

Búsqueda de estrellas variables extragalácticas usando algoritmos de Machine Learning

Javier Alejandro Acevedo Barroso
201422995

Director: Alejandro Garcia Varela

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 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 y número de las observaciones se vuelven ineficientes, e incluso arbitrarios. Entonces, es necesaria la exploración de diferentes métodos para automatizar la búsqueda y mejorar los catálogos de estrellas variables.

En este proyecto se busca entrenar un clasificador de estrellas variables que reciba series de tiempo y devuelva candidatos a estrella variables. 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, 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 un catálogo de estrellas variables para la galaxia NGC 300.

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.