abajo.

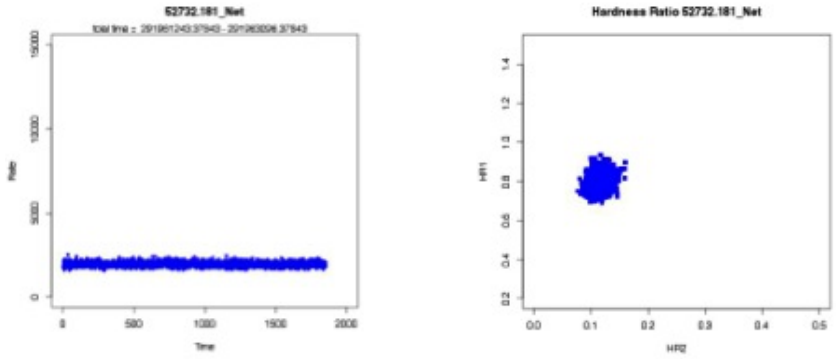


Figura 4.26: Ejemplo grupo 9

Grupo 10. Descripción muy similar al grupo 9, se diferencian en el *PeakPeriod* encontrado es más bajo.

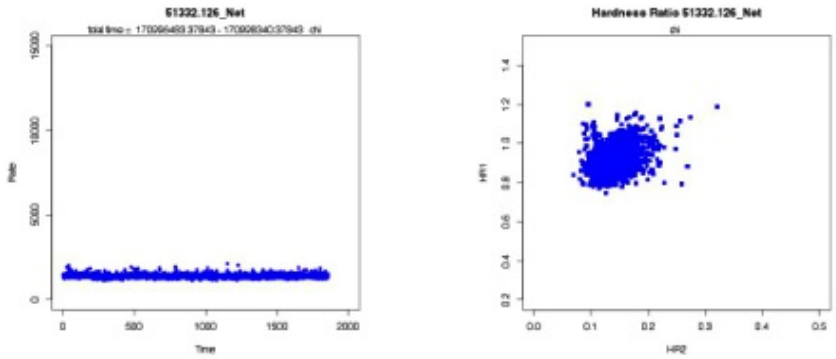


Figura 4.27: Ejemplo grupo 10

Grupo 11. Descripción muy similar al grupo 9 y 10, se diferencian en el *PeakPeriod* se encuentra en un valor intermedio.

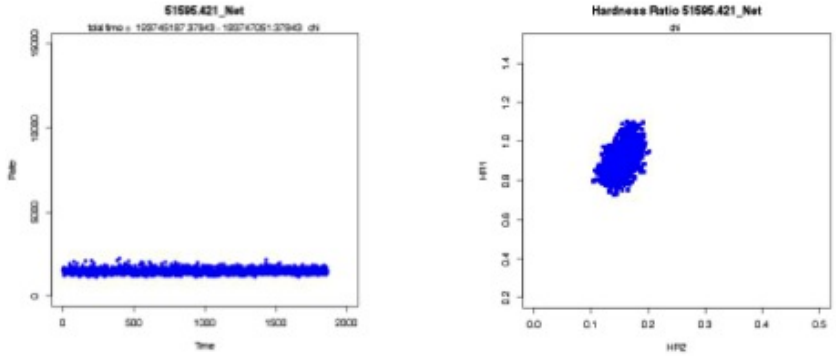


Figura 4.28: Ejemplo grupo 11

Grupo 12. Al igual que el grupo 1, se encuentra formado mayoritariamente por clases pertenecientes a ρ, siendo aparentemente un subgrupo de esta clase, con mayor frecuencia (*PeakPeriod*) y mayor amplitud *PeakAmplitude*.

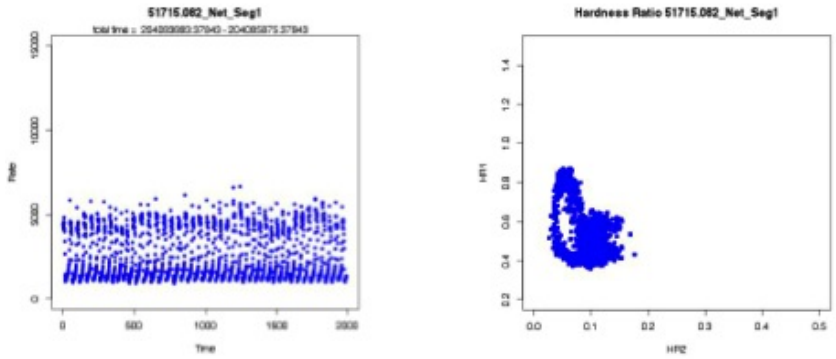


Figura 4.29: Ejemplo grupo 12

Al analizar los datos vemos que existen grupos cuya señal es prácticamente ruido, como son los grupos 9, 10 y 11, formados en su mayor parte por muestras clasificadas como χ. Grupos que son casi ruido o con pocas frecuencias significativas, como son los grupos 7 y 8, compuestas en su mayoría por muestras clasificadas como ϕ. Y grupos que contienen frecuencias significativas, resto de los grupos.

Destacan los grupos 1 y 12, por contener subgrupos de muestras pertenecientes a la clase ρ, y el grupo 5 cuya descripción resulta difícil por la gran diferencia entre las muestras que la componen.

4.4.5. Transición entre Grupos

Si bien se han obtenido grupos, que representan un conjunto de muestras que siguen un mismo patrón, la señal no es una entidad estática, sino que varía a lo largo del tiempo. Un grupo representa un estado de la señal, que se puede mostrar estable a lo largo del tiempo, o cambiar a otro grupo distinto en un momento dado.

Estas transacciones son de gran interés a la hora de analizar los grupos, por lo que se han obtenido distintas representaciones de ellas que faciliten su análisis.

En las muestras originales en las que la longitud era más larga que nuestros segmentos es posible observar estas transiciones entre estados. En ellas se puede observar como existen muestra, cuyo estado se mantiene estable a lo largo del tiempo. Ejemplos.

$50229{,}729_Net = G6 \rightarrow G6 \rightarrow G6 \rightarrow G6 \rightarrow G6 \rightarrow G6$

$52738{,}876_Net = G9 \rightarrow G9 \rightarrow G9 \rightarrow G9 \rightarrow G9$

En otros casos se producen transiciones dentro de la misma muestra, es decir se detectan estados intermedios que si no se hubiera segmentado la muestra se habrían perdido. Ejemplos.

$50232{,}531_Net = G7 \rightarrow G3 \rightarrow G3 \rightarrow G7 \rightarrow G7$

$53640{,}390_Net = G8 \rightarrow G8 \rightarrow G6 \rightarrow G3 \rightarrow G4 \rightarrow G4 \rightarrow G5 \rightarrow G4$

Existen varias formas de representar la transición entre estados.

Matriz de adyacencia . Nos indica, en términos globales, como se produce una transición de un estado a otro.

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
G1	139	0	0	5	2	0	0	0	0	0	0	1
G2	1	105	14	20	0	7	4	0	1	1	4	0
G3	0	6	259	27	1	1	66	2	2	15	2	0
G4	3	21	25	252	7	26	3	0	0	5	0	1
G5	1	3	3	5	9	2	2	0	0	0	1	0
G6	0	4	1	31	1	360	0	0	0	0	0	0
G7	0	1	51	4	0	0	170	1	27	7	37	0
G8	0	0	1	0	0	1	1	56	0	0	0	0
G9	0	0	1	0	0	0	23	0	106	10	0	0
G10	0	1	18	2	0	0	11	0	9	181	37	0
G11	0	1	2	0	1	0	27	0	0	28	191	0
G12	0	0	0	1	0	1	0	0	0	0	0	127

Cuadro 4.4: Matriz de adyacencia entre grupos

Probabilidad de transición . Nos indica la probabilidad de, estando en un grupo, saltar a otro diferente o quedarse estable en el mismo grupo. Siendo el $\sum_{i=1}^{N} P(G_j \rightarrow G_i) = 1$ siendo G_j el grupo actual, G_i cada uno de los grupos, N el número totales de grupos.

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
G1	0,95	0	0	0,03	0,01	0	0	0	0	0	0	0,01
G2	0,01	0,67	0,09	0,13	0	0,04	0,03	0	0,01	0,01	0,03	0
G3	0	0,02	0,68	0,07	0	0	0,17	0,01	0,01	0,04	0,01	0
G4	0,01	0,06	0,07	0,73	0,02	0,08	0,01	0	0	0,01	0	0
G5	0,04	0,12	0,12	0,19	0,35	0,08	0,08	0	0	0	0,04	0
G6	0	0,01	0	0,08	0	0,91	0	0	0	0	0	0
G7	0	0	0,17	0,01	0	0	0,57	0	0,09	0,02	0,12	0
G8	0	0	0,02	0	0	0,02	0,02	0,95	0	0	0	0
G9	0	0	0,01	0	0	0	0,16	0	0,76	0,07	0	0
G10	0	0	0,07	0,01	0	0	0,04	0	0,03	0,7	0,14	0
G11	0	0	0,01	0	0	0	0,11	0	0	0,11	0,76	0
G12	0	0	0	0,01	0	0,01	0	0	0	0	0	0,98

Cuadro 4.5: Matriz de probabilidad transicional entre grupos

Grafo dirigido . Una forma gráfica y menos abstracta de representar estas probabilidades es mediante el grafo dirigido.

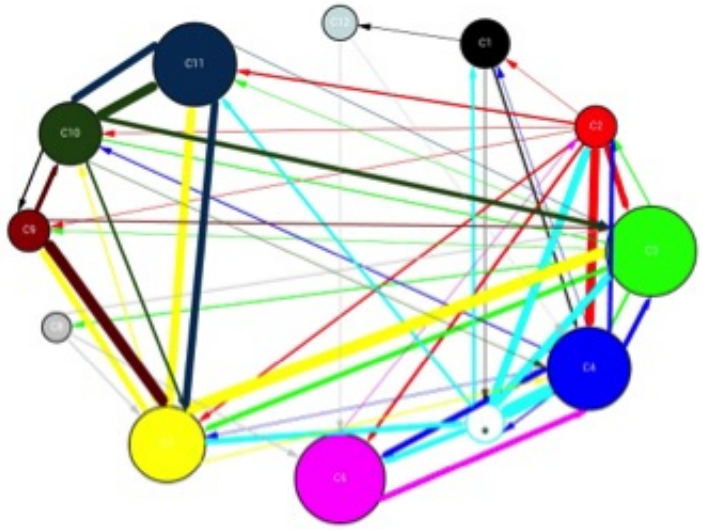


Figura 4.30: Grafo dirigido entre grupos

En el grado dirigido 4.30 que se muestra a continuación se ha dado

- **Nodos**. Cada nodo tiene un volumen directamente relacionado a la probabilidad inicial, el grupo 5 está rodeado adicionalmente por una corona para que resulte visible debido a su escaso tamaño. A cada grupo se le ha asignado un color diferente.
- **Arcos**. Cada arco muestra un groso directamente relacionado a la probabilidad de que se produzca la transición que representa el arco. El color viene marcado por el grupo origen del arco para que resulte más sencillo seguirlos.

Análisis de las transiciones . En el grafo dirigido se puede observar de manera sencilla las siguientes circunstancias.

Los dos grupos que representan la clase ρ, no tienen conexiones significativas entre ellas. Hay una transición poco significativa $P(G_1 \to G_2) = 0{,}1$

La grupo 5 es un grupo muy inestable, en contraposición con la mayoría de los grupos, nada más llegar al estado representado por el grupo 5 hay muchas probabilidades de saltar a otro estado distinto siendo solo $P(G_5 \to G_5) = 0{,}35$.

Los grupos que contienen muestras pertenecientes a la clase χ tiene transiciones bastante abundantes entre ellas y los grupos que contienen la clase ϕ. Por ejemplo $P(G_7 \rightarrow G_3) = 0{,}17$. Es decir, el grupo 7 puede ser un estado de transición a estados que contiene χ.

4.4.6. Resumen otras ejecuciones

En las otras ejecuciones el análisis de los datos no nos ha indicado un resultado radicalmente distinto, en este apartado vamos a destacar algunas de las cosas observadas tras analizar los mismo datos proporcionados para la ejecución principal.

Conjunto de atributo seleccionados 2. Eleva un poco el número de grupos, y sigue teniendo un grupo con muy pocas muestras. Si examinamos este grupo sí que aparentemente las muestras guardan más relación que el conjunto 1.

Conjunto de atributo 2. Existe un grupo con un gran número de muestras. Estas se encuentran repartidas entre varias clases.

Conjunto de atributo 3. Existe un grupo con un único elemento.

Todas estas ejecuciones tienen características comunes. ρ aparece dividido en dos grupos principales. χ y ϕ aparecen en similar número de grupos o δ θ guardan cierta relación en todas ellas.

4.5. Evaluación de Resultados

La evaluación de un modelo de agrupamiento resultado complicado, aunque existen distintas estrategias. En nuestro caso podemos basarnos.

Evaluación basada la verosimilitud. El algoritmo utilizado, EM, nos proporciona $P(D|h)$. Genera funciones de densidad para los distintos grupos, maximizando el *likehood*, de entre todos los modelos posibles selecciona aquel que se ajusta mejor a los datos.

Una estrategia de evaluación para comparar distintos modelos es basarnos en estos datos. A la hora de comparar nos puede indicar cuál es el que mejor se ajusta al modelo. Si el modelo es bueno, $p(D|h)$ en los puntos observados será elevada.

