Clasificación de Series de Tiempo Astronómicas

Muriel Pérez
201011755

2 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; y dado el volumen de datos disponible, es 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.

Debido a limitaciones en el tiempo de observación, fallas técnicas, periodos de matenimiento de los instrumentos utilizados y 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. Como el tiempo durante el cual cada estrella no es observada no es predecible y en ellos no se observan características importantes de las

[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 en escala logarítmica (ver el capítulo 4 de [2]).

curvas de luz, estas no pueden ser analizadas con técnicas clásicas de análisis de series de tiempo.

En estudios previos(citas!) se le ha asignado a cada curva de luz un vector de características y, basado en este vector, se ha hecho la clasificación automática. 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 capturar la información que hace a cada clase de variabilidad diferente de las demás. Usualmente este vector de características está conformado por coeficientes de Fourier de la curva de luz, que son calculados mediante métodos como el periodograma de Lomb-Scargle (cita) o minimización de entropía (cita).

Esta elección de caracterísicas no es del todo conveniente porque presenta problemas computacionales 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 todos 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. Por un lado estos 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 (citar) y limita el estudio a fuentes periódicas.

En (cita Bayron) los autores notaron que algunas variables descriptivas de la serie de magnitudes de una curva de luz (como su sesgo o su curtosis) servían para clasificar ciertos tipos de estrellas con clasificadores lineales. En este trabajo retomamos esa idea y construimos un vector de características basadas en variables descriptivas. El uso de este tipo de variables tiene las ventajas de que se pueden calcular 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 de los resultados de la tercera fase del *Optical Gravitational Lensing Experiment* (OGLE III) 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 2.1).

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 atribu-

tos. Estos clasificadores fueron elegidos porque representan 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.

Este documento está organizado de la siguiente forma. En el capítulo 2 damos una descripción del conjunto de datos utilizado en este trabajo. En el capítulo 3 presentamos y discutimos los resultados de la clasificación automática utilizando los diferentes algorítmos así como la elección de atributos. En los apéndices damos una descripción matemática del problema de aprendizaje en general y una descripción de cada uno de los algorítmos utilizados en el trabajo.

Capítulo 2

El conjunto de Datos de OGLE III

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 observatorio de Las Campanas, Chile[17]. Uno de los principales resultados de OGLE-III es la reducción[18] y publicación 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 2.1) 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 fueron clasificadas por tipo de variabilidad estelar en un porceso que que involucró la inspección manual de las curvas de luz (ver referencias en la tabla 2.1). También se encuentra disponible información adicional sobre las curvas de luz como sus periodos o algunos coeficientes de Fourier (ver referencias en la tabla 2.1) pero dado que para futuras bases de datos sin clasificar el cálculo de estas cantidades es computacionalmente intensivo y proponemos hacer la clasificación utilizando variables tomadas de estadística descriptiva,

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

Tipo de variabilidad y origen	Número de Objetos
RR Lyrae - Bulbo Galáctico[5]	16836
RR Lyrae - Nube Menor de Magallanes [10]	2475
RR Lyrae - Nube Mayor de Magallanes [12]	24906
Cefeidas - Bulbo Galáctico [8]	32
Cefeidas - Nube Menor de Magallanes [7]	4630
Cefeidas - Nube Mayor de Magallanes [6]	3361
Variables de Largo Periodo - Bulbo Galáctico [16]	232406
Variables de Largo Periodo - Nube Menor de Magallanes [15]	19384
Variables de Largo Periodo - Nube Mayor de Magallanes [13]	91995
Binaria Eclipsante - Nube Menor de Magallanes [3]	6138
Binaria Eclipsante - Nube Mayor de Magallanes [1]	26121
δ-Scuti - Nube Mayor de Magallanes[4]	2786
Cefeidas Tipo II - Bulbo Galáctico [9]	335
Cefeidas Tipo II - Nube Menor de Magallanes [14]	43
Cefeidas Tiplo II - Nube Mayor de Magallanes [11]	197
BeSC - Vía Láctea (cita!)	475

Cuadro 2.1: Conjunto de datos utilizados (faltan las citas de los catálogos)

no utilizamos esta información. En este trabajo utilizamos las curvas de luz registradas en el filtro I [2]

[2]Los objetos observados emiten radiación en una parte amplia del espectro electromagnético. Los telescopios utilizan filtros para recoger solo la radiación emitida en ciertas partes del espectro electromagnético. El sistema UBVRI (*Ultraviolet, Blue, Visual, Red, Infrared)*

Capítulo 3

Clasificación

3.1. Atributos Seleccionados

Capítulo 4

Apéndices

4.1. El Problema del Aprendizaje

4.2. K Vecinos Más cercanos

4.3. Árboles de Clasificación y Regresión

4.4. Máquinas de Soporte Vectorial

4.5. Bosques Aleatorios

4.6. Estimación del Error de Clasificación

Capítulo 5

Cosas que la evolución se llevó

5.1. vieja introducción

En este trabajo abordamos el problema de clasificar curvas de luz de estrellas variables por su tipo de variabilidad [1] como un problema de aprendizaje supervisado. Para esto utilizamos una parte de los resultados de la tercera fase del *Optical Gravitational Lensing Experiment* (OGLE III) 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 capítulo 2) (ver cuadro 2.1).

