Aplicaciones en astronomía: las Hyades

Informe

Nicolás Abuchar
Franco Betteo
Milena Dotta
Francisco Valentini

Octubre de 2018

Índice

CHEQUEAR ESTA ESTRUCTURA con el crisp posta

1. Comprensión del dominio

1.1. Información general del dominio

En el presente informe se presenta un análisis realizado con base en (describir datos y cosas de astronomia que esten en la consigna).

1.2. Recursos

El análisis se llevó a cabo a partir de los datos disponibles en tres catálogos de estrellas: (describirlos brevemente) (ver Sección 2.1)

Se usó hardware de uso personal con un procesador XXX y XXX GB de memoria RAM.

(¿acá va R, Rstudio y librerias de R? Los pibes no lo pusieron)

Cuadro 1: Descripción del catálogo Hipparcos

Nombre	Descripción	Tipo
HIP	Identificador	character
RA_J2000	xxx	double
DE_J2000	xxx	double
Plx	Paralaje	double
pmRA	xxx	double
pmDE	xxx	double
Vmag	xxx	double
B-V	xxx	double

1.3. Objetivo de data mining

El objetivo principal del trabajo es (identificar Hyades potenciales, evaluar la bondad de los candidatos obtenidos, etc, ver consigna).

1.4. Plan del proyecto

No corresponde.

2. Comprensión de los datos

2.1. Recolección inicial de los datos

Los datos analizados provienen de mediciones realizadas por (describir con un poco mas de detalle fuentes de los datos - formato del archivo - libreria de R para leer archivos Excel)

2.2. Descripción de los datos

El catálogo Hipparcos cuenta con 2655 estrellas, sobre las cuales se midieron los atributos descritos en el Cuadro 1. Por su parte, Tycho tiene 16258 registros definidos por los atributos presentados en el Cuadro 2. 2402 de las estrellas de Hipparcos se encuentran identificadas en Tycho.

Para el análisis exploratorio y de clustering, en el caso de Hipparcos no se tuvo en cuenta la variable HIP por tratarse del atributo identificador. En cambio, los atributos TYCID1, TYCID2, TYCID3, HD y HIP de Tycho fueron descartados porque (xxx). Cabe destacar que el campo identificador HIP fue usado posteriormente para realizar una identificación cruzada que permitió descartar de Tycho los candidatos ya identificados como resultado del clusering sobre Hipparcos.

2.3. Exploración de los datos

2.4. Verificación de la calidad de los datos

El análisis de la calidad de los datos consistió en la detección y análisis de datos faltantes. En Hipparcos únicamente la variable 'B-V' cuenta con registros faltantes, en particular en 15 de los 2655 casos.

(aca meter lo de faltantes en tycho si lo hacemos)

Cuadro 2: Descripción del catálogo Tycho

Nombre	Descripción	Tipo
recno	Identificador	character
TYCID1	ID???	character
TYCID2	ID???	character
TYCID3	ID???	character
RA_J2000_24	ra	double
DE_J2000	de	double
pmRA	pmra	double
pmDE	pmde	double
BT	bt	double
VT	vt	double
V	v	double
B-V	bv	double
HD	Identificador de Hyades (?)	character
HIP	Identificador de Hipparcos	character
Plx	Paralaje	double

3. Preparación de los datos

3.1. Selección de los datos

(acá va lo de las variables que no usamos? los pibes lo pusieron antes)

3.2. Limpieza de datos

La principal tarea de limpieza fue el tratamiento de los datos faltantes. En el caso de la variable 'B-V' del catálogo Hipparcos se optó por reemplazar los datos faltantes por la mediana, debido a que solo representaban el 0.56 % de los casos.

En cambio, la variable Plx de Tycho fue omitida antes de realizar el análisis de clustering por la alta proporción de casos faltantes que presentó (86.14 %).

3.3. Construcción de los datos

En ambos datasets se normalizaron las variables a partir de una transformación en z-scores. No se generaron nuevos registros ni atributos.

3.4. Integración de los datos

No corresponde ya que se trabajó con cada catálogo por separado. Sin embargo, cabe destacar que para evitar obtener candidatos de Tycho que ya hayan sido identificados en los agrupamientos de Hipparcos, se omitieron las estrellas correspondientes en Tycho usando el campo de identificación cruzada HIP disponible en el dataset.

3.5. Formateo de datos

No corresponde.

4. Modelado

Antes de llevar adelante los agrupamientos, se computó el índice de Hopkins de tendencia a la clusterización en ambos catalógos, fijando una cantidad de puntos en el espacio al azar equivalente al 10 % de cada dataset. El indicador arrojó un valor de 0.0639 en Hipparcos y 0.0834 en Tycho, lo cual indica tendencias significativas al agrupamiento, y entonces que intentar clusterizar las estrellas es razonable.

