Clasificación de Series de Tiempo Astronómicas

Muriel Pérez
201011755

26 de abril de 2015

Índice general

Capítulo 1

Introducción

Con los avances en técnicas de observación astronómica que han sucedido en los últimos años, hay grandes cantidades de datos disponibles. Por ejemplo se espera que el *VISTA Variables in the Via Lactea* (VVV) del *European Southern Observatory* (ESO) produzca del orden de 10^9 curvas de luz[1] de fuentes puntuales en el infrarojo cercano con hasta 100 observaciones en diferentes épocas de alta calidad. De la misma forma estudios como la misión Kepler de la *National Aeronautics and Space Administration* (NASA), cuyo objetivo principal es la detección de exoplanetas, tienen como subproductos gran cantidad de curvas de luz.

Para que estos datos sean útiles para la comunidad científica es necesario clasificarlos y extraer sus características. Aunque los métodos automáticos muchas veces no pueden igualar la inspección manual por parte de un experto, la cantidad de datos disponible hace que esta tarea no sea posible en corto tiempo y hace necesario utilizar técnicas de minería de datos. Este interés se manifiesta en proyectos como el *VVV Templates Project* que tiene como objetivo consolidar una base bien definida de curvas de luz de estrellas variables en el infrarojo cercano para ser utilzadas como referencia para la clasificación automática de curvas de luz.

Las curvas de luz no pueden ser analizadas con técnicas de análisis de series de tiempo porque, debido a limitaciones en el tiempo de observación, fallas técnicas, periodos de mantenimiento de los instrumentos utilizados y

[1] La curva de luz de una estrella es el resultado de medir su magnitud como función del tiempo. La magnitud de una estrella es el flujo de energía observado en una parte del espectro electromagnético (una banda), delimitada por un filtro, en escala logarítmica (ver el capítulo 4 de [5]).

la imposibilidad de abservar todas las regiones del cielo durante todo el año, las curvas de luz no constan del mismo número de observaciones y éstas no son hechas en intervalos regulares por lo que el tiempo durante el cual cada estrella no es observada es impredecible y algunas características importantes de las curvas de luz no son observadas.

Dependiendo de la serie de magnitudes observadas, una estrella puede ser clasificadas como variable o no variable; periódicas o no periódicas; y en diferentes clases de variabilidad estelar que depende de la morfología de su curva de luz. La forma de la curva de luz depende de las condiciones físicas de la estrella por lo que conocer a qué tipo de variabilidad pertenece cada estrella es de vital importancia para el estudio de las estrellas variables. A su vez, el estudio de las estrellas variables ha sido importante para el estudio de la evolución estelar, la determinación de distancias cósmicas y la búsqueda de exoplanetas, entre otras.

En estudios previos [3, 13, 9] se le ha asignado a cada curva de luz un vector, llamado vector de características, y, basado en él, se ha hecho la clasificación automática. Este proceso consiste en entrenar un clasificador basado en una muestra clasificada previamente, la muestra de entrenamiento, utilizando el vector de características escogido. La escogencia de el vector de características es crucial para el proceso de clasificación porque con él se debe poder clasificar cada curva de luz, es decir, debe logar que, en el espacio de características, las clases se superpongan lo menos posible. Para la conformación de este vector se han elegido coeficientes de Fourier de la curva de luz[3, 13, 9], que son calculados mediante métodos como el periodograma de Lomb-Scargle [14] o la minimización de la entropía de Shannon de la gráfica de la curva [2].

Esta elección de caracterísicas no es del todo conveniente porque requiere de gran poder computacional y limita el tipo de objetos que pueden ser clasificados. El cálculo del periodogramas como el de Lomb-Scargle para curvas de luz, y en general el de los métodos utilizados en la literatura, requiere de intentar una gran cantidad de periodos candidatos a ser el periodo de la curva de luz para luego elegir el mejor y de la inspección manual de las curvas de luz. Los periodos de los objetos observados varía entre desde unos pocos minutos y varios años por lo cual se requiere probar una gran cantidad de periodos. Por un lado este es un proceso es computacionalmente intensivo, lo que limita su uso en conjuntos grandes de curvas de luz; y por otro lado no es seguro que dé como resultado el periodo real de una curva de luz, por lo que a menudo éste debe ser revisado manualmente. Además el resultado

de la clasificación puede ser sensible a la calidad de las curvas de luz que sean elegidas como muestra de entrenamiento [3] y limita el estudio a fuentes periódicas.

En [10, 12], los autores notaron que algunas variables descriptivas de la serie de magnitudes de una curva de luz (como su sesgo o su curtosis) sirven para clasificar ciertos tipos de estrellas con clasificadores lineales. En este trabajo retomamos esa idea y construimos un vector de características basadas en variables tomadas de estadística descriptiva. El uso de este tipo de variables tiene las ventajas de que puede ser calculadas de manera rápida y da como resultado un vector de caracterísicas que sive para realizar clasificación con una taza de éxito alta. Para evaluar esta aproximación al problema utilizamos una parte del Catálogo de Estrellas Variables de la tercera fase del *Optical Gravitational Lensing Experiment* (OGLE III)[16, 21, 23, 19, 18, 17, 27, 26, 24, 7, 4, 8, 20, 25, 22] que contiene curvas de luz de estrellas previamente clasificadas en seis tipos de variabilidad estelar y curvas de luz de estrellas candidatas a ser clasificadas como Be (ver cuadro 3.1).

En este trabajo utilizamos k-vecinos más cercanos, árboles de clasificación, máquinas de soporte vectorial y bosques aleatorios para realizar la clasificación automática de las curvas de luz basada en nuestra elección de características. Asímismo, estimamos la probabilidad de que una nueva curva de luz sea clasificada correctamente por uno de estos clasificadores utilzando validación cruzada de 10 iteraciones. Estos clasificadores fueron elegidos porque son aproximaciones muy distintas al problema de clasificación, por su naturaleza no lineal y no paramétrica; y por el hecho de que han mostrado ser efectivos en gran cantidad de aplicaciones prácticas. Para todo el análisis utilzamos el paquete estadístico de fuente abierta *R* (cita de R). Para cada tarea utilizamos paquetes específicos que son referenciados a lo largo del documento.

Este documento está organizado de la siguiente forma. En el capítulo 3 damos una descripción del conjunto de datos utilizado en este trabajo. En el capítulo 4 abordamos el problema de clasificación de manera informal, presentamos y discutimos la elección de atributos y evaluamos el desempeño de los clasificadores mediante validación cruzada de 10 iteraciones. En los apéndices abordamos formalmente el problema de aprendizaje en general y damos una descripción de cada uno de los algorítmos utilizados en el trabajo.

