Clasificación automatizada de sistemas binarios eclipsantes detectados en el relevamiento VVV

I. Daza[1,2,3], L.V. Gramajo[1,4], M. Lares[1,2,4], C.E. Ferreira Lopes[5], J.J. Clariá[1,4], T. Palma[1,4] & D. Minniti[6,7,8]

[1] *Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina*
[2] *Instituto de Astronomía Teórica y Experimental, CONICET–UNC, Argentina*
[3] *Facultad de Matemática, Astronomía, Física y Computación, UNC, Argentina*
[4] *Observatorio Astronómico de Córdoba, UNC, Argentina*
[5] *National Institute For Space Research, Brazil*
[6] *Departamento de Ciencias Físicas, Facultad de Ciencias Exactas, Universidad Andrés Bello, Chile*
[7] *Instituto Milenio de Astrofísica, Chile*
[8] *Vatican Observatory, Vatican City State, Italy*

Contacto / vanessa.daza@gmail.com.ar

Resumen / Con el advenimiento de grandes relevamientos sin precedentes realizados en el cielo, la ciencia moderna está siendo testigo del amanecer de la Astronomía de las grandes bases de datos, en la cual el manejo y el descubrimiento automático resultan esenciales e indispensables. En este contexto, las tareas de clasificación se encuentran entre las capacidades más requeridas que debe poseer una tubería de reducción de datos para compilar conjuntos de datos confiables, de manera que su procesamiento pueda lograrse con una eficiencia imposible de alcanzar mediante un tratamiento detallado y la intervención humana. El relevamiento VVV (VISTA variables in the Vía Láctea), en la parte meridional del disco Galáctico, incluye datos fotométricos de varias épocas necesarios para el posible descubrimiento de estrellas variables, incluidos los sistemas binarios eclipsantes (SBE). En este estudio, utilizamos un catálogo recientemente publicado de un centenar de SBE de la región d040 del VVV, clasificados de acuerdo a modelos teóricos como sistemas de contacto, separados o semi-separados. Describimos el método implementado para obtener dos modelos de aprendizaje automático, capaces de clasificar los SBE usando información extraída de las curvas de luz de los candidatos a objetos variables en el espacio de fase. Discutimos también la eficiencia de los modelos, la importancia relativa de sus características y las perspectivas futuras para la construcción de una extensa base de datos de SBE en el relevamiento VVV.

Abstract / With the advent of unprecedentedly large surveys of the sky, modern science is witnessing the dawn of big data astronomy where automatic handling and discovery are essential and indispensable. In this context, classification tasks are among the most required skills a data reduction pipeline must possess to compile reliable datasets, so as to accomplish data processing with an efficiency impossible to achieve by means of detailed processing and human intervention. The VISTA Variables of the Vía Láctea (VVV) Survey, in the southern part of the Galactic disc, comprises multi-epoch photometric data necessary for the potential discovery of variable objects, including eclipsing binary systems (EBSs). In this study, we use a recently published catalogue of one hundred EBSs, classified by fine-tuning theoretical models according to contact, detached or semi-detached classes, belonging to the tile d040 of the VVV. We describe the method implemented to obtain two supervised machine learning models, capable of classifying EBSs using information extracted from the light curves of variable object candidates in the phase space. We also discuss the efficiency of the models, the relative importance of the features and future prospects to construct an extensive database of EBSs in the VVV survey.

Keywords / binaries: eclipsing — surveys — catalogue — methods: statistical

1. Introducción

La producción de grandes volúmenes de datos en la Astronomía generados mediante simulaciones numéricas o incluidos en catálogos en las últimas décadas, ha crecido exponencialmente (Szalay et al., 2002). En este contexto, actualmente se diseñan herramientas automáticas que imitan las metodologías y resultados obtenidos mediante la intervención humana. Tal es el caso del estudio de los sistemas binarios eclipsantes (SBE). Usualmente, en el estudio de estos sistemas, se usan metodologías clásicas tales como el método plegado de fase y el de análisis armónico, entre otros, los cuales permiten la determinación del período y algunas características de la forma de la curva de luz (Lomb, 1976; Scargle, 1982; Lafler & Kinman, 1965). Esta información es frecuentemente usada, en forma manual o automática, para determinar la clase del SBE estudiado.

Los modelos de aprendizaje automático supervisado son herramientas útiles para la automatización de la clasificación de SBE. El éxito de estos modelos predictivos depende fuertemente del conjunto de datos disponibles, es decir, del tamaño, la información contenida y la precisión de los mismos, además de las estrategias elegidas

