Búsqueda de estrellas variables extragalácticas usando algoritmos de Machine Learning

Javier Alejandro Acevedo Barroso
201422995

Director: Alejandro García

30 de octubre de 2019

Resumen

La clasificación de estrellas de acuerdo a las variaciones de su brillo es una de las actividades astronómicas más importantes desde finales del siglo XIX. Esta ha llevado a la detección de estrellas binarias, al mejoramiento de la escala de distancias, y a fuertes avances en astrofísica estelar. Por lo anterior, existen numerosos proyectos recolectando datos, en cantidades cada vez más grandes, con el fin de encontrar y clasificar estrellas variables. Los métodos tradicionales de búsqueda de estas estrellas se vuelven ineficientes ante ese tamaño de datos. Entonces, es necesaria la exploración de diferentes técnicas para automatizar la búsqueda y tener una clasificación fiable las estrellas variables.

En este proyecto se busca entrenar un clasificador de estrellas variables que reciba series de tiempo y devuelva candidatos a estrellas variables. Se procesarán datos públicos del proyecto Araucaria de la galaxia NGC 55, NGC247 y NGC7793 para obtener series de tiempo y utilizar el clasificador sobre ellas. Se reducirán observaciones en los filtros B y V para 25 a 30 épocas tomadas con el instrumento Wide Field Imager del telescopio MPG/ESO en La Silla. Se hará fotometría PSF y crossmatch de las observaciones utilizando la suite de software astronómico DAO de Peter Stetson, y se obtendrán series de tiempo. Posteriormente, se usará el clasificador ya entrenado sobre las series y se generará un catálogo de estrellas candidatas. Por último, se revisarán las candidatas y se reportarán las estrellas variables. El objetivo final del proyecto es generar catálogos de estrellas variables en cada galaxia.

Como muestra de entrenamiento se utilizará las series de tiempo del proyecto OGLE (Optical Gravitational Lensing Experiment). Para el clasificador se usarán algoritmos de vanguardia como: Random forest, Gradient boosted forest y diferentes arquitecturas de redes neuronales, entre otros. El código se escribirá principalmente en Python 3 haciendo uso de librerías libres como Numpy, Scikit-learn, Astropy, etc. Dado el alto volumen de datos, se usará el Cluster de cómputo de alto rendimiento de la Facultad de Ciencias.

1. Introducción

La clasificación de estrellas de acuerdo a sus propiedades ópticas ha sido una de las tareas más útiles de la astronomía y astrofísica moderna. El proceso permite segregar estrellas y luego estudiar los mecanismos propios de cada categoría de forma independiente. Por ejemplo, las primeras estrellas variables se registraron durante el siglo XV, pero no fue sino hasta principios del siglo XX que se clasificó sus curvas de luz y se estudiaron las propiedades de las diferentes clases; en particular, esto llevó al descubrimiento de la relación periodo-luminosidad en las variables Cefeidas [1] y la formulación del mecanismo κ.

Adicionalmente, usando la relación periodo-luminosidad de una población de estrellas Cefeidas se puede medir su distancia a la tierra. Esto se usa de la mano con calibraciones basadas en paralaje estelar para calcular distancias a galaxias cercanas y es parte fundamental de la escala de distancias. Por lo anterior, todas las mediciones que impiquen distancias mayores a 10 Mpc dependen fuertemente del cálculo de distancias usando variables Cefeidas, en particular, el parámetro de Hubble. Así, se vuelve esencial el mejoramiento de la precisión en la escala de distancias. En este contexto, nace el «Araucaria Project».

El Proyecto Araucaria es una colaboración iniciada en el año 2000 entre astrónomos de instituciones chilenas, estadounidenses y europeas; con el fin de mejorar la precisión de la escala de distancias. El proyecto hizo seguimiento durante al menos un año y medio a diferentes galaxias cercanas con el fin de generar curvas de luz de sus poblaciones estelares, y usar las curvas para el cálculo de distancia. Para el cálculo final de la distania se usó diferentes métodos dependiendo de las poblaciones obtenidas; en particular, si se encontró una población de estrellas Cefeidas, se usó el método de relación periodo-luminosidad. Adicionalmente, un año después de cada toma de datos, estos se publican en el catálogo de ESO para uso de parte de la comunidad astronómica internacional.

