Búsqueda de estrellas variables extragalácticas usando algoritmos de Machine Learning

Javier Alejandro Acevedo Barroso
201422995

Director: Alejandro García

2 de noviembre 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: Bosques Aleatorios, otros métodos de ensambles, y diferentes arquitecturas de redes neuronales. El código se escribirá principalmente en Python 3 haciendo uso de librerías libres como Numpy, Scikit-learn, Astropy, Pytorch, entre otras. 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 [**1908AnHar..60...87L**] 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 impliquen 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 distancia 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) [**1992AcA....42..253U**]. 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 [**1997AcA....47..319U**]; 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 Artificial, 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 usan-

do 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 [**1938BHarO.907....6S**], a inicios de los sesenta para NGC55 [**1961ApJ...133..405D**, **1966AuJPh..19..111R**], y finales de los años setenta para NGC247 [**1978ApJ...224..710D**, **1979ApJ...227..729D**, **1980ApJ...239..783D**]. 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 [**1982ApJ...253L..73G**, **1985ApJS...58..107C**, **1987ApJ...323...79P**, **1990AJ....100..641C**, **1995AAS...187.4809W**, **1997IrAJ...24...45Z**, **1998ApJ...496...39Z**]

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 [**2006AJ....132.2556P**, **2008AJ....136.1770G**, **2010AJ....140.1475P**]. Además, ha realizado seguimientos en infrarrojo para obtener mediciones de distancia con precisión del 1 % [**2008ApJ...672..266G**, **2009ApJ...700.1141G**, **2017ApJ...847...88Z**].

Por otro lado, el proyecto OGLE ha publicado catálogos de estrellas variables para las nubes de Magallanes [**2015AcA....65..233S**, **2015AcA....65..297S**, **2016AcA....66..131S**, **2016AcA....66..421P**], el bulbo galáctico [**2014AcA....64..177S**, **2016AcA....66..405S**], y otras regiones de la Vía Láctea [**2008AcA....58...69U**, **2015AcA....65....1U**]. 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 [**alejandroThesis**].

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 [**1995AAS...187.8805N**, **2006ApJ...650..497B**]. 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) [**2011ApJ...733...10R**, **2017A&A...605A.123P**, **2018MNRAS.475.2326P**]. Los algoritmos utilizados fueron principalmente regresiones logísticas, Bosques Aleatorios, K-vecinos más cercanos y Support Vector Machine. No obstante, se han desarrollado metodologías alternativas

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

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 recurrentes para la clasificación [**2018NatAs...2..151N**]; o utilizar un esquema de meta-clasificación para evitar problemas de grano fino y mejorar la recolección del clasificador, para luego clasificar los elementos de la meta-clase en las categorías finales [**2016ApJ...819...18P**].

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 [**2006MmSAI..77..239P**].

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

Las imágenes a utilizar fueron tomadas con el instrumento Wide Field Imager (WFI) montado en el telescopio MPG/ESO de 2.2 m. La cámara es un mosaico de 4x2 CCDs cada una con una resolución de 2k por 4k. Debido al espacio entre chips, se tomó cinco imágenes seguidas por observación en cada filtro cambiando un poco la posición del telescopio, de forma que se puede llenar el espacio entre las CCDs. Este proceso se conoce como «dithering». Para juntar las cinco imágenes de cada noche y realizar la calibración de Flat y Bias se puede utilizar los paquetes de IRAF: ESOWFI y MSCRED, diseñados específicamente para procesar imágenes de campo amplio como las del instrumento WFI.

Para hacer fotometría de campo denso se utiliza fotometría PSF. En particular, se puede usar el software astronómico DAOPHOT de Peter Stetson [**1987PASP...99..191S**]

para todas las etapas del proceso. Además, para generar las curvas de luz es necesario identificar las estrellas entre las diferentes observaciones. Este proceso se conoce como crossmatch. Una de las aproximaciones al problema es encontrar la transformación de coordenadas entre cada observación con una imagen de referencia, y luego generar catálogos de magnitud contra tiempo de las estrellas que (hasta cierta precisión) ocupen la misma posición. Esto se hace tradicionalmente con los programas DAOMATCH y DAOMASTER, obteniendose como producto final un archivo con las curvas de luz de todas las estrellas detectadas. Esas curvas de luz son las que permitirán detectar variabilidad. Por lo tanto, es esencial obtener el mayor número de estrellas correctamente asociadas entre imágenes, pues de ahí depende la calidad de las curvas de luz y de la búsqueda de estrellas variables.

2.3. Clasificación usando Machine Learning

Machine Learning (ML) nació en los años cincuenta como una rama de la inteligencia artificial profundamente relacionada con la estadística y se refiere a la creación de modelos utilizables por una máquina para predicción o clasificación a partir de un conjunto de datos. Desde los años noventa tomó su propia dirección como ciencia propia gracias al mejoramiento de los algoritmos, el rápido crecimiento de los conjuntos de datos y el mejoramiento de los computadores. El uso de algoritmos y técnicas de ML en la astronomía comenzó tan temprano como 1990 [**1993VA.....36..141M**] con redes neuronales artificiales.