EM, usa la verosimilitud como medida de calidad de los modelos que va generando, y selecciona aquel con mayor verisimilitud. Como resultado nos indica que probabilidad de que una muestra pertenezca a cada uno de los grupos, siendo la hipótesis la que tenga mayor probabilidad. De esta forma obtenemos la probabilidad media de los puntos conocidos,

Siendo para:

$$\overline{P(D_1|h_1)} = 0{,}983 \tag{4.7}$$

Y para el resto de las ejecuciones con el resto de conjunto de atributos seleccionados.

$$\overline{P(D_2|h_2)} = 0{,}9715 \tag{4.8}$$

$$\overline{P(D_3|h_3)} = 0{,}949 \tag{4.9}$$

$$\overline{P(D_4|h_4)} = 0{,}966 \tag{4.10}$$

En todos los casos se trata de probabilidades son altas, si eliminamos la fase baja algo, y parece que el conjunto de datos extraídos de los colores obtiene mayor verosimilitud que el conjunto formado por los atributos de la señal.

¿Cómo podemos saber si mejora el agrupamiento con respecto a la clasificación manual? El modelo de clasificación seleccionado también utiliza el grado de verisimilitud, por ello esta cuestión será abordada en la generación del modelo de clasificación.

Evaluación basada en distancia. En el aparto 4.5 uno de las herramientas que hemos utilizado para el análisis de resultados, es la distancia de Mahalanobis.

Una mayor separación entre grupos nos da una medida de la calidad del modelo de agrupación. A mayor distancia entre grupos mayor será la calidad de la separación. En el apartado usamos la distancia de Mahalanobis para el análisis de los resultados, y vimos que la separación entre grupos es aproximadamente equidistante.

La distancia de Mahalanobis media entre grupos vendría marcada por:

$$\overline{m} = \frac{\sum_{1=j}^{N-1} dm_{j,j+1}}{\sum_{1=j}^{N-1} 1} \tag{4.11}$$

Siendo $\overline{dm}$ la distancia media de Mahalanobis, N el número de grupos, $dm_{j,j+1}$ distancia de Mahalanobis entre el grupo j y el grupo $j+1$.

En nuestro caso los grupos obtenidos en la ejecución principal con el agrupamiento, conjunto de atributos 1, tienen una distancia media de:

$$\overline{dm_1} = 20 \tag{4.12}$$

Para el resto de los conjuntos.

$$\overline{dm_2} = 15,75 \tag{4.13}$$

$$\overline{dm_3} = 10,89 \tag{4.14}$$

$$\overline{dm_4} = 11,45 \tag{4.15}$$

Aparentemente la distancia del conjunto principal obtiene grupos con mayor separación, pero hay que tener en cuenta la eliminación de atributos. Para utilizar esta medida como medida de calidad sería necesario aplicarla sobre el mismo conjunto de atributos.

Evaluación basada en modelo. La estrategia de evaluación consistiría en comparar los resultados con otro modelo. Solo hemos abordado un modelo de agrupamiento, y aun no contamos con un modelo de clasificación, es por ello que este punto también será abordado en la sección 5.6.

En nuestro caso, se cuenta con una clasificación manual de las muestras. Una comparación con los resultados de estas pueden darnos información sobre la eficacia del modelo.

Por ejemplo, un modelo de agrupamiento cuyos grupos coincidan con las clases, nos indica que el modelado ha funcionado correctamente, y es un buen modelo, pero no aportara nueva información a la ya conocida. Un modelo cuyos grupos no coincidan en nada con la clasificación manual, puede ser indicativo de que el modelo no funciona correctamente, aunque también puede indicar lo errónea de la clasificación manual.

La comparativa entre muestras y clases ya la obtuvimos e 4.4.2 y ya fue utilizada en el análisis de resultados. Existen coincidencias que hacer pensar que el método de agrupamiento ha funcionado correctamente, es decir encuentra algunos patrones parecidos. Ahora bien no sigue una relación de un grupo una clase, a pesar de coincidir en el número de clases de la clasificación manual y los grupos el modelo de agrupamiento principal, lo que indica la existencia de patrones diferentes aunque con algunas semejanzas.

Resulta difícil y poco útil una descripción forma de esta relación entre grupo y las clases, ya que no pretendemos que los modelos se ajusten siquiera a la representación de las clases. Pero parece estar situada en un punto intermedio entre ambas descripciones para todas las ejecuciones hechas.

Evaluación basada en la complejidad de la hipótesis. A la hora de comparar distintos modelos esto ha quedado fuera. La razón se ha utilizado el mismo método de modelado, jugando siempre con la selección de atributos de las muestras. Modelos con menos atributos darán siempre favorecidos, pero no

necesariamente serán mejor. En cualquier caso, dentro de las mismas ejecuciones, al usar un criterio BIC para medir los parámetros óptimos, se premia el que requiera menos complejidad.

Evaluación basada en análisis. Los grupos obtenidos en el agrupamiento inicial han sido presentados a los expertos, junto a todas las herramientas de análisis. No se trata de una evaluación formal, sino una valoración de que los resultados obtenidos son correctos y no son radicalmente distintos a los esperados.

CAPÍTULO 5

Modelo de Clasificación

5.1. Selección método de modelado

En el apartado 2.4.3 ya adelantábamos que no todos las técnicas son válidas para cualquier tipo de problema.

En el trabajo se corresponde a la clasificación de datos temporales, ya que el **objetivo de este capítulo es generar un modelo de clasificación**, tomando como entradas los atributos generados en la sección 3.2 y como salida los grupos obtenidos en el capítulo anterior con el agrupamiento. El **problema es de clasificación secuencial** ya que contamos con datos secuenciales tal y como se explica en el apartado 4.4.6.

Por la naturaleza del problema las técnicas normalmente utilizadas en la clasificación no pueden usarse sin más, existen alternativas mencionadas también en el apartado 2.4.3.

Se ha **elegido el modelo oculto de Márkov** por ser una técnica de clasificación muy conocida para el tratamiento de textos, voz, criptoanálisis [Huang et al., 2001], y tener buenos resultado tanto en aprendizaje supervisado secuencial como en problemas de clasificación de secuencias [Rabiner and Juang, 1986].

Por otra parte, nos permitirá realizar una evaluación del modelo de agrupamiento, proporcionando $P(D|h)$, como veremos en este mismo capítulo.

5.2. Modelo oculto de Márkov

El **modelo oculto de Márkov o HMM** *Hidden Markov Model* es un modelo estadístico en el que se asume que el sistema a modelar es un proceso de Márkov de parámetros desconocidos [Petrie and Baum, 1966] [Huang et al., 2001]

Cadena de Márkov. Un modelo de Márkov o cadena Márkov es un tipo especial de modelo estocástico [1] discreto en el que la probabilidad de que ocurra un evento depende del evento inmediatamente anterior, con lo que se consigue que estos modelos tengan memoria o historia. El conocer la historia hasta el instante actual se conoce como la **propiedad de Márkov.**

La aplicación a los grupos obtenidos en el apartado anterior vendría dado por:

$$P(G_{n+1} = G_{n+1}|_n = g_n, G_{n-1} = g_{n-1}, ..., G_2 = g_2, G_1 = g_1) = P(G_{n+1} = g_{n+1} = G_n = g_n) \tag{5.1}$$

Donde g_1, g_2, g_3,.... son cada uno de los grupos, representados como una variable aleatoria.

Elemento de HMM. En un HMM los estados del modelo no son directamente observables, como ocurre en un modelo de Márkov. Solo se puede observar una cierta salida, la cual dependerá del estado en que se encuentre el modelo [Petrie and Baum, 1966].

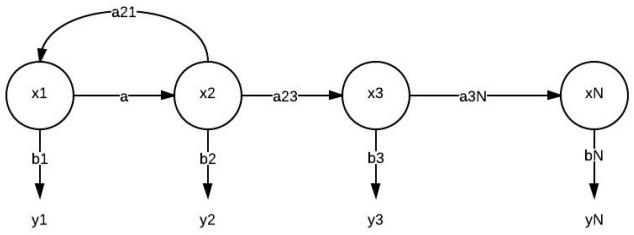


Figura 5.1: Gráfico Modelo Oculto de Márkov

Los elementos que componen un HMM son:

- **Estado Ocultos**. Son los estados del modelo que no son directamente observables $S = \{S_1, S_2, S_3, .., S_N\}$ se denomina con X_t al ser un valor

[1] Un proceso estocástico es un concepto matemático que sirve para caracterizar una sucesión de variables aleatorias (estocásticas) que evolucionan en función de otra variable, generalmente el tiempo. Cada una de las variables aleatorias del proceso tiene su propia función de distribución de probabilidad y, entre ellas, pueden estar correlacionadas o no.

desconocido en un tiempo t. Dependiendo del problema a resolver los estados ocultos representados en el modelo pueden variar como veremos.

- **Símbolos Observable.** Observaciones distintas para cada estados, M, representada por Yt, en un modelo discreto $O = \{o_1, o_2, ..., o_M\}$. En nuestro caso vendría a esta representada por la señal y los atributos extraídos en la sección 3.2.
- **Probabilidades Iniciales**. Probabilidad de cada estado.

$$\pi = Pr(Xi = Si) \tag{5.2}$$

Siendo $1 \leq i \leq N$.

- **Probabilidades de las observaciones**. Probabilidad de que para cada símbolo observable se dé un estado oculto. Vendría definida por:

$$B = b_j(k) \tag{5.3}$$

para cuando en un estado j se cumple:

$$b_j(k) = Pr(y_k, ent|X_t = Sj), siendo 1 \leq j \leq N y 1 \leq k \leq M \tag{5.4}$$

- **Probabilidad de transición.** Es la probabilidad de que se produzca un cambio de estado oculto a otro estado oculto,

$$A = a_{ij} \tag{5.5}$$

Siendo

$$a_{ij} = Pr(X_{t+1} = S_j|X_t = S_i), \forall i \geq 1, j \leq N \tag{5.6}$$

No todas las transiciones son posibles, es común que $a_{ij} = 0$, ver la tabla 4.5.

Procesar secuencia con HMM. Teniendo una secuencia de observaciones:

$$O = O_1, O_2, .., O_T \tag{5.7}$$

Donde T es el número total de observaciones.

HMM puede ser utilizada bien para analizar, bien para clasificar, siempre que se cuente con N, M, A, B y π. Los pasos serían los siguientes:

1. Se elige un estado inicial $X_1 = S_1$, de la distribución inicial.
2. Se inicia $T = 1$.
3. Se elije $O_t = y_k$ atendiendo a la distribución de probabilidad B.
4. Se cambia a un estado nuevo $Xt + 1 = X_j$ atendiendo a la distribución de probabilidad A.
5. Se actualiza $t = t + 1$ si $t < T$ y se vuelve al paso 3. Si $t \geq T$ se finaliza.

Modelo de Oculto de Márkov con atributos continuos. Hasta ahora hemos estado hablamos de HMM cuyos símbolos se corresponden con variables discretas. Ahora bien, esta situación no se da en nuestro problemas donde los atributos extraídos en el apartado 3.2 se corresponde con variables continuas.

El funcionamiento de un HMM con variables continuas es muy similar. Solo es necesaria cambiar la **probabilidad de las observaciones que vendría definidas como funciones de densidad** [Levinson, 1986]. De tal forma que:

$$B = b_j(k) = F(b_j(k)) \tag{5.8}$$

Así el gráfico que representa el HMM quedaría como:

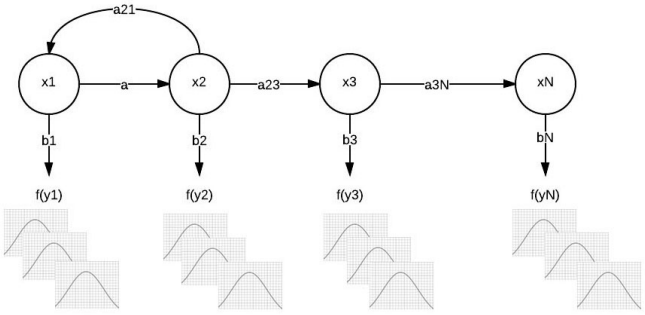


Figura 5.2: Gráfico Modelo Oculto de Márkov

5.3. Algoritmos

En los modelos ocultos de Márkov, existen tres preguntas fundamentales a resolver:

1. ¿Cómo calcula la probabilidad de una secuencia observada?
2. ¿Cuál es la secuencia óptima S de estados dada una secuencia de observaciones O?
3. ¿Cómo podemos estimar los parámetros del modelo $\mu = (\pi, A, B)$ para maximizar $P(O|\mu)$?

Para resolver estas cuestiones hemos utilizaremos tres algoritmos diferentes que se complementan.

5.3.1. Algoritmo de avance-retroceso

Algoritmo de avance-retroceso, o **forward-backward**, permite el cálculo de probabilidad de una secuencia de observaciones O dado un modelo $\mu = (\pi, A, B)$, respondiendo a la cuestión 1. Siendo el objetivo el calcular eficientemente $P(O|\mu)$. En condiciones normales, para 10 estados y 10 observaciones sería necesario realizar del orden de 10^{11} operaciones avance-retroceso permite reducir la complejidad [Rabiner and Juang, 1986].

Procedimiento hacia adelante. Considerando la variable $\alpha_t(i)$

$$\alpha_t(i) = P(o_1, o_2, ..., o_t, q_t = i|\mu) \tag{5.9}$$

Dado el modelo μ, $\alpha_t(i)$ es la probabilidad de observación de $o_1, o_2, ..., o_t$ y estar en un instante t en el estado i.

