Minería de datos en la misión Gaia: visualización del catálogo, optimización del procesado y parametrización de estrellas

Autor: Marco Antonio Álvarez González
(marco.antonio.agonzalez@udc.es)
Tesis doctoral dirigida por: José Carlos Dafonte Vázquez y Minia Manteiga Outeiro
Centro: Universidade da Coruña
Fecha de lectura: 16 de septiembre de 2019

El trabajo realizado en esta tesis se enmarca en el proyecto *Gaia* de la Agencia Espacial Europea (ESA), cuyo objetivo es medir y procesar los datos sobre posiciones y brillos de más de mil millones de estrellas para generar el catálogo estelar más grande conocido hasta la actualidad, lo que lo convierte en un gran reto para toda la comunidad científica tanto desde el punto de vista computacional como astrofísico.

Para llevar a cabo el procesado y análisis de los datos de *Gaia* se formó un consorcio internacional, denominado *Data Processing and Analysis Consortium* (DPAC), destinado a diseñar e implementar los mecanismos que permiten explotar la ingente cantidad de información que el satélite está obteniendo en la actualidad y que, se estima, será del orden de un Petabyte al final de la misión. Está formado por más de 400 científicos e ingenieros entre los que nos incluimos los miembros del grupo de investigación en el que desarrollé esta tesis.

Nuestro trabajo se basa principalmente en la aplicación de técnicas de Inteligencia Artificial sobre los datos proporcionados por *Gaia*, así como en la elaboración de herramientas que permitan a la comunidad científica utilizar esas técnicas para analizar la información astrofísica que contiene el catálogo.

Concretamente, los objetivos que se pretenden con este trabajo son los siguientes:

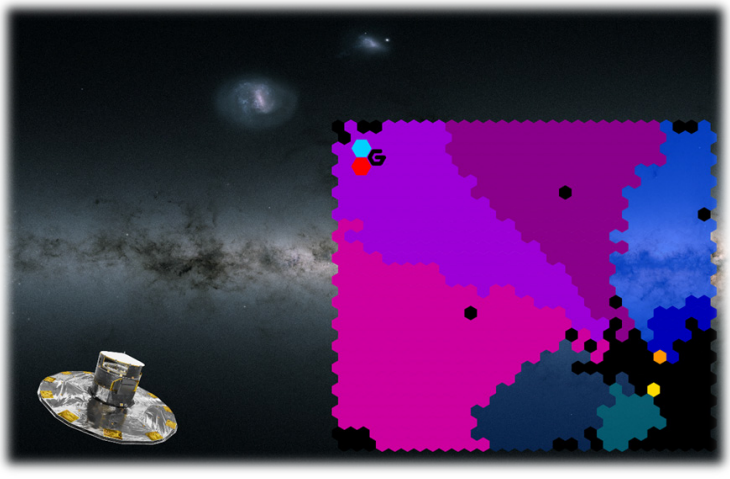
- Aplicar técnicas de aprendizaje supervisado para la estimación de los principales parámetros atmosféricos para las estrellas en las que el instrumento RVS de *Gaia* medirá espectros con suficiente relación señal a ruido: temperatura efectiva, gravedad superficial, metalicidad y abundancia de elementos alfa respecto al hierro. En esta tesis hemos podido validar satisfactoriamente el funcionamiento de nuestro algoritmo con medidas del instrumento RVS a disposición para el consorcio DPAC.
- Proporcionar a la comunidad científica una herramienta útil para la búsqueda y análisis de conjuntos de datos homogéneos, en el sentido de similitud de sus distribuciones espectrales de energía (SED) obtenidas con *Gaia*, mediante la aplicación de un algoritmo de aprendizaje no supervisado. Esta herramienta permite clasificar volúmenes gigantescos de datos por lo que la optimización del tiempo de cómputo del algoritmo es un factor esencial. Se explican las técnicas que permiten a esta herramienta procesar millones de datos en un tiempo reducido.
- Desarrollar una herramienta que facilita el análisis de los resultados obtenidos por la técnica de clasificación sobre millones de objetos estelares, permitiendo visualizar desde diferentes perspectivas las diferentes clases de estrellas, lo que posibilita llevar a cabo un análisis estadístico y, en general, explorar sus características. Dado que esta herramienta trabaja en un entorno *Big Data* el tratamiento de los datos adquiere un papel primordial.

En todos los casos, la gran cantidad de datos a tratar sugiere la necesidad de aplicar técnicas de procesamiento distribuido para evitar un consumo de recursos excesivo (tiempo de ejecución y uso de memoria) que puede llegar a impedir una ejecución satisfactoria de los métodos propuestos. Procesar toda esta información en el marco del proyecto *Gaia* requiere una capacidad de cómputo importante, por lo que para reducir estos tiempos se realizan optimizaciones mediante técnicas de computación distribuida, como es *Apache Spark,* y mediante técnicas de procesado gráfico, como es CUDA.

Otro aspecto importante es que el software resultante debe ser integrado dentro de las cadenas de ejecución existentes en DPAC y desplegado en los centros de procesado asociados, lo que requiere de un proceso de adaptación del *software* original para la plataforma de destino.

Por último, se demuestra la utilidad de la técnica de aprendizaje no supervisado en otros ámbitos diferentes al de la astrofísica, demostrándose por ejemplo que es capaz de mejorar la detección de intrusiones en tráfico de redes de comunicaciones o que puede servir de ayuda en la generación de perfiles de usuarios para mejorar aplicaciones de marketing online.

Tesis disponible en:
https://ruc.udc.es/dspace/handle/2183/23974



Mapa SOM representando una clasificación estelar por tipo espectral generada con nuestra herramienta GUASOM, sobre un fondo de la Vía Láctea generado con los datos de la DR2 de Gaia.