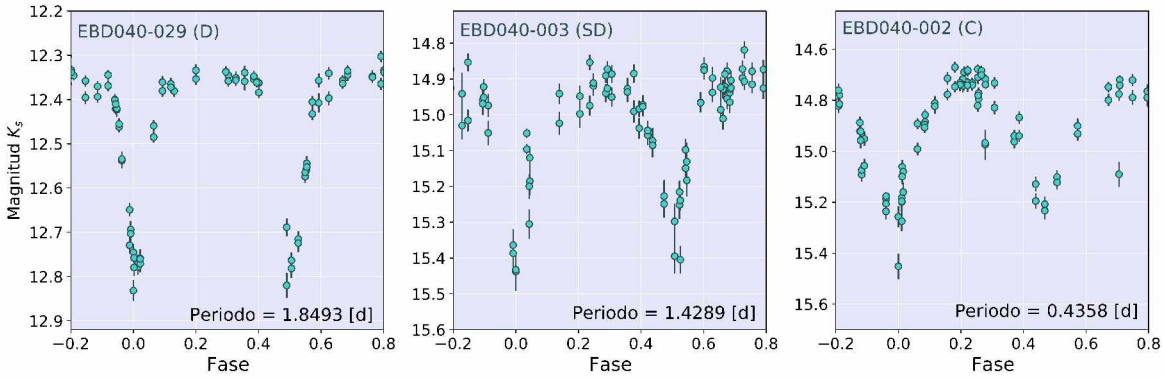

Figura 1: Ejemplos de curvas de luz en la muestra D3: D (separados), SD (semi-separados) y C (de contacto)

en la selección de las características más relevantes de los modelos.

Aquí elaboramos dos modelos de aprendizaje automático supervisado para la clasificación del tipo de SBE. Para la construcción de los mismos, usamos un catálogo de 100 SBE de Gramajo et al. (2020), recientemente descubiertos en el relevamiento VVV (VISTA Variables in the Vía Láctea, Minniti et al. 2010; Saito et al. 2012). Este catálogo contiene información de la posición de los sistemas, series de tiempo, parámetros físicos de las componentes y la clase de los sistemas: D, SD y C, para los tipos separados, semi-separados y de contacto, respectivamente. Esta clase de sistema se determina, en cada caso, a partir de los lóbulos de Roche. Además, para complementar la información de este catálogo, incluimos los períodos provistos por Ferreira Lopes et al. (2020).

2. Muestra de sistemas binarios eclipsantes clasificados

Dado que la determinación de la clase de un SBE puede realizarse a través de la inspección de las curvas de luz de las dos estrellas en el espacio de fase y queremos un modelo que sea capaz de reproducir o superar la clasificación manual, elegimos como datos de base las series de tiempo y los períodos disponibles. Curvas de luz de ejemplos propuestos al clasificador se muestran en la Fig. 1. Seleccionamos del centenar de SBE antes mencionado, aquéllos con períodos en el intervalo 0.4–15.2 días. Esta restricción disminuye el número de sistemas a 96, muestra que en adelante denominaremos D3.

Dado que sólo el 12.5 % de los sistemas de la muestra total son SBE semi-separados y esta pequeña fracción disminuye el rendimiento del modelo en la clasificación de los 3 tipos de sistemas, decidimos usar los mismos 96 SBE de la muestra D3, aunque sólo con dos clases definidas como D + SD y C, que denominamos D2. Es decir, la primera clasificación representa la combinación de los sistemas D y SD, en tanto que la segunda coincide con los sistemas de contacto. En la Fig. 2 mostramos con diagramas de barras el desbalance existente entre las distintas clases en ambas muestras, D3 y D2.

3. Valoración de dos métodos de aprendizaje automático supervisado

En esta sección describimos los criterios que se adoptaron en el entrenamiento de los modelos y la selección de las características de los mismos, usando las dos muestras previamente definidas en la Sec. 2

3.1. Extracción de características

Usamos las series temporales para extraer información implícita que caracteriza a las distintas clases de SBE. Esta manera de proceder es la que usualmente se utiliza en el estudio de estrellas variables, dado que los modelos de aprendizaje automático mejoran cuando se les ingresa información relevante de cada clase. La extracción de dicha información se realizó a través del paquete denominado FEETS de PYTHON (Cabral et al., 2018). Este paquete toma magnitud y tiempo como entrada y devuelve características, tanto estadísticas como de la forma de la curva de luz (media, dispersión, pendiente, entre otras). En nuestro caso, obtuvimos 63 características de las series de tiempo de cada sistema binario. A este conjunto, le agregamos los períodos provistos por Ferreira Lopes et al. (2020), logrando así generar un conjunto de 70 características a seleccionar para obtener los modelos de aprendizaje automático supervisado con mejor rendimiento.

3.2. Selección de características

Previamente a la elección del modelo de aprendizaje automático supervisado sobre los datos de las muestras D3 y D2, usamos los métodos de selección de características usuales en datos cuantitativos con etiqueta de clase cualitativa: información mutua, análisis de la varianza (ANOVA por sus sigloides en inglés, ANalysis Of VAriance) y chi cuadrado (χ^2).

