Análisis de curvas de luz y estructuración de los resultados en HDF 5

Valentina Contreras Rojas
Departamento de Astronomía
Universidad de La Serena, La Serena, Chile.
Herramientas Computacionales para la Astrofísica
Profesor Cristian Vega Martínez

Diciembre, 2023

1. Introducción

En el ámbito de la investigación astronómica, el análisis de series temporales de magnitudes de luz estelar, también conocidas como curvas de luz, desempeña un papel esencial para desentrañar fenómenos físicos como pulsaciones estelares, explosiones de supernovas, entre otros. Estas curvas de luz representan gráficamente la intensidad de la luz emitida por un objeto celeste o región en función del tiempo. Por lo general, se construyen utilizando datos fotométricos de una región específica del espectro electromagnético.

Estas son comúnmente empleadas en el estudio de objetos celestes variables y se encuentran almacenadas en bases de datos especializadas, tales como OGLE y VSX. Sin embargo, el manejo de estas extensas bases de datos puede resultar desafiante debido a su tamaño, lo cual impacta tanto en el tiempo de procesamiento como en el espacio ocupado en la memoria. En este contexto, este trabajo propone la utilización de un formato de archivo más eficiente en términos de tamaño, como es el formato HDF5, para abordar estas limitaciones y permitir un análisis más efectivo de las curvas de luz estelar.

2. Contexto científico

El proyecto se fundamenta en el trabajo de tesis de magíster titulado "Búsqueda de nuevas estrellas simbióticas en la Vía Láctea: utilizando técnicas de aprendizaje profundo aplicadas a datos S-PLUS". El objetivo principal consiste en incrementar el conocimiento sobre estrellas simbióticas en nuestra galaxia, caracterizándolas a partir de un catálogo conocido y correlacionándolas con algunas base de datos conocidas. Sin embargo, los primeros datos enviados, que originalmente iban a ser utilizados para este proyecto, no mostraron ninguna coincidencia con estas estrellas, por lo que fueron descartados.

En el mismo contexto, se solicitó tiempo de telescopio en el T-80 para capturar 30 imágenes de campo, cada una abarcando aproximadamente 1^o, centradas en una estrella simbiótica. Estas imágenes se obtuvieron utilizando el filtro de banda angosta F660, equivalente a la línea espectral H_α. El propósito es utilizar estas imágenes para construir una curva de luz en dicho filtro, la cual se comparará posteriormente con registros públicos en bandas ópticas e infrarrojas. Con el fin de facilitar esta investigación, se necesita desarrollar un algoritmo capaz de extraer de manera homogénea ciertas características de la onda, asegurando su consistencia para permitir una clasificación automática en momentos posteriores.

3. Problemas principales

En un principio, el desafío consistía en la lectura y manipulación de archivos de gran tamaño sin depender de supercomputadoras, con el objetivo de minimizar el tiempo de procesamiento. No obstante, los datos involucrados en esta instancia fueron descartados, ya que contenían información poco relevante para la investigación. Como resultado, el proyecto se amplió para abordar el análisis de las curvas de luz, y estructurar los resultados en un archivo de fácil manipulación, lectura y con un tamaño reducido, siendo este el objetivo principal.

4. Herramientas seleccionadas

Para realizar este proyecto, se necesitaron principalmente 3 herramientas que formaron parte del contenido del curso:

4.1. Librerías de python

- **Numpy:** es un paquete fundamental para la programación científica en python, provee de arreglos multidimensionales, objetos derivados y una variedad de rutinas para operaciones rápidas en matrices, incluidas matemáticas, lógicos, manipulación de formas, clasificación, selección y transformadas discretas de Fourier. En este trabajo, principalmente se ocupa para comparar los resultados de las funciones que posteriormente serán descritas como por ejemplo el promedio de las magnitudes de una curva de luz, la desviación estándar, etc; así como también para hacer operaciones matemáticas sencillas y buscar máximos.
- **Scipy:** es una colección de algoritmos y funciones de conveniencia creadas en la extensión de Numpy diseñados para la optimización, integración, interpolación, resolución de problemas de valores propios, ecuaciones algebraicas, etc. En esta ocasión se utiliza para obtener frecuencias de los armónicos de la curva de luz.
- **Pandas:** es una biblioteca de código abierto que proporciona estructuras de datos y herramientas de análisis de alto rendimiento fácil de usar en python. Esta solo se utiliza en el proyecto para leer los archivos de entrada por comodidad de la autora.

