Clasificación de Series de Tiempo Astronómicas

Muriel Pérez ¹ Adolfo Quiroz ¹ Alejandro García²

¹Universidad de los Andes, Departamento de Matemáticas

²Universidad de los Andes, Departamento de Física

28 de mayo de 2015

Plan

Clasificación de Series de Tiempo Astronómicas

Curvas de Luz

El problema de Aprendizaje

Aprendizaje Supervisado

Metodología de Aprendizaje

Características Escogidas

Clasificadores y Resultados

Curvas de Luz



Figura : Telescopio de 2,2m utilizado por el *Optical Gravitational Lensing Experiment* (OGLE) localizado en Las Campanas, Chile.

Curvas de Luz

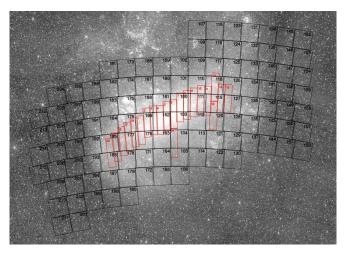


Figura : Campos observados por OGLE-III en la Gran Nube de Magallanes

Curvas de Luz

Se mide la magnitud, que está asociada a la densidad de flujo F ($[F] = Wm^{-2}$) por

$$m = -2.5 \log \frac{F}{F_0} \tag{1}$$

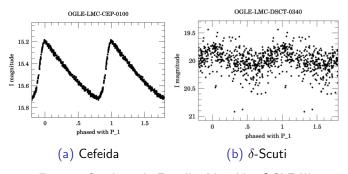


Figura : Catalogo de Estrellas Variables OGLE-III

Curvas de luz

Con avances en instrumenación hay disponibles gran cantidad de curvas de luz. Existen proyectos como:

- Kepler Mission, NASA;
- VISTA Variables in the Via Lactea Survey (VVV), ESO;
- Panoramic Survey Telescope & Rapid Response System (PANSTARRS), Universidad de Hawaii;

que obtendrán entre resultados del orden de 10^9 curvas de luz. Es necesario un sistema de clasificación automática.

Datos

	Total
de Objetos	TOTAL
16836	
2475	44217
24906	
32	
4630	8004
3361	
232406	
19384	343782
91995	
6138	32259
26121	
2786	2788
335	
43	603
197	
475	475
	2475 24906 32 4630 3361 232406 19384 91995 6138 26121 2786 335 43

Cuadro : Conjunto de datos utilizados. GB hace referencia al Bulbo Galáctico; SMC, a la Pequeña Nube de Magallanes y LMC, a la Gran Nube de Magallanes.

Tipos de Variabilidad

Dos objetos de tipo RR Lyrae

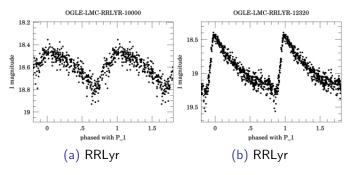


Figura: Catalogo de Estrellas Variables OGLE-III

Tipos de Variabilidad

Dos Cefeidas

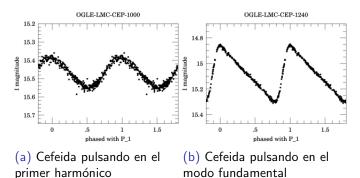


Figura: Catalogo de Estrellas Variables OGLE-III

Tipos de Variabilidad

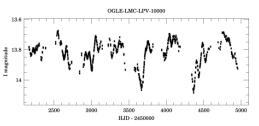


Figura : Variable de Largo Periodo

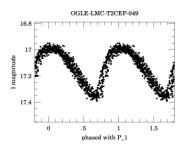


Figura : Cefeida Tipo 2

Características

- Cada curva de luz es un objeto complicado (diferentes números de mediciones hechas en intervalos irregulares de tiempo).
- ▶ A cada curva de luz podemos asignarle un vector $\vec{x_i} \in \mathbb{R}^n$ de cantidades calculadas a partir de los valores de magnitud y los instantes en que fueron medidos. Cada uno de esos vectores tiene una etiqueta $j \in J = \{RRLyrae, ..., Be\}$, que corresponde al tipo de variabilidad de la estrella observada.
- ▶ El vector $\vec{x_i} \in \mathbb{R}^n$ es llamado vector de características.