La metodología estándar para trabajar con curvas de luz es crear un espacio de características en el cual proyectar las curvas y alimentar los algoritmos con los datos proyectados en tal espacio. Esto está fuertemente motivado por la irregularidad del muestreo en curvas de luz, pues las condiciones de observación son muy erráticas. El espacio de parámetros puede estar compuesto por muchos parámetros con alta correlación entre sí [**2018MNRAS.475.2326P**], o por pocos parámetros con baja correlación y naturaleza robusta para dar cuenta del comportamiento global de los datos [**2017A&A...605A.123P**]. A continuación, se presentan los métodos de ML más usados en variabilidad estelar:

- Bosques Aleatorios: la idea del algoritmo es convertir el espacio de características en un conjunto de combinaciones pequeñas. Por ejemplo, a partir de un espacio con 20 características, hacer 50 grupos con 4 características aleatorias cada uno. Luego, con las características de cada grupo se crea un árbol de clasificación, la clasificación final se elige con alguna regla de selección entre los árboles. Un árbol de clasificación es un clasificador particularmente bueno para separar regiones no lineales pero sufre de sobreajuste a los datos de entrenamiento, ahí brilla la principal ventaja de Bosques Aleatorios: debido a que cada árbol de regresión ajustará solo una pequeña parte de la información, la clasificación se mantiene regularizada y se evita el sobreajuste. Dada la alta dimensionalidad del espacio de parámetros,

el sobreajuste es una de las principales problemáticas en la clasificación de nuevos datos, llevando a la alta eficacia del método en clasificación estelar.

- Métodos de ensambles: Bosques Aleatorios es un ejemplo particular de un tipo de métodos más generales llamados Métodos de ensambles. La idea principal es que al tener muchos clasificadores «simples», combinarlos usando alguna regla de selección puede llevar a mejores resultados. Además de Bosques Aleatorios, existen métodos de «Boosted Gradient». Estos métodos en vez de combinar los clasificadores con una simple regla de mayoría, asignan un peso a cada clasificador que se van actualizando a través de un proceso de optimización multivariable («Gradient descent»). La ventaja de estos métodos es que, al igual que Bosques Aleatorios, evitan el sobreajuste usando clasificadores que individualmente tiene poco poder predictivo, e igualmente, se han probado con éxito en clasificación estelar [**2019arXiv190606628K**]. Al combinar los clasificadores usando pesos y un algoritmo de optimización, se está mejorando el comportamiento del ensamble completo y maximizando el poder predictivo sin reducir mucho la varianza, de forma que el clasificador final es preciso y no pierde mucha eficacia al pasar al conjunto de prueba.

- Redes Neuronales: continuando con la idea de clasificadores de ensamble que además actualizan los pesos de cada clasificador interno. Una red neuronal es un clasificador de ensamble donde cada neurona corresponde a una regresión con una función no lineal y sus conexiones a los pesos de las regresiones. Las Redes Neuronales, incluso las más simples, tienen considerablemente mayor poder predictivo comparado con una única regresión de una función no lineal (como una regresión logística). Para evitar el problema del sobreajuste existen diferentes métodos de regularización como penalizar los pesos muy grandes (L1,L2), o apagar aleatoriamente conexiones de la red (dropout). Además, se han desarrollado arquitecturas de redes muy variadas para fines ajenos a la astronomía, como visión de computadores, predicción en series de tiempo, reconocimiento de imágenes, entre otras. Estas arquitecturas pueden terminar siendo particularmente útiles en problemas de clasificación estelar y vale la pena probarlas en datos de alto volumen y alta dimensionalidad.

  Las diferentes arquitecturas de redes neuronales cambian aspectos en la conección de las neuronas, las operaciones entre secciones de la red, las funciones de activación, y hasta el algoritmo de optimización. Las arquitecturas más interesantes para el problema de variabilidad estelar son las Redes Neuronales Convolucionales (CNN por sus siglas en inglés) y las Redes Neuronales Recurrentes (RNN). Las CNN se han probado con éxito en reconocimiento de imágenes unidimensionales. Para aplicarlas a curvas de luz, es necesario primero reconstruir los puntos faltantes en las curvas para que estas sean de tamaños regulares y poderlas interpretar como imágenes, este problema se puede abordar con Procesos Gaussianos. Por

otro lado, la arquitectura de las redes RNN usualmente empieza con la transformación de la serie de tiempo en un vector de parámetros con dimensión fija. Luego, se alimenta ese vector a una red cuyo resultado vuelve a alimentarse a la red numerosas veces (de ahí la parte recurrente) y al final se optiene una clasificación similar al caso CNN. Este enfoque se ha probado con éxito en clasificación de curvas de luz de estrellas variables [**2018NatAs...2..151N**]

3. Objetivo general

Crear catálogos de estrellas variables para las galaxias NGC55, NGC247 y NGC7793 con las observaciones del proyecto Araucaria, y utilizando algoritmos de Machine Learning para la búsqueda y clasificación estelar.

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, métodos de ensambles, 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 las tres galaxias.
- 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 es principalmente 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 del proyecto.

El principal costo computacional viene del almacenamiento de los datos a utilizar. Se espera tener datos para al menos veintiocho noches y tres galaxias. El proyecto toma imágenes en los filtros B, V, R e I. El total de las imágenes ciencia para una galaxia ocupa alrededor de 40 Gigabytes. Al incluir las imágenes para la corrección de Bias y Flat, se estima unos 70 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 400 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 astronómico IRAF [**Tody86theiraf**], en particular las tareas ESOWFI y MSCRED, pues fueron escritas y optimizadas para este tipo de datos. El proceso de crossmatch se hará con los programas DAOMATCH y DAOMASTER de Peter Stetson. Adicionalmente, se explorará la posibilidad de usar STILTS [**2006ASPC..351..666T**] y se compará resultados.

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.

- 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.

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 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 Pietrzyńsky (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.

Dr. Alejandro García Director.	Javier Alejandro Acevedo Barroso Estudiante 201422995