4.2. HDF 5 (Hierarchical Data Format)

HDF 5 es un formato que ordena los datos jerárquicamente en un conjunto de archivos diseñados para almacenar y organizar grandes cantidades de datos. Este formato es compatible con muchas plataformas de software y lenguajes de programación completamente gratuito. A lo largo del proyecto, se ocupan su herramienta de visualización $HDF5view$ y su respectiva librería en python $h5py$, que no fue nombrada en el apartado anterior. Ambas herramientas son parte fundamental de proyecto ya que garantizan un formato jerárquicamente ordenado, manipulable y de poco espacio.

4.3. Doxygen:

La última herramienta seleccionada es llamada Doxygen, cuya finalidad es generar documentación a partir de un codigo fuente en diferentes formatos (HTML, Latex, RTF, pdf, etc). Este es útil para orientarse en las distribuciones de un código grande mediante los gráficos de dependencia que genera. En el proyecto actual, produjo la documentación correspondiente a las tres clases principales que se ocupan para analizar la curva de luz, de las cuales se señaló sus atributos mediante la palabra clave @var, la descripción de la clase con @brief y la entrada de datos con @param. El resultado es un archivo en Látex que más adelante se detallará.

5. Hoja de ruta

Como se nombró anteriormente, el trabajo en un principio se trataba de un conversor de tablas del tipo csv a formato HDF 5 con el fin de poder manipularlas sin tener que exigirle al equipo utilizar recursos excesivos. Sin embargo, al comenzar a trabajar por bloques estos archivos (idr4 del S-PLUS), estos datos debieron ser desechados ya que no contenían ninguna información de interés.

Posteriormente, se converso con el profesor del curso la idea de expandir el proyecto con la idea de trabajar ingeniería de datos, sin embargo, al comenzar a indagar sobre lo mismo (leer en varias página web e intentar entender) y encontrar una conexión con la idea anterior fue sumamente complicado considerando que no se tenían

datos para que este ambito fuera trabajado, por lo que se opto por analizar las curvas de luz, que es lo que describe el actual proyecto. Esta decisión también fue impulsada por la continuidad del trabajo de tesis que se esta llevando a cabo.

Este se describe en los siguientes pasos:

- **Selección de características:** Antes de iniciar el desarrollo del programa, se llevó a cabo una cuidadosa selección de las características que se pretendían extraer de la curva de luz. Esta elección se basó principalmente en la utilización de la librería feets de Python, la cual proporciona una variedad de parámetros automáticos extraíbles de cualquier serie temporal, y además, se tomó en cuenta el conocimiento adquirido durante un taller de investigación realizado por la misma estudiante, donde se clasificaron las curvas de luz y se priorizaron las características según su relevancia.

  El conjunto final de características seleccionadas fue una combinación equilibrada entre parámetros estadísticos y funciones de onda. Entre estos se incluyen el promedio de las magnitudes, la mediana, la varianza, la desviación estándar, la amplitud, la curtosis, la asimetría, los percentiles 20, 35, 50, 65 y 80 de las magnitudes, la frecuencia fundamental y su periodo asociado, la frecuencia de los primeros cuatro armónicos, así como también el respectivo periodograma.

- **Programar las funciones:** Luego de la selección, se comenzaron a programar diferentes funciones para analizar la curva de luz en el lenguaje de mas bajo nivel posible con el fin de no ignorar la matemática que existe detrás de todos estos parámetros nombrados anteriormente. Además todas ellas fueron comparadas con librerías de mas alto nivel para saber si funcionaban correctamente o requerían de algún tipo de corrección.