Junto al proyecto Araucaria, está el proyecto OGLE (Optical Gravitational Lensing Experiment) [2]. OGLE busca encontrar evidencia de materia oscura a partir de su efecto de microlente gravitacional sobre estrelals de fondo. Para ello, construyeron en 1997 el telescopio de 1.3-m de Varsovia en el observatorio «Las Campanas» en Chile [3]; y desde entonces han mantenido un monitoreo fotométrico constante. Entre los resultados del proyecto se encuentra un catálogos de estrellas variables con sus correspondientes curvas de luz.

Paralelamente, en los años noventa resurge el Machine Learning (aprendizaje de máquinas) como principal línea de investigación dentro de la Inteligencia Artifical, lo que llevó a un rápido avance en algoritmos y técnicas. Sin embargo, los análisis de los proyectos mencionados anteriormente hacen uso de métodos más tradicionales de la astronomía para la búsqueda de estrellas variables, y no de los novedosos algoritmos de su época. Con todo lo anterior, se vuelve interesante implementar un clasificador de estrellas variables usando algoritmos de Machine learning, entrenar el clasificador usando el catálogo de estrellas variables de OGLE, y utilizar el clasificador para encontrar

estrellas variables en los datos públicos del proyecto Araucaria.

Estado del arte

Los estudios fotométricos de las galaxias de interés se pueden rastrear a finales de los años 30 para NGC7793 [4], a inicios de los sesenta para NGC55 [5, 6], y finales de los años setenta para NGC247 [7-9]. Desde entonces hasta los años noventa se caracterizó su composición química, distancia, perfil de luminosidad, perfil cinemático, metalicidad, regiones de formación estelar y hasta polvo intergaláctico [10-16]

Adicionalmente, el Proyecto Araucaria empieza a operar en el año 2000 y publica sus primeros resultados sobre las galaxias de interés durante la misma década. El Proyecto encontró variables Cefeidas en las tres galaxias y calculó su distancia usando la relación Periodo-Luminosidad [17-19]. Además, ha realizado estudios adicionales en infrarrojo cercano para confirmar las distancias [20-22].

Por otro lado, el proyecto OGLE ha publicado catálogos de estrellas variables para las nubes de Magallanes [23-26], el bulbo galáctico [27, 28], y otras regiones de la Vía Láctea [29, 30]. Los catálogos se encuentran disponibles bajo el catálogo general «OGLE Collection of Variable Stars»[1].

Sumado a esto, la detección de estrellas variables se hace tradicionalmente estudiando la tendencia de la curva desviación-magnitud de la población para generar una lista más reducida de estrellas candidatas. Luego, estudiar las curvas de luz y los periodogramas de tales candidatas y clasificarlas [31].

Sin embargo, desde los noventa y en particular en la última década se han trabajado nuevas técnicas de clasificación haciendo uso de métodos de Machine Learning para sistematizar la búsqueda y mejorar los resultados en la selección de estrellas candidatas [32, 33]. La metodología usual durante principios de la década fue proyectar las curvas de luz en un espacio de características, y alimentar los algoritmos con las proyecciones. Las características deben ser seleccionadas de forma inteligente para conservar la información importante y descartar la superflua (como número de puntos en la curva de luz) [34-36]. Los algoritmos utilizados fueron principalmente regresiones logísticas, Random Forest, k-vecinos más cercanos y Support Vector Machine. No obstante, se han desarrollado metodologías alternativas tales como: en vez de proyectar las curvas de luz en el espacio de parámetros, usar la curva completa y métodos basados en redes neuronales para la clasificación [37]; o utilizar un esquema de meta-clasificación para evitar problemas de grano fino y mejorar el recall del clasificador, para luego clasificar los elementos de la meta-clase en las categorías finales [38].

