Clasificación de Series de Tiempo Astronómicas

Muriel Pérez
201011755

9 de marzo de 2015

Índice general

1. Introducción **5**

2. El conjunto de Datos de OGLE III **9**
2.1. Descripción de los Datos 9
2.2. Atributos Seleccionados 9

3. El Problema del Aprendizaje **11**
3.1. Estimación del Error de Clasificación 11

4. Aprendizaje Supervisado **13**
4.1. K Vecinos Más cercanos 13
4.2. Árboles de Clasificación y Regresión 13
4.3. Bosques Aleatorios 13

Capítulo 1

Introducción

Con los avances en técnicas de observación astronómica que han sucedido en los últimos años, hay grandes cantidades de datos disponibles. Estudios como la misión Kepler de la *National Aeronautics and Space Administration* (NASA), o el *VISTA Variables in the Via Lactea* (VVV) del *European Southern Observatory* (ESO) tienen como productos gran cantidad de curvas de luz [1](necesito información sobre más estudios y dar números sobre la cantidad de estrellas observadas) de alta calidad.

Para que estos datos sean útiles en la comunidad científica es necesario clasificarlos y extraer sus características. Si se quiere lograr esto en poco tiempo, es necesario utilizar técnicas de minería de datos debido a los volúmenes que deben ser procesados. (este párrafo necesita más palabras)

En este trabajo abordamos el problema de clasificar curvas de luz de estrellas variables por su tipo de variabilidad [2] como un problema de aprendizaje supervisado [3]. Para esto utilizamos una parte de los resultados de la tercera fase del *Optical Gravitational Lensing Experiment* (OGLE III) que contiene

[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]).

[2]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]).

[3]Se busca, a partir de los datos, inferir una regla que le asigne a cada curva de luz un tipo de variabilidad. Esta regla debe poder aplicarse a curvas de luz que estén fuera de la muestra y la probabilidad de errror es estimada utilizando la muestra original. (ver capítulo 3)

Tipo de variabilidad y origen	Número de Objetos
RR Lyrae - Bulbo Galáctico[10]	16836
RR Lyrae - Nube Menor de Magallanes [7]	2475
RR Lyrae - Nube Mayor de Magallanes [14]	24906
Cefeidas - Bulbo Galáctico [11]	32
Cefeidas - Nube Menor de Magallanes [6]	4630
Cefeidas - Nube Mayor de Magallanes [5]	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 [8]	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 [12]	335
Cefeidas Tipo II - Nube Menor de Magallanes [9]	43
Cefeidas Tiplo II - Nube Mayor de Magallanes [13]	197
BeSC - Vía Láctea (cita!)	475

Cuadro 1.1: Conjunto de datos utilizados (faltan las citas de los catálogos)

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 (¿De dónde vienen estos datos de las Be?)(ver capítulo 2) (¿Por qué elegimos este conjunto de datos?) (ver cuadro 1.1). Estas curvas de luz fueron clasificadas por personas y tomaremos esta clasificación como verdadera.

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 2.2) que intenten describir los tipos de variabilidad. Como

los elementos de la muestra han sido clasificados previamente, le asignamos a cada curva de luz c_i una etiqueta $j_i \in J = \{\text{RR Lyr}, \ldots, \text{BeSC}\}$ (ver tabla 1.1) 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 3). 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 2.2.

El siguiente problema será el de inferir 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 3). 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 3.1).En el capítulo 4 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 3 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 4 describimos los métodos de clasificación utilizados, damos los estimados del error de clasificación y comparamos los resultados con otros valores dados en la literatura.

Capítulo 2

El conjunto de Datos de OGLE III

2.1. Descripción de los Datos

2.2. Atributos Seleccionados

Capítulo 3

El Problema del Aprendizaje

A cada curva de luz $c_i = \{(t_n^i, m_n^i)\}_n$ le asignamos un vector de atributos $\vec{x}_i$ (ver sección 2.2) y con estos esperamos construir una regla de desición $\hat{g} : \mathbb{R}^n \to J = \{\text{RR Lyr}, \ldots, \text{BeSC}\}$(ver capítulo 4) que sea la mejor posible. ¿Pero qé significa que una regla sea la mejor posible?

3.1. Estimación del Error de Clasificación

Capítulo 4

Aprendizaje Supervisado

4.1. K Vecinos Más cercanos

4.2. Árboles de Clasificación y Regresión

4.3. Bosques Aleatorios

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. Soszynski, R. Poleski, A. Udalski, M. K. Szymanski, M. Kubiak, G. Pietrzynski, 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[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. II.Type II Cepheids and Anomalous Cepheids in the Large Magellanic Cloud. *Acta Astronomica*, 58:293, December 2008.

[14] 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.

[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.