1. **Inicialización**

$$\begin{gathered} \alpha_1(i) = \pi_i b_i(o_1), \\ 1 \leq i \leq N \end{gathered} \tag{5.10}$$

2. **Recurrencia**

$$\begin{gathered} \alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1}, \\ t = 1, 2, ...T-1, 1 \leq j \leq N \end{gathered} \tag{5.11}$$

3. **Terminación**

$$P(O|\mu) = \sum_{i=1}^{N} \alpha_T(i) \tag{5.12}$$

Procedimiento hacia atrás. Consideramos la variable $\beta_t(i)$ definida por:

$$\beta_t(i) = P(o_1, o_2, ..., o_t, |q_t = i, \mu) \tag{5.13}$$

Dado el modelo μ , $B_T(i)$ es la probabilidad de la secuencia de observación desde el instante de tiempo $t+1$ hasta el final, cuando el estado en el instante de tiempo t es i.

1. **Inicialización**

$$\begin{gathered} \beta_T(i) = 1, \\ 1 \leq i \leq N \end{gathered} \tag{5.14}$$

2. **Recurrencia**

$$\begin{gathered} \beta_t(i) = \sum_{j=1}^{N} a_{ij} B_{t+1}(j) b_j(o_{t+1}, \\ t = T-1, T-2, ..,1, 1 \leq j \leq N \end{gathered} \tag{5.15}$$

3. **Terminación**

$$P(O|\mu) = \sum_{i=1}^{N} \beta_1(i)\pi_i b_i(o_1) \tag{5.16}$$

5.3.2. Algoritmo de Viterbi

Este algoritmo permite contestar a la cuestión 2, permitiendo encontrar las secuencias de estados más probables en un HMM [Forney Jr, 2005].

Obtiene un $S = (q_1, q_2, ..., q_T$ que mejor explique las observaciones $O = (o_1, o_2, ..., o_t)$.

Consideremos la variable $\delta_t(i)$ definida por:

$$\delta_t(i) = \max_{q_1, q_2, \ldots, q_{t-1}} P(q_1, q_2, ..., q_t = 1, o_1, o_2, ...o_t|\mu) \tag{5.17}$$

$\delta_t(i)$ es la probabilidad del mejor camino hasta el estado i haciendo visto las t primeras observaciones. Esta función se calcula para todos los estados e instantes de tiempo.

$$\delta_{t+1} = \left[\max_{1 \leq i \leq N}\right] b_j(ot + 1) \tag{5.18}$$

El objetivo es encontrar la secuencia de estados más probables por l, que será necesario almacenar, el argumento que hace máxima la ecuación anterior en cada instante de tiempo t y para cada estado j. Para ello utilizaremos la variable $\delta_t(j)$.

El proceso completo sería:

1. **Inicialización**

$$\begin{gathered} \delta_1(i) = \pi_i b_i(o_1), \\ 1 \leq i \leq N \end{gathered} \tag{5.19}$$

2. **Recurrencia**

$$\begin{gathered} \delta_{t+1}(j) = \left[max_{1 \leq i \leq N} \delta_t(i) a_{ij}\right] b_j(o_{t+1}) \\ t = 1, 2, ...T - 1, 1 \leq j \leq N \end{gathered} \tag{5.20}$$

$$\begin{gathered} \delta_{t+1}(j) = argmax_{1 \leq i \leq N} \delta_t(i) a_{ij} \\ t = 1, 2, ...T - 1, 1 \leq j \leq N \end{gathered} \tag{5.21}$$

3. **Terminación**

$$q_T^* = arg \max_{1 \leq i \leq N} \delta_T(i) \tag{5.22}$$

Como se observar algunas de las secuencias son familiares al algoritmo adelante, explicado en el punto anterior. Una de las diferencias es la sustitución de la suma por $argmax$ para el cálculo de estados más probables.

5.3.3. Algoritmo de Baum-Welch

Responde a la última cuestión, la 3 permitiendo un modelo μ que maximice la probabilidad de unas secuencia de observaciones $O = (o_1, o_2, ...o_t)$, es decir, determinar el modelo que mejor explique tal secuencia maximizando $P(O|\mu)$ [Baum et al., 1970].

Siendo $\xi_t(i,j)$ la probabilidad de estar en un estado i en un instante t y en un estado j en el instante $t+1$, dado una observación O y el modelo μ.

$$\begin{aligned} \xi_t(i,j) &= P(q_t = i, q_{t+1} = j|O, \mu) \\ \xi_t(i,j) &= \frac{P(q_t=i,q_{t+1}=j,O|\mu)}{P(O|\mu} = \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}}{P(O|\mu} \\ \xi_t(i,j) &= \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}}{\sum_{k=1}^{N}\sum_{l=1}^{N}\alpha_t(k)a_{kl}b_l(o_{t+1})\beta_{t+1}(l)} \end{aligned} \tag{5.23}$$

Donde $\alpha_t(i)$ viene definida en por la formula 5.9, y $\delta_t(i)$ refretroceso

1. **Inicialización** Se parte de un modelo inicial μ, que puede ser seleccionado aleatoriamente.
2. **Calculo de transición de símbolos de emisión** que son más probables según el modelo inicial escogido.
3. **Construcción de nuevo modelo** en el que se incrementa la probabilidad de las transiciones y símbolos determinados en el paso anterior. Para la secuencia de observables en cuestión, el modelo tendrá ahora una probabilidad mayor que el modelo anterior.

El proceso se repite hasta que no exista mejora entre un modelo y el siguiente revisado.

La probabilidad de estar en el estado i en un instante $t = 1$:

$$\begin{gathered} \overline{\pi}_i = \gamma_1(i) \\ 1 \leq i \leq N \end{gathered} \tag{5.24}$$

Reestimación de probabilidades de transición. El numerador representa el número esperado de transición de i a j, y el denominador representa el número esperado de transiciones desde i:

$$\begin{gathered} \overline{a}_{ij} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}, \\ 1 \leq i \leq N, 1 \leq j \leq N \end{gathered} \tag{5.25}$$

Reestimación de las probabilidades de emisión. El numerador representa el número esperado de veces que se pasa por el estado j y se observa o_k, y el denominador representa el número esperado de veces que se pasa por el estado j.

$$\begin{gathered}\overline{b}_j(o_k) = \frac{\sum_{t=1:o_t=o_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \\ 1 \leq i \leq N, 1 \leq j \leq N\end{gathered} \tag{5.26}$$

5.4. Ejecución del Método de Clasificación

El modelo oculto de Márkov puede ser de utilidad en el estudio de GRS1915+105 en distintos aspectos. Por ejemplo.

- **Detección de Sistemas Similares**. Permitiendo detectar dentro del Espacio objetos similares. En este modelo existen dos estados ocultos.
  - *Objeto Similar*. El sistema detectado es similar a GRS1915+105.
  - *Objeto No Similar*. El sistema detectado no es similar a GRS1915+105.

  El problema queda fuera del alcance del proyecto, al no disponer de las probabilidades de las observaciones, ni de las probabilidades iniciales.

- **Detección de Estado del sistema**. Por lo que sabemos hasta ahora del sistema, GRS1915+105 es un sistema binario compuesto por una estrella y un agujero negro, que a su vez es un microquasar. Este sistema binario pasa por dos estados básicos conocidos.
  - *Expulsa materia*.
  - *No expulsa materia*.

  Sobre estos dos estados no hemos trabajado a lo largo del proyecto. Sin embargo sí que se han definido grupos. Estos grupos tienen entre sus características información que nos pueden indicar que está haciendo el sistema en cada momento. Por ejemplo, grupos que prácticamente son ruido y la señal se mantiene baja son probablemente estados donde el sistema no está expulsando materia. Grupos con frecuencias significativas, altas, y señales similares a latidos de corazón, tienen mucha probabilidad de tratarse de un estado donde el sistema expulsa materia.

  El problema vuelve a ser cual es la probabilidad inicial y cuál es la probabilidad de las observaciones, aunque en este caso podemos llegar a una aproximación gracias al análisis de los grupos. Esta labor queda fuera del proyecto dada la necesidad de un análisis más profundo de los grupos por expertos en la materia.

- **Detección de grupos**. En el apartado anterior se han obtenido distintos grupos correspondiente a patrones de distintos estados de la señal. El HMM puede ayudarnos a la detección de dichos estados por medio de los símbolos visibles. Esto permitirá la evaluación de los grupos, y por tanto del modelo de agrupamiento. Es la utilidad que en estos momentos se pretende dar al modelo de clasificación.

  Se cuentan con los datos necesarios para abordarla al tener la **probabilidad inicial de cada grupo**.

En la sección anterior del trabajo se han nombrado una sería de elementos que es necesario definir.

Estados Ocultos . En el caso utilizar grupos $S = (s_1, s_2, ...s_3) = G = (g_1, g_2, ..., g_3)$, siendo G el conjunto de grupos obtenidos en el agrupamiento.

Símbolos Observables. En todo modelo oculto de Markov tenemos estados ocultos y símbolos observarles. Son los atributos extraídos de 3.1.3. Siendo $O = (o_1, o_2, ..., o_n)$, cada observación o_i viene definida por el conjunto total de atributos $PeakPeriod$, $PeakAmplitude$, etc.

Probabilidades iniciales de cada grupo. Definida por en la tabla.

Probabilidad Inicial	
Grupo 1	0,07
Grupo 2	0,06
Grupo 3	0,13
Grupo 4	0,12
Grupo 5	0,01
Grupo 6	0,13
Grupo 7	0,11
Grupo 8	0,04
Grupo 9	0,06
Grupo 10	0,09
Grupo 11	0,12
Grupo 12	0,05

Cuadro 5.1: Probabilidades iniciales

Probabilidades de las observaciones. Siendo estos variables continuas deberán ser tratadas mediante funciones de densidad [Levinson, 1986].

Ejecuciones realizadas. El objetivo del modelo de clasificación es principalmente tener un modelo con el que poder comparar el agrupamiento. Es por ello que tanto la ejecución, como el análisis como la evaluación están orientados a poder evaluar mejor el agrupamiento.

1. Con la agrupación obtenida del primer conjunto de atributos, ejecutando el modelo con cada atributo de manera individual. Con ello podremos medir

la capacidad de clasificación de cada atributo mediante la generación de un modelo [Hernández Orallo et al., 2004].

2. Grupos obtenidos del conjunto de atributos seleccionados 1. Nos permitirá la evaluación de los grupos obtenido por medio del modelo de agrupamiento, y por tanto también el modelo de agrupamiento.

3. Clasificación manual. Ejecutando el HMM con la clasificación manual con dudas, esto es con un bajo $P(D|h)$, podemos estimar cuanto de buena o mala es el agrupamiento automático con respecto a la clasificación manual.

4. Grupos obtenidos del conjunto de atributos seleccionados 2. Permitirá evaluar el poder de clasificación de la fase, atributo del que se tienen dudas sobre su utilidad, conjuntamente con el resto de los atributos.

5. Grupos obtenidos del conjunto de atributos seleccionados 3. Permite evaluar los atributos extraídos de la señal como clasificadores.

6. Grupos obtenidos del conjunto de atributos seleccionados 4. Permite evaluar los atributos extraídos de los Hardness Ratios como clasificadores.

5.5. Análisis de Resultados

5.5.1. Ejecución sobre los atributos

Esta ejecución no tenía como objetivo un modelo estable. Sino valorar los atributos como clasificadores sobre el conjunto total de datos proporcionados. Destacan varias cosas tras la ejecución.

- Ningún atributo da buenos resultados por sí solo. Como cabría esperar.
- $HR1.Median$ y $SignFreq1$. Son menos significativos de los que esperábamos.
- $PeakPhaseShift$. Presenta una tasa de error de muestreo muy elevado, prácticamente no clasifica nada bien como se aprecia en la figura refhmm.peakphaseshift

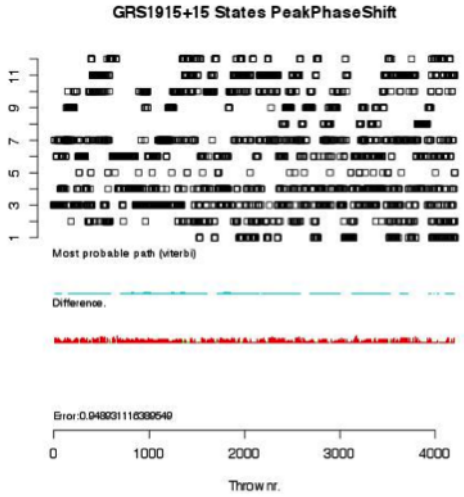


Figura 5.3: Resultado HMM para PeakPhaseShift

- Visualmente también se puede valorar como cuando un grupo resulta bien clasificado, su probabilidad de que pertenezca a ese grupo es alta.

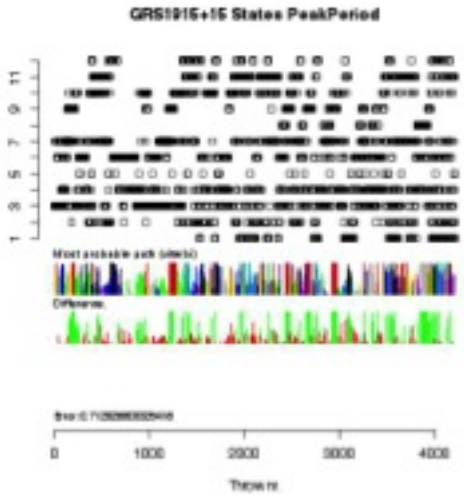


Figura 5.4: Resultado HMM para PeakPeriod

5.5.2. Ejecución sobre conjuntos de atributos

Con el conjunto total de datos sobre cada unos de los conjuntos de atributos seleccionados el número de muestras correctamente seleccionados es elevado.