Características

Escogimos algunas variables descriptivas de la densidad de magnitudes, la variación cuadrática y los valores Abbe.

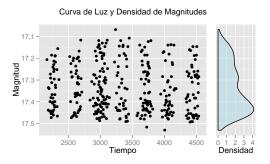


Figura : Curva de luz de OGLE-LMC-CEP-0503 y su densidad de magnitudes

Características

Debido a que en algunas curvas existen puntos atípicos, es necesario utilizar medidas robustas. Estas medidas están típicamente basadas en cuantiles de la distribución de las magniudes.

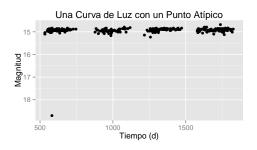


Figura : Curva de luz de una candidata a Be

Clasificadores

El problema: Encontrar una función $g: \mathbb{R}^n \to J$ que se equivoque lo menos posible.

- Suponemos que hay una medida de probabilidad P sobre $\mathbb{R}^n \times J$ tal que $P(\vec{x}, j)$ es la probabilidad de observar el vector \vec{x} con la etiqueta j.
- La probabilidad de que nuestro clasificador g se equivoque es $P(g(\vec{x}) \neq j)$ que queremos minimizar.

El clasificador de Bayes

¿Cuál es el mejor clasificador posible?

► El clasificador

$$g(\vec{x}) = \operatorname{argmax}_{j} P(\vec{x}|j) P(j)$$
 (2)

es llamado el **clasificador de Bayes**. Es el mejor clasificador posible.

En general no se conocen las distribuciones marginales $P(\vec{x}|j)$.

Metodología de aprendizaje

- Utilizamos un algorítmo de aprendizaje para escoger una regla de clasificación de un conjunto de hipótesis basado en la muestra de entrenamiento.
- Estimamos el error de clasificación usando validación cruzada de 10 iteraciones.

Caracteristicas escogidas

Las medidas escogidas debían ser robustas ante la presencia de datos atípicos.

Tipo de variables	Variable
Variables de localización	Media
Variables de escala	Rango Intercuartiles (IQR)
	Desviación Absoluta Mediana (MAD)
Medidas de Sesgo	Medcouple
	Medidas de peso de colas
Medidas de forma	Entropía diferencial
	Valores Abbe
	Variación Cuadrática

Cuadro: Variables escogidas

Parámetros de Escala

La desviación estándar muestral es sensible a la presencia de puntos atípicos. Utilizamos la desviación absuluta mediana (MAD)

$$\sigma = mediana_i(|m_i - mediana_j(m_j)|). \tag{3}$$

Y el rango intercuartiles (IQR)

$$IQR = Q_{0,75} - Q_{0,25} (4)$$

Parámetros de Localización

Utilizamos un estimador robusto del promedio que hace parte de los llamados M-estimadores. Huber [5] propuso escoger μ al resolver resolver el problema

$$\sum_{i} \psi\left(\frac{m_{i} - \mu}{\sigma}\right) = 0 \tag{5}$$

donde σ es la MAD y

$$\psi(x) = \begin{cases} -c & \text{si } x < -c \\ x & \text{si } |x| < c \\ c & \text{si } x > c \end{cases}$$
 (6)

Medidas de Sesgo

Para calcular el sesgo es necesario calcular $(m_i - \mu)^3$ por lo que es sensible a la presencia de datos atípicos. Utilizamos medidas de sesgo de la forma

$$\frac{(Q_{1-p}-Q_{0,5})-(Q_{0,5}-Q_p)}{Q_{1-p}-Q_p} \tag{7}$$

para p = 0.125, 0.25. Además consideramos el *medcouple* propuesta por Brys, Hubert y Struyf [3]

$$MC = mediana_{x_i \leq Q_{0,5} \leq x_j} h_1(x_i, x_j)$$
 (8)

con

$$h_1(x_i, x_j) = \frac{(x_{(j)} - Q_{0,5}) - (Q_{0,5} - x_{(i)})}{x_{(j)} - x_{(i)}}$$
(9)