3.3. Validación de los modelos

Elaboramos tres subconjuntos formados por 80 %, 16 % y 4 % de la muestra total para entrenamiento, validación y prueba, respectivamente. Los dos primeros son usados para la selección de las características con alta entropía

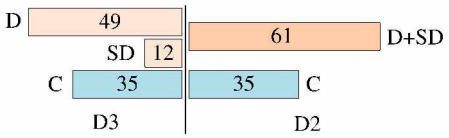

Figura 2: Balance de las muestras de sistemas binarios para realizar la clasificación en dos (D2) o tres (D3) clases.

y selección de parámetros e hiperparámetros de los modelos, mientras que el tercer subconjunto es usado en el período de evaluación de los modelos. Al igual que en la selección de características, en este paso comparamos el rendimiento de los modelos en el subconjunto de validación de los modelos de aprendizaje automático supervisado. Los métodos de clasificación supervisada usados fueron: k vecinos más cercanos, bosque aleatorio, gradiente descendente estocástico y arboles de decisión, todos implementados en la librería SKLEARN de PYTHON.

3.4. Métricas

Dado que el conjunto de datos presenta un desbalance inherente en clases tendiendo a encontrar más SBE de tipo D, conservamos ese desbalance en el desarrollo y evaluación. Por esa razón, en nuestra selección de métricas, no incluimos la métrica exactitud y, en su reemplazo, usamos como métrica decisiva el micro promedio del valor-F1 ($F1_{mic}$, ver Ecu. 1), la cual sí contempla el mencionado desbalance de clases, además de las métricas precisión (P), exhaustividad (R), valorF1 (F1) y soporte (S)*, a excepción de S que cuantifica la cantidad de aciertos, las restantes métricas indican un buen rendimiento del modelo cuando sus valores son cercanos a 1.

$$F1_{mic} = 2 * \frac{P_{micro} * R_{micro}}{P_{micro} + R_{micro}} \quad (1)$$

Donde micro hace referencia al calculo de P y R teniendo en cuenta todos los valores de las clases y no por cada clase como son usualmente calculados.

4. Resultados

Presentamos a continuación los resultados obtenidos para las muestras D3 y D2, respectivamente, a partir de la metodología expuesta en la Sec. 3.1

4.1. Modelo - D3

Elegimos 10 de las 70 características con el método de ANOVA, aun cuando el rendimiento del algoritmo en el conjunto de validación no varía mayormente con la cantidad de características que se seleccione. El modelo con mejor rendimiento para los datos con las tres clases de SBE fue el bosque aleatorio, superando en centésimas el rendimiento de los otros modelos. El valor de sus métricas se presenta en la Tabla 1.

*Estas métricas se conocen mejor por sus nombres en inglés: *precision*, *recall*, *F1score* y *support*.

4.2. Modelo - D2

En el caso del conjunto de datos con sólo dos clases, observamos un mejor rendimiento del modelo cuando se usa como método de selección ANOVA, con un conjunto de características de tamaño 35. Sin embargo, se obtiene un rendimiento similar usando la misma cantidad de características con alta entropía pero seleccionadas con χ^2. Para esta muestra, el modelo con mejor rendimiento resultó ser gradiente descendente estocástico y al igual que en la muestra D3 el rendimiento de este modelo solo supera en centésimas a los otros modelos. Al igual que la muestra D3, la métrica $F1_{mic}$ alcanza un valor de 0.75. Sin embargo, sus valores en el resto de las métricas son más cercanos a 1 (ver Tabla 1).

Tabla 1: Rendimiento de los mejores modelos en las muestra D2 y D3. Tanto para las muestras D3 como D2 el valor de $F1_{mic}$ resultó 0.75 en ambas muestras de prueba.

Clases	P	R	F1	S
Bosque aleatorio - D3:				
D	0.75	0.82	0.78	11
SD	0.00	0.00	0.00	3
C	0.75	1.00	0.86	6
Gradiente descendente estocástico - D2:				
D + SD	0.76	0.93	0.84	14
C	0.67	0.33	0.44	6

5. Conclusiones

La clasificación de SBE en tres clases no exhibe un buen rendimiento para los sistemas semi-separados, independientemente del modelo predictivo o del método de selección de características que se utilice. Observamos que si tratamos a los SBE con sólo dos tipos (D + SD y C), las predicciones de los modelos son más acertadas. Como trabajo a futuro esperamos poder evaluar el rendimiento de una tubería de clasificación jerárquica, sujeto a nuevos métodos de selección de características y al uso de redes neuronales.

Agradecimientos: Este trabajo ha sido financiado por el Consejo Nacional de Investigaciones Científicas y Técnicas de la República Argentina (CONICET) y por la Secretaría de Ciencia y Técnica (SECYT) de la Universidad Nacional de Córdoba.

Referencias

Cabral J.B., et al., 2018, Astron. Comput., 25, 213
Ferreira Lopes C.E., et al., 2020, MNRAS, 496, 1730
Gramajo L.V., et al., 2020, arXiv e-prints, arXiv:2011.02530
Lafler J., Kinman T.D., 1965, ApJS, 11, 216
Lomb N.R., 1976, Ap&SS, 39, 447
Minniti D., et al., 2010, NewA, 15, 433
Saito R.K., et al., 2012, A&A, 537, A107
Scargle J.D., 1982, ApJ, 263, 835
Szalay A.S., Gray J., Vandenberg J., 2002, *American Astronomical Society Meeting Abstracts*, vol. 201, 134.06