Búsqueda de estrellas variables extragalácticas usando algoritmos de Machine Learning

Javier Alejandro Acevedo Barroso
201422995

Director: Alejandro Garcia Varela

27 de septiembre de 2019

Resumen

La clasificación de estrellas de acuerdo a las variaciones de su brillo es una de las actividades astronómicas más importantes desde el siglo XIX. Esta ha llevado a la detección de estrellas binarias, al mejoramiento de la escalera de distancias cosmológicas, y 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 se han probado efectivos, pero al aumentar el tamaño de la muestra se vuelven ineficientes e incluso arbitrarios. Entonces, es necesaria la exploración de diferentes nuevos métodos para automatizar la búsqueda y mejorar los catálogos de estrellas variables.

En este proyecto se busca gentrenar un clasificador de estrellas variables que reciba series de tiempo y devuelva candidatos a estrellas variables. Adicionalmente, se procesarán datos públicos del proyecto Araucaria de la galaxia NGC 300 para obtener series de tiempo y utilizar el clasificador sobre ellas. Se reducirán observaciones en los filtros BVRI para 28 noches 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.

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», 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

- Realizar fotometría PSF usando los datos públicos de alguna de las galaxias del proyecto Araucaria tales, como NGC 300, 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.

1. 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.

2. 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.

3. 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)

4. Consideraciones éticas

Todos los datos que se planea usar son públicos y se encuentran disponibles en la página de ESO. Todo el software utilizado para el desarrollo del proyecto es software Libre. En caso de hacer uso de software propietario, se tendrá la respectiva licencia. 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.

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

Javier Alejandro Acevedo Barroso.
Estudiante 201422995