- **Construcción de las clases:** a partir de las funciones anteriores, y del concepto de herencia en programación, las diferentes funciones fueron agrupadas en tres diferentes clases llamadas 'estadisticaestelar', 'estadisticaestelarextendida' y 'analisisdefrecuencias', a las cuales se le agregó su respectiva descripción, su función de inicialización, se describieron los parámetros de entrada y sus respectivos atributos o parámetros de salida. La agrupación de estas clases se baso principalmente en dos conceptos claves: la naturaleza de los atributos (si eran estadisticos o parámetros que describen una onda) y la herencia que tenían, por ejemplo, la desviación estándar necesitaba la varianza de la clase anterior para poder ser calculada, por ende, su función se agrupo en la clase 'estadisticaestelarextendida'.

- **Importar a HDF 5:** Para finalizar con la construcción del código se creó una función que conducía a la creación de un archivo en hdf 5 diseñado para guardar información con los mismos tres grupos que las clases. En esta función se describieron los respectivos atributos y se exportaron los datos de una sola curva de luz elegida azaharozamente desde el entrecruzamiento de los datos de H_α y vsx echo con TOPCAT previo a la realización de este proyecto.

- **Documentación:** El código previamente mencionado fue documentado utilizando una herramienta llamada *Doxygen*, la cual se detalló en la sección anterior. Para emplear esta herramienta, se accede a través de la terminal de los sistemas, y se debe ajustar el documento de configuración. En dicho documento, se especificó que la salida debía estar en formato LaTeXy que el idioma de la documentación sería el inglés. Previamente a esta documentación, se incorporaron al código las palabras reservadas como brief, var y param, según las indicaciones de Doxygen. Estas palabras se utilizaron para describir la clase, las variables de salida (atributos) y los parámetros de entrada, respectivamente.

6. Resultados

A continuación se enumeran los resultados del avance del proyecto:

- **Código en python:** el resultado principal del proyecto es el código en python que permite analizar las curvas de luz de manera homogénea. Este consta de tres clases en las que relaciona la serie temporal con las magnitudes de la luz estelar. A continuación se detallan las partes del codigo:

  - Clase estadistica Luz Estelar: en esta clase se da como parámetro de entrada las lista que contiene las magnitudes de cada curva y se obtienen como atributos el promedio, la mediana y la varianza a través de tres funciones, mas la función que inicializa el proceso.

- Clase estadistica de luz estelar extendido: en esta clase es una extension de la anterior que añade los atributos de amplitud, kurtosis, autocorrelacion, asimetria, desviación estandar y percentiles.

  Cabe destacar que esta clase necesita de algunos parámetros de la función anterior como por ejemplo el promedio para que sus atributos sean calculados, es por ello que aunque ambas traten la estadística, están separadas.

- Clase Analisis Frecuencia: esta clase ocupa como entrada un arreglo de datos que contiene las magnitudes estelares y los tiempos asociados medidos en días juliano.

  El único atributo que esta devuelve es la frecuencia fundamental que complementa otros métodos adicionales como lo son el calculo de la frecuencia de los armónicos, el calculo del periodo y del periodograma respectivo a la curva de luz.

- Función guardar_en_hdf5: esta función guarda las estadísticas y frecuencias en un archivo HDF5. Toma las instancias de las clases EstadisticasLuzEstelar, EstadisticasLuzEstelarExtendido, y AnalisisFrecuencias, y guarda sus resultados en un archivo hdf5 llamado resultadosza sea creando el mismo o actualizándolo.

  NOTA: el código se puede ver con mayor detalle en:
  https://github.com/Valeign/HCAIA2023/blob/main/Proyecto/Código%20funciones.ipynb

- **Tabla de datos:** Como se describió en el ítem anterior, el código genera una tabla la cual contiene los siguientes tres grupos enumerados con los respectivos datos que guarda:

Grupo `EstadisticasLuzEstelar`

- `promedio`: Promedio de las magnitudes de luz estelar.
- `mediana`: Mediana de las magnitudes de luz estelar.
- `varianza`: Varianza de las magnitudes de luz estelar.

Grupo `EstadisticasLuzEstelarExtendido` (hereda de `EstadisticasLuzEstelar`)

- `amplitud`: Amplitud de las magnitudes de luz estelar.
- `kurtosis`: Kurtosis de las magnitudes de luz estelar.
- `autocorrelacion`: Coeficiente de autocorrelación de las magnitudes de luz estelar.
- `asimetria`: Asimetría de las magnitudes de luz estelar.
- `desviacion_estandar`: Desviación estándar de las magnitudes de luz estelar.
- `percentiles`: Tupla que contiene los percentiles 20, 35, 50, 65, 80 de las magnitudes de luz estelar.