Capítulo 2

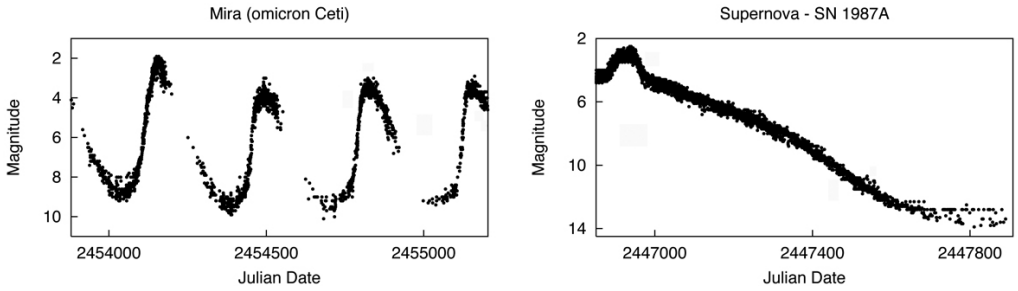
El Problema del Aprendizaje

Aquí diré cómo está organizado el capítulo.

2.1. Lulu

Las personas reconocemos con facilidad las letras en manuscritos, las caras de otras personas, las palabras que alguien nos dice o el estado de la comida basado en su olor. La capacidad de agrupar los estímulos que recibimos en categorías, por ejemplo el olor de la comida en buen o mal estado,y la capacidad para actuar en respuesta a ellos ha sido de vital importancia para nuestra supervivencia . Por ello hemos desarrollado complejos sistemas para llevar a cabo estas tareas.

Con la popularización de computadores electrónicos, la construcción máquinas que aprendan de la experiencia ha sido objeto de estudio. La habilidad de crear estas máquinas tiene una importancia estratégica puesto que existen tareas que no pueden ser llevadas a cabo utilizando técnicas de porgramación clásicas porque no existe un modelo matemático para ellas. En el caso de la clasificaciónd ecurvas de luz, por la forma en que se hacen las observaciones y el hecho de que la identificación de una curva de luz se hace con base en su forma, es difícil hacer un modelo matemático que capture estas diferencias. A pesar de esto existe gran gantidad de ejemplos de curvas de luz disponibles, por lo que es natural preguntarse si se puede entrenar un computador para identificar estas diferencias de la misma forma en que una persona puede ser entrenada para reconocerlas. En la figura 2.1 se observan dos curvas de luz, una pulsante y una eruptiva, que pueden ser distinguidas utilizando única-



(a) Curva de luz de una estrella Mira (b) Curva de luz de la Supernova 1987A

Figura 2.1: Las estrellas pueden ser clasificadas en grupos basado en la forma de sus curvas de luz. Estas clasificación puede ser hecha con base en la forma de las curvas de luz, sin embargo es difícil crear un modelo matemático que capture estas diferencias. Imágenes tomadas de [1]

mente esta información. La pregunta de si es posible entrenar un sistema basado en datos disponibles puede ser hecha para otras tareas, como el reconocimiento de textos en manuscritos, la detección e identificación de caras y objetos en imágenes o la identificación de genes en secuencias de ADN.

El reconocimiento de patrones es una disciplina científica cuya meta es la clasificación de objetos en clases. Existen situaciones en las cuales existe una gran cantidad de objetos previamente clasificados en clases predefinidas y la tarea es encontrar, o aproximar lo mejor posible, la dependencia funcional entre objetos y clases. Podemos precisar esto de la siguiente forma. Llamemos al espacio de los objetos que queremos clasificar Xy $\{1, \ldots, M\}$ es el conjunto de las posibles clases a las que pueden pertenecer los elementosd e X. En el caso de la clasificación de curvas de luz, X consta de todas las curvas de luz y $\{1, \ldots, M\}$ representa los posibles tipos de variabilidad estelar. Contamos con una muestra aleatoria, llamada muestra de entrenamiento, $\mathcal{L} = \{(x_i, j_i), \ldots, (x_N, j_N)\}$ con $x_l \in X$ y $j_l \in \{1, \ldots, M\}$, es decir, una muestra de X previamente clasificada. Nuestra tarea es entonces proponer una función $g : X \to \{1, \ldots, M\}$ a partir de la información contenida en $\mathcal{L}$, que representa nuestra predicción de la clase a la que pertenece cada elemento de X. La función g se llama clasificador y, para un elemento $x \in X$ cuya clase j es desconocida, el clasificador falla si $g(x) \neq j$.

El espacio X puede ser complejo o no estar matemáticamente bien de-

finido, por lo cual con frecuencia se representan los objetos con vectores, llamados de características, en $\mathbb{R}^n$. Por ejemplo si queremos realizar detección de rostros, X consiste de todos los posibles rostros, por lo que es más conveniente representar cada rostro con un conjunto de números como la separación de los ojos, el ángulo que forma las líneas que unen los ojos con la barbilla, etcétera; lo mismo sucede con las curvas de luz, por lo que representamos cada una con un vector. Estos vectores de características pueden, en principio, ser una combinación de variables continuas, discretas y categóricas, sin embargo esto no afecta en gran medida la teoría. Así las cosas, la elección de un clasificador puede ser una función $g : \mathbb{R}^n \to \{1, \ldots, M\}$.

Se debe utilizar un marco probabilístico para modelar la dependencia entre características y clases. Puede suceder que dos observaciones con un mismo vector de características pertenezcan a clases diferentes. Esto puede suceder en escenarios en los que pertenencia a una u otra clase no so sea completamente explicada por diferencias en los vectores de características, o porque la dependencia real entre características y clases sea no determinista. En este orden de ideas suponemos que existe una medida de probabilidad P sobre $\mathbb{R}^n \times \{1, \ldots, M\}$ tal que $P(\vec{x}, j)$ es la probabilidad de observar un vector de características $\vec{x} \in \mathbb{R}^n$ cuyo objeto representado pertenece a clase j. Así definimos la probabilidad de error del clasificador g, $P_e(g)$, como

$$P_e(g) = P(g(\vec{x}) \neq j). \tag{2.1}$$

Surge entonces la pregunta de qué tan bueno puede ser un clasificador. El mejor clasificador posible es llamado el clasificador de Bayes.

2.2. El Clasificador de Bayes

Decimos que un clasificador $g_B : \mathbb{R}^n \to \{1, \ldots, M\}$ es de Bayes si minimiza la probabilidad de error, es decir, que si g es otro clasificador entonces

$$P_e(g_B) \leq P_e(g). \tag{2.2}$$

Llamaremos P_e^* a $P_e(g_B)$.