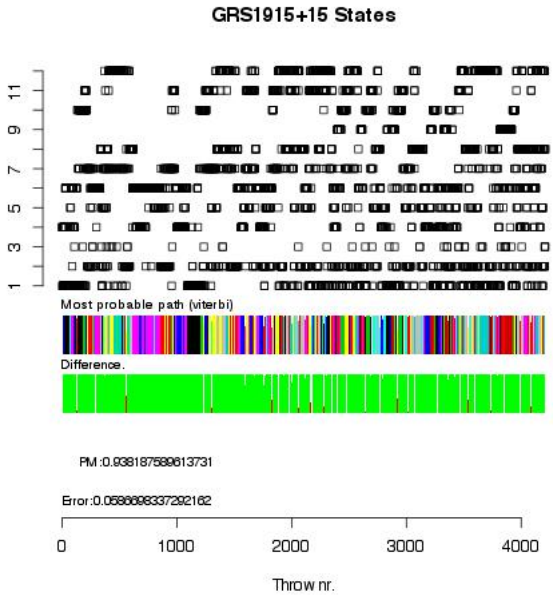


Figura 5.5: Resultado HMM conjunto atributos 1

Además se puede observar que la probabilidad de que la muestra pertenezca al grupo donde fue clasificada es muy alta. Por ejemplo en el gráfico 5.5 nos muestra de arriba a abajo:

- Distribución del grupo al que pertenece.
- Hipótesis seleccionada por Viterbi, y probabilidad obtenida para ella según avance-retroceso, representada por el ancho. Cada color representaría un grupo diferente [2].
- Diferencias en las hipótesis de EM y HMM: verde en caso de acierto; rojo en caso de ser diferentes, y la probabilidad de Viterbi para la hipótesis de EM.
- Hipótesis

Con los conjuntos de atributos seleccionados 2, 3 y 4 .

[2]La relación de color-grupo la podemos observar en el grafo dirigido 4.30.

5.5.3. Ejecución sobre la clasificación original

Generamos un modelado con la clasificación inicial, con el conjunto de datos 2, que contiene tanto clasificaciones dudosas como no dudosas. Se puede observar a simple vista que el modelo generado no es valido, con una tasa de error muy elevada.

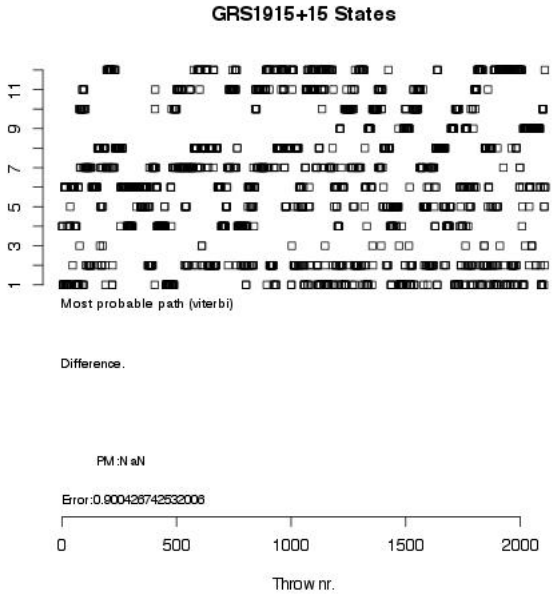


Figura 5.6: Resultado HMM clasificación

Esto puede ser debido a varios factores. La clasificación manual proporcionada no muestra ninguna transición entre estados. Las muestras en la clases no siguen un distribución normal. Existen un porcentaje importante de clases mal clasificadas, falsos χ.

5.6. Evaluación de Resultados

La evaluación se ha hecho atendiendo a los errores de muestreo y a verosimilitud. La evaluación se realiza mediante la comparación con los resultados obtenidos en el agrupamiento.

5.6.1. Evaluación atributos individuales

Si atendemos al error de muestreo los atributos menos significativos son $PeakPhaseShift$ y $SignFreq1$.

PeakPeriod	0,71
SignFreq1	
Median	0,65
Mad	0,62
kurtosis	0,70
skewness	0,73
PeakAmplitude	0,67
PeakPhaseShift	0,95
HR1.Median	0,87
HR1.Mad	0,73
HR1.kurtosis	0,63
HR1.skewness	0,68
HR2.Median	0,72
HR2.Mad	0,76
HR2.kurtosis	0,78
HR2.skewness	0,73

Cuadro 5.2: Error de muestreo para los atributos

$PeakPhaseShift$ no es algo nuevo, ya que no salió en ninguna selección de atributos, y tampoco sale en buenas posiciones en el rankink. $SignFreq1$ Por sí solo no es un atributo significativo.

5.6.2. Evaluación de modelos

Si atendemos al error de muestreo de las distintas selecciones de atributos.

Selección Atributos 1	0,058
Selección Atributos 2	0,126
Selección Atributos 3	0,071
Selección Atributos 4	0,072

Cuadro 5.3: Error de muestreo con los distintos conjuntos de atributos

El error de muestreo es bajo, indicando la coincidencia entre los resultados del modelo de agrupamiento y el modelo de clasificación, y de forma indirecta confirmando el buen funcionamiento de ambos sistemas.

La eliminación de $PeakPhase$ hace aumentar al doble el error, lo que nos indica que finalmente si es útil al modelo, aunque hay que valorar si merece la pena el aumento en la complejidad de las reglas antes de determinar que sea un mejor modelo.

Eliminar tanto los atributos derivados de la señal, como los derivados de los colores, aumentan ligeramente el error de la muestra, ni mucho menos tanto como solo eliminar la $PeakPhase$, simplificando las reglas de clasificación.

Si nos basamos en la verosimilitud, tenemos que atender a $P(D|h)$, para el modelo de agrupamiento ya se resolvió en el apartado 4.5. Para los modelos generados en la clasificación realizamos la misma acción tomando como hipótesis más plausible la obtenida en EM, y comprobando el resultado de avance-retroceso para esa hipótesis.

	P(D|h) EM	P(D|h) HMM
Selección Atributos 1	0,983	0,938
Selección Atributos 2	0,9715	0,825
Selección Atributos 3	0,949	0,9153
Selección Atributos 4	0,966	0,923

Cuadro 5.4: Probabilidades medias de los distintos conjuntos de atributos

Resultados muy parecidos, siendo el conjunto de atributos seleccionados 2 el que cuanta con mayor diferencia entre el modelo de agrupamiento y el modelo de clasificación.

Las conclusiones son las mismas que al examinar el error de muestreo.

5.6.3. Evaluación de los grupos frente a las clases

Se realizo ha generado un modelo de HMM con las muestras clasificadas con las que se cuenta. Esto dio como resultado un modelo que no clasifica nada bien. Siendo:

$$E_m = 0{,}9 \tag{5.27}$$

Y

$$\overline{P(D_c|h_c)} = NaN \tag{5.28}$$

Es decir un error muy elevado, y una probabilidad desconocida. Las razones ya se comentaron en el análisis. Esto quiere decir.

- Los grupos generado siguen una distribución normal, frente a las muestras clasificadas en la que no se encuentra la función de densidad.
- En la agrupación se ha encontrado estados intermedios, frente a las muestras clasificadas donde una muestra pertenece a una única clase con independencia de cómo vaya variando la señal.
- La clasificación manual cuenta con clases mal clasificadas, falsos χ que finalmente tiene frecuencias significativas.

Estas son las tres ventajas que aporta el agrupamiento frente a la clasificación ya conocida.

Parte IV

Conclusiones

CAPÍTULO 6

Conclusiones

6.1. Conclusiones generales

El trabajo tenía varios objetivos que a continuación pasaremos a valorar si se ha cumplido.

Por un lado los astrónomos y expertos que acudieron buscaban patrones en la variabilidad del sistema GRS1915+105. El modelo principal concluye que existen 12 grupos distintos de muestras. Y por los resultados obtenidos en el agrupamiento y en la clasificación se puede concluir que dicho objetivo ha sido satisfecho.

El objetivo del trabajo es seleccionar algoritmos, técnicas y métodos para la generación de modelos que se ajusten al sistema y problema planteado. Este objetivo general se desgloso en distintos objetivos específicos.

La probabilidad de $P(D|h) > 0{,}9$ con todos los conjuntos de atributos utilizados para evaluar, además el error de muestreo en tres de los cuatro es $E_m < 0{,}1$, siendo en el cuarto $E_m < 0{,}15$. Además los grupos guardan cierta relación con las clases, y entre ellos si realizamos una exploración manual.

- **Determinar tareas.** Son dos las tareas abordadas en el trabajo, el agrupamiento, y la clasificación. La primera nos ha servido a la búsqueda de patrones, la segunda a la evaluación de dichos patrones, complementando los resultados obtenidos en la primera. Ambas tareas han cumplido con las expectativas esperadas para el trabajo.

  Cotas de probabilidad similares, error de clasificación bajo, distancias obtenidas y el análisis de los grupos sugieren esto.

- **Transformación de los datos.** Se ha generado un conjunto de atributos significativos y útiles tanto para el agrupamiento como para la clasificación, se he realizado una medición de estos tanto al principio del en la sección 3.2, como en la clasificación.

  Los atributos seleccionados son suficiente significativos como para obtener un modelo. Finalmente la aplicación de filtros fue eliminada como sugería la selección y evaluación, la fase tiene una baja utilidad como clasificadora individual. Si es verdad que, cuándo se deja dentro del conjunto, consigue grupos más separados, con mayor verosimilitud, y con menos error de muestreo. Para valorar cual es el mejor conjunto de datos habría que tener en cuenta la complejidad de la hipótesis, a mayor número de datos a valorar mayor será la complejidad de obtener los datos.

  La segmentación también consigue su objetivo, en el rango buscada se encuentran patrones. Ahora bien pueden resultar útiles otro tipo de aplicaciones.

- **Seleccionar métodos adecuados para el modelado.** Tras concluir que se abordarían las tareas de agrupamiento y clasificación, se han seleccionado dos métodos, uno para cada tarea, en la sección 4.1 y la sección 5.1, y que cumplen con todos los requisitos encontrados.

- **Analizar los resultados.** Se proporcionan las herramientas necesarias para el análisis de los resultados.

  - Distancia de Mahalanobis. Nos permite medir la separación entre grupos y la búsqueda de las clases más representativas de cada grupo.
  - Comparación entre clases y grupos. Nos ha permitido facilitar el análisis de los patrones encontrados.
  - Generación de gráficos por cada muestra, que facilitan el análisis los resultados obtenidos.
  - Covarianza, varianza y desviación típica, nos permite saber el comportamiento de cada atributo dentro de un grupo.
  - Matriz de adyacencia, probabilidad transicional y grafo dirigido, nos permite el análisis de la transición entre estados.

  La información completa fue proporcionada para el análisis detallado de los grupos, resultando de utilidad.

- **Evaluación de los resultados.** También se han definido estrategias de evaluación y proporcionado herramientas, algunas comunes al análisis de los resultados, para la evaluación tanto del modelo de agrupamiento (sección 4.5) como el modelo de clasificación (sección 5.6). Estas herramientas son:

- Basada en la verosimilitud. Mediante la $\overline{P(D|h)}$ de cada modelo generado.
- Basada en la distancia. Mediante la distancia de Mahalanobis entre grupos.
- Basada en el modelo. Proporcionando un modelo de clasificación, en el que obtenemos el error de muestreo, o la verosimilitud.

No hemos valorado la complejidad de las hipótesis, siendo una tarea pendiente.

- **Obtener futuras lineas de trabajo.** Este último punto aún no ha sido resuelto, pero lo hará a lo largo de este capítulo, en la sección 6.6

A lo largo del trabajo nos hemos encontrado con distintos requisitos que todo método o algoritmo utilizado han tenido que cumplir. Y en todos los casos se ha encontrado una alternativa eficaz, no siendo necesario la generación de nuevos métodos y algoritmos, sino la implementación de ya existentes.

6.2. Conclusiones Selección de Atributos

Los atributos generados a partir de las curvas de luz buscan dos cosas: frecuencias significativa en la variabilidad; y describir la forma de la señal. Las muestras se encontraban incompletas y con distintas longitudes de tiempo, la señal que contenía presentaba ruido y cortes. Las muestras incompletas han tenido que ser eliminadas.

El ruido y los cortes hemos dejado que se encargue el algoritmo de búsqueda de frecuencias significativas.

Lomb-Scargle. Ha sido la base en la búsqueda de frecuencias significativas, su tolerancia al ruido y a los cortes ha quedado patente a la hora de analizar las muestras. Además, no es necesario la aplicación de ningún filtro para el tratamiento de la señal.

El resumen estadístico. Parece suficiente para describir la forma de la señal y los hardness ratios. Esto queda patente tanto el selección de atributos, como en la evaluación del modelo de clasificación. Sin embargo a raíz de la lectura [Belloni et al., 2000] es posible que se les pueda dar más relevancia a la forma. Una posibilidad sería la utilización de componentes principales.

Selección de atributos. Ya entonces no todos los atributos tenían la misma importancia, siendo confirmados los resultados más adelante mediante la evaluación de los atributos con el modelo de clasificación.

Los atributos derivados de aplicar los filtros pueden ser eliminados, ya que

Lomb-Scargle se muestra suficientemente eficaz para encontrar todas las frecuencias, ademas, estos atributos han resultado linealmente dependientes en la sección 4.3.