Entropía Diferencial

Utilizamos un estimado de la densidad f por núcleos \hat{f} de las magnitudes que se encuentren entre la mediana y $\pm 6\sigma$ y estimamos la entropía diferencial

$$H(M) = -\int f(m)\log f(m)dm \tag{10}$$

con

$$\hat{H}(M) \approx -\frac{1}{n} \sum_{i} \log \hat{f}(m_i)$$
 (11)

Valor Abbe

El valor Abbe fue propuesto por Mowlavi [6] para detectar curvas de luz con fenómenos transientes. Se define como

$$A = \frac{n}{2(n-1)} \frac{\sum_{i} (m_i - m_{i-1})^2}{\sum_{i} (m_i - \mu)^2}$$
 (12)

y puede ser calculado en subintervalos de tiempo. Si $\mathcal{A}_{t,i}$ es el valor abbe calculado en $[m_i - \Delta t/2, m_i + \Delta t/2]$, tomamos

$$\bar{\mathcal{A}}_t = \frac{1}{n} \sum_{i=1}^n \mathcal{A}_{t,i} \tag{13}$$

para $\Delta t = 5d, 10d, 20d, 50d, 100d, 200d, 500d, 750d$.

Classificación

Evaluamos el desempeño diferentes álgorítmos de aprendizaje:

- k vecinos más cercanos
- Árboles de clasificación
- Bosques Aleatorios
- Máquinas de soporte vectorial
- Un algorítmo híbrido

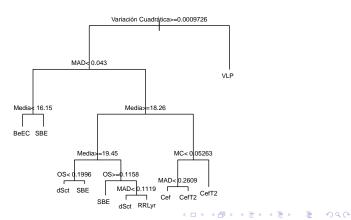
k Vecinos Más Cercanos

A un punto a clasificar se la asigna la clase a la cual pertenece la mayoría entre sus k vecinos más cercanos. Se sabe que este método es consistente, es decir, que tiende al clasificador de Bayes cuando el tamaño de la muestra tiende a infinito. Existe una implementación abierta en el paquete FNN para R.

Referencia	BeEC	Cef	δ Scuti	SBE	VLP	RRLyr	CefT2
BeEC	0.762	0.001	0.000	0.001	0.000	0.000	0.002
Cef	0.061	0.832	0.004	0.005	0.000	0.041	0.166
$\delta Scuti$	0.000	0.001	0.648	0.014	0.000	0.010	0.000
SBE	0.080	0.016	0.175	0.915	0.001	0.027	0.078
VLP	0.076	0.007	0.003	0.022	0.998	0.001	0.090
RRLyr	0.021	0.139	0.169	0.041	0.000	0.920	0.289
CefT2	0.000	0.004	0.000	0.001	0.000	0.002	0.376
Sensitividad	0.762	0.832	0.648	0.915	0.998	0.920	0.376
Error	0.038	0.008	0.018	0.003	$\sim 10^4$	0.003	0.039

Árboles de Clasificación

Arbol de clasificación construido con todos los datos y todas las variables. Solo se utilizan 5 de las 17 variables originales y el estimado del error por resubstitución es 21 %



Árboles de clasificación

Fue propuesta por Breiman, Friedman, Stone y Olshen [2]. Creamos un árbol de decisión para clasificar los datos. Para clasificar un nuevo dato se hacen preguntas binarias de tipo $x_i \leq \alpha_i$. Si la respuesta es sí se procede al siguiente nodo de la izquierda y si es no, a la derecha. En los nodos terminales se le asigna al nuevo dato una etiqueta.

Existe una implementación libre en el paquete rpart para R.

