UNIVERSIDAD DE LA REPÚBLICA
Facultad de Ingeniería

Juego de datos aplicado a la astronomía

Introducción a la ciencia de datos

Tarea Final - Curso 2024

Autoría
Matias López

Montevideo, 2024

Contenido

Introducción

El objetivo de este proyecto es poder determinar de la forma más acertada posible la membresía de distintas estrellas en diferentes grupos estelares jóvenes y móviles en el entorno solar (NYMG por su siglas en inglés: *Nearby Young Moving Groups*). Los resultados que surjan a partir de este trabajo son de gran utilidad para la búsqueda de exoplanetas, estrellas de baja masa, estudiar el proceso de formación y evolución estelar, comprender la formación planetaria en torno a discos circumestelares, entre otros.

Los NYMG son grupos de estrellas co-móviles que poseen posiciones y velocidades similares, se ubican típicamente a menos de $\sim$ 200 pc del Sol y poseen edades menores a $\sim$ 100 Myr. Actualmente se conoce muy poco acerca de poblaciones jóvenes, por lo que estos grupos, dada su juventud y proximidad, presentan una gran oportunidad para estudiar la evolución estelar a edades tempranas.

Este juego de datos contiene información cinemática y espectroscópica acerca de un gran número de estrellas recopilada por distintos sondeos principalmente GAIA[1], GALAH[2], APOGEE[3] y LAMOST[4]. Estos sondeos fueron desarrollados con el objetivo principal de estudiar la estrucutra y evolución de la Vía Lactea por lo que, dado su porte e importancia, poseen una gran cantidad de información estelar, siendo las velocidades radiales y las abundancias químicas las más relevantes para este trabajo.

[1] https://sci.esa.int/web/gaia
[2] https://www.galah-survey.org
[3] https://www.sdss4.org
[4] https://www.lamost.org

Calidad de datos

Este conjunto de datos no presenta grandes problemas en cuanto a su calidad debido a que la información recopilada por los sondeos mencionados en la Introducción fue realizada con instrumentos de muy alta calidad por lo que representan un alto nivel de precisión, exactitud y consistencia. Aún así, dadas las limitaciones, bajas pero existentes, de estos sondeos, no podemos evitar la presencia de valores nulos dado que no todas las estrellas presentan información cinemática y/o química debido al alcance de estos proyectos. Como mencionamos en la Introducción, estas dos cantidades resultan ser las más importantes en nuestro análisis por lo que los casos faltantes deberán tratarse con cuidado. Sin pérdida de generalidad, se cuenta con información suficiente como para poder clasificar aquellas estrellas para las que sí se tiene información.

Dado que la precisión en este trabajo es de suma importancia, el tratamiento de valores nulos no debería resolverse a la ligera. El escenario ideal sería poder obtener estos valores mediante observaciones espectroscópicas directas[1], una tarea compleja que no entra dentro de los objetivos de este proyecto. Una alternativa posible sería aplicar técnicas de aprendizaje automático para inferir estas cantidades dada la información ya existente, poniendo particular atención al arrastre de los errores a la hora de aplicar nuevas técnicas de aprendizaje automático encima de éstas.

Independientemente de la incompletitud, no solo en los grupos estelares antes mencionados, sino también en los datos, la calidad de los mismos permite el análisis exploratorio sin mayores dificultades y da lugar a elaborar preguntas que pueden resultar de gran interés para este proyecto.

[1] Objetivo de tesis de grado de Matías López: "Diseño de un protocolo y propuesta de observación para establecer membersías y candidaturas en los grupos estelares jóvenes del entorno solar"; Facultad de Ciencias, UdelaR; Montevideo, 2022.

Métodos de aprendizaje

Este proyecto evidencia una clara oportunidad para aplicar alguna técnica de aprendizaje automático supervisado, y en particular, un método de clasificación. Hablamos de aprendizaje supervisado dado que hoy en día ya contamos con información de estrellas miembros de ciertos grupos, las cuales serán de gran utilidad para el entrenamiento, validación y testeo de nuestro modelo. Aún así, es posible que existan agrupaciones de las que no tenemos conocimiento y por ende, en ese caso estaríamos frente a un caso de aprendizaje no supervisado. En nuestro caso en particular, nos interesa tratar la completitud de agrupaciones conocidas por lo que utilizaremos una técnica de aprendizaje supervisado. Es importante aclarar que la clasificación implica un problema de multi-clase, por lo que el modelo a desarrollar debe ser capaz de clasificar entre diferentes clases posibles, y no de forma binaria.