La fase "*PeakPhaseShift*", no fue seleccionado dentro del conjunto óptimo de atributos, y tampoco resulta útil en la clasificación de forma individual. Sin embargo su eliminación del modelo no hace que este mejore sino lo contrario. $E_{m-SA1} < E_{m-SA2}$ y $P(D_{SA1}|h_{SA1}) > P(D_{SA2}|h_{SA2})$, además la distancia de Mahalanobis parece mayor. Ahora bien para decidir si quitarlo o dejarlo hay que valorar el aumento en la complejidad de las reglas.

Se han generado algunos modelados más, solo con la señal o solo con los colores, se obtienen valores similares, y grupos similares.

Segmentación. Las distintas longitudes se han realizado una estandarización de las muestras más largas. Está ha resultado adecuada, y en las nuevas muestras se han podido encontrar patrones, en el apartado 3.2.4 se justificaba la elección de como se había realizado.

Sin embargo, tras presentar los datos a la ESAC, se han pedido distintos experimentos con el fin de encontrar distinta información, que explicaremos más adelante en las futuras líneas de trabajo.

6.3. Conclusiones Modelo de Agrupamiento

Los requisitos encontrados para el agrupamiento era encontrar un método de agrupamiento que no tuviera que tener el número de grupos predefinidos, debía trabajar con grandes volúmenes de datos y atributos continuos.

EM cumplía con estos requisitos, y con los resultados obtenidos se puede decir que ha cumplido con su objetivo. Los grupos guardan algunas semejanzas con la clasificación manual, siendo distintas las agrupaciones, la verosimilitud es alta, compararlo con el modelo de clasificación este confirma los resultados por sus semejanzas, y la valoración de la ESAC ha sido positiva.

Hay que destacar en la ejecución principal.

- La clase ρ ha quedado repartida en distinto grupos, sin transiciones (ver 3.1. Lo que indica la existencia de dos patrones distintos donde se creía que había uno.
- La clase χ es la más abundante, siendo prácticamente ruido, se ha dividido en distintos grupos. Analizando algunos de estos patrones vemos que se basan en $PeakPeriod$, el cual no se corresponde con una frecuencia significativa.

- La clase ϕ es una clase de transición, una secuencia de grupos cuya clase principal sea ϕ suele venir de un grupo con clases χ o saltar a ellas.
- Existen clases tan cercanas entre ellas, que por los datos se pueden considerar parte de un mismo grupo. Es el caso de υ y α o δ y θ
- Ademas se han encontrado abundantes falsos χ con bastantes frecuencias significativas.

En otras ejecuciones se consiguen datos similares, aunque variando ligeramente el número de grupos.

6.4. Conclusiones Modelo de Clasificación

En la clasificación nos hemos encontrados nuevos requisitos. Junto a los ya existentes en el agrupamiento hay que añadir que los grupos se encuentran en una secuencia.

La utilización del modelo de clasificación es doble, por un lado obtener un modelo predictivo es de utilidad en la comprensión de los patrones, pero sobre todo para poder evaluar los resultados obtenidos en el agrupamiento, y por tanto de forma indirecta el modelo de agrupamiento que obtiene dichos resultados.

También se ha hecho una utilización de la clasificación para realizar una nueva evaluación de los atributos basadas en el modelo.

- En base a los resultados obtenidos en el agrupamiento se puede concluir que atributos no tienen valor o muy poco valor por si solos. No existe ninguno que destaque como clasificador individual.
- El bajo error del modelo a la hora de predecir los grupos, y la alta verosimilitud con probabilidades similares a EM, también confirman que el modelo de agrupamiento ha funcionado correctamente. El error hubiera sido mayor en el caso de haber funcionado mal alguno de los dos modelos, dando falsas hipótesis como resultado. Si estos resultados los comparamos con los obtenidos del modelo generado con la clasificación inicial, el resultado es que se produce una mejora sustancial, las causas están descritas en la sección. 5.6.

6.5. Recomendaciones

En base a los resultados obtenidos podemos realizar las siguientes recomendaciones.

1. Introducir mecanismos para no calcular *PeakPeriod* en caso de que la frecuencia no sea suficientemente significativa.

2. Aumentar en todo lo posible la precisión en la búsqueda de estados intermedios. Esto se consigo disminuyendo el desplazamiento que aplicamos al aplicar una segmentación de tiempo en las muestras.

3. Buscar atributos que permitan una mejor definición de la forma de la señal y de HR.

4. Introducir mecanismos de evaluación basados en coste, como matrices de confusión o ROC. El objetivo sería no perder grupos con poco representativos.

5. Introducir mecanismo para la evaluar la complejidad de la hipótesis.

6.6. Lineas de trabajo futuro

Tras la evaluación de los resultados, y presentarlos, surgen varias lineas de trabajo. Algunas de ellas inmediatas.

Se han realizado algunas ejecuciones alternativas de los métodos de aprendizaje, otras han quedado pendientes de ejecutarse, el objetivo es centrarse en la búsqueda de patrones sobre algunos elementos a petición de los expertos.

- Analizar más profundamente el modelo basado en la señal completa, compararlo con el obtenido con la clasificación, con la ejecución original y con el modelo basado en los colores.

- Analizar más profundamente el modelo basado en los colores, compararlo con el obtenido con la clasificación, con la ejecución original y con el modelo basado en la señal completa.

- Se cree que existen clases con estados intermedios difíciles de detectar, esto se concluye tras una entrevista con los expertos. Para ello se disminuirá específicamente el desplazamiento en estas muestras. Generando un modelo de agrupación y un modelo de clasificación solo con muestras de clases como χ, ν o α, y disminuyendo el desplazamiento en la segmentación.

Inclusión de mecanismo de evaluación basada en el coste. No todos los errores tienen el mismo coste, para manejar esto es necesario generar una matriz de confusión [Hernández Orallo et al., 2004], ROC, o mecanismos similares.

Existen métodos para la eliminación de ruidos, algunos pertenecientes a los métodos bio-inspirados. Aunque lomb-scargle funciona bien con ruido, esto puede mejorar el análisis de las clases. Existían abundantes muestras clasificadas como clase χ que no eran solo ruido, algunas se pueden corresponder a estados intermedios sobre la muestra principal, en cualquier caso si se consiguiera eliminar el ruido este tipo de errores o estados se podría ver visualmente.

Para tomar una decisión sobre cuál es el mejor modelo falta un último elemento, que es el cálculo de la complejidad en la hipótesis. Es por ello que es necesario medir complejidad, e introducir mecanismos que nos permitan encontrar un equilibrio entre el modelo con mejor resultados y menos complejidad en las hipótesis.

Anexos

ANEXOS A

Listado Completo de Atributos

ID	Atributos	Descripción
		Número de fila.
	ID	Identificador de la muestra
	N.ALL	Longitud de la muestra sin tratar.
	Nyquist	Resultado de Nyquist
	MinPeriod	Periodo mínimo donde se buscan frecuencias significativas.
	MaxPeriod	Periodo máximo donde se buscan frecuencias significativas.
	MaxFrequency	Frecuencia máxima donde se buscan frecuencias significativas.
	Nindependent	Número de frecuencias independientes.
	M	Iteraciones de búsqueda de frecuencias significativas.
	lengthTestFrequencies	Número de frecuencias buscadas en el intervalo.
	PeakIndex	Índice de la frecuencia significativa entre de las frecuencias de test.
1	PeakSPD	Valor de Lomb-Scargle de la frecuencia seleccionada.
2	PeakPeriod	Periodo de la frecuencia seleccionada.
3	PeakPvalue	Probabilidad de falsa alarma de la frecuencia seleccionada.
	N	N.ALL después de eliminar tramos cortados.
	SignFreq2	Número de frecuencias con probabilidad de falsa alarma menor de 0,5.
4	SignFreq2	Número de frecuencias con probabilidad de falsa alarma menor de 0,2.

Cuadro A.1: Tabla de atributos seleccionados

ID	Atributos	Descripción
5	SignFreq1	Número de frecuencias con probabilidad de falsa alarma menor de 0,1.
	MaxSignPeriod2	Frecuencia Significativa con falsa alarma menor de 0,2 y máximo periodo.
	MaxSingPeriod1	Frecuencia Significativa con falsa alarma menor de 0,1 y máximo periodo.
6	Mean	Media de la señal.
7	MeanDeviation	Desviación media de la señal.
8	Variance	Varianza de la señal.
9	Deviation	Desviación típica de la señal.
10	Median	Median de la señal.
11	Mad	Desviación Absoluta Mediana de la señal.
12	Quantile000	Cuantiles 0,00 de la señal.
13	Quantile025	Cuantiles 0,25 de la señal.
14	Quantile050	Cuantiles 0,50 de la señal.
15	Quantile075	Cuantiles 0,75 de la señal.
16	Quantile100	Cuantiles 1 de la señal.
17	highCut5	Corte cinco veces por encima de MAD.
18	lowCut5	Corte cinco veces por debajo de MAD.
19	highCut2	Corte dos veces por encima de MAD.
20	lowCut2	Corte dos veces por debajo de MAD.
21	Num5MAD	Puntos de la muestra que salen del corte 17 y 18.
22	Num2MAD	Puntos de la muestra que salen del corte 19 y 20.
23	kurtosis	Curtosis de la señal.
24	skewness	Skewness de la señal.
25	jarque.statistic	Jarque-Bera de la señal.
26	jarque.p.value	Jarque-Bera valor probabilístico de la señal.
	jarque.alternative	Alternativa a Jarque.
	jarque.method	Método de Jarque utilizado.
27	PeakAmplitude	Amplitud de la frecuencia seleccionada por Lomb-Scargle.
28	PeakPhaseShift	Fase de la frecuencia seleccionada por Lomb-Scargle.
	Index2	Índice de la segunda frecuencia más significativa entre de las frecuencias de test.
	SPD2	Valor de Lomb-Scargle de la segunda frecuencia seleccionada.
	Period2	Periodo de la segunda frecuencia seleccionada.
	Frequency2	Frecuencia de la segunda frecuencia seleccionada.
	Pvalue2	Probabilidad de falsa alarma de la segunda frecuencia seleccionada.
	Amplitude2	Amplitud de la segunda frecuencia seleccionada por Lomb-Scargle.

Cuadro A.1: Tabla de atributos seleccionados

ID	Atributos	Descripción
	PhaseShift2	Fase de la segunda frecuencia seleccionada por Lomb-Scargle.
	Index2	Índice de la segunda frecuencia más significativa entre de las frecuencias de test.
	SPD3	Valor de Lomb-Scargle de la tercera frecuencia seleccionada.
	Period3	Periodo de la tercera frecuencia seleccionada.
	Frequency3	Frecuencia de la tercera frecuencia seleccionada.
	Pvalue3	Probabilidad de falsa alarma de la tercera frecuencia seleccionada.
	Amplitude3	Amplitud de la tercera frecuencia seleccionada por Lomb-Scargle.
	PhaseShift3	Fase de la tercera frecuencia seleccionada por Lomb-Scargle.
29	HR1.Mean	Media de HR1.
30	HR1.MeanDeviation	Desviación media de HR1.
31	HR1.Variance	Varianza de HR1.
32	HR1.Deviation	Desviación tipica de HR1.
33	HR1.Median	Median de HR1.
34	HR1.Mad	Desviación Absoluta Mediana de HR1.
35	HR1.Quantile000	Cuantiles 0,00 de HR1.
36	HR1.Quantile025	Cuantiles 0,25 de HR1.
37	HR1.Quantile050	Cuantiles 0,50 de HR1.
38	HR1.Quantile075	Cuantiles 0,75 de HR1.
39	HR1.Quantile100	Cuantiles 1 de HR1.
40	HR1.kurtosis	Curtosis de HR1.
41	HR1.skewness	Skewness de HR1.
42	HR1.jarque.statistic	Jarque-Bera de HR1.
43	HR1.jarque.p.value	Jarque-Bera valor probabilístico de HR1.
	HR1.jarque.alternative	Alternativa a Jarque.
	HR1.jarque.method	Método de Jarque utilizado.
44	HR2.Mean	Media de HR1.
45	HR2.MeanDeviation	Desviación media de HR2.
46	HR2.Variance	Varianza de HR2.
47	HR2.Deviation	Desviación tipica de HR2.
48	HR2.Median	Median de HR2.
49	HR2.Mad	Desviación Absoluta Mediana de HR2.
50	HR2.Quantile000	Cuantiles 0,00 de HR2.
51	HR2.Quantile025	Cuantiles 0,25 de HR2.
52	HR2.Quantile050	Cuantiles 0,50 de HR2.
53	HR2.Quantile075	Cuantiles 0,75 de HR2.
54	HR2.Quantile100	Cuantiles 1 de HR2.
55	HR2.kurtosis	Curtosis de HR2.