En el caso de que existan densidades condicionales f_j tales que para cada $A \subset \mathbb{R}^n$ medible se cumple

$$P(A|j) = \int_A f_j(\vec{x}) d\vec{x} \tag{2.3}$$

podemos dar una expresión explícita para el clasificador de Bayes. Para un clasificador g podemos escribir

$$\begin{aligned} P_e(g) &= 1 - P(g(\vec{x}) = j) \\ &= 1 - \sum_{j=1}^{M} P(g(\vec{x}) = j|j)P(j) \\ &= 1 - \sum_{j=1}^{M} \left(\int_{\{g(\vec{x})=j\}} f_j(\vec{x}) d\vec{x} \right) P(j) \\ &= 1 - \int \sum_{j=1}^{M} \chi_{\{g(\vec{x})=j\}} f_j(\vec{x}) P(j) d\vec{x}. \end{aligned} \tag{2.4}$$

Donde $P(j)$ es la probabilidad *a priori* de encontrar un objeto de clase j. Ahora, para cada $\vec{x}$

$$\sum_{j=1}^{M} \chi_{\{g(\vec{x})=j\}} f_j(\vec{x}) P(j) \leq \max_j [f_j(\vec{x}) P(j)] \tag{2.5}$$

entonces

$$P_e(g) \geq \int \max_j [f_j(\vec{x}) P(j)] d\vec{x}. \tag{2.6}$$

Como la desigualdad 2.6 es igualdad cuando g le asigna a cada $\vec{x}$ la clase j para la cual $f_j(\vec{x})P(j)$, podemos concluir que éste es el clasificador de Bayes, es decir,

$$g_B(\vec{x}) = \arg\max_{j \in \{1,\ldots,M\}} f_j(\vec{x}) P(j) \tag{2.7}$$

y

$$P_e^* = \int \max_j [f_j(\vec{x}) P(j)] d\vec{x}. \tag{2.8}$$

2.3. Clasificadores

2.3.1. K Vecinos Más Cercanos

2.3.2. Máquinas de Soporte Vectorial

2.3.3. Árboles de Clasificación y Regresión

2.3.4. Bosques Aleatorios

Capítulo 3

El conjunto de Datos

Los datos utilizados en este trabajo provienen de la tercera fase del *Optical Gravitational Lensing Experiment* (OGLE-III). OGLE es un proyecto de larga duración cuyo objetivo principal es la búsqueda de materia oscura mediante el aprovechamiento de lentes gravitacionales. La tercera fase del protecto comenzó en 2001 y hace uso de un telescopio de 1,3m de diámetro localizado en el observatorio de Las Campanas, Chile[28]. Uno de los principales resultados de OGLE-III es la reducción y publicación [29] de las curvas de luz de objetos en el bulbo de la Galaxia, la Gran Nube de Magallanes y la Pequeña Nube de Magallanes. En este trabajo utilizamos las curvas de luz de 431653 objetos del catálogo de estrellas variables de OGLE-III de seis tipos de variabilidad (ver tabla 3.1) al cual se puede acceder en la página del proyecto [1] y 475 curvas de luz de estrellas candidatas a ser clasificadas como Be (ESCRIBIR DE DÓNDE FUERON TOMADAS ESTAS).

Las curvas de luz tomadas del catálogo de estrellas variables de OGLE-III se encuentran clasificadas por tipo de variabilidad estelar en un proceso que que involucró, en una etapa, la inspección manual de las curvas de luz (ver referencias en la tabla 3.1) por lo cual tomaremos esta clasificación como verdadera. En este trabajo utilizamos únicamente las curvas de luz registradas en la banda I [2] a pesar de que también se encuentra disponible información adicional sobre las curvas de luz como sus periodos y algunos coeficientes

[1] http://ogle.astrouw.edu.pl/

[2] Los objetos observados emiten radiación en una parte amplia del espectro electromagnético. Los telescopios utilizan filtros para recoger solo la radiación emitida por estos objetos en ciertas partes del espectro electromagnético. El filtro I (infrarojo) tiene un ancho de banda de 149nm y una longitud de onda efectiva de 797nm (ver [5])

Tipo de variabilidad y origen	Número de Objetos
RR Lyrae - BG[16]	16836
RR Lyrae - PNM [21]	2475
RR Lyrae - GNM [23]	24906
Cefeidas - BG [19]	32
Cefeidas - PNM [18]	4630
Cefeidas - GNM [17]	3361
Variables de Largo Periodo - BG [27]	232406
Variables de Largo Periodo - PNM [26]	19384
Variables de Largo Periodo - GNM [24]	91995
Sistema Binario Eclipsante - PNM [7]	6138
Sistema Binario Eclipsante - GNM [4]	26121
δ-Scuti - Nube Mayor de Magallanes[8]	2786
Cefeidas Tipo II - BG [20]	335
Cefeidas Tipo II - PNM [25]	43
Cefeidas Tiplo II - GNM [22]	197
Cadidata a Be - Vía Láctea (cita!)	475

Cuadro 3.1: Conjunto de datos utilizados. BG hace referencia al Bulbo Galáctico; PNM, a la Pequeña Nube de Magallanes y GNM, a la Gran Nube de Magallanes.

de Fourier (ver referencias en la tabla 3.1). Esta elección se debe a que el cálculo de estas cantidades es computacionalmente intensivo, no siempre se encuentran disponible y proponemos hacer la clasificación utilizando variables tomadas de estadística descriptiva.

Agrupamos los 432128 objetos disponibles en siete clases de variabilidad estelar (ver tabla3.2). Esta elección de clases puede ser refinada puesto que en cada una de estas clases existen subclases. Por ejemplo entre las Cefeidas se puede distinguir entre aquellas que pulsan en su modo fundamental, en su primer sobretono (segundo armónico) o en su segundo sobretono (tercer armónico) (ver figura **??**). Sin embargo conocer a qué clase de variabilidad estelar pertence un objeto facilita considerablemente su clasificación en subclases y análisis subsecuentes.

En el Catálogo de Estrellas Variables de OGLE-III, cada curva de luz está disponible en un archivo que contiene tres columnas con los valores de