	BeEC	Cef	δ Scuti	SBE	VLP	RRLyr	CefT2
BeEC	0.903	0.008	0.002	0.026	0.007	0.001	0.005
Cef	0.038	0.824	0.005	0.009	0.041	0.259	0.214
$\delta Scuti$	0.000	0.003	0.907	0.165	0.000	0.079	0.002
SBE	0.002	0.011	0.054	0.624	0.000	0.034	0.020
VLP	0.011	0.013	0.009	0.058	0.890	0.002	0.005
RRLyr	0.000	0.002	0.022	0.002	0.000	0.509	0.007
CefT2	0.046	0.139	0.002	0.116	0.060	0.116	0.748
Sensitividad	0.903	0.824	0.907	0.624	0.890	0.509	0.748
Error	0.027	0.008	0.011	0.005	0.001	0.005	0.035

Bosques aleatorios

Fue propuesto por Breiman [1]. Se crean árboles de clasificación que son clasificadores débiles pero que están poco correlacionados. Se toma la decisión utilizando la regla de la mayoría. Los árboles son creados utilizando un subconjunto aleatorio de las características en cada nodo y solo se construyen árboles poco profundos.

Referencia	BeEC	Cef	δ Scuti	SBE	VLP	RRLyr	CefT2
BeEC	0.842	0.000	0.000	0.000	0.000	0.000	0.000
Cef	0.002	0.810	0.002	0.001	0.000	0.018	0.138
δ Scuti	0.000	0.000	0.738	0.004	0.000	0.004	0.000
SBE	0.103	0.013	0.146	0.959	0.000	0.014	0.108
VLP	0.048	0.009	0.004	0.020	1.000	0.001	0.111
RRLyr	0.004	0.166	0.110	0.016	0.000	0.963	0.299
CefT2	0.000	0.002	0.000	0.000	0.000	0.000	0.345
Sensitividad	0.842	0.810	0.738	0.959	1.000	0.963	0.345
Error	0.033	0.009	0.016	0.002	0.000	0.002	0.038

Máquinas de Soporte Vectorial (SVM)

Consideremos un problema de dos clases, es decir, $J=\{-1,1\}$. Queremos una regla de decisión

$$g(\vec{x}) = sign(\langle \vec{w}, \vec{x} \rangle + b) \tag{14}$$

que maximice la distancia de los puntos al plano perpendicular a \vec{w} En el caso en que los datos sean linealmente separables, podemos encontrar el plano resolviendo el problema de optimización

minimizar
$$\langle \vec{w}, \vec{w} \rangle$$
 sujeto a $j_i(\langle \vec{w}, \vec{x_i} \rangle + b) \geq 1, \ i = 1, \ldots, N.$

Máquinas de Soporte Vectorial (SVM)

Si los datos no son linealmente separables, podemos introducir variables de holgura ξ_i y resolver

$$\begin{split} & \underset{\vec{w},b,\vec{\xi}}{\text{minimizar}} & \langle \vec{w},\vec{w} \rangle + C \sum_i \xi_i^2, \\ & \text{sujeto a} & j_i (\langle \vec{w},\vec{x_i} \rangle + b) \geq 1 - \xi_i, \\ & \xi_i > 0, \\ & i = 1,\dots,N, \end{split}$$

cuyo problema dual es

$$\begin{split} & \underset{\vec{\alpha}}{\text{maximizar}} & & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,k=1}^{N} j_i j_k \alpha_i \alpha_k \left(\langle \vec{x}_i, \vec{x}_k \rangle + \frac{1}{C} \delta_{ik} \right) \\ & \text{sujeto a} & & \sum_{i=1}^{N} j_i \alpha_i = 0, \\ & & \alpha_i > 0, i = 1, \dots, N, \end{split}$$

Máquinas de Soporte Vectorial (MSV)

El problema de optimización solo depende de la matriz $\langle \vec{x_i}, \vec{x_j} \rangle$. Podemos encontrar una función $K(\vec{x}, \vec{y})$ que cumpla $K(\vec{x}, \vec{y}) = \langle \phi(\vec{x}), \phi(\vec{y}) \rangle_H$ para cierta función $\phi: \mathbb{R}^n \to H$ siendo H algún espacio con producto interno y encontrar el plano separador en H resolviendo el problema de optimización.

