

Universidad Internacional de Valencia

MASTER EN BIG DATA Y DATA SCIENCE

PRÁCTICA 1: CLASIFICADOR DE ESTRELLAS



Universidad
Internacional
de Valencia

Minería de datos

Autor:
Adrián Hernández Padrón
Junio 2022

Índice

1. Descripción del proceso KDD	2
2. Elección de la base de datos y extracción de los datos	2
2.1. Base de datos	2
2.2. Extracción de datos	2
2.3. Data set seleccionado para el estudio	3
3. Salida de los Datos	3
4. Tratamiento de los datos	4
4.1. Limpieza de datos	4
4.2. Detección de estrellas	5
5. Conclusión y discusión de resultados	5

1. Descripción del proceso KDD

El objetivo de este proceso KDD va a ser crear una herramienta que sea un identificador de estrellas. Se va a obtener los datasets de una base de datos astronómica y se aplicará un criterio sobre uno de los datos para poder identificar si los objetos descargados de la base de datos son estrellas o no.

Se usará un programa de creación propia para obtener el dataset de la base de datos astronómica GAIA, una vez se tenga dicho conjunto de datos y hayan limpiado estos datos procederemos a determinar si son estrellas o no. La manera de determinar si un objeto es una estrella o no es mirando su paralaje y su error. Si un objeto cumple lo siguiente:

$$\text{Paralaje} > 3 * \text{error_paralaje} \quad (1)$$

podemos afirmar que este objeto es una estrella.

Se intentará que este proceso KDD sea lo más automático posible, es decir, que el usuario tenga que manipular lo menos posible. La idea es crear un conjunto de herramientas que dada una región del espacio deseada, obtenga todas las estrellas contenidas dentro de la región y que la respuesta a esto sea independiente de la región.

2. Elección de la base de datos y extracción de los datos

2.1. Base de datos

La base de datos con la que se va a trabajar va a ser GAIA, GAIA es un proyecto de la Agencia Espacial Europea (ESA) que tiene como objetivo crear un mapa tridimensional la vía láctea para poder responder preguntas sobre la formación y evolución de la galaxia. Esta base de datos tiene diferentes datasets, puesto que con el tiempo han ido recogiendo datos de diferentes regiones y también mejorando algunos de los datos ya existente.

El conjunto de datos con los que se trabajará serán los correspondientes a los últimos que GAIA ha recogido, el Data Release 3. Este Data Release 3 no está disponible a día de hoy puesto que será lanzado oficialmente el 13 de Junio de 2022.

2.2. Extracción de datos

Para extraer los datos de esta base de datos tenemos que acceder a la interfaz de GAIA, en la cual GAIA permite al usuario acceder a la información mediante el uso queries para solicitar los datos deseados. La forma en la que funciona es bastante sencilla, tiene dos opciones principales:

- Búsqueda por nombre: De esta manera se busca por el nombre de galaxias o objetos conocidos.
- Búsqueda por coordenadas(Equatorial): De esta manera introducimos el centro de la región del espacio de donde queremos obtener los datos.

Figura 1: Interfaz de GAIA para solicitar información.

Ambas opciones funcionan de la misma manera, damos un centro y después seleccionamos un radio determinado, es decir, estamos pasando un rango del espacio y la respuesta de GAIA serán los datos de los objetos existentes dentro de ese radio. La búsqueda por coordenadas está en coordenadas equatoriales que son las llamadas RAJ2000 y DECJ2000.