[1] Disponible en `http://ogledb.astrouw.edu.pl/~ogle/OCVS/` .

2. Marco Teórico

A continuación se presenta brevemente los conocimientos necesarios para el desarrollo del proyecto.

2.1. El Proyecto Araucaria

Nace en el año 2000 con el objetivo de mejorar la calibración de la escala de distancia en el universo local, a través del Esto principalmente a través de estudiar y caracterizar los efectos de la edad y la metalicidad en la determinación de distancias usando poblaciones estelares [39].

El proyecto hace uso del telescopio de Varsovia de 1.3 m en el Observatorio de Las Campanas (LCO) y el telescopio de 2.2 m MPG/ESO en el Observatorio de la Silla. Ambos telescopios cuentan cámaras de campo amplio.

Procedimentalmente, el proyecto observa durante largos periodos de tiempo a galaxias del Grupo Local y el Grupo del Escultor. Las imágenes se toman principalmente en los filtros V e I, pero también hay noches con imágenes en los filtros B y R. Para el cálculo de distancia el proyecto utiliza diferentes métodos como la relación periodo-luminosidad de las variables Cefeidas, tip of the red giant branch, red clump, y binarias eclipsantes. Por último, las galaxias estudiadas hasta ahora son: LMC, SMC, Carina, Fornax, Sculptor, IC1613, M33, M81, NGC55, NGC247, NGC300, NGC3109, NGC6822, NGC7793, WLM.

2.2. Generación de curvas de luz

TODO: hablar del proceso de reducción de las imagenes, de la fotometría psf, del crossmatch

2.3. Clasificación usando Machine Learning

TODO: hablar de los algoritmos a usar: Random Forest, redes neuranales, redes LSTM. Hablar del espacio de features y estadistica robusta.

3. Objetivo general

Crear catálogos de estrellas variables sobre galaxias del proyecto Araucaria usando algoritmos de Machine Learning para automatizar la búsqueda y clasificación.

4. Objetivos específicos

- Realizar fotometría PSF usando los datos públicos de las galaxias del proyecto Araucaria NGC55, NGC247 y NGC7793, con el fin de generar series de tiempo.

- Definir un espacio de características significativas de las curvas de luz, y proyectar las curvas en este espacio.
- Diseñar y entrenar un clasificador de estrellas variables utilizando como muestra de entrenamiento el catálogo de series de tiempo del proyecto OGLE (Optical Gravitational Lensing Experiment); y métodos como Random Forest, Support Vector Machine, y diferentes arquitecturas de redes neuronales.
- A partir del clasificador, generar un catálogo de estrellas variables con los datos del Proyecto Araucaria.
- Reencontrar las variables Cefeidas previamente reportadas para estas galaxias.
- Generar un catálogo final de estrellas variables para la galaxia elegida y diagramas de magnitud-color y color-color.
- Generar los diagramas magnitud-color y color-color para todas las estrellas detectadas en las galaxias, así como relación periodo-luminosidad de las variables Cefeidas.

5. Metodología

El proyecto tiene una parte fuerte computacional. Se requiere el uso del Cluster de cómputo de alto rendimiento tanto para la reducción de datos, como para entrenar el clasificador. A continuación se presentan los requerimientos computacionales del proyecto.

Se descargaran imagenes para una galaxia del proyecto Araucaria. Se espera tener datos para al menos veintiocho noches. El proyecto toma imágenes en los filtros B, V, R e I. El total de las imágenes ciencia para una galaxia ocupa alrededor de 35 a 40 Gigabytes. Al incluir las imágenes para la corrección de Bias y Flat, se estima unos 50 Gigabytes. Adicionalmente, durante la reducción se crean archivos temporales de tamaño considerable, por lo que se requiere espacio extra disponible. Por último, para entrenar el clasificador se utilizaran los datos del proyecto OGLE, que pesan menos de 10 Gigabytes. En total, se estima un requisito total de almacenamiento de 120 Gigabytes.