maximizar
$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,k=1}^{N} j_i j_k \alpha_i \alpha_k K\left(\vec{x}_i, \vec{x}_k\right)$$
sujeto a
$$\sum_{i=1}^{N} j_i \alpha_i = 0,$$
$$\alpha_i \geq 0, i = 1, \dots, n.$$

Máquinas de Soporte Vectorial (MSV)

Referencia	BeEC	Cef	δ Scuti	SBE	VLP	RRLyr	CefT2
BeEC	0.891	0.000	0.000	0.001	0.000	0.000	0.000
Cef	0.002	0.869	0.002	0.002	0.000	0.016	0.128
δ Scuti	0.000	0.000	0.769	0.006	0.000	0.005	0.002
SBE	0.048	0.008	0.117	0.963	0.001	0.013	0.075
VLP	0.053	0.004	0.005	0.016	0.999	0.001	0.070
RRLyr	0.006	0.115	0.107	0.012	0.000	0.964	0.265
CefT2	0.000	0.004	0.000	0.001	0.000	0.001	0.461
Sensitividad	0.891	0.869	0.769	0.963	0.999	0.964	0.461
Error	0.028	0.007	0.016	0.002	$\sim 10^{-5}$	0.002	0.040

Cuadro: Resultados de la clasificación con MSV

Propuesta

Podemos utilizar diferentes métodos para clasificar diferentes clases. Mostramos los resultados utilizar árboles de clasificación para separar primero Cefeidas y Cefeidas Tipo 2 del resto de tipos de variabilidad. Luego usamos MSV para distinguir entre clases en estos dos grupos.

Referencia	BeEC	Cef	δ Scuti	SBE	VLP	RRLyr	CefT2
BeEC	0.815	0.000	0.000	0.001	0.000	0.000	0.000
Cef	0.065	0.976	0.023	0.098	0.009	0.359	0.411
δ Scuti	0.000	0.000	0.769	0.006	0.000	0.005	0.000
SBE	0.036	0.004	0.103	0.784	0.000	0.009	0.013
VLP	0.051	0.007	0.003	0.013	0.987	0.001	0.078
RRLyr	0.006	0.010	0.102	0.007	0.000	0.608	0.023
CefT2	0.027	0.002	0.001	0.092	0.004	0.018	0.474
Sensitividad	0.815	0.976	0.769	0.784	0.987	0.608	0.474
Error	0.035	0.003	0.016	0.004	$\sim 10^{-4}$	0.005	0.040

Cuadro : Estimados para las tasas de clasificación al clasificar con un método que combina árboles de clasificación y MSV

Conclusiones



Conclusiones

- Es posible utilizar variables descriptivas de la densidad de magnitudes para clasificar curvas de luz.
- Este tipo de clasificadores puede ser utilizado como primera aproximación a una base de datos nueva. Así, se puede reducir el tiempo humano empleado al clasificar curvas de luz.
- La mejor abordar este problema de clasificación es utilizar una combinación de clasificadores.

Preguntas y Respuestas

¡Gracias!

Referencias I



Leo Breiman.

Random forests.

Machine learning, 45(1):5-32, 2001.



Leo Breiman, Jerome Friedman, Charles J. Stone, and R. A. Olshen.

Classification and Regression Trees.

Chapman and Hall/CRC, New York, N.Y., 1 edition edition, January 1984.



G. Brys, Mia Hubert, and A. Struyf.

A robust measure of skewness.

Journal of Computational and Graphical Statistics, 13(4), 2004.

Referencias II



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.



Peter J. Huber.

Robust statistics.

Springer, 2011.



N. Mowlavi.

Searching transients in large-scale surveys. A method based on the Abbe value.

Astronomy and Astrophysics, 568:78, 2014.

Referencias III



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.



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.

Referencias IV



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.



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.

Referencias V



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.



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.

Referencias VI



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.



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.

Referencias VII



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.



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.

Referencias VIII



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.



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.

Referencias IX



- 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.



- 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.