Búsqueda de estrellas variables extragalácticas usando algoritmos de Machine Learning

Javier Alejandro Acevedo Barroso
201422995

Director: Alejandro Garcia Varela

20 października 2019

Streszczenie

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, NGC 247 y NGC 7793 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 precisón 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) [4]. 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

TODO: aquí irían todas las citas a los papers del proyecto araucaria en las galaxias de interés, las citas al proyecto OGLE en sus catálogos de estrellas variables y proceso de clasificación, y también las citas a los papers de clasificación estelar con y sin machine learning.

2 Marco Teórico

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

2.1 Araucaria Project

El proyecto Araucaria nace en 2001 con el objetivo principal de mejorar la calibración de la escala de distancia en el universo local. Ello a través de estudiar los efectos de la edad y la metalicidad en la determinación de distancias usando poblaciones estelares [2].

TODO: hablar de la toma de datos, la cámara, las galaxias, las bandas observadas y el telescopio.

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 alguna de las galaxias del proyecto Araucaria tales, como NGC 55, NGC 247 y NGC 7793, con el fin de generar series de tiempo de magnitud para su población estelar.
- Definir un espacio de características significativas de las curvas de luz, y proyectar las curvas en ese espacio.
- Diseñar y entrenar un clasificador de estrellas variables utilizando el catálogo de series de tiempo del proyecto OGLE (Optical Gravitational Lensing Experiment).
- Reencontrar las variables Cefeidas previamente reportadas para esas galaxias.
- Generar un catálogo de estrellas candidatas a estrellas variables con los datos del proyecto Araucaria utilizando el clasificado.
- Encontrar las estrellas de tipo Cefeida previamente detectadas en las galaxias de interés.
- Generar un catálogo final de estrellas variables para la galaxia elegida y diagramas de magnitud-color y color-color.
- Generar diagramas de magnitud-color y color-color para las estrellas variables de las galaxias, así como relación periodo-luminosidad de las variables Cefeidas.

5 Metodología

El proyecto es 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 descargará imagenes para una galaxia del proyecto Araucaria. Se espera tener datos para al menos veintiocho noches. El proyecto toma cinco imágenes por noche por filtro 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, Flat y Dark; 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 utilizará los datos del proyecto OGLE, que pesan menos de 10 Gigabytes. En total, se estima un requisito total de almacenamiento de 120 Gigabytes.

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.

Se descargará unas pocas noches de algunas galaxias del proyecto Araucaria con el fin de elegir la galaxia a trabajar. Una vez decidida la galaxia, se descargará todas las observaciones del proyecto en los diferentes filtos. Siguiente, se realizará las correcciones de astronómico IRAF, en particular las tareas ESOWFI y MSCRED, pues fueron escritas particularmente para el tipo de datos a utilizar.

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 Eugenia Sabogal Martínez (Universidad de los Andes)
- Dr. Ronnald Mennickent (Universidad de Concepción)
- Dr. Grzegorz Pietrzynsky (Universidad de Concepción)
- Dr. Igor Soszyński (Universidad de Varsovia)

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.

Literatura

[1] Henrietta S. Leavitt. 1777 variables in the Magellanic Clouds. *Annals of Harvard College Observatory*, 60:87–108.3, Jan 1908.

[2] G. Pietrzyński and W. Gieren. The Araucaria Project . Mem. Soc. Astron. Italiana, 77:239, Jan 2006.

[3] A. Udalski, M. Kubiak, and M. Szymanski. Optical Gravitational Lensing Experiment. OGLE-2 – the Second Phase of the OGLE Project. Acta Astron., 47:319–344, July 1997.

[4] A. Udalski, M. Szymanski, J. Kaluzny, M. Kubiak, and M. Mateo. The Optical Gravitational Lensing Experiment. Acta Astron., 42:253–284, 1992.

Dr. Alejandro Garcia Varela.
Directo de trabajo de grado.

Javier Alejandro Acevedo Barroso.
Estudiante 201422995