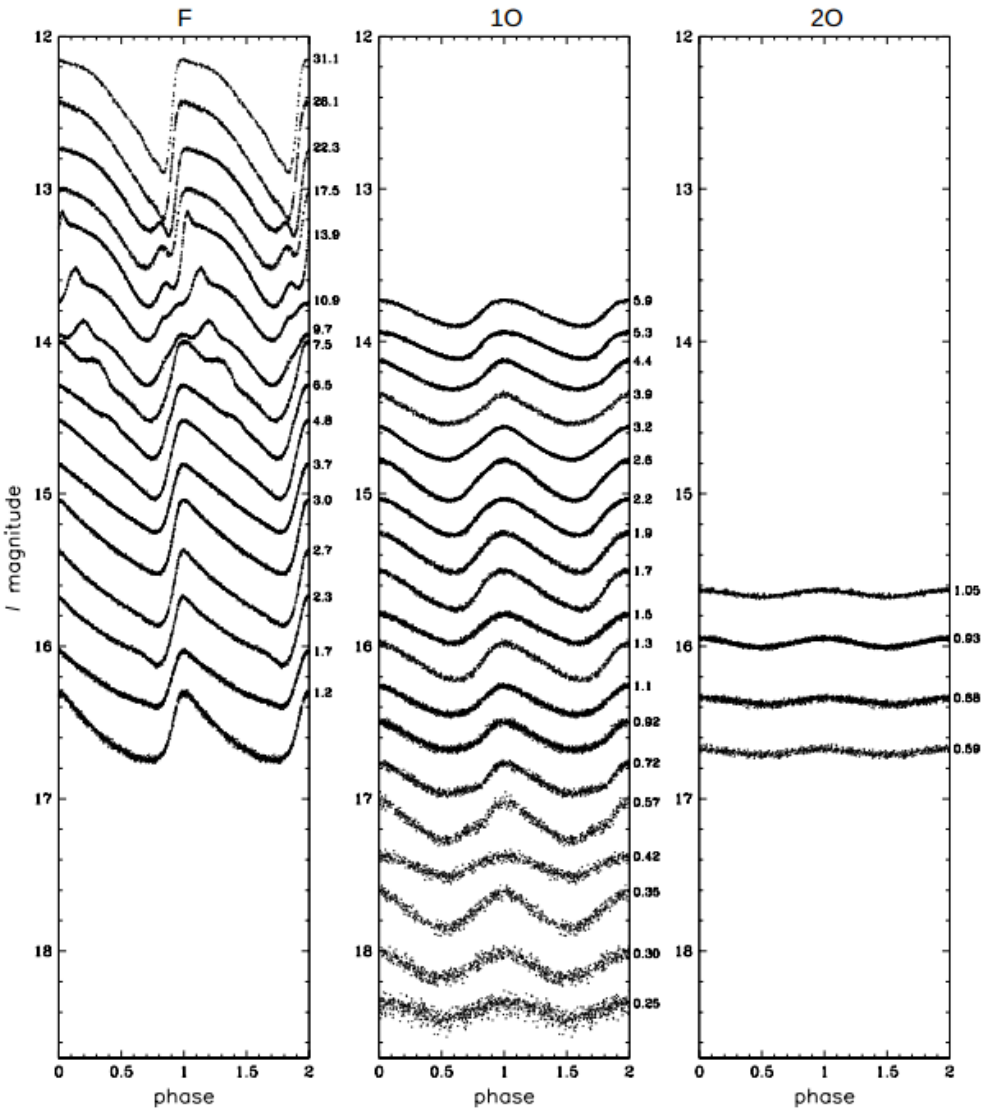


Figura 3.1: Curvas de luz ilustrativas de Cefeidas en modo fundamental (izquierda), primer sobretono (mitad), segundo sobretono (derecha). Los números pequeños a la derecha de cada recuadro muestran los periodos redondeados de las curvas de luz presentadas en los recuadros. Tomado de [19]

Tipo de Variabilidad	Cantidad
Variables de Largo Periodo (VLP)	343782
RR Lyrae (RRLyr)	44217
Cefeida (Cef)	8004
Sistema Binario Eclipsante (SBE)	32259
δ-Scuti (δSct)	2788
Cefeida Tipo II (CefT2)	603
Candidata a Be (BeEC)	475
Total	432128

Cuadro 3.2: Cantidad de datos por tipo de variabilidad

magnitud, fecha juliana [3] en la que fue tomada cada medida y error en la medida de la magnitud. El número de medidas para cada objeto y la separación temporal varía ampliamente. La separación mínima dos mediciones en toda la muestra es de 0.00147d, la máxima es 2156.9d y en promedio están separadas por 5.1d; por su parte el número promedio de observaciones por objeto es 759; el máximo, 5173; y el mínimo, 11. El 75 % de los objetos cuenta con más de 386 observaciones. Para todos los objetos estas observaciones están repartidas en los (número de años) años en que estuvo activo OGLE-III. En la figura **??** se puede observar una curva de luz del catálogo de estrellas variables de OGLE-III.

[3] La fecha Juliana es el tiempo medido en días desde el 1 de enero de 4713 a. C.

