Boletín de Estadística e Investigación Operativa
Vol. 29, No. 1, Febrero 2013, pp. 3-20

Estadística

Implementation of statistical classification methods in stellar catalogs [1]

Ángel Berihuete, Andrés Jiménez, Francisco Álvarez and J.M. Gutiérrez

Departamento de Estadística e IO
Universidad de Cádiz

✉ angel.berihuete@uca.es, andres.jimenez@uca.es, francisco.alvarez@uca.es

Abstract

One of the present challenges in Astronomy is to establish efficient statistical methods to perform suitable classification of stellar objects within large databases. In this paper we study the behavior of different statistical methods to perform stellar classification based on several measurements of the star.

Specifically, and using JHC stelar catalogue, we show that: (i) there is a relationship between color of the star and the statistical central moments of its spectrum, (ii) we can fit a regression model between star temperature and spectrum shape; (iii) we make an stellar classification based on kernel methods (iv) it is possible to construct an adaptive stelar classification using Kohonen maps.

Keywords: Astrostatistics, Classification, Kohonen maps.
AMS Subject classifications: 62P35, 62H30.

1. Introducción

Los avances tecnológicos para la adquisición de datos astronómicos en las distintas misiones espaciales de las dos últimas décadas, han impulsado la creación de grandes bases de datos cuyo tratamiento hace necesario el trabajo conjunto de científicos en campos tan distintos como la Astronomía, la Estadística o la Computación. Conscientes de éste hecho, Babu y Feigelson (1993) hicieron un primer llamamiento a la comunidad científica a través del congreso *New challenges in modern Astronomy*, en el cual se presentaba a la Estadística como una de las disciplinas esenciales para dar respuesta a grandes retos de la Astronomía

[1] This research has made use of the Spanish Virtual Observatory supported from the Spanish MEC through grants AyA2008-02156, AyA2011-24052.

actual. Recientemente, trabajos como los de Babu y Feigelson (1996), Jiménez, Berihuete y Gutiérrez (2008), Starck, Llebaria y Loredo (2008), van consolidando la palabra *Astroestadística* como una nueva disciplina en el panorama científico.

En la actualidad existen aproximadamente unos noventa eventos anuales (congresos, talleres, reuniones científicas, etc.) sobre la aplicación de nuevas técnicas estadísticas al campo de la Astronomía. Cabe destacar de entre todas ellas la realizada recientemente en nuestro país con el nombre *Astrostatistics and Data Mining in massive astronomical databases*[2] (La Palma, 2011), en la que se abordaron temas tales como el análisis multivariante, inferencia bayesiana, series temporales, tratamiento de imágenes o tratamiento computacional de grandes bases de datos.

En este sentido, la aplicación de métodos estadísticos para la reducción de la dimensión y la selección de características son parte fundamental del análisis eficiente de bases de datos de gran tamaño, cobrando especial interés el desarrollo de sistemas híbridos que aprovechen las bondades de las redes neuronales artificiales (RNA) y su capacidad de adaptación a nuevos conjuntos de datos. Estas líneas de investigación han sido estudiadas, entre otros, por Jiménez (2003), Sarro y Berihuete (2008) y Berihuete (2010).

Este trabajo estudia el comportamiento de algunas técnicas de reducción de la dimensión y clasificación en su aplicación al catálogo estelar publicado por Jacoby, Hunter y Christian (1984), también conocido como catálogo JHC. Específicamente, analizaremos la clasificación del tipo espectral basada en núcleos, así como su contrapartida mediante redes de Kohonen, obteniéndose resultados novedosos en cuanto a la asociación de parámetros físicos relativos al color de una estrella y la morfología de su espectro físico.

En la Sección 2 se describen el conjunto de datos tratados y un estudio descriptivo básico de los mismos. La Sección 3 propone los momentos estadísticos centrales como descriptores de forma del espectro físico de una estrella, utilizándose posteriormente en la clasificación basada en núcleos de la Sección 4, y en las redes de Kohonen de la Sección 5. En la Sección 6 se detallan los resultados más relevantes del estudio. Por último en el Apéndice 6 se ofrecen varias direcciones a páginas web que permitirán profundizar en varios de los conceptos astrofísicos utilizados en el trabajo.