Una vez decidida la galaxia, se descargaran todas las observaciones del proyecto en los diferentes filtros. Posteriormente, se realizaran las calibraciones usando el software IRAF, en particular las tareas ESOWFI y MSCRED, pues fueron escritas particularmente para el tipo de datos a utilizar.

Los algoritmos se escribirán en Python usando librerías de alta eficiencia y optimización como Pytorch, Scikit-learn, Numpy, entre otras. El entrenamiento del clasificador se hará en paralelo usando multiples CPUs y cuando sea posible, multiples GPUs. Para el entrenamiento paralelo en GPUs se utilizará Nvidia CUDA.

Los requisitos de memoria no son tan rígidos porque se puede entrenar el clasificador usando «batches» de datos en vez de la muestra completa; y la reducción de imágenes astronómicas está optimizada para usar poca memoria, pues los programas a usar fueron escritos cuando la memoria RAM disponible era ordenes de magnitud menor. Por lo tanto, los cuatro Gigabytes de memoria por CPU y GPU del Cluster es suficiente.

6. Cronograma

A continuación se presenta el cronograma del proyecto. Los periodos tienen una duración de dos semanas cada uno. Dado que se debe entregar la primera versión del documento final en la semana 11 del segundo semestre de ejecución del proyecto, se diseñó el cronograma con 13 periodos, o 26 semanas.

Tareas \ Periodo	1	2	3	4	5	6	7	8	9	10	11	12	13
1	X	X	X	X	X	X	X	X					
2	X	X	X	X									
3				X	X	X	X						
4					X	X	X	X	X				
5	X	X	X	X	X	X	X						
6	X	X		X	X	X	X						
7			X	X	X	X	X	X	X	X			
8						X	X	X	X	X			
9								X	X	X	X	X	X
10							X	X			X	X	X
11	X	X	X	X	X	X	X	X	X	X	X	X	X

- Tarea 1: revisión bibliográfica.
- Tarea 2: descargar las galaxias del repositorio público de ESO correspondientes al proyecto Araucaria, así como las imágenes de calibración y realizar el correspondiente procesamiento.
- Tarea 3: realizar fotometría PSF sobre las imágenes procesadas y obtener catálogos de magnitud y coordenadas.
- Tarea 4: realizar el cross-matching de las estrellas en los catálogos de fotometría para obtener las series de tiempo.
- Tarea 5: definir un espacio de características en el que se pueda proyectar las curvas de luz reteniendo la mayor cantidad de información para la implementación del método supervisado.

- Tarea 6: construir la muestra de entrenamiento con las estrellas clasificadas del proyecto OGLE y proyectarlas al espacio de características.
- Tarea 7: diseñar un clasificador usando algoritmos de Machine Learning y explorar el espacio de hiperparámetros para optimizar los resultados.
- Tarea 8: usar el clasificador sobre las curvas de luz generadas y formar un catálogo de estrellas candidatas.
- Tarea 9: Inspeccionar las estrellas candidatas, determinar periodos, y reportar el catálogo final de estrellas variables.
- Tarea 10: preparar presentaciones del proyecto.
- Tarea 11: escribir el documento.

7. Personas Conocedoras del Tema

- Dra. Beatriz Sabogal (Universidad de los Andes)
- Dr. Ronnald Mennickent (Universidad de Concepción, Chile)
- Dr. Grzegorz Pietrzyński (Instituto Copérnico, Polonia)
- Dr. Igor Soszyński (Universidad de Varsovia, Polonia)

8. Consideraciones éticas

Todos los datos que se planea usar son públicos y se encuentran disponibles en el catálogo del Observatorio Europeo Austral (ESO, por sus siglas en inglés). Todo el software utilizado para el desarrollo del proyecto es software Libre. No se modificará ninguna muestra de datos. En caso de hacer uso de algoritmos ya propuestos, se incluirá la debida referencia y citación en el documento final.