Cuadro A.1: Tabla de atributos seleccionados

ID	Atributos	Descripción
56	HR2.skewness	Skewness de HR2.
57	HR2.jarque.statistic	Jarque-Bera de HR2.
58	HR2.jarque.p.value	Jarque-Bera valor probabilístico de HR2.
	HR2.jarque.alternative	Alternativa a Jarque.
	HR2.jarque.method	Método de Jarque utilizado.
	PeakIndex.low	Índice de la frecuencia significativa entre de las frecuencias de test tras aplicar un filtro paso bajo.
	PeakSPD.low	Valor de Lomb-Scargle de la frecuencia seleccionada tras aplicar un filtro paso bajo.
59	PeakPeriod.low	Periodo de la frecuencia seleccionada tras aplicar un filtro paso bajo.
	PeakPvalue.low	Probabilidad de falsa alarma de la frecuencia seleccionada tras aplicar un filtro paso bajo.
	SignFreq5.low	Número de frecuencias con probabilidad de falsa alarma menor de 0,5 tras aplicar un filtro paso bajo.
60	SignFreq2.low	Número de frecuencias con probabilidad de falsa alarma menor de 0,2 tras aplicar un filtro paso bajo.
61	SignFreq1.low	Número de frecuencias con probabilidad de falsa alarma menor de 0,1 tras aplicar un filtro paso bajo.
62	SignFreq2.low.boolean	Indica si hay frecuencias con probabilidad de falsa alarma menor de 0,2 tras aplicar un filtro paso bajo.
63	SignFreq1.low.boolean	Indica si hay frecuencias con probabilidad de falsa alarma menor de 0,1 tras aplicar un filtro paso bajo.
	PeakAmplitude.low	Amplitud de la frecuencia seleccionada por Lomb-Scargle tras aplicar un filtro paso bajo.
	PeakPhaseShift.low	Fase de la frecuencia seleccionada por Lomb-Scargle tras aplicar un filtro paso bajo.
	PeakIndex.high	Índice de la frecuencia significativa entre de las frecuencias de test tras aplicar un filtro paso alto.
	PeakSPD.high	Valor de Lomb-Scargle de la frecuencia seleccionada tras aplicar un filtro paso alto.
64	PeakPeriod.high	Periodo de la frecuencia seleccionada tras aplicar un filtro paso alto.
	PeakPvalue.high	Probabilidad de falsa alarma de la frecuencia seleccionada tras aplicar un filtro paso alto.
65	PeakPvalue.high	Probabilidad de falsa alarma de la frecuencia seleccionada tras aplicar un filtro paso alto.
	SignFreq5.high	Número de frecuencias con probabilidad de falsa alarma menor de 0,5 tras aplicar un filtro paso alto.
66	SignFreq2.high	Número de frecuencias con probabilidad de falsa alarma menor de 0,2 tras aplicar un filtro paso alto.

Cuadro A.1: Tabla de atributos seleccionados

ID	Atributos	Descripción
67	SignFreq1.high	Número de frecuencias con probabilidad de falsa alarma menor de 0,1 tras aplicar un filtro paso alto.
68	SignFreq2.high.boolean	Indica si hay frecuencias con probabilidad de falsa alarma menor de 0,2 tras aplicar un filtro paso alto.
69	SignFreq1.high.boolean	Indica si hay frecuencias con probabilidad de falsa alarma menor de 0,1 tras aplicar un filtro paso alto.
	PeakAmplitude.high	Amplitud de la frecuencia seleccionada por Lomb-Scargle tras aplicar un filtro paso alto.
	PeakPhaseShift.high	Fase de la frecuencia seleccionada por Lomb-Scargle tras aplicar un filtro paso alto.
	DOUBT	Indica si hubo dudas en la clasificación manual.
	CLASS_2	Segunda clase a la que puede pertenecer la muestra.
	CLASS_3	Tercera clase a la que puede pertenecer la muestra.
	CLASS_4	Cuarta clase a la que puede pertenecer la muestra.
	CLASS	Clase a la que pertenece la muestra.
	1	Probabilidad de pertenecer al grupo 1.
	2	Probabilidad de pertenecer al grupo 2.
	3	Probabilidad de pertenecer al grupo 3.
	4	Probabilidad de pertenecer al grupo 4.
	5	Probabilidad de pertenecer al grupo 5.
	6	Probabilidad de pertenecer al grupo 6.
	7	Probabilidad de pertenecer al grupo 7.
	8	Probabilidad de pertenecer al grupo 8.
	9	Probabilidad de pertenecer al grupo 9.
	10	Probabilidad de pertenecer al grupo 10.
	11	Probabilidad de pertenecer al grupo 11.
	12	Probabilidad de pertenecer al grupo 12.
	Cluster	Grupo con la probabilidad más alta de contener a la muestra.
70	CLASS	Clasificación de la muestra.

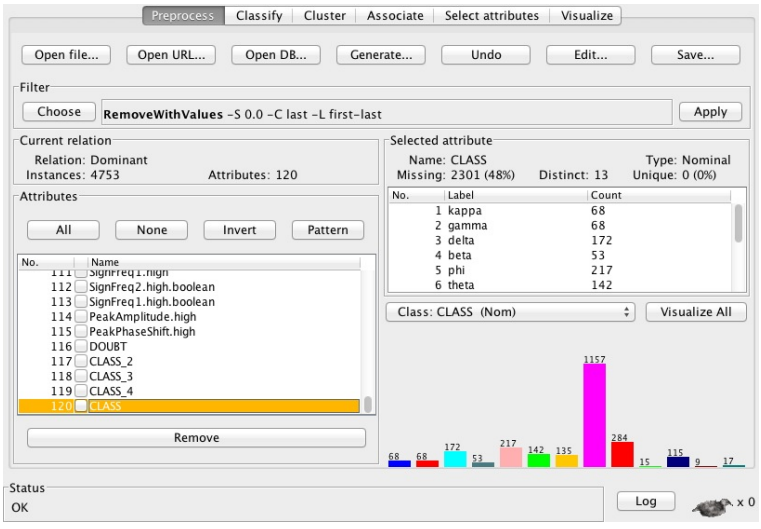
Cuadro A.1: Tabla de atributos seleccionados

ANEXOS B

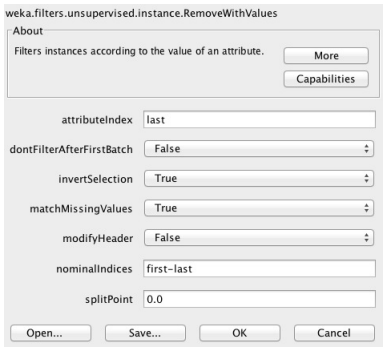
Ejecución de Selección de Atributos

Para métodos de selección y evaluación de atributos se ha utilizado la herramienta WEKA (http://www.cs.waikato.ac.nz/ml/weka/).

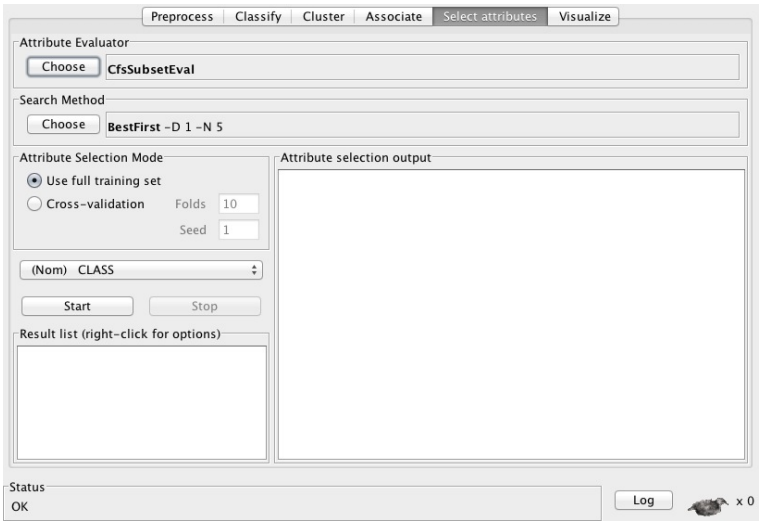
Como entrada se ha utilizado un conjunto de entrenamiento donde ya se han eliminado instancias con algún atributo sin valor. Sobre este conjunto nos quedamos con las instancias clasificadas.



Para ello tras abrir el archivo con la opción *"Open File..."*, seleccionamos la opción *"Choose"*, que permite seleccionar distintos tipos de filtrado. Seleccionando el filtro *"filters.unsupervised.instance. RemoveWithValues"*. Y lo aplicamos.



Para terminar vamos a la pestaña *"SelectAttributes"* .



Para la ejecución de los distintos métodos, combinamos *"AttributeEvaluator"* (métodos de evaluación) con *"SearchMethod"* métodos de búsqueda.

- *AttributeEvaluator = CfsSubSetEval - SearchMethod = BestFirst.*
- *AttributeEvaluator = WrapperSubSetEval - SearchMethod = BestFirst.* En este caso pulsando a la derecha del botón "Choose" en la palabra "WrapperSubSetEval". Saldrá una nueva ventana, el botón "Choose" a la derecha de "classifier" nos permite seleccionar J48 y Bayes Net como clasificador.
- *AttributeEvaluator" = RefieFAttributeEval- "SearchMethod" = Ranker.*
- *AttributeEvaluator" = InfoGainAttribute- "SearchMethod" = Ranker.*
- *AttributeEvaluator" = GainRatioAttributeEval - "SearchMethod" = Ranker.*

ANEXOS C

Ejecución de Agrupamiento y Clasificación

Existen varios scripts desarrollados para la ejecución de los métodos y algoritmos implementados en el trabajo. En este anexo se realiza una breve descripción de como ejecutarlos y librerías en las que se basan.

Librerías Utilizadas

- **Ficheros FITs light curve.** Para el manejo y lectura de estos ficheros se ha usado el paquete "FITSio".
- **Selección de atributos.** Para el tratamiento de la señal y los colores.
  - "LombScargle.R". Calculo de Lomb-Scargle, modificado para que genere ademas atributos como amplitud, fase, número de frecuencias significativas, etc.
  - Paquete "signal". Filtros paso alto y filtro paso bajo.
  - Paquete "moments". Resumen estadístico, incluyendo curtosis y skewness.
- **Modelado Agrupamiento.** Se ha utilizado el paquete "mclust". Para el análisis de las distancias se ha utilizado el paquete "ecodist".
- **Modelado Clasificación.** Se ha utilizado el paquete "RHmm", y para el calculo de las funciones de densidad necesarias se ha utilizado el paquete "mvnmle".

Análisis de datos iniciales

Para el análisis inicial de los datos es necesario lanzar el comando.

Analiza las muestras proporcionadas

```
summaryFitsSamples( PATH_LC,
                PATH_RESULT,
                FILE_CLASS )
```

Entradas:

- **PATH_LC**: Ruta donde se guardan los ficheros de curvas de luz.
- **PATH_RESULT**: Ruta donde guarda los resultados.
- **FILE_CLASS**: Fichero con la relación muestra-classes a las que pertenecen.

Salidas:

- **Fichero %_PATH_RESULT_%/HistogramTime.jpg**. Imagen con el histograma de los tiempos.
- **Fichero %_PATH_RESULT_%/HistogramLostSignal.jpg**. Imagen con el histograma de los puntos de muestreo con valor nulo.
- **Fichero %_PATH_RESULT_%/summary.txt** con la siguiente información.
  - **num_lc_samples**: Número de muestras totales.
  - **num_completed_lc_samples**: Número de muestras con todos los ficheros, tres de color más el total.
  - **num_lc_samples_class**: Número de muestras clasificadas sobre el conjunto total.
  - **distribution_class**: Distribución de las muestras totales en las clases sobre el conjunto total.
  - **num_lc_samples_class_no_doubt**: Número de muestras clasificadas sin dudas sobre el conjunto total.
  - **distribution_class_no_doubt**: Distribución de las muestras totales en las clases sin dudas.
  - **num_complete_lc_samples_class**: Número de muestras clasificadas sobre el conjunto con de muestras completas.
  - **distribution_complete_class**: Distribución de las muestras completas en las clases.

- **num_complete_lc_samples_class_no_doubt**: Número de muestras clasificadas sin dudas sobre el conjunto completo.
- **distribution_complete_class_no_doubt**: Distribución de las muestras completas en las clases sin dudas.
- **total_time**: Tiempo en segundos de muestro.
- **total_points**: Número total de puntos de muestreo
- **total_lost**: Número total de puntos de muestreo cuyo valor es nulo.

Ejemplo de ejecución.

```
PATH_LC <- '/Users/Uned/GRS1915_data_lc'
PATH_RESULT_SUMMARY <- '/Users/Uned/GRS1915_RESULT_SUMMARY'
FILE_CLASS <- '/Users/Uned/GRS1915_data/samples.csv'

summary <- summaryFitsSamples(
                PATH_LC = PATH_LC,
                PATH_RESULT = PATH_RESULT_SUMMARY,
                FILE_CLASS = FILE_CLASS )
```

Extracción de Información

Existen dos funciones creadas para el tratamiento de los atributos.

Generación de atributos.

```
processLC( PATH_LC,
                PATH_RESULT,
                FILE_CLASS,
                PointSegment,
                LengthSegment,
                CHARTS )
```

Entradas:

- **PATH_LC**: Ruta donde se guardan los ficheros de curvas de luz.
- **PATH_RESULT**: Ruta donde guarda los resultados.
- **FILE_CLASS**: Fichero con la relación muestra-classes a las que pertenecen.
- **PointSegment**: Puntos de segmentación, cada cuantos segundos se realiza un corte.
- **LengthSegment**: Longitud de la segmentación.
- **CHARTS**: Booleano que indica si se crea o no las gráficas de cada muestra.

Salidas:

- **%_PATH_RESULT_%/Dominant.csv**: Cuyo contenido viene descrito en el anexo A.
- **%_PATH_RESULT_%/HistogramPeriod.csv**: Con un listado de las muestras con sus periodos más significativos.

Limpieza de datos.

```
cleanData(PATH_LC,
                FILE_NAME,
                FILE_RESULT,
                LengthSegment)
```

Entradas:

- **PATH_LC**: Ruta donde se guardan los ficheros de curvas de luz.
- **FILE_NAME**: Fichero con los atributos extraídos.
- **FILE_RESULT**: Fichero con la limpieza de datos.
- **LengthSegment**: Longitud de la segmentación. Eliminara segmentos muy pequeños.