4.1. Selección de las técnicas de modelado

Las siguientes técnicas de clustering fueron implementadas sobre los dos catálogos para identificar candidatos: K-Medias, clustering difuso y DBSCAN. En el caso de Tycho se removieron del dataset las estrellas ya identificadas como candidatas en Hipparcos antes de implementar los algoritmos.

4.2. Construcción de los modelos

4.2.1. K-Medias

El algoritmo K-Medias se corrió usando su implementación *stats::kmeans()* disponible en R base. (REFERENCIA??) En cada ejecución del algoritmo se escogieron 20 asignaciones distintas de los centroides, seleccionandóse aquella que minimizara la suma de cuadrados dentro de los grupos.

Para definir la cantidad de agrupamientos óptima en Hipparcos se usaron dos criterios: la maximización del Silhouette promedio (S) y la búsqueda de un punto de quiebre en el scree-plot de la Suma de Cuadrados Dentro de los grupos (SCD). Para tal motivo, se generaron agrupamientos de K-Medias para 10 valores posibles de K, entre 1 y 10. Los resultados se visualizan en la figura 1. Como se observa, la cantidad óptima de agrupamientos bajo estos criterios es 2.

Una vez fijado el K en su valor óptimo, se obtuvo la representación del modelo en lenguaje PMML (ver Anexo). La misma fue importada y ejecutada en Python, obteniendo los mismos resultados que en R.

4.2.2. Clustering difuso

xxx

4.2.3. DBSCAN

xxx

4.3. Evaluación de los modelos

4.3.1. K-Medias

```
## % latex table generated in R 3.5.1 by xtable 1.8-2 package
## % Thu Oct 18 20:40:22 2018
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrr}
##   \hline
##  & FALSE & TRUE \\
##   \hline
## 1 & 325 &  49 \\
```

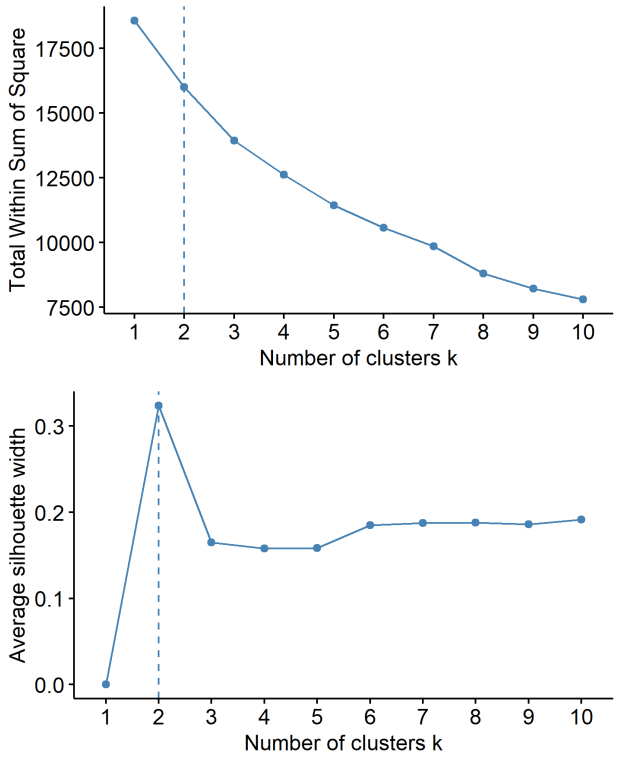

Figura 1: Selección de K para K-Medias (Hipparcos)

```
##   2 & 2280 &   1 \\
##    \hline
## \end{tabular}
## \caption{Distribución de estrellas Hyades en clusters de K-Medias (Hipparcos)}
## \end{table}
```

En el Cuadro **??** se observan la distribución de los clusters generados en función de la pertenencia al grupo de estrellas Hyades. Bajo esta técnica fue posible identificar 325 estrellas candidatas en Hipparcos con un nivel de confianza alto ya que 49 de las 50 Hyades se hallan agrupadas junto a ellas.

(Acá meter silhouette?)

4.3.2. Clustering difuso

xxx

4.3.3. DBSCAN

xxx

5. Evaluación

5.1. Evaluación de resultados

5.2. Proceso de revisión

5.3. Futuras etapas

6. Implementación

No corresponde.

7. Bibliografía (???)

8. Anexo (???)

PONER EL PMML EN UNA PÁGINA ACÁ?