2.3. Data set seleccionado para el estudio

La idea principal en la extracción de datos de este proceso KDD, es la creación de un programa (Usando Python, Bash o c) que acceda a esta interfaz de GAIA, de manera que se solicite la información del query desde el propio programa. Este programa aceptará unos valores de entrada que serán la construcción del query y después accediendo a la web de GAIA pasaremos este query de manera que la respuesta sea una tabla igual a la que obtendríamos desde la web

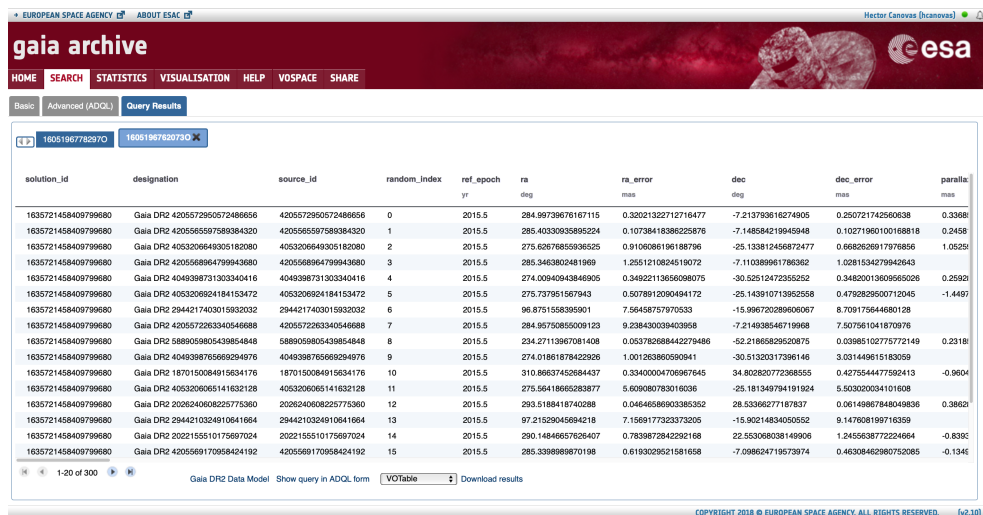
Si no se consigue desarrollar el programa se va a trabajar con un dataset específico o se usarán algunas herramientas ya creada para el acceso a los datos de GAIA.

3. Salida de los Datos

Aunque los datos puedan variar dependiendo de con que región del espacio con la que se va a trabajar, el número de columnas será siempre el mismo:

- source.id: Es el número identificativo único del objeto correspondiente.
- random.index: Es un índice que sigue el orden de la tabla respuesta.
- ref_epoch: Año de creación de los datos.

- ra: Ascensión recta, junto con la declinación forman la coordenada del objeto.
- ra.error: Error en la ascensión recta.
- dec: Declinación, junto con la ascensión recta forman la coordenada del objeto.
- dec.error: Error en la declinación.
- parallax: El paralaje es el cambio aparente observado en la posición de un objeto con respecto al cambio en la posición del observador.
- parallax.error: Error en el paralaje.