Referencias

1. Leavitt, H. S. 1777 variables in the Magellanic Clouds. *Annals of Harvard College Observatory* **60,** 87-108.3 (ene. de 1908).
2. Udalski, A. *et al.* The Optical Gravitational Lensing Experiment. Acta Astron. **42,** 253-284 (1992).
3. Udalski, A., Kubiak, M. y Szymanski, M. Optical Gravitational Lensing Experiment. OGLE-2 – the Second Phase of the OGLE Project. Acta Astron. **47,** 319-344 (jul. de 1997).

4. Shapley, H. y Mohr, J. Photometry of Stars and Clusters in the Southern Spiral, NGC 7793. *Harvard College Observatory Bulletin* **907,** 6-12 (ene. de 1938).

5. de Vaucouleurs, G. Southern Galaxies. I. Luminosity, Rotation, and Mass of the Magellanic System NGC 55. ApJ **133,** 405 (mar. de 1961).

6. Robinson, B. J. y van Damme, K. J. 21 cm observations of NGC 55. *Australian Journal of Physics* **19,** 111 (feb. de 1966).

7. de Vaucouleurs, G. The extragalactic distance scale. IV - Distances of nearest groups and field galaxies from secondary indicators. ApJ **224,** 710-717 (sep. de 1978).

8. de Vaucouleurs, G. The extragalactic distance scale. VI - Distances of 458 spiral galaxies from tertiary indicators. ApJ **227,** 729-755 (feb. de 1979).

9. de Vaucouleurs, G. y Davoust, E. Southern galaxies. VIII - Surface photometry of the SD spiral NGC 7793. ApJ **239,** 783-802 (ago. de 1980).

10. Graham, J. A. y Lawrie, D. G. A large shell nebula in NGC 55. ApJ **253,** L73-L75 (feb. de 1982).

11. Carignan, C. Surface photometry of the sculptor group galaxies - NGC 7793, NGC 247, and NGC 300. ApJS **58,** 107-124 (mayo de 1985).

12. Pritchet, C. J. *et al.* The late-type stellar content of NGC 55. ApJ **323,** 79-90 (dic. de 1987).

13. Carignan, C. y Puche, D. H I studies of the Sculptor group galaxies. IV - NGC 247. AJ **100,** 641-647 (sep. de 1990).

14. Wyse, R. F. G. *et al. Diffuse Ionized Gas In Sculptor Group Spirals* en *American Astronomical Society Meeting Abstracts* **27** (dic. de 1995), 1354.

15. Zang, Z., Warwick, R. S. y Meurs, E. J. A. ROSAT PSPC observations of three sculptor group galaxies: NGC 55, NGC 247 and NGC 300. *Irish Astronomical Journal* **24** (ene. de 1997).

16. Zabludoff, A. I. y Mulchaey, J. S. The Properties of Poor Groups of Galaxies. I. Spectroscopic Survey and Results. ApJ **496,** 39-72 (mar. de 1998).

17. Pietrzyński, G. *et al.* The Araucaria Project: The Distance to the Sculptor Group Galaxy NGC 55 from a Newly Discovered Abundant Cepheid Population. AJ **132,** 2556-2565 (dic. de 2006).

18. García-Varela, A. *et al.* The Araucaria Project: the Distance to the Sculptor Group Galaxy NGC 247 from Cepheid Variables Discovered in a Wide-Field Imaging Survey. AJ **136,** 1770-1777 (nov. de 2008).

19. Pietrzyński, G. *et al.* The Araucaria Project: First Cepheid Distance to the Sculptor Group Galaxy NGC 7793 from Variables Discovered in a Wide-field Imaging Survey. AJ **140,** 1475-1485 (nov. de 2010).

20. Gieren, W. *et al.* The Araucaria Project: Near-Infrared Photometry of Cepheid Variables in the Sculptor Galaxy NGC 55. ApJ **672,** 266-273 (ene. de 2008).