2. Conjunto de datos

El catálogo estelar JHC está compuesto por varias medidas físicas de 161 objetos estelares. Específicamente, los *colores* $U-B$ y $B-V$, los *espectros* en el rango 3510 Å a 7427 Å con una resolución espectral de 4.5 Å, el *tipo espectral* (rango `O-M`) y la *clase de luminosidad* (tipos `I, III` y `V`).

Los espectros se obtuvieron en 26 noches diferentes (diciembre de 1980 a

[2]Página web del taller `http://www.iwinac.uned.es/Astrostatistics/`

diciembre de 1981) utilizando el instrumento *Intensified Reticon Scanner* a partir de tres filtros (azul, verde, rojo) combinados posteriormente unos sobre otros. Específicamente, *Azul*, rejilla No. 35, cubriendo de 3430 Å a 4950 Å. *Verde*, rejilla No. 56, cubriendo de 4760 Å a 6220 Å. *Rojo*, rejilla No. 36, cubriendo de 6000 Å a 7450 Å.

2.1. Estudio descriptivo del catálogo

Desde un punto de vista estadístico, el catálogo contiene variables continuas unidimensionales (colores $U-B$ y $B-V$) y multidimensionales (espectro físico de la estrella), así como variables categóricas (tipo espectral y luminosidad). Para una descripción detallada de las variables pueden visitarse los enlaces del Apéndice. En la Tabla 1 se muestran las frecuencias absolutas para las variables categóricas *tipo espectral* y *luminosidad* del catálogo en estudio. Además, el tipo espectral se subdivide en una nueva categoría mediante los numerales (0–9). Por ejemplo, *A*0 denota una estrella más caliente dentro de la clase A, mientras que *A*9 denota la más fría dentro de la misma clase. El Sol está clasificado como *G*2. En la Figura 1 se representan los espectros de cuatro objetos estelares contenidos en el catálogo, y en la Figura 2 un diagrama de dispersión para las variables color.

Tipo espectral	n_i
O	19
A	23
B	34
F	29
G	28
K	14
M	12
MPF	2

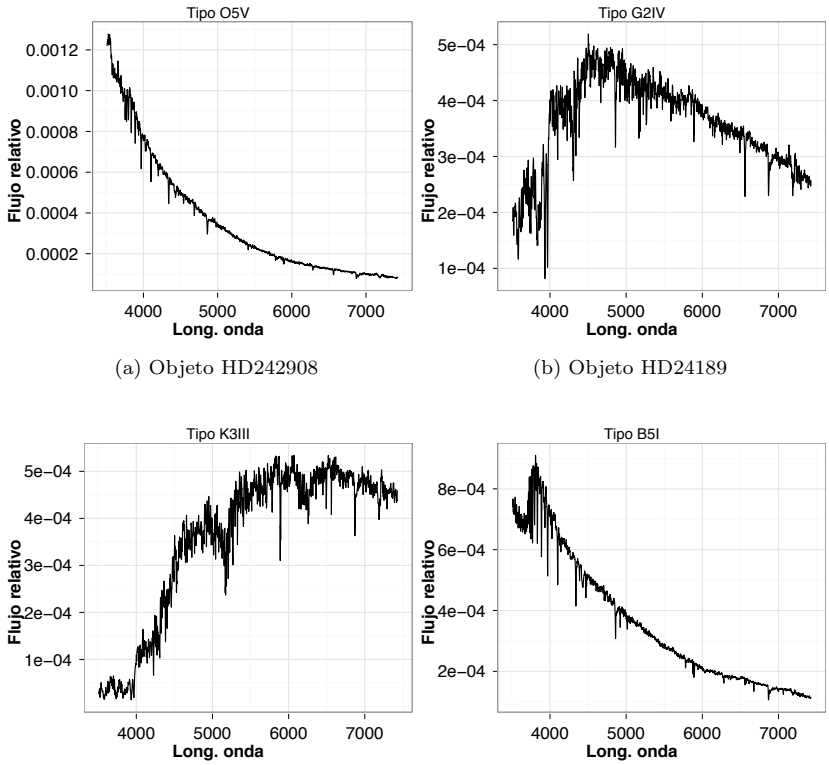
(a)

Luminosidad	n_i
EV	1
I	45
II	8
III	44
IV	4
V	59

(b)

Tabla 1: Frecuencias acumuladas para las variables categóricas de los objetos en el catálogo JHC.