Salidas:

- **FILE_RESULT**: Fichero con la limpieza de datos.

Ejemplo de ejecución.

```
PATH_LC <- '/Users/Uned/GRS1915_data_lc'
PATH_RESULT <- '/Users/Uned/GRS1915_RESULT_2000_1000'
FILE_CLASS <- '/Users/Uned/GRS1915_data/samples.csv'
PointSegment = 1000,
LengthSegment <- 2000,


# Generacion de atributos.
processLC(PATH_LC = PATH_LC,
                PATH_RESULT = PATH_RESULT,
                FILE_CLASS = FILE_CLASS,
                PointSegment = PointSegment,
                LengthSegment = LengthSegment,
                CHARTS = TRUE)

FILE_RESULT <- "Domaint.clean.csv"
```

```
# Limpieza de datos.
cleanData(PPATH_LC = PATH_LC,
                FILE_NAME = FILE_NAME,
                FILE_RESULT = FILE_RESULT,
                LengthSegment = LengthSegment)
```

Ejecución del modelo de Agrupamiento.

Existe una función principal para la ejecución y varias funciones encargadas de analizar los resultados.

Generación de grupos.

```
clustering( FILE_ATTRIBUTE,
                FILE_CLUSTERS,
                FILE_MEAN_CLUSTERS,
                FILE_EM,
                Min_NClusters,
                Max_NClusters,
                ModelNames,
                SELECTED_ATTRIBUTE,
                DEL_ROW_1 )
```

Entradas:

- **FILE_ATTRIBUTE**: Fichero obtenido de la selección de atributos.
- **Min_NClusters**: Mínimo de grupos buscados.
- **Max_NClusters**: Máximo de grupos buscados.
- **ModelNames**: Forma de la función buscada.
- **SELECTED_ATTRIBUTE**: Atributos seleccionados, dentro de FILE_ATTRIBUTE para el agrupamiento.
- **DEL_ROW_1**: Booleano que indica si es necesario eliminar la primera columna, en caso que el fichero contenga un identificador de file.

Salidas:

- **FILE_CLUSTERS**: Fichero de salida con los grupos obtenidos.
- **FILE_MEAN_CLUSTERS**: Fichero de salida con la media de los atributos de cada grupo.
- **FILE_EM**: Fichero de salida con los grupos, tras la segunda ejecución.

Validación de los grupos.

```
validateClustering( FILE_CLUSTERS,
                    FILE_OUTPUT,
                   SELECTED_ATTRIBUTE,
                   DEL_ROW_1 )
```

Entradas:

- **FILE_CLUSTERS**: Fichero con los grupos obtenidos.
- **SELECTED_ATTRIBUTE**: Atributos seleccionados, dentro de FILE_ATTRIBUTE para el agrupamiento.
- **DEL_ROW_1**: Booleano que indica si es necesario eliminar la primera columna, en caso que el fichero contenga un identificador de file.

Salidas:

- **FILE_OUTPUT**: Fichero de salida con distintas medidas de validación.

Validación de los grupos.

```
validateClustering( FILE_CLUSTERS,
                    FILE_OUTPUT,
                   SELECTED_ATTRIBUTE,
                   DEL_ROW_1 )
```

Entradas:

- **FILE_CLUSTERS**: Fichero con los grupos obtenidos.
- **SELECTED_ATTRIBUTE**: Atributos seleccionados, dentro de FILE_ATTRIBUTE para el agrupamiento.
- **DEL_ROW_1**: Booleano que indica si es necesario eliminar la primera columna, en caso que el fichero contenga un identificador de file.

Salidas:

- **FILE_OUTPUT**: Fichero de salida con distintas medidas de validación.

Relación grupo-clase (incluyendo dudosas).

```
analyzeClusterDoubt( FILE\_CLUSTERS,
                    FILE\_OUT)
```

Entradas:

- **FILE_CLUSTERS**: Fichero con los grupos obtenidos.

Salidas:

- **FILE_OUTPUT**: Fichero de salida con la relación entre grupos y clases.

Relación grupo-clase.

```
analyzeCluster( FILE\_CLUSTERS,
                FILE\_OUT)
```

Entradas:

- **FILE_CLUSTERS**: Fichero con los grupos obtenidos.

Salidas:

- **FILE_OUTPUT**: Fichero de salida con la relación entre grupos y clases.

Obtener probabilidad media.

```
analyzeVerisimilitude ( FILE_CLUSTERS,
                FILE_OUT )
```

Entradas:

- **FILE_CLUSTERS**: Fichero con los grupos obtenidos.

Salidas:

- **FILE_OUTPUT**: Fichero de salida con la relación entre grupos y clases.

Distancia de Mahalanobis de las muestras.

```
getMatchPerfectSamples( FILE_CLUSTERS,
                PATH_RESULT,
                FILE_MEAN_CLUSTERS,
                SELECTED_ATTRIBUTE,
                DEL_ROW_1 )
```

Entradas:

- **FILE_CLUSTERS**: Fichero con los grupos obtenidos.
- **FILE_MEAN_CLUSTERS**: Fichero con la media de los atributos de cada grupo.
- **SELECTED_ATTRIBUTE**: Atributos seleccionados, dentro de FILE_ATTRIBUTE para el agrupamiento.
- **DEL_ROW_1**: Booleano que indica si es necesario eliminar la primera columna, en caso que el fichero contenga un identificador de file.

Salidas:

- **PATH_RESULT**: Path donde se guardaran los ficheros con las muestras ordenadas por distancia de Mahalanobis, con el nombre *Mahalanobis_N_.csv*.

Distancia de Mahalanobis entre grupos.

```
ProcessMahalanobis( FILE_MEAN_CLUSTERS ,
                FILE_MAHALANOBIS_CLUSTER )
```

Entradas:

- **FILE_MEAN_CLUSTERS**: Fichero con la media de los atributos de cada grupo.
- **DEL_ROW_1**: Booleano que indica si es necesario eliminar la primera columna, en caso que el fichero contenga un identificador de file.

Salidas:

- **FILE_MAHALANOBIS_CLUSTER**: Fichero donde se guarda la distancia de Mahalanobis entre cada grupo.

Lista de arcos.

```
createArches ( FILE_CLUSTERS,
                FILE_RESULT ,
                DEL_ROW_1)
```

Entradas:

- **FILE_MEAN_CLUSTERS**: Fichero con la media de los atributos de cada grupo.
- **DEL_ROW_1**: Booleano que indica si es necesario eliminar la primera columna, en caso que el fichero contenga un identificador de file.

Salidas:

- **FILE_RESULT**: Fichero de salida con los arcos.

Lista grupos por las que pasa cada muestra.

```
createPatterns( FILE_CLUSTERS ,
                FILE_RESULT ,
                DEL_ROW_1 )
```

Entradas:

- **FILE_MEAN_CLUSTERS**: Fichero con la media de los atributos de cada grupo.
- **DEL_ROW_1**: Booleano que indica si es necesario eliminar la primera columna, en caso que el fichero contenga un identificador de file.

Salidas:

- **FILE_RESULT**: Fichero de salida con las muestras y los grupos que contiene.

Obtiene matriz de adyacencia.

```
createAdjacencyMatrix( FILE_ARCHES ,
                FILE_RESULT_ADJACENCY,
                FILE_RESULT_PADJACENCY,
                FILE_RESULT_TRANS_PROBS,
                DEL_ROW_1 )
```

Entradas:

- **FILE_ARCHES**: Fichero con los arcos.
- **DEL_ROW_1**: Booleano que indica si es necesario eliminar la primera columna, en caso que el fichero contenga un identificador de file.

Salidas:

- **FILE_RESULT_ADJACENCY**: Fichero salida con la matriz de adyacencia.
- **FILE_RESULT_PADJACENCY**: Fichero salida con la matriz de adyacencia traspuesta.
- **FILE_RESULT_TRANS_PROBS**: Fichero salida con la matriz de probabilidad de transacciones.

Ordena los gráficos por grupo y distancia de Mahalanobis.

```
getBestSamplesClusters ( FILE_CLUSTER,
                PATH_RESOURCE,
                PATH_CHARTS,
                PATH_OUTPUT )
```

Entradas:

- **FILE_CLUSTERS**: Fichero con los grupos obtenidos.
- **PATH_RESOURCE**: Path donde se encuentran los ficheros con las distancia de Mahalanobis.
- **PATH_CHARTS**: Path donde se guardan las gráficas.
- **DEL_ROW_1**: Booleano que indica si es necesario eliminar la primera columna, en caso que el fichero contenga un identificador de file.

Salidas:

- **PATH_OUTPUT**: Carpeta donde se guarda los resultados.

Ejemplo de ejecución.

```
FILE_ATTRIBUTE_CLEAN <- paste(PATH_RESULT,
                 '/Dominant_clean.b.csv', sep="")
FILE_CLUSTERS <- paste(PATH_RESULT,
                 '/Dominant_clean.b_MClust_SA1.csv', sep="")
FILE_MEAN_CLUSTERS <- paste(PATH_RESULT,
                 '/MeanClusters_SA1.csv', sep="")
FILE_EM <- paste(PATH_RESULT,
                 '/Dominant_clean_EM_SA1.csv', sep="")

SELECT_ATRIBUTE_1 <- c("ID", "PeakPeriod", "SignFreq1" ,
                 "Median", "Mad", "kurtosis", "skewness",
                 "PeakAmplitude", "PeakPhaseShift",
                 "HR1.Median", "HR1.Mad", "HR1.kurtosis", "HR1.skewness",
                 "HR2.Median", "HR2.Mad", "HR2.kurtosis", "HR2.skewness")
SELECT_ATRIBUTE_1_CLUSTER <- c(SELECT_ATRIBUTE_1, "Cluster")

clustering(FILE_ATTRIBUTE = FILE_ATTRIBUTE_CLEAN,
                 FILE_CLUSTERS = FILE_CLUSTERS,
                 FILE_MEAN_CLUSTERS = FILE_MEAN_CLUSTERS,
                 FILE_EM = FILE_EM,
                 Min_NClusters = 1,
                 Max_NClusters = 15,
                 ModelNames = "VVV",
                 SELECTED_ATTRIBUTE = SELECT_ATRIBUTE_1,
                 DEL_ROW_1 = TRUE)

FILE_OUTPUT_VALIDATE_CLUSTER <- paste(PATH_RESULT,
                '/validate_cluster_SA1.txt', sep="")

validateClustering( FILE_CLUSTERS = FILE_CLUSTERS,
```

```
                FILE_OUTPUT = FILE_OUTPUT_VALIDATE_CLUSTER,
                SELECTED_ATTRIBUTE = SELECT_ATRIBUTE_1,
                DEL_ROW_1 = TRUE )

# distribucion de las clases, incluidas dudosas, dentro de los grupos.
analyzeClusterDoubt( FILE_CLUSTERS = FILE_CLUSTERS,
                FILE_OUT = FILE_CLASS_CLUSTER_D)

# Analiza distribucion de las clases dentro de los grupos.
analyzeCluster( FILE_CLUSTERS = FILE_CLUSTERS,
                FILE_OUT = FILE_CLASS_CLUSTER)

FILE_VERI <- paste(PATH_RESULT_MAHALANOBIS, '/Verisimilitude.txt', sep="")
analyzeVerisimilitude ( FILE_CLUSTERS = FILE_CLUSTERS,
               FILE_OUT =  FILE_VERI )

getMatchPerfectSamples( FILE_CLUSTERS = FILE_CLUSTERS,
                PATH_RESULT = PATH_RESULT_MAHALANOBIS,
                FILE_MEAN_CLUSTERS = FILE_MEAN_CLUSTERS,
                SELECTED_ATTRIBUTE = SELECT_ATRIBUTE_1_CLUSTER,
                DEL_ROW_1 = TRUE )

ProcessMahalanobis( FILE_MEAN_CLUSTERS = FILE_MEAN_CLUSTERS ,
                FILE_MAHALANOBIS_CLUSTER = FILE_MAHALANOBIS_CLUSTER)


FILE_ARCHES <- paste(PATH_RESULT_MAHALANOBIS, "/SA1Arches.csv", sep="")

# Crea fichero con los arcos.
createArches ( FILE_CLUSTERS = FILE_CLUSTERS ,
                FILE_RESULT = FILE_ARCHES ,
                DEL_ROW_1 = TRUE)


# Crea secuencia de estados para cada muestras
FILE_PATTERNS <- paste(PATH_RESULT_MAHALANOBIS, "/SA1Patterns.csv", sep="")
createPatterns( FILE_CLUSTERS = FILE_CLUSTERS ,
                FILE_RESULT = FILE_PATTERNS ,
                DEL_ROW_1 = TRUE)

# Crea matriz de adyacencia
FILE_ADJACENCY <- paste(PATH_RESULT_MAHALANOBIS,
                 "/SA1AdjacencyMatrix.csv", sep="")
FILE_PADJACENCY <- paste(PATH_RESULT_MAHALANOBIS,
                "/SA1PAdjacencyMatrix.csv", sep="")
FILE_TRANS_PROBS <- paste(PATH_RESULT_MAHALANOBIS,
```

```
                "/SA1transProbs.csv", sep="")
createAdjacencyMatrix( FILE_ARCHES = FILE_ARCHES ,
                FILE_RESULT_ADJACENCY = FILE_ADJACENCY,
                FILE_RESULT_PADJACENCY = FILE_PADJACENCY,
                FILE_RESULT_TRANS_PROBS = FILE_TRANS_PROBS,
                DEL_ROW_1 = TRUE)


# Calcula la distancia de mahalanobis entre clusters.
getBestSamplesClusters ( FILE_CLUSTER = FILE_CLUSTER,
                PATH_RESOURCE = PATH_RESULT_MAHALANOBIS,
                PATH_CHARTS = PATH_CHARTS,
                PATH_OUTPUT = paste(PATH_RESULT_MAHALANOBIS, "/charts", sep="") )
```

Ejecución del modelo de Clasificación

Para la clasificación se han generado las siguientes funciones.