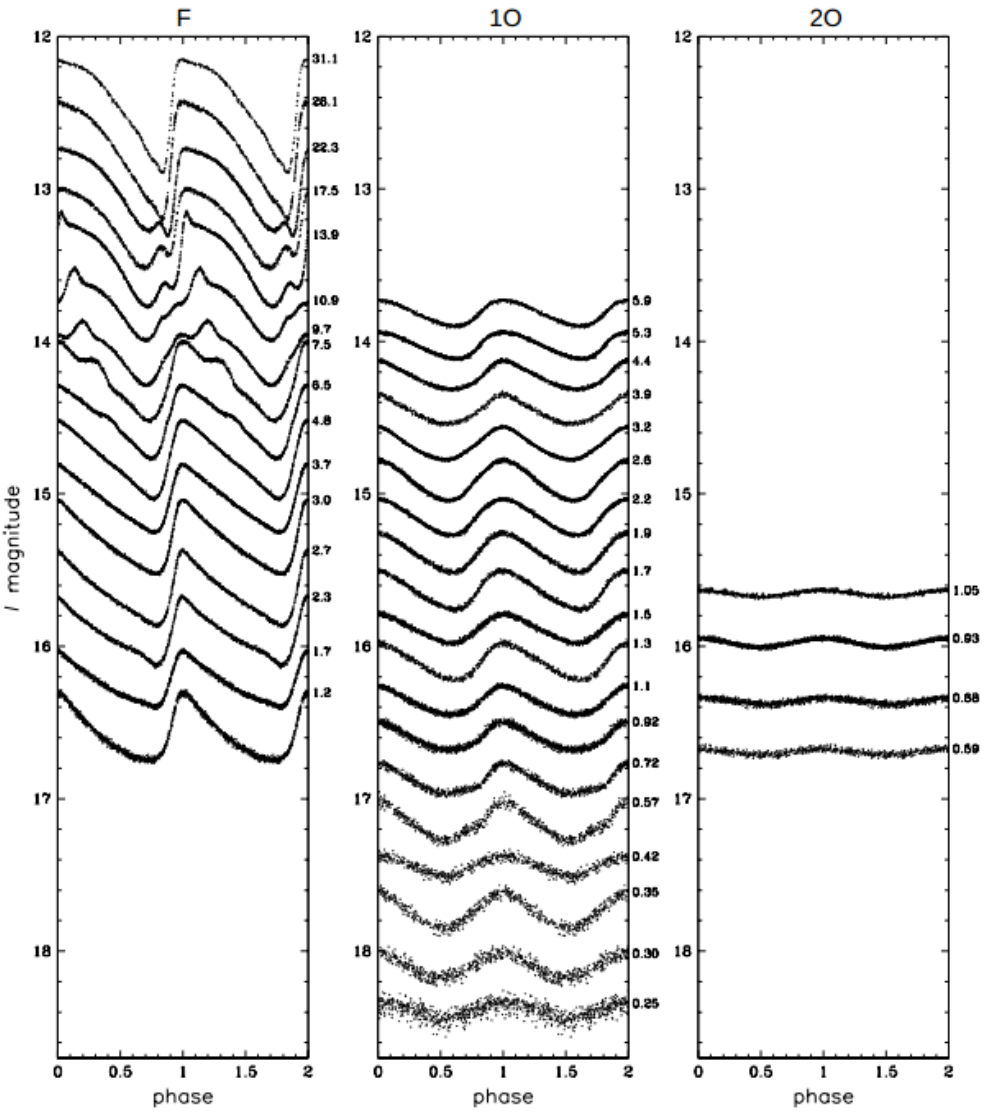


Figura 3.2: (figura pendiente de imagen) Curva de luz (nombre del archivo) del catálogo OGLE-III. Los periodos en los que no hay mediciones corresponden a los momentos del año en los que la zona en la que se encuentra el objeto no puede ser observada debido a la posición relativa entre el Sol y la Tierra.

Capítulo 4

Clasificación

Conociendo la curva de luz de un objeto podemos clasificarlo según su tipo de variabilidad estelar, sin embargo, esta relación no puede ser programada en un computador de manera sencilla. Cuando decimos que a cada curva de luz le corresponde un tipo de variabilidad estelar, queremos decir que existe una función, llamada función objetivo, cuya entrada es una curva de luz y cuya salida es un tipo de variabilidad estelar. El objetivo general del aprendizaje supervidado es aproximar esta función utilizando la experiencia previa. Esta experiencia previa es, en este caso, nuestro conjunto de datos (ver cuadro 3.1) y la estimación de esta regla, o función de decisión, es encontrada mediante un algorítmo de aprendizaje. Un algorítmo de aprendizaje escoge una función de decisión de un conjunto de funciones, llamado conjunto de hipótesis, utilizando un criterio que usualmente consiste en la minimización de una función de costo asociada a las clasificaciones erróneas. En nuestro caso utilizamos como función de costo la probabilidad de clasificación incorrecta.

Cada curva de luz en nuestra muestra es una tríada que consta de una sucesión de mediciones de magnitud, una suseción de fechas y un tipo de variabilidad estelar. Como cada curva de luz tiene un número de mediciones diferentes que están repartidas en diferentes intervalos de tiempo, esto dificulta la implementación de algorítmos para entrenar una regla de desisión. En consecuencia, a cada curva de luz le asignamos un vector de dimensionalidad fija, llamado vector de características. Este vector puede ser, en principio, una combinación de variables categóricas y numéricas; en este trabajo le asignamos únicamente variables numéricas. La función de decisión divide el espacio de caracterísitcas en regiones tales que a cada elemento del espacio

de características le asigna un tipo de variabilidad estelar basado en qué región se encuentra. Así, para clasificar una curva cuyo tipo de variabilidad es desconicido, calculamos su vector de caracterísicas y le asignamos la clase de variabilidad dada por la regla de decisión previamente entrenada. Por lo tanto la elección de características es crucial puesto que si los vectores de características de diferentes clases se superponen, no podrán ser distinguidos por la regla de decisión.

Subsecuentemente llamaremos $g : \mathbb{R}^n \to \{VLP, \ldots, BeEC\}$ a la función de decisión en cuestión que le asigna a cada vector de características un tipo de variabilidad (ver cuadro 3.2). Cada dato es representado por una pareja $(\vec{x}, i)$, con $\vec{x} \in \mathbb{R}$ siendo el vector de características y $i \in \{VLP, \ldots, BeEC\}$ la clase a la que pertenece. La regla de decisión se equivoca si $g(x) \neq i$. Suponemos que existe una distribución de probabilidad $p(\vec{x}, i)$ que representa la probabilidad de observar el vector de características $\vec{x}$ con el tipo de variabilidad i.

Para estimar la probabilidad de clasificación correcta de la función de decisión entrenada por un algorítmo de aprendizaje utilizamos validación cruzada de v iteraciones. Para esto dividimos la muestra $\mathcal{L}$ en v muestras de prueba $\mathcal{L}_k$, $k = 1, \ldots, v$ con el mismo número de elementos (o lo más próximo posible) y definimos la k-ésima muestra de entrenamiento como $\mathcal{L}^k = \mathcal{L} \setminus \mathcal{L}_k$. Utilizando cada una de las v muestras de entrenamiento $\mathcal{L}^k$ entrenamos una regla de decisión utilizando el algorítmo de aprendizaje en cuestión, con ella clasificamos los elementos de la muestra de prueba $\mathcal{L}^k$ y calculamos N_{ij}^k el número de elementos de de la clase j clasificado como i. Definimos $N_{ij} = \sum_k N_{ij}^k$ el número total de elmentos de la clase j clasificado como i. Estimamos la probabilidad de que un elemento de la clase j sea clasificado como i, $p^{VC}(g(\vec{x}) = i|j)$, con N_{ij}/N_j, donde N_j es el número de elementos pretenecientes a la clase j en la muestra $\mathcal{L}$. Intuitivamente, si la muestra es grande tendremos aproximadamente el mismo poder para clasificar con la muestra completa que con una fracción $\frac{v-1}{v}$ de ella, por lo cual p^{VC} será una buena aproximación a la probabilidad real de clasificación. Tomamos $v = 10$ siguiendo la popularidad de este valor en la literatura.

La estimación de la probabilidad de que un elmento cualquiera sea clasificado correctamente, llamada precisión, será $\sum_i p^{VC}(g(\vec{x} =)i|i)p(i)$. $p(i)$ es la probabilidad *a priori* de encontrar un objeto del tipo de variabilidad i. Como nuestra muestra no es representativa de las poblaciones de estrellas observadas y no existen estudios al respecto en la literatura para todos los tipos de variabilidad, tomamos $p(i)$ uniforme, es decir, $p(i) = 1/7$ para cada

i (hay 7 tipos de variabilidad estelar en la muestra).

Utilizamos la maximización de la precisión como criterio para elegir la mejor función de decisión producida por cada algorítmo de aprendizaje. Adicionalmente analizamos para cada clase la sensibilidad $p(g(\vec{x}) = i|i)$ (tasa de verdaderos positivos), la especificidad $p(g(\vec{x}) \neq i|i^c)$ (tasa de verdaderos negativos), el poder de predicción positiva $p(i|g(\vec{x}) = i)$ (probabilidad de que una vez clasificado, la clasifición sea correccta) y el poder de predicción negativa $p(i^c|g(\vec{x}) \neq i)$. El poder de predicción positiva juega un papel importante en este análisis puesto que, dada una nueva base de datos cuya clasificación no se conoce, si aplicamos el clasificador entrenado con nuestra muestra, esta es la estimación de la probabilidad de que esa clasificación sea correcta, lo cual corresponde a las situaciones reales que se encontrarán una vez se hagan públicos nuevas curvas de luz de estrellas variables sin clasificar.