21. Gieren, W. *et al.* The Araucaria Project: The Distance to the Sculptor Galaxy NGC 247 from Near-Infrared Photometry of Cepheid Variables. ApJ **700,** 1141-1147 (ago. de 2009).

22. Zgirski, B. *et al.* The Araucaria Project. The Distance to the Sculptor Group Galaxy NGC 7793 from Near-infrared Photometry of Cepheid Variables. ApJ **847,** 88 (oct. de 2017).

23. Soszyński, I. *et al.* The OGLE Collection of Variable Stars. Anomalous Cepheids in the Magellanic Clouds. Acta Astron. **65,** 233-250 (sep. de 2015).

24. Soszyński, I. *et al.* The OGLE Collection of Variable Stars. Classical Cepheids in the Magellanic System. Acta Astron. **65,** 297-312 (dic. de 2015).

25. Soszyński, I. *et al.* The OGLE Collection of Variable Stars. Over 45 000 RR Lyrae Stars in the Magellanic System. Acta Astron. **66,** 131-147 (jun. de 2016).

26. Pawlak, M. *et al.* The OGLE Collection of Variable Stars. Eclipsing Binaries in the Magellanic System. Acta Astron. **66,** 421-432 (dic. de 2016).

27. Soszyński, I. *et al.* Over 38000 RR Lyrae Stars in the OGLE Galactic Bulge Fields. Acta Astron. **64,** 177-196 (sep. de 2014).

28. Soszyński, I. *et al.* The OGLE Collection of Variable Stars. Over 450 000 Eclipsing and Ellipsoidal Binary Systems Toward the Galactic Bulge. Acta Astron. **66,** 405-420 (dic. de 2016).

29. Udalski, A. *et al.* The Optical Gravitational Lensing Experiment. Final Reductions of the OGLE-III Data. Acta Astron. **58,** 69-87 (jun. de 2008).

30. Udalski, A., Szymański, M. K. y Szymański, G. OGLE-IV: Fourth Phase of the Optical Gravitational Lensing Experiment. Acta Astron. **65,** 1-38 (mar. de 2015).

31. Varela, J. A. G. *Estrellas cefeidas en la galaxia espiral NGC 247 del grupo de Sculptor mejorando la calibración de la escala de distancias* Tesis doct. (Universidad de Concepción, Facultad de Ciencias Físicas y Matemáticas Universidad de Concepción, Concepción Chile, ene. de 2008).

32. Naim, A. *The Application of Artificial Neural Networks to Astronomical Classification* en *American Astronomical Society Meeting Abstracts* **187** (dic. de 1995), 88.05.

33. Ball, N. M. *et al.* Robust Machine Learning Applied to Astronomical Data Sets. I. Star-Galaxy Classification of the Sloan Digital Sky Survey DR3 Using Decision Trees. ApJ **650,** 497-509 (oct. de 2006).

34. Richards, J. W. *et al.* On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data. ApJ **733,** 10 (mayo de 2011).

35. Pérez-Ortiz, M. F. *et al.* Machine learning techniques to select Be star candidates. An application in the OGLE-IV Gaia south ecliptic pole field. A&A **605,** A123 (sep. de 2017).

36. Pashchenko, I. N., Sokolovsky, K. V. y Gavras, P. Machine learning search for variable stars. MNRAS **475,** 2326-2343 (abr. de 2018).

37. Naul, B. *et al.* A recurrent neural network for classification of unevenly sampled variable stars. *Nature Astronomy* **2,** 151-155 (nov. de 2018).

38. Pichara, K., Protopapas, P. y León, D. Meta-classification for Variable Stars. ApJ **819,** 18 (mar. de 2016).

39. Pietrzyński, G. y Gieren, W. The Araucaria Project . Mem. Soc. Astron. Italiana **77,** 239 (ene. de 2006).

Dr. Alejandro García
Director.

Javier Alejandro Acevedo Barroso
Estudiante 201422995