Proyectos de Minería de Datos, Descubrimiento de Conocimiento y Estadística en grandes archivos astronómicos: el grupo de astroestadística del Observatorio Virtual Español.

L.M. Sarro, M. García Torres, M. López, A. Berihuete, M.J. Márquez, F. García Sedano

Abstract Parte del trabajo realizado por el SVO consiste en desarrollar y probar técnicas de descubrimiento de conocimiento en grandes bases de datos astronómicas. Los archivos que la tecnología VO pone a disposición de la comunidad astronómica contienen gran cantidad de información que, analizada con las herramientas adecuadas, puede dar lugar a nuevos descubrimientos científicos. En el grupo de astroestadística del SVO trabajamos en la aplicación de algunas de estas técnicas tomadas de la Estadística y la Inteligencia Artificial a grandes bases de datos astronómicas. He aquí algunos ejemplos...

1 Explotación del archivo del instrumento SUMER abordo de SOHO

En esta línea de investigación combinamos el Análisis de Componentes Independientes y la descomposición wavelet de imágenes del archivo del instrumento SUMER de SOHO (ver figura 1) con los siguientes objetivos:

L.M. Sarro
L.M. Sarro, Universidad Nacional de Educación a Distancia, e-mail: lsb@dia.uned.es

M. García
M. García, Laboratorio de Astrofísica Espacial y Física Fundamental

M. López
M. López, Laboratorio de Astrofísica Espacial y Física Fundamental

A. Berihuete
A. Berihuete, Universidad de Cádiz

M.J. Márquez
M.J. Márquez, Universidad Nacional de Educación a Distancia

F. García
Universidad Nacional de Educación a Distancia

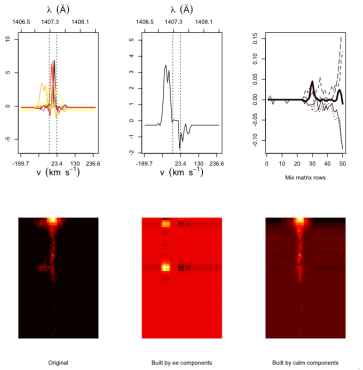

Fig. 1 Ejemplo de análisis y descomposición ICA de un espectro espacial de SUMER.

1. Crear un catálogo de episodios explosivos en espectros bidimensionales del Sol
2. Separar las series temporales de espectros bidimensionales en sus componentes independientes. Analizar la evolución de la distribución espacial de brillo, desplazamientos Doppler, componentes y retardos.
3. Analizar las propiedades estadísticas de la muestra en función de las diferentes temperaturas de formación de las líneas y posición en el disco solar (efecto de perspectiva y geometría).
4. Determinar la geometría del proceso de reconexión magnética que tiene lugar en la atmósfera solar.

2 Participación en DPAC (Data Processing and Analysis Consortium) de GAIA

El SVO se ha implicado en dos unidades de coordinación del DPAC de Gaia: CU7 dedicada al estudio de los datos de variabilidad del catálogo de Gaia y CU8 dedicado a la obtención de parámetros estelares. En CU8, se dirigimos el paquete de trabajo encargado del desarrollo de software para el análisis de agrupamiento de la base de datos. En CU7 se dirigen los paquetes de trabajo de clasificación no supervisada y estudios de variabilidad global y se participa en el de clasificación supervisada. Dentro del grupo de estudios de variabilidad global se dirige el paquete de trabajo de evaluación de calidad por comparación estadística de muestras procedentes de diferentes instrumentos y se participa en el de estimación de sesgos.

Los objetivos concretos de la participación se pueden resumir en:

1. (CU7 y CU8) Desarrollo de software Java para el cálculo distribuido de soluciones de agrupamiento para la base de datos de Gaia (109 y 108 casos respectivamente). Evaluación de algoritmos y mejora. El objetivo es disponer de técnicas apropiadas para la detección de nuevas clases/subclases de objetos. Con ello, se está en disposición de adaptar la cadena de procesamiento de datos durante la misión si Gaia descubre nuevas clases de objetos.