4.1. Características Seleccionadas

Para una curva de luz denotaremos con $(m_i)_{1\leq i\leq n}$, $(t_i)_{1\leq i\leq n}$ y j a su serie de magnitudes, tiempos y tipo de variabilidad respectivamente.

Idealmente, el vector de características debe ser fácil de cálcular y debe capturar las diferencias entre los tipos de variabilidad estelar. En la literatura [3, 13, 9] se han utilizado coeficientes de Fourier para este propósito. Suponiendo que los pares (t_i, m_i) provienen de una versión corrupta de la magnitud verdadera es posible encontrar estimadores de mpinimos cuadrados para los coeficientes con el periodograma de Lomb-Scargle. Sea $m(t)$ la magnitud verdadera de la estrella observada, $y(t) = m(t) + \epsilon$ la medición que es una versión corrupta de $m(t)$ con ϵ siendo una variable aleatoria. Los autores de [3] encuentran los parámetros a_{ls}, f_l y b_{ij} que mejor se ajustan a los datos de forma tal que y es estimada por $\tilde{y}$

$$\tilde{y}(t) = \sum_{l=1}^{3}\sum_{s=1}^{4}(a_{ls}\,\mathrm{sen}\,2\pi f_l st + b_{ls}\cos 2\pi f_l st) + b_0$$

Luego los autores utilizan estos coeficientes para dar una descripción de $y(t)$ que es independiente de traslaciones temporales. Lo importante no es entrear en los detalles de esta elección de parámetros sino resaltar que la búsqueda de estos es computacionalmente intensiva. Los autores de [3] utilizan el periodograma de Lomb-Scargle [14] con el cual se obtiene una potencia para

cada periodo posible. Predefinir los periodos posibles es un reto si no se tiene más información que la curva de luz de un objeto. Por ejemplo para clasificar las curvas de luz de Cefeidas Clásicas para el catálogo de OGLE-III [17] los autores probaron frecuencias entre 0.0 y 24.0 ciclos por día en aumentos de frecuencias de 0.0001 para 32 millones de objetos, para lo cual utilizaron supercomputadores del *Centre for Mathematical and Computational Modeling* de la Universidad de Varsovia, siguido de un análisis que llevó a la inspección manual de decenas de miles de curvas de luz. En este trabajo, basado en los hallazgos de [10] y [12], proponemos utilizar en lugar de estos coeficientes, variables descriptivas de la serie de magnitudes (ver tabla 4.1) que pueden ser calculadas en tiempos abrumadoramente menores, con menos poder computacional y sin intervención manual.

Cantidad	Fórmula
Media	$\mu = \frac{1}{n}\sum_i m_i$
Desviación estándar	$\sigma = \sqrt{\frac{1}{n}\sum_i (m_i - \mu)^2}$
Sesgo	$\frac{1}{n}\sum_i \left(\frac{m_i-\mu}{\sigma}\right)^3$
Curtosis	$\frac{1}{n}\sum_i \left(\frac{m_i-\mu}{\sigma}\right)^4$
Rango	$\text{máx}_i\, m_i - \text{mín}_i\, m_i$
Variación cuadrática	$\frac{1}{n}\sum_i (m_i - m_{i-1})^2$
Valor Abbe [6]	$\mathcal{A} = \frac{n}{2(n-1)} \frac{\sum_i (m_i - m_{i-1})^2}{\sum_i (m_i-\mu)^2}$
Abbe promedio [6]	$\bar{\mathcal{A}}_t$
Entropía de Shannon [15]	$\sum_i -p_i \log_2 p_i$
Entropía de Rényi[11]	$\frac{1}{1-\alpha} \log_2 \sum_x p_i^\alpha$

Cuadro 4.1: Variables utilizadas

Bajo este punto de vista, las magnitudes son vistas como una variable aleatoria independiente del tiempo y las cantidades de la tabla 4.1 son variables descriptivas de su densidad. En la figura 4.1 se observa una curva de luz y la densidad estimada de sus magnitudes. Al utilizar la distribución de la serie de magnitudes se asume que el número de observaciones es lo suficientemente grande, que estas son hechas en intervalos que evitan el aliasing y que son hechas durante más de un periodo del objeto observado. Es de esperar que las curvas que tienen formas similares, es decir, que pertenecen al mismo tipo de variabilidad estelar, tengan densidades de magnitudes similares y que, por ende, los parámetros descriptivos utilizados también sean similares.

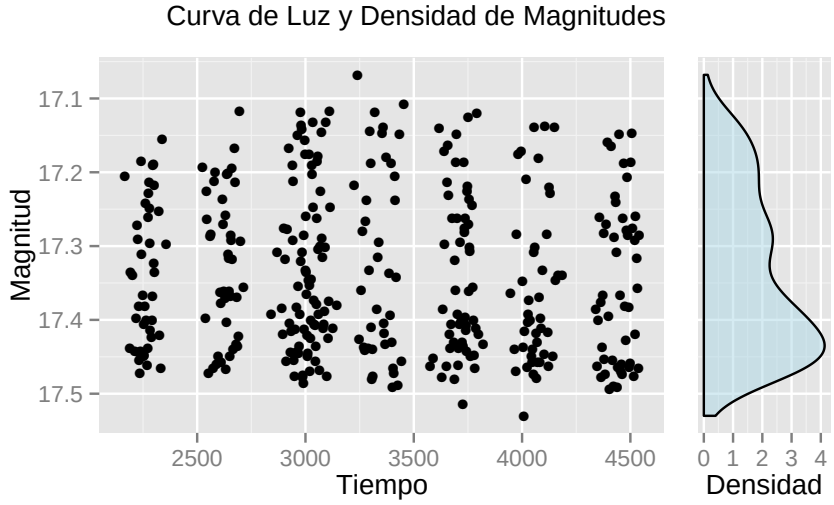


Figura 4.1: Curva de luz OGLE-LMC-CEP-0503 y densidad estimada de las magnitudes.

La media μ y la desviación estándar σ (ver cuadro 4.1) son variables descriptivas bien conocidas. En este caso la media es el valor al rededor del cual la serie de magnitudes oscila y la desviación una medida de la amplitud de estas oscilaciones. (Dar argumentos astronómicos). La figura 4.2 muestra la densidad de cada una de las clases en el plano μ-σ. Aunque las diferentes clases se superponen en este plano, hay pares de clases que pueden ser distinguidas como δSct y RRLyr.

