Aplicaciones de Data Mining en Ciencia y Tecnología

TP1. Aplicaciones en astronomía: Las *Hyades*

Un *cluster* abierto de estrellas es un grupo de estrellas que se originaron a partir de la misma nube molecular y que actualmente conservan interacciones gravitatorias entre ellas, pero débiles. Los *cluster*s abiertos pueden llegar a tener varios miles de estrellas, en general son jóvenes, menos de algunos cientos de millones de años, y en ellos todavía ocurren procesos de formación de estrellas. Debido a estas características, las estrellas de un *cluster* tienden a tener características químicas y edades similares.

Las *Hyades* es un *cluster* abierto, el más cercano al Sistema Solar, se ubica a unos 151 años-luz, y probablemente por su cercanía es uno de los *cluster*s abiertos mejor estudiados. Tiene unas 300 a 400 miembros, con un núcleo de 17,6 años luz de diámetro con estrellas cercanas entre sí, rodeada por otro grupo de estrellas más separadas entre sí que se extienden en una región de 130 años luz. Por fuera de este límite exterior se contaron cerca de un tercio de las estrellas de la *Hyades*, en un halo extendido, donde la atracción gravitatoria del *cluster* se debilita, lo que facilita el escape de sus miembros.

Objetivo del trabajo práctico

El objetivo del trabajo práctico es encontrar en los catálogos *Hipparcos* y *Tycho* -producidos ambos por la misión *Hipparcos*-, estrellas que podrían pertenecer a las *Hyades*. Para esto se cuenta con una lista de estrellas que los astrónomos asignaron a las *Hyades*, a partir de los datos de ellas se buscarán estrellas con características similares.

Datos

En la página web del curso hay un archivo Excel llamado “Hyades_source.xlsx” con tres hojas. La primera, “*Symbad*”, es un listado de estrellas que pertenecen a las *Hyades*, indicando el nombre y las coordenadas. La siguiente hoja contiene un listado de estrellas del catálogo *Hipparcos* correspondiente a la región del espacio donde se encuentran las *Hyades*. La tercera hoja es un listado de estrellas del catálogo *Tycho*, también de la misma región del espacio.

El segundo archivo necesario para realizar este trabajo es “id_cruzada_Symbad_Hipparcos.xlsx”, en éste se encuentra un listado de estrellas pertenecientes a las *Hyades* que tienen asignado un identificador en *Hipparcos*.

Procedimiento

1. El listado *Symbad* de miembros de las *Hyades* sólo incluye un identificador de la estrella, su posición indicada por la ascensión recta (*RA*) y la declinación (*DE*) y en algunos pocos casos el tipo espectral (*Sp_type*). Estos datos son insuficientes para poder postular candidatos por similitud. La misión *Hipparcos* recolectó más datos, tales como movimiento propio y mediciones precisas de la magnitud. Varias de las estrellas del listado *Symbad* de las *Hyades* fueron relevadas por *Hipparcos*, por lo que se pueden agregar los datos de este segundo catálogo al primer listado.
2. Sin embargo, en la mayoría de los casos no hay identificadores cruzados entre ambos listados. Por otra parte, las coordenadas registradas en ambos listados para las mismas estrellas no son necesariamente las mismas, debido a diferente instrumentación y errores de medición. Para solucionar esto se pueden calcular las distancias entre los registros de *Symbad* e *Hipparcos* para las estrellas del archivo "id_cruzada_Symbad_Hipparcos.xlsx", que son aquellas *Hyades* que sí tienen un ID confirmado en *Hipparcos*. De esta forma, se puede estimar cuál es el error de localización máximo entre ambos listados. Para completar este paso analizar el código en el archivo "busqueda_estrellas.R".
3. Al co-localizar las estrellas de ambos catálogos obtenemos un listado de estrellas *Hyades* confirmadas enriquecidas con información obtenida de *Hipparcos*.
4. El siguiente paso es realizar agrupamientos de todas las estrellas del listado *Hipparcos* utilizando las variables numéricas. Como método de agrupamiento utilizar K-medias y/o PAM. Evaluar la calidad de los agrupamientos y determinar en qué casos se obtienen *cluster*s que tengan un elevado porcentaje de *Hyades* confirmadas. Estos últimos *cluster*s contendrán las estrellas candidatos a ser consideradas *Hyades*.
5. Repetir la búsqueda del paso anterior para el listado de *Tycho*. Como se mencionó antes este catálogo fue realizado por la misma plataforma que compiló los datos para *Hipparcos*. La diferencia es que las mediciones son menos precisas, por ejemplo, las determinaciones de paralaje son parciales. La mayoría de los registros de *Hipparcos* también se encuentran en *Tycho*, la identificación correspondiente en *Hipparcos* se indica el campo HIP. Estas estrellas no deben incluirse en este segundo análisis porque ya fueron consideradas en el paso anterior.

Pre-informe

Presentar un pre-informe con la identificación cruzada entre las estrellas confirmadas de las *Hyades* y las correspondientes de *Hipparcos*. Indicar cómo se seleccionó el umbral para considerar que registros de diferentes catálogos se corresponden a la misma estrella.

Actividades optativas

Cada grupo puede elegir algunas de las actividades de acuerdo a sus intereses (para aprobar el TP deberán sumar al menos 10 créditos).

1. Para los que les guste programar o investigar algoritmos. Investigar las diferencias en cuanto a complejidad y tiempo de búsqueda de la estrategia de búsqueda propuesta por grillas y una búsqueda por cálculo de distancias de todos contra todos. *Puntaje: 4 créditos.*
2. Agregar al informe final un análisis sobre los datos faltantes de paralaje en *Tycho* ¿Presentan algún patrón? ¿Son todos imputables? Sugerir estrategias de imputación. *Puntaje: 3 créditos.*
3. Determinar qué características diferenciales en cuanto a posición, desplazamiento y características espectrales presentan las estrellas candidatas a *Hyades* obtenidas a partir de *Hipparcos* con respecto a las obtenidas desde *Tycho*. *Puntaje: 3 créditos.*
4. Repetir los agrupamientos y la búsqueda de estrellas candidatas utilizando métodos de *cluster*ing difuso y por densidad. *Puntaje: 4 créditos.*
5. Generar un procedimiento de agrupamiento de estrellas, exportarlo siguiendo la especificación PMML e importarlo en otra plataforma de análisis y ejecutarlo. Por ejemplo, crear el modelo en R e importarlo desde SPSS, o cualquier otra combinación. *Puntaje: 3 créditos.*

Indicaciones para la realización del TP:

- Cada grupo deberá tener hasta **cuatro** integrantes.
- Enviar por mail a **soria@agro.uba.ar** / **juank@dc.uba.ar** e incluir los nombres de todos los miembros del grupo.
- La fecha de entrega del informe es **Domingo 21 de Octubre a las 24hs**. Los trabajos que ingresen después no serán considerados.
- La fecha de entrega del pre-informe es **Domingo 30 de Septiembre a las 24hs**. Los trabajos que ingresen después no serán considerados.
- La presentación del informe final se deberá realizar siguiendo los lineamientos de la **metodología CRISP.**

Aprobación del trabajo práctico

Para aprobar el trabajo práctico deberán entregar el informe indicando los procedimientos utilizados, la validación de los *cluster*s obtenidos, un listado de las estrellas candidatas obtenidas a partir de *Hipparcos* y otro listado con las candidatas adicionales obtenidas a partir de *Tycho* (sin olvidar el ID y catálogo de origen).

Además cada grupo deberá sumar al menos 10 créditos realizando algunas de las actividades optativas.