Grupo `AnalisisFrecuencias`

- `frecuencia_fundamental`: Frecuencia fundamental de la serie temporal de magnitudes de luz estelar.
- `frecuencias_armonicos`: Frecuencias de los primeros cuatro armónicos.
- `periodo_asociado`: Período asociado a la frecuencia fundamental.
- `frecuencias_periodograma`: Frecuencias del periodograma de la serie temporal.
- `potencia_periodograma`: Potencia asociada a cada frecuencia en el periodograma.

Este archivo en HDF 5 solo se puede observar en algún visualizador de dicha distribución, por lo tanto no esta disponible en ningún link.

- **Documentación:** como se dijo anteriormente, se obtuvo el documento con la información apropiada para comprender el código el cual se divide en 3 secciones incluyendo la explicación de jerarquía de clases, el listado de las mismas y la explicación de cada clase junto a su respectiva funcionalidad. El archivo se puede encontrar en el siguiente link: https://www.overleaf.com/read/scmqvbjtqjtn1d8021

7. Conclusión

En resumen, el proyecto ha alcanzado con éxito la extracción y análisis de características de curvas de luz astronómicas. La cuidadosa selección de parámetros, basada en la librería feets de Python y en la experiencia adquirida durante el taller de investigación, ha resultado en la construcción de un conjunto robusto de características, que abarcan tanto aspectos estadísticos como funciones de onda. Estas clases y características serán implementadas en la investigación asociada a la tesis de magíster titulada "Búsqueda de nuevas estrellas simbióticas en la Vía Láctea: utilizando técnicas de aprendizaje profundo aplicadas a datos S-PLUS". El objetivo de esta investigación es establecer correlaciones entre las curvas de luz capturadas en fotometría de banda angosta H_α y los datos ópticos o infrarrojos recopilados de diversas bases de datos relacionadas con VSX.

Adicionalmente, se quiere destacar que el hecho de programar utilizando clases y agregar sus respectiva descripción y atributos ha conducido a comprender mejor los diagramas UML, los cuales se utilizaran en un futuro para el diseño y la implemetación efectiva de futuros proyectos. Esta comprensión fortalece la calidad y la coherencia de la estructura del código, lo que agrega valor al producto de este trabajo.

Por último, se espera que la estructura jerárquica proporcionada por HDF5 facilite de manera eficiente la organización, el almacenamiento y la manipulación de los datos, incluyendo las diversas características extraídas de las curvas de luz. Este enfoque contribuirá significativamente a la eficiencia general del análisis astronómico realizado en el marco de este proyecto.

8. Trabajos futuros

A partir del trabajo descrito anteriormente, se espera que se pueda cumplir con los siguientes requerimientos para que se de por completados:

- **Expansión de resultados:** en primer lugar, se espera que se pueda replicar el mismo análisis para todas las curvas de luz involucradas en la investigación, incluyendo aquellas que se obtendrán con la fotometría de banda angosta, y que al igual que la curva de luz utilizada en el proyecto, se guarden los datos en la misma tabla de HDF 5.
- **Input de datos:** considerando que los datos de las curvas de luz en bandas ópticas e infrarrojas provienen de diferentes bases de datos, proponer un input que pueda resolver las columnas a utilizar de manera independiente y automática, reconociendo las columnas que contengan el tiempo y las magnitudes es vital para darle continuidad al proyecto.
- **Comparaciones:** en los últimos pasos de este trabajo, se requiere medir el tiempo que se demora el algoritmo creado versus uno que funcione con un mayor nivel para saber si el tiempo de procesamiento efectivamente se ve reducido con este algoritmo y confirmar que los datos realmente son certeros.
- **Machine Learning:** por último, se quiere comprobar la selectividad de las características usando un clasificador de machine learning creado anteriormente que clasifique las diversas curvas de luz, e identifique posibles objetos de interés para la tesis anteriormente mencionada, o sea, estrellas variables tipo Mira.