Basado en el trabajo de [10] utilizamos el sesgo y la curtosis como características. El sesgo es el tercer momento central estandarizado [1] y es una medida de de la asimetría de una distribución. Una distribución es simétrica si su sesgo es 0, su cola izquierda es más larga si su sesgo es positivo y su cola derecha es más larga si su sesgo es negativo. Por su parte la curtosis es el cuarto momento central estndarizado. Es una medida de qué tan concentrada está la distribución al rededor de la media. La curtosis de una distribución normal es 3 y con frecuencia se estudia una cantidad llamada exceso de curtosis que es el resultado de restarle 3 a la curtosis. Los autores de [10] encontraron que algunos tipos de variabilidad estelar podían ser

[1] El k-ésimo momento centrado de una variable aleatoria X (o de su distribución) es $\mu_k = E[(X - \mu)^k]$, siendo μ su media. Su k-ésimo momento central estandarizado es $\frac{\mu_k}{\sigma^k}$, siendo σ la desviación estándar.

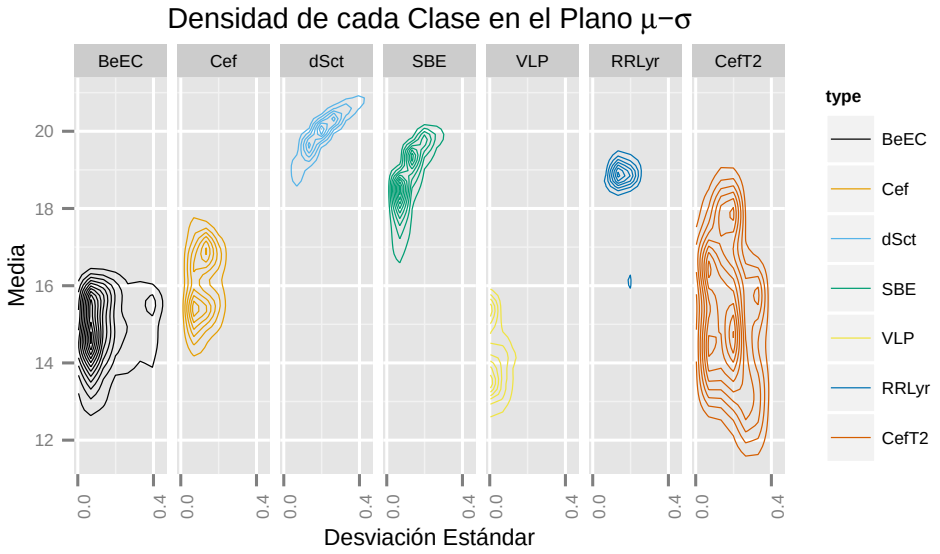


Figura 4.2

distinguidos utilizando clasificadores lineales en el plano sesgo-curtosis.

Dado que la estimación de el sesgo y la curtosis con las fórmulas del cuadro requiere de calcular las potencias $(\mu - m_i)^3$ y $(\mu - m_i)^4$, son propensas a dar estimaciones erróneas en el caso de que existan datos atípicos. Calculamos también la l-curtosis y el l-sesgo[2] y reemplazando el sesgo y la curtosis por estas cantidades no encontramos dierencias importantes en el poder para clasificar de los clasificadores que utlizamos. Esto puede ser un indicador de que no existe una proporción grande de datos atípicos en las curvas de luz. Por su simplicidad utilizamos el sesgo y la curtosis.

[2]Los l-momentos son combinaciones lineales de los estadísticos de orden. Son robustos, toman valores entre 0 y 1, y la interpretación de sus valores es análoga a la de los momentos. De la misma manera en que se define el sesgo y la curtosis muestral, es posible definir la l-curtosis y el l-sesgo. Fuero propuestos en (cita l-momentos) y su cálculo fue realizado utilizando el paquete lmoments(cita paquete l-moments) para R

4.2. Clasificación

4.2.1. K Vecinos Más Cercanos

Cuadro 4.2: k = 1

	becand	cep	dcst	ebs	lpv	rrlyr	t2cep
becand	381	1	0	50	41	0	0
cep	1	6595	4	172	54	986	87
dcst	0	7	2064	340	13	151	0
ebs	51	184	536	30129	573	668	29
lpv	42	187	19	944	342740	371	158
rrlyr	0	971	165	595	308	41911	217
t2cep	0	59	0	29	53	130	112

Cuadro 4.3: k=1, validación cruzada de 10 iteraciones

	becand	cep	dcst	ebs	lpv	rrlyr	t2cep
becand	0.80	0.00	0.00	0.00	0.00	0.00	0.00
cep	0.00	0.82	0.00	0.01	0.00	0.02	0.14
dcst	0.00	0.00	0.74	0.01	0.00	0.00	0.00
ebs	0.11	0.02	0.19	0.93	0.00	0.02	0.05
lpv	0.09	0.02	0.01	0.03	1.00	0.01	0.26
rrlyr	0.00	0.12	0.06	0.02	0.00	0.95	0.36
t2cep	0.00	0.01	0.00	0.00	0.00	0.00	0.19

4.2.2. Árboles de clasificación y regresión

4.2.3. Máquinas de Soporte Vectorial

4.2.4. Bosques Aleatorios

Cuadro 4.4: Matriz de confusión para CART

	becand	cep	dcst	ebs	lpv	rrlyr	t2cep
becand	470	12	6	1039	8906	47	18
cep	0	6210	12	18	320	3098	92
dcst	0	35	2511	5538	91	1615	1
ebs	0	1	148	21346	346	64	1
lpv	5	107	24	1762	304810	100	3
rrlyr	0	648	82	552	13880	34313	92
t2cep	0	991	5	2004	15429	4980	396

Cuadro 4.5: Tasas de clasificación estimadas por validación cruzada de 10 iteraciones

	becand	cep	dcst	ebs	lpv	rrlyr	t2cep
becand	0.99	0.00	0.00	0.03	0.03	0.00	0.03
cep	0.00	0.78	0.00	0.00	0.00	0.07	0.15
dcst	0.00	0.00	0.90	0.17	0.00	0.04	0.00
ebs	0.00	0.00	0.05	0.66	0.00	0.00	0.00
lpv	0.01	0.01	0.01	0.05	0.89	0.00	0.00
rrlyr	0.00	0.08	0.03	0.02	0.04	0.78	0.15
t2cep	0.00	0.12	0.00	0.06	0.04	0.11	0.66

Cuadro 4.6: $\gamma = 0{,}1$, costo = 16

	becand	cep	dcst	ebs	lpv	rrlyr	t2cep
becand	297	0	2	26	16	0	0
cep	0	4622	0	55	85	788	69
dcst	0	0	1683	163	4	47	0
ebs	89	48	865	29439	198	563	5
lpv	89	490	22	1380	342823	1113	255
rrlyr	0	2844	216	1196	656	41706	274
t2cep	0	0	0	0	0	0	0

Cuadro 4.7: $\gamma = 0{,}1$, costo = 16, validación cruzada de 10 iteraciones

	becand	cep	dcst	ebs	lpv	rrlyr	t2cep
becand	0.63	0.00	0.00	0.00	0.00	0.00	0.00
cep	0.00	0.58	0.00	0.00	0.00	0.02	0.11
dcst	0.00	0.00	0.60	0.01	0.00	0.00	0.00
ebs	0.19	0.01	0.31	0.91	0.00	0.01	0.01
lpv	0.19	0.06	0.01	0.04	1.00	0.03	0.42
rrlyr	0.00	0.36	0.08	0.04	0.00	0.94	0.45
t2cep	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Bibliografía

[1] BSJ. Types of Variables, June 2012.