Modelo Oculto de Markov.

```
processRHMM(
                FILE_PATTERNS = FILE_PATTERNS,
                FILE_TRANS = FILE_TRANS_PROBS,
                FILE_CLUSTERS = FILE_CLUSTERS,
                FILE_RESULT_INIT_PROB = FILE_INIT_PROB,
                CHART_RESULT = CHART_RESULT,
                DEL_ROW_1 = TRUE)
```

Entradas:

- **FILE_PATTERNS**: Fichero con las muestras y los grupos que contiene.
- **FILE_TRANS**: Fichero con la matriz de probabilidad de transacciones.
- **FILE_CLUSTERS**: Fichero con los grupos obtenidos.
- **DEL_ROW_1**: Booleano que indica si es necesario eliminar la primera columna, en caso que el fichero contenga un identificador de file.

Salidas:

- **FILE_RESULT_INIT_PROB**: Fichero con las probabilidades iniciales de cada grupo.
- **CHART_RESULT**: Carpeta donde guarda el gráfico con los resultados de la ejecución. El gráfico incluye error de muestreo y probabilidad media de que la hipótesis se cumpla.

Métodos similares. Otros métodos para crear HMM con idénticas entradas y salidas.

- **processRHMM_S2** utilizando el conjunto seleccionados de atributos 2.
- **processRHMM_S3** utilizando el conjunto seleccionados de atributos 3.
- **processRHMM_S4** utilizando el conjunto seleccionados de atributos 4.
- **processRHMMClass** utilizando las clases y el conjunto de atributos principal.
- **processRHMM.PeakPeriod** utilizando periodo.
- **processRHMM.SignFreq1** utilizando Frecuencias Significativas.
- **processRHMM.Median** utilizando la media.
- **processRHMM.Mad** utilizando MAD.
- **processRHMM.kurtosis** utilizando la curtosis.
- **processRHMM.skewness** utilizando skewness.
- **processRHMM.PeakAmplitude** utilizando la amplitud.
- **processRHMM.PeakPhaseShift** utilizando la fase.
- **processRHMM.HR1.Median** utilizando la media de HR1.
- **processRHMM.HR1.Mad** utilizando MAD de HR1.
- **processRHMM.HR1.kurtosis** utilizando la curtosis de HR1.
- **processRHMM.HR1.skewness** utilizando skewness de HR1.
- **processRHMM.HR2.Median** utilizando la media de HR2.
- **processRHMM.HR2.Mad** utilizando MAD de HR2.
- **processRHMM.HR2.kurtosis** utilizando la curtosis de HR2.
- **processRHMM.HR2.skewness** utilizando skewness de HR2.

Ejemplo de ejecución.

```
FILE_PATTERNS <- paste(PATH_RESULT_MAHALANOBIS,
                 "/SA1Patterns.csv", sep="")
FILE_TRANS_PROBS <- paste(PATH_RESULT_MAHALANOBIS,
                 "/SA1transProbs.csv", sep="")

FILE_INIT_PROB <- paste(PATH_RESULT_MAHALANOBIS,
                 "/SA1initProv.csv", sep="")

CHART_RESULT <- paste(PATH_RESULT_MAHALANOBIS,
                 "/SA4_Chart_Result.jpg", sep="")


processRHMM(
                 FILE_PATTERNS = FILE_PATTERNS,
                 FILE_TRANS = FILE_TRANS_PROBS,
                 FILE_CLUSTERS = FILE_CLUSTERS,
                 FILE_RESULT_INIT_PROB = FILE_INIT_PROB,
                 CHART_RESULT = CHART_RESULT,
                 DEL_ROW_1 = TRUE)
```

Bibliografía

[NAS, 2012] (2012). *Kepler: A Search for Terrestrial Planets. Kepler Archive Manual*. NASA.

[Bakiri and Dietterich, 2002] Bakiri, G. and Dietterich, T., editors (2002). *.ᴬchieving high-accuracy text-to-speech with machine learningïn Data Mining Techniques in Speech Synthesis.*

[Baum et al., 1970] Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). *A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains*.

[Beech, 2007] Beech, M. (2007). The classification of meteor light curves: an application of hat theory. *Beech, Martin*, 380:1649–1655.

[Belloni et al., 2000] Belloni, T., Klein-Wolt, M., Méndez, M., Van der Klis, M., and Van Paradijs, J. (2000). A model-independent analysis of the variability of gpr 1915+105.

[Birney et al., 2006] Birney, D. S., Gonzalez, G., and Oespe, D. (2006). *Observational Astronomy*. Birney,D. Scott and Gonzalez, Guillermo and Oespe, David.

[Bishop., 2006] Bishop., C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[Blomme et al., 2007] Blomme, J., Debosscher, J., et al. (2007). *Automated classification of variable stars in the asteroseismology program of the Kepler space mission*.

[Box et al., 1994] Box, G., Jenkins, G. M., and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*. Prentice-Hall.

[Cartwright et al., 2012] Cartwright, K. V., Russell, P., and Kaminsky, E. J. (2012). *Finding the maximum magnitude response (gain) of second-order filter without calculus.*

[Chapman et al., 2000] Chapman, P. N., Clinton, J. S., et al. (2000). *CRISP-DM 1.0SPSS: Modeling: CRISP DM 1.0 - Step-by-step data mining guide.* SPSS.

[Debosscher et al., 2011] Debosscher, J., Blomme, J., et al. (2011). *Global stellar variability study in the field-of-view of the Kepler satellite.* Debosscher, J. and Blomme, J. and others.

[Dietterich, 1998] Dietterich, T. (1998). Approximate statistical test for comparing supervised classification learning algorithms. *Neuronal Computation*, 10(7):1895–1924.

[Dietterich, 2002] Dietterich, T. (2002). Machine learning for sequential data: A review. *In Proc. of Structural, Syntactic, and Statistical Pattern Recognittion, Joint IAPR International Workshops SSPR SSPR/SPR 2002*, pages 15–30.

[Dodge, 2003] Dodge, Y. (2003). *The Oxford Dictionary of Statistical Terms.* OUP.

[Dodge and Rousson, 1997] Dodge, Y. and Rousson, V. (1997). *The Complications of the Fourth Central Moment.* The American Statistician.

[ESO, 2001] ESO (2001). *A Very Massive Stellar Black Hole in the Milky Way Galaxy.* ESO.

[Fawcett, 1997] Fawcett, T; Provost, F. (1997). *Data Mining and Knowledge Discovery*, pages 291–316.

[Finley, 1994] Finley, D. (1994). Micro-quasar within our galaxy!

[Forney Jr, 2005] Forney Jr, G. D. (2005). *The Viterbi Algorithm: A Personal History.*

[Garre et al., 2005] Garre, M., Cuadrado, J. J., et al. (2005). Segmented parametric software estimation models: Using the em algorithm with the isbsg 8 database. *Information Technology Interfaces.*

[Garre et al., 2007] Garre, M., Cuadrado, J. J., et al. (2007). Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software.

[Graymer, 1998] Graymer, R. (1998). *The Least Squares Spectrum, Its inverse trasnform and autocorrelation function: Theory and some aplications in geodesy.* Graymer, Ruthven.

[Greiner, 2001] Greiner, J. (2001). Grs 1915+105.

[Hand and Kamber, 2006] Hand, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques.* Morgan Kaufmann, 2nd edition edition.

[Hernández Orallo et al., 2004] Hernández Orallo, J., Ramírez Quintana, M. J., and Ferri Ramirez, C. (2004). *Introducción a la Minería de Datos*. Pearson.

[Hoaglin et al., 1983] Hoaglin, D. C., Tukey, F. M., and W., J. (1983). *Understanding Robust and Exploratory Data Analysis.* John Wiley & Sons.

[Huang et al., 2001] Huang, X., Acero, A., and Hon, H. (2001). *Statistical Methods for Speech Recognition*. Prentice-Hall.

[Ibanogammalu, 1998] Ibanogammalu, C. (1998). *Variable Stars As Essential Astrophysical Tools.* Springer.

[Irani and Fayyad, 1993] Irani, K. B. and Fayyad, U. M. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. *Proc10th IntConfMachine Learning,*, pages 194–201.

[Jarque and Bera, 1980] Jarque, C. M. and Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, (6):255–259.

[Jelinek, 1997] Jelinek, F. (1997). *Statistical Methods for Speech Recognition.* MIT Press.

[Kendall and Stuart, 1969] Kendall, M. and Stuart, A. (1969). *The Advanced Theory of Statistics, Volume 1: Distribution Theory, 3rd Edition.* Griffin.

[Kira K., 1992] Kira K., R. L. (1992). *A practical approach to feature selection.* In: Proc. Intern., Aberdeen.

[Kononenko, 1994] Kononenko, I. (1994). Estimating attributes: Analysus abd extensions of relieff.

[Larose, 2005] Larose, D. T. (2005). *Discovering Knowledge in Data. An Introduction to Data Mining*. John Wiley & Sons.

[Law et al., 2009] Law, N. M., Kulkarni, S. R., et al. (2009). The palomar transient factory: System overview, performance, and first results. *Publications of the Astronomical Society of the Pacific*, 121:1395–1408.

[Levinson, 1986] Levinson, S. E. (1986). *Continuously Variable Duration Hidden Markov Models for Speech Analysis.* AT&T Bell Laboratories.

[Liu and Dash, 1997] Liu, M. and Dash, H. (1997). *Feature Selection for Classification. Intelligent Data Analysis,*, pages 131–156. Number 1. Elsevier.

[M. Ball and Brunner, 2010] M. Ball, N. and Brunner, R. J. (2010). Data mining and machine learning in astronomy.

[MacDonald and Ian, 2009] MacDonald, W. Z. and Ian (2009). *MacDonald, Walter Zucchini and Ian.* Hidden Markov Models for Time Series, An Introduction Using R.

[Mahalanobis, 1936] Mahalanobis, P. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Science of India*, 12.

[NASA, 1997] NASA (1997). *Lightcurves: An Introduction by NASA's Imagine the Universe*. NASA.

[NASA, 2008] NASA (2008). *The RXTE Cook Book Recipes for Data Analysis and Reduction*. NASA.

[NASA, 2011] NASA (2011). *Overview of Kepler Light Curve Format Changes. How To Transition from Format v1.0 to v2.0*. NASA.

[Negueruela et al., 2005] Negueruela, I., Smith, D. M., Reig, P., et al. (2005). *Supergiant Fast X-ray Transients: A new class of high mass X-ray binaries unveiled by INTEGRAL*.

[Pérez, 2003] Pérez, L. C. (2003). *Hidden Markov Models and the Baum–Welch Algorithm*.

[Petrie and Baum, 1966] Petrie, L. and Baum, E. (1966). *Statistical Inference for Probabilistic Functions of Finite State Markov Chains*, volume 37. The Annals of Mathematical Statistics.

[Proakis and Manolakis, 1998] Proakis, J. G. and Manolakis, D. G. (1998). *Tratamiento digital de señales. Principios, algoritmos y aplicaciones*. Prentice Hall.

[Pyle, 1999] Pyle, D. (1999). *Data Preparation for Data Mining*. Morgan Kaufmann Publishers, Inc.

[Quinlan, 1993] Quinlan, J. R. (1993). C4.5: Programs for machine learning. morgan kaufmann publishers.

[Rabiner and Juang, 1986] Rabiner, L. and Juang, B. (1986). An introduction to hidden markov models. *IEEE ASSP Magazine*, pages 4–15.

[Rebbapragada et al., 2012] Rebbapragada, U., Lo, Kitty; Wagstaff, K. L., et al. (2012). *Classification of ASKAP Vast Radio Light Curves*. NASA.

[Ricci, 2005] Ricci, V. (2005). *Fitting Distributions with R*.

[Rissanen, 1978] Rissanen, J. (1978). *Modeling by shortest data description.*, volume 14. Automatica.

[S. Sumathi and Sivanandam, 2006] S. Sumathi, S. and Sivanandam, S. (2006). *Introduction to Data Mining and its Applications (Studies in Computational Intelligence)*. Springer-Verlag.

[Sáez Castillo, 2010] Sáez Castillo, D. (2010). *Métodos Estadísticos con R y R Commander Versión 2.1.*

[Team, 2000] Team, R. D. C. (2000). *Introducción a R. Notas sobre R: Un entorno de programación para Análisis de Datos y Gráficos.*

[Wehrle et al., 2009] Wehrle, A. E., Zacharias, N., et al. (2009). What is the structure of relativistic jets in agn on scales of light days? *Astro2010 : The Astronomy and Astrophysics Decadal Survey, Science White Papers*, 309.

[Weidenspointner et al., 2009] Weidenspointner, G., Skinner, G., et al. (2009). An asymmetric distribution of positrons in the galactic disk revealed by gamma-rays. *Nature*, (05):159–162.

[Zasche and Uhlar, 2009] Zasche, P. and Uhlar, R. (2009). A detailed light curve analysis of ty delphini. *Revista Mexicana de Astronomía y Astrofísica*, pages 205–214.