Una técnica que podríamos utilizar es KNN, o K vecinos más cercanos, una técnica muy utilizada en problemas de clasificación y que nos permitiría clasificar aquellas estrellas en algunos de los grupos ya conocidos dadas sus características individuales. Otras técnicas que pueden resultar interesantes en este problema son: SVM, o Support Vector Machines, Random Forests o modelos de boosting de gradiente como XGBoost.

Cualquiera sea el modelo escogido, el objetivo será clasificar nuevas estrellas en alguna de las clases existentes, es decir las agrupaciones de estrellas ya conocidas, a partir de un conjunto de variables a seleccionar. Algunas de las variables/features a utilizar son: posición, velocidad, velocidad radial, abundancia de litio, fotometría, edad, entre otras cantidades físicas de interés como la temperatura, radio, distancia, densidad, etc. Cabe destacar que muchas de estas cantidades son medidas en un sistema de referencia dado, por lo que suelen ser tridimensionales, lo que aumenta la dimensionalidad de nuestro problema.

Obtención de resultados

Para responder las preguntas planteadas en este proyecto, creemos que el proceso de trabajo tentativo debería seguir los siguientes pasos:

1. Preprocesamiento de datos: colección de datos de diferentes fuentes donde se encuentra almacenada la información y realizar su limpieza correspondiente, tratando datos faltantes, considerando formatos y tratamiento de posibles outliers
2. Análisis exploratorio de datos: observar distribuciones y estadísticas de variables, reducción de dimensionalidad por componentes principales, interpretar posibles patrones y diferenciar grupos, graficar distribuciones de velocidades o mapas de densidad
3. Entrenamiento del modelo: separar un porcentaje de los datos para entrenamiento y testeo, y posible validación para evaluar performance, considerando posible estratificación y estandarización de datos que pertenencen a diferentes dominios
4. Evaluación del modelo: analizar las métricas del modelo como la precisión, el recall o el f1-score, aplicar validación cruzada entre variables considerando diferentes combinaciones de parámetros para obtener el mejor resultado
5. Visualización de resultados: construir una matriz de confusión para analizar los resultados de aciertos entre las clases consideradas, graficar la clasificación resultante de forma bidimensional para facilitar la interpretabilidad
6. Aplicación del modelo a datos desconocidos para inferir la clase a la que pertenecen

Conclusiones

Vimos la importancia que tienen los NYMG para entender diversos eventos astronómicos en diferentes áreas de la astronomía, desde la astrofísica y las ciencias planetarias, hasta la astronomía galáctica y extragaláctica. Es por eso que la completitud de estos grupos estelares es sumamente necesaria y demandada por muchos autores dedicados al estudio de estos fenómenos.

La calidad de datos de este conjunto no presenta demasiadas complicaciones, salvo la existencia de datos faltantes cuya información podría marcar la diferencia en caso de existir. El tratamiento de nulos debe ser tratado con cuidado para no afectar los resultados posteriores, dada la sensibilidad que presentan las cantidades desconocidas (cinemáticas y espectroscópicas).

Encontramos que los problemas relacionados a los NYMG pueden ser atacados mediante la aplicación de técnicas de aprendziaje automático, tanto supervisado como no supervisado. Dado que contamos con catálogos ya disponibles de estrellas miembros de diferentes agrupaciones, las técnicas supervisadas paracen ser las mejores candidatas en este proyecto, en particular el uso del algoritmo KNN por su simpleza e intepretabilidad.

Existen otros métodos de clasificación que podrían ser de ayuda dada la alta dimensionalidad de este problema como por ejemplo SVM, Random Forests, XGBoost, y quizas también la construcción de arquitecturas de Redes Neuronales. Aún así, creemos que se pueden alcanzar resultados muy interesantes a partir de métodos más simples y no tan complejos.

Finalmente, los resultados que obtengamos a partir de este trabajo, en conjunto con el protocolo y la propuesta de observación desarrollada en la tesis de grado de Matias López, serían de gran ayuda para comprender aún más estas agrupaciones, fortalecer la clasificación de nuevos candidatos a miembros y generar nuevas discusiones que pueden desencadenar nuevos trabajos a futuro.