Para abordar el problema de clasificación adoptamos el siguiente punto de vista. Cada curva de luz $c_i = \{(t_n^i, m_n^i)\}_n$ es una sucesión de parejas donde la primera es el tiempo y la segunda es la magnitud medida en ese instante. Debido a limitaciones en el tiempo de observación, fallas técnicas, periodos de matenimiento de los instrumentos utilizados y el hecho de que no todas las regiones del cielo son observables durante todo el año y solo se puede observar una región limitada en cada oportunidad, las curvas de luz no constan del mismo número de observaciones y éstas no son hechas en intervalos regulares ($t_k - t_{k+1}$ no es constante). Una forma de hacer frente a esto es asignarle a cada curva de luz c_i un vector de atributos $\vec{x_i} = \vec{x_i}(c_i) \in \mathbb{R}^n$ calculados a partir de de c_i (ver sección 3.1) que intenten describir los tipos de variabilidad. Como los elementos de la muestra han sido clasificados previamente, le asignamos

[1] Las estrellas variables son estrellas cuya magnitud cambia en el tiempo (ver nota 1). Pueden ser periódicas o no periódicas y se pueden clasificar como pulsantes, eruptivas o variables eclipsantes aunque existen subclases de variabilidad estelar. Una estrella puede ser clasificada en estas subclases conociendo su curva de luz (ver el capítulo 13 de [2]).

a cada curva de luz c_i una etiqueta $j_i \in J = \{\text{RR Lyr}, \ldots, \text{BeSC}\}$ (ver tabla 2.1) que corresponde al tipo de variabilidad estelar de la estrella observada. Dicha etiqueta, a su vez, es heredada por el vector de atributos $\vec{x}_i$.

Si nuestra elección de atributos es acertada, podremos utilizar la representación de las curvas de luz en el espacio de atributos para realizar la clasificación, esto es, existirá una función $g : \mathbb{R}^n \to J$ que, de alcanzar la mejor tasa de clasificación correcta posible para esos atributos, le asigna a cada curva de luz el tipo de variabilidad correcto con probabilidad alta (ver capítulo 4.1). Puede suceder que, si los atributos no caracterizan los diferentes tipos de variabilidad, incluso utilizando el mejor clasificador posible (la mejor función g) no sea posible alcanzar errores de clasificación bajos. De esto se sigue que la elección de atributos es crucial para lograr una buena clasificación. La elección de los atributos utilizados se discute en la sección 3.1.

El siguiente problema será el de inferir (aprender) de los datos una función $\hat{g} : \mathbb{R}^n \to J$ que se aproxime tanto como sea posible a la mejor regla posible en cuanto a que maximice la probabilidad de clasificación correcta. Para encontrar esta regla existen diferentes aproximaciones y algorítmos que permiten la división del espacio de atributos en zonas a cada una de las cuales se le asigna un tipo de variabilidad. En la práctica no se conoce la mejor regla posible porque esto requeriría el conocimiento de la distribución exacta de los atributos (ver capítulo 4.1). Tampoco se conoce el mínimo error de clasificación posible por lo que, para la selección del mejor clasificador entre los posibles, se utilizan estimaciones del error de clasificación que utilizan la muestra disponible, en este caso validación cruzada (ver sección 4.6).En el capítulo **??** utilizamos k vecinos más cercanos, arboles de clasificación y regresión; y maquinas de soporte vectorial para inferir la función $\hat{g}$. Asímismo analizamos los estimados de la probabilidad de error al usar cada algorítmo y comentamos las ventajas comparativas de cada uno. Estos tres métodos son muy diferentes en su naturaleza, fueron elegidos porque han mostrado ser efectivos en gran variedad de aplicaciones y por su carácter no paramétrico y no lineal.

Así la clasificación de una curva de luz correspondiente a una estrella cuyo tipo de variabilidad es desconocido será un proceso de dos pasos. El primero será la extracción de los atributos. El segundo paso será la clasificación basada en los atributos utilizando la función $\hat{g}$ que fue encontrada con ayuda de la muestra disponible. Esta clasificación será correcta con cierta probabilidad, estimada con validación cruzada.

Este documento está organizado de la siguiente manera. En el capítulo 2 damos un análisis descriptivo del conjunto de datos que consideramos y discutimos la elección de los atributos para realizar la clasificación. En el capítulo 4.1 discutimos brevemente el problema de clasificación en general, describimos el mejor clasificador posible (el clasificador de Bayes), discutimos la imposibilidad de utilizarlo en la mayoría de aplicaciones complejas y describimos el método que usamos para estimar la probabilidad de error de los clasificadores. En el capítulo **??** describimos los métodos de clasificación utilizados, damos los estimados del error de clasificación y comparamos los resultados con otros valores dados en la literatura.

Bibliografía

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.