En la Tabla 2 se recoge un resumen de datos con las medias, desviaciones típicas e intervalos de confianza (para la media), de los colores $U-B$ y $B-V$ de cada uno de los tipos espectrales. Se analiza el modelo de regresión lineal para cada tipo espectral atendiendo a los colores $B-V$ y $U-B$, que denotaremos con las variables C_{B-V} y C_{U-B} respectivamente. Los distintos ajustes para el modelo junto con los residuos obtenidos se presentan en las Tablas 3a y 3b. No se incluye el tipo espectral `MPF` al disponer únicamente de dos observaciones. Puede observarse que el tipo espectral O tiene el mayor coeficiente de determinación, es decir, un 98,4 % de la variación observada en el color $B-V$ puede explicarse por los cambios del color $U-B$ a través de la relación lineal.



(a) Objeto HD242908 (b) Objeto HD24189

(c) Objeto HD249384 (d) Objeto BD51710

Figura 1: Distintos espectros del catálogo JHC.

Las imágenes de la Figura 3 representan la comparación del espectro solar con las estrellas recogidas en el catálogo JHC. En la Figura 3b se representan los distintos valores para la suma de cuadrados del error (SCE) entre el espectro del Sol y los diferentes elementos del catálogo. Se observa claramente, como era de esperar, que los objetos de tipo espectral G tienen un error menor.

3. Caracterización de espectros mediante momentos centrales

Los espectros dentro del catálogo JHC pueden considerarse como una muestra de un vector aleatorio de dimensión 880 (luz muestreada en 880 longitudes de onda diferentes). Sin embargo, esta dimensión alta genera problemas en la aplicación directa de las técnicas de clasificación como las que se abordan en este trabajo. Una forma de superar esta limitación es considerar los momentos estadísticos centrales como descriptores de la morfología del espectro físico, pu-

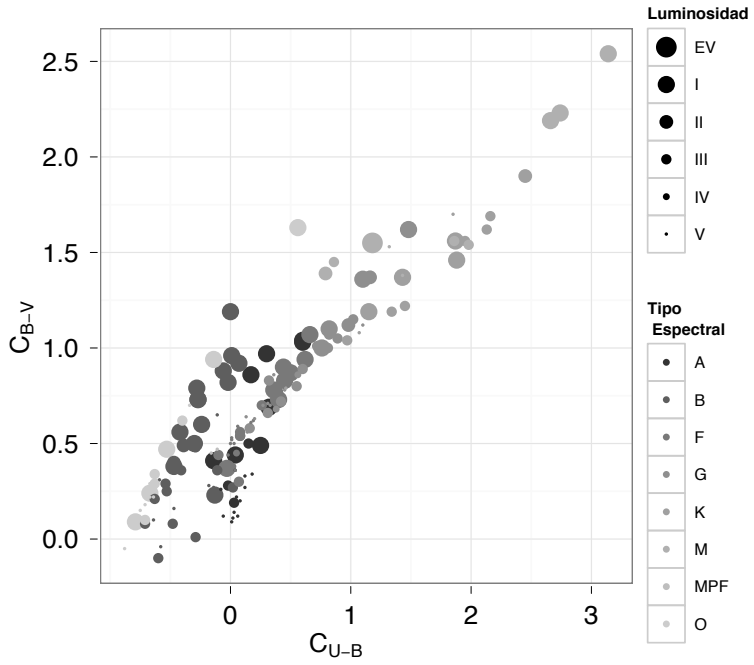


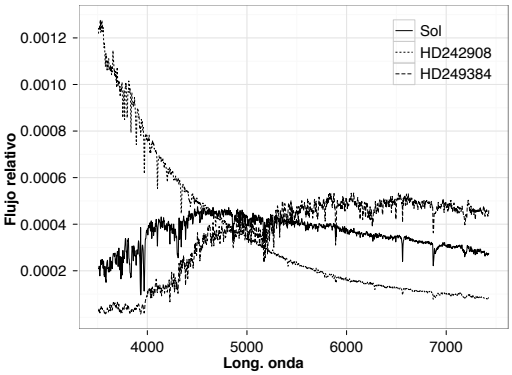
Figura 2: Gráfico color-color. Cada punto representa una estrella, especificando su tipo espectral y su luminosidad.

diendo caracterizar la estrella no por sus 880 valores de intensidad lumínica, sino por un conjunto reducido de momentos estadísticos (ver Tabla 4).