solution_id	designation	source_id	random_index	ref_epoch yr	ra deg	ra_error mas	dec deg	dec_error mas	parallax mas
1635721458409799680	Gaia DR2 4205572950572486656	4205572950572486656	0	2015.5	284.98739676187115	0.32051322712716477	-7.213793616274905	0.250721742560638	0.3368
1635721458409799680	Gaia DR2 4205565597598384320	4205565597598384320	1	2015.5	285.40330855895224	0.10726418386226878	-7.146584219845848	0.1327196010016818	0.2458
1635721458409799680	Gaia DR2 4205306649305182080	4205306649305182080	2	2015.5	275.62876855586525	0.910696196198796	-25.133813456873477	0.680626917976985	1.0529
1635721458409799680	Gaia DR2 4205568964799543680	4205568964799543680	3	2015.5	283.3483802481989	1.2551210824519072	-7.110389961786352	1.0281534279542643	
1635721458409799680	Gaia DR2 4049398731303340416	4049398731303340416	4	2015.5	274.0094943848905	0.34022119558086075	-30.52512472355252	0.34820013606565028	0.2592
1635721458409799680	Gaia DR2 4053208924184153472	4053208924184153472	5	2015.5	275.737951587943	0.5078912090494172	-25.143910713962558	0.4782829500712045	-1.4497
1635721458409799680	Gaia DR2 2944217403015932032	2944217403015932032	6	2015.5	96.875155835901	7.55458757970333	-15.986732028960607	8.70917564480128	
1635721458409799680	Gaia DR2 4205572263340546688	4205572263340546688	7	2015.5	284.95750855009123	9.238430029402958	-7.214638548719968	7.507561041876978	
1635721458409799680	Gaia DR2 5889059805439854848	5889059805439854848	8	2015.5	234.27113967081408	0.053782688442279486	-52.21865829520875	0.0398510277572149	0.2318
1635721458409799680	Gaia DR2 404939875698294076	404939875698294076	9	2015.5	274.01861878422926	1.001263860590941	-30.51320317396146	3.031449615183059	
1635721458409799680	Gaia DR2 1870150084915834176	1870150084915834176	10	2015.5	310.88637452684437	0.33400004706967645	34.802820772368555	0.4275544477592413	-0.9604
1635721458409799680	Gaia DR2 4053206065141632128	4053206065141632128	11	2015.5	275.56418665283877	5.609080783016036	-25.181348794191924	5.503020034101608	
1635721458409799680	Gaia DR2 202624068225775360	202624068225775360	12	2015.5	293.5188418740288	0.04646586903385352	28.53366277187837	0.06149867844049836	0.3862
1635721458409799680	Gaia DR2 2944210324910641864	2944210324910641864	13	2015.5	97.21529045694218	7.1569177323373205	-15.902148340550552	9.147608198716359	
1635721458409799680	Gaia DR2 2022155510175697024	2022155510175697024	14	2015.5	290.14846657626407	0.7839872842292168	22.553068038149906	1.2455638772224664	-0.8392
1635721458409799680	Gaia DR2 4205569170958424192	4205569170958424192	15	2015.5	285.3388886870198	0.6193029521581658	-7.098624719573874	0.46308462860752085	-0.1346

Figura 2: Tabla respuesta de GAIA

Los datos mas importantes en este dataset son las coordenadas del objeto (ra y dec) y el paralaje, puesto que este último es necesario para identificar las estrellas. Al tener la posibilidad de elegir los datos de entrada con los queries, se eligieran aquellos que tengan interés práctico según el ejercicio. Los datos de salida por tanto serán los mismos mas una columna añadida que dará información sobre si es una estrella o no.

4. Tratamiento de los datos

Una vez se tenga el dataset deseado en forma de tabla, se comenzará con la manipulación de este. Para ello se va trabajar con él en Python haciendo uso de librerías como Pandas o Numpy para su tratamiento y matplotlib para el uso de gráficas.

4.1. Limpieza de datos

El primer paso será limpiar nuestros datos, se tiene que eliminar cualquier objeto que posea celdas vacías en algún identificador(coordenadas o id) y también eliminar

los objetos que no posean información en el paralaje o en su error puesto que esto podría inducir a errores de cómputo. Así mismo eliminaremos cualquier fila vacía que este dataset pudiera contener.

4.2. Detección de estrellas

Después se determinarán que objetos son estrellas , para ello se tendrá que aplicar el criterio del paralaje mencionado anteriormente al dataset y se va a generar una respuesta que identifique los objetos como estrellas. En cuanto a los outliers que se pudieran encontrar, estos serían aquellos objetos que posean un paralaje anómalo o no acorde con el objeto. Como estamos trabajando con el paralaje y su error, al aplicar el criterio se filtrarán prácticamente todos los objetos.

5. Conclusión y discusión de resultados

Para terminar haremos un análisis completo de todo el proceso KDD realizado identificando los problemas encontrados y proponiendo algunas mejoras que se pudieran realizar en dicho proceso.

Este proceso KDD se puede considerar como una introducción de un proceso que puede ser aún mas potente, lo ideal sería no solo tener un clasificador de estrellas sino, usando tecnicas de machine learning, tener un clasificador completo de los objetos astronómicos. Es decir, teniendo una región del espacio determinada, poder clasificar todos los objetos posibles existentes en ella(estrellas, quásares, galaxias...).