[2] Pablo M. Cincotta, Mariano Mendez, and Josue A. Nunez. Astronomical time series analysis. I. A search for periodicity using information entropy. *The Astrophysical Journal*, 449:231, 1995.

[3] Jonas Debosscher, L. M. Sarro, Conny Aerts, J. Cuypers, Bart Vandenbussche, R. Garrido, and E. Solano. Automated supervised classification of variable stars-I. Methodology. *Astronomy & Astrophysics*, 475(3):1159–1183, 2007.

[4] D. Graczyk, I. Soszyński, R. Poleski, G. Pietrzyński, A. Udalski, M. K. Szymański, M. Kubiak, Ł. Wyrzykowski, and K. Ulaczyk. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. XII. Eclipsing Binary Stars in the Large Magellanic Cloud. *Acta Astronomica*, 61:103–122, June 2011.

[5] Hannu Karttunen, Pekka Kröger, Heikki Oja, Markku Poutanen, and Karl Johan Donner, editors. *Fundamental Astronomy*. Springer, Berlin ; New York, 5th edition edition, August 2007.

[6] N. Mowlavi. Searching transients in large-scale surveys. A method based on the Abbe value. *Astronomy and Astrophysics*, 568:78, 2014.

[7] M. Pawlak, D. Graczyk, I. Soszyński, P. Pietrukowicz, R. Poleski, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, S. Kozłowski, and J. Skowron. Eclipsing Binary Stars in the OGLE-III Fields of the Small Magellanic Cloud. *Acta Astronomica*, 63:323–338, September 2013.

[8] R. Poleski, I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, O. Szewczyk, and K. Ulaczyk. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. VI. Delta Scuti Stars in the Large Magellanic Cloud. *Acta Astronomica*, 60:1–16, March 2010.

[9] Joseph W. Richards, Dan L. Starr, Nathaniel R. Butler, Joshua S. Bloom, John M. Brewer, Arien Crellin-Quick, Justin Higgins, Rachel Kennedy, and Maxime Rischard. On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data. *The Astrophysical Journal*, 733(1):10, May 2011.

[10] Bayron Stevenson Rodríguez Feliciano and José Alejandro García Varela. *Análisis estadístico en poblaciones de estrellas variables*. Tesis (Físico). Universidad de los Andes. Bogotá : Uniandes, 2012., 2012.

[11] Alfréd Rényi and others. On Measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[12] B. E. Sabogal, A. García-Varela, and R. E. Mennickent. Search for Southern Galactic Be Star Candidates. *Publications of the Astronomical Society of the Pacific*, 126:219–226, 2014.

[13] L. M. Sarro, Jonas Debosscher, M. López, and Conny Aerts. Automated supervised classification of variable stars-II. Application to the OGLE database. *Astronomy & Astrophysics*, 494(2):739–768, 2009.

[14] Jeffrey D. Scargle. Studies in astronomical time series analysis. II-Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 263:835–853, 1982.

[15] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal, The*, 27(3):379–423, July 1948.

[16] I. Soszyński, W. A. Dziembowski, A. Udalski, R. Poleski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, S. Kozłowski, and P. Pietrukowicz. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. XI. RR Lyrae Stars in the Galactic Bulge. *Acta Astronomica*, 61:1–23, March 2011.

[17] I. Soszyński, R. Poleski, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, L. Wyrzykowski, O. Szewczyk, and K. Ulaczyk. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. I. Classical Cepheids in the Large Magellanic Cloud. *Acta Astronomica*, 58:163–185, September 2008.

[18] I. Soszyński, R. Poleski, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, O. Szewczyk, and K. Ulaczyk. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. VII. Classical Cepheids in the Small Magellanic Cloud. *Acta Astronomica*, 60:17–39, March 2010.

[19] I. Soszyński, A. Udalski, P. Pietrukowicz, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, R. Poleski, and S. Kozłowski. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. XIV. Classical and TypeII Cepheids in the Galactic Bulge. *Acta Astronomica*, 61:285–301, December 2011.

[20] I. Soszyński, A. Udalski, P. Pietrukowicz, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, R. Poleski, and S. Kozłowski. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. Type II Cepheids in the Galactic Bulge - Supplement. *Acta Astronomica*, 63:37–40, March 2013.

[21] I. Soszyński, A. Udalski, M. K. Szymański, J. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, and R. Poleski. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. IX. RR Lyr Stars in the Small Magellanic Cloud. *Acta Astronomica*, 60:165–178, September 2010.

[22] I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, O. Szewczyk, K. Ulaczyk, and R. Poleski. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. II.Type II Cepheids and Anomalous Cepheids in the Large Magellanic Cloud. *Acta Astronomica*, 58:293, December 2008.

[23] I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, O. Szewczyk, K. Ulaczyk, and R. Poleski. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable

Stars. III. RR Lyrae Stars in the Large Magellanic Cloud. *Acta Astronomica*, 59:1–18, March 2009.

[24] I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, O. Szewczyk, K. Ulaczyk, and R. Poleski. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. IV. Long-Period Variables in the Large Magellanic Cloud. *Acta Astronomica*, 59:239–253, September 2009.

[25] I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, and R. Poleski. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. VIII. Type II Cepheids in the Small Magellanic Cloud. *Acta Astronomica*, 60:91–107, June 2010.

[26] I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, R. Poleski, S. Kozłowski, and P. Pietrukowicz. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. XIII. Long-Period Variables in the Small Magellanic Cloud. *Acta Astronomica*, 61:217–230, September 2011.

[27] I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, K. Ulaczyk, R. Poleski, S. Kozłowski, P. Pietrukowicz, and J. Skowron. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. XV. Long-Period Variables in the Galactic Bulge. *Acta Astronomica*, 63:21–36, March 2013.

[28] A. Udalski. The Optical Gravitational Lensing Experiment. Real Time Data Analysis Systems in the OGLE-III Survey. *Acta Astron.*, 53(astro-ph/0401123):291, 2004.

[29] A. Udalski, M. K. Szymanski, I. Soszynski, and R. Poleski. The Optical Gravitational Lensing Experiment. Final Reductions of the OGLE-III Data. *Acta Astronomica*, 58:69–87, 2008.