2. (CU7) Diseño e implementación del sistema de evaluación de la calidad de la cadena de procesado de CU7: análisis estadístico de surveys de variabilidad (OGLE, ASAS, MACHO...) Desarrollo de técnicas de comparación estadística de muestreos. Cálculo de funciones de densidad de probabilidad en espacios de alta dimensión.
3. (CU7) Diseño e implementación de clasificadores automáticos basados en descomposición jerárquica de esquemas de clasificación. Empleo de técnicas de selección de variables y evaluación estadística de los clasificadores. Técnicas Clasificadores basados en redes bayesianas, Máquinas de Vectores Soporte, Combinación bayesiana de clasificadores, etc... técnicas de filtrado y envoltura para la selección de atributos. Algoritmos de clustering basados en densidad, computación distribuida, ajuste de curvas principales y clustering o agrupamiento paramétrico Estimación kernel de funciones de densidad de probabilidad optimizadas para grandes bases de datos, detección de correlaciones en espacios multidimensionales, estimación y corrección de sesgos, test de hipótesis en múltiples dimensiones.

3 Desarrollo de sistemas de clasificación automática de objetos variables para COROT

El grupo de astroestadística del SVO participa en el procesado de datos procedentes de COROT en colaboración con el grupo de variabilidad estelar de la Universidad de Leuven empleando técnicas de clasificación supervisada y no supervisada, estimación no paramétrica de funciones de densidad de probabilidad y técnicas de caracterización en frecuencia para la realización de las siguientes tareas:

1. Etiquetar todos los objetos variables de cada campo COROT con un vector de probabilidades de pertenencia a cada una de las clases de variabilidad preestablecidas.
2. Analizar estadísticamente las distribuciones de probabilidad de objetos variables en un espacio multidimensional que caracteriza las propiedades de los objetos (espacio de frecuencias, amplitudes armónicas, cocientes de amplitudes y diferencias de fase)

La figura 2 muestra una proyección de la distribución de objetos variables de la Nube de Magallanes en el plano de logaritmo de la frecuencia fundamental frente a logaritmo de la amplitud. Se ven claramente las secuencias correspondientes a las variables pulsantes clásicas (Mira, Semirregulares, RR Lyrae, Cefeidas; casi todas monoperiódicas). La figura 3, muestra la calidad sin precedentes de las series temporales obtenidas con COROT: se trata de una pulsante multiperiódica (en batimiento o beat) en un sistema eclipsante.

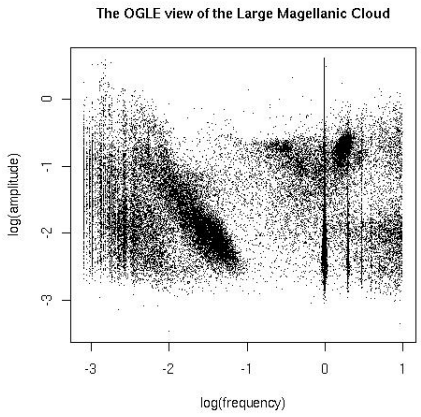

Fig. 2 Variables en la Gran Nube de Magallanes observadas en el proyecto OGLE.

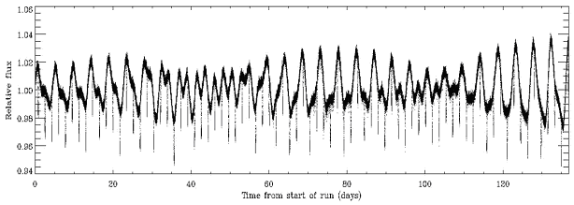

Fig. 3 Ejemplo de la calidad de los datos obtenidos con COROT.

4 Técnicas de visión artificial para el análisis de campos profundos

Los objetivos de esta línea de trabajo, realizada en colaboración con los grupos de investigación participantes en el proyecto ASTRID, tiene como objetivo desarrollar un sistema experto para el análisis de campos profundos cosmológicos en cubos de imágenes multirrango. Para ello, se pretende implementar un sistema de etiquetado de fuentes (aisladas, solapadas o con posible contaminación); un sistema de asignación probabilista de contrapartidas (cross-match) múltiple bayesiano en cubos multirrango: definición de un árbol que describa las posibles contrapartidas de un objeto dado (nodo raíz) en sucesivas imágenes de igual o mejor resolución; un sistema automático de definición de aperturas óptimas para fotometría teniendo en cuenta la información obtenida anteriormente; y la definición de un proceso iterativo que refine las etapas anteriores teniendo en cuenta los resultados de ciclos anteriores. En particular, mejorar la asignación probabilista de contrapartidas a partir de las distribuciones espectrales de energía resultantes mediante el factor de Bayes fotométrico.

Para ello, se están utilizando teselaciones de Voronoi (Delaunay) de imágenes, definición de aperturas basadas en soluciones de contornos activos (snakes) e inferencia bayesiana basada en información astrométrica y fotométrica (SEDs) para la asignación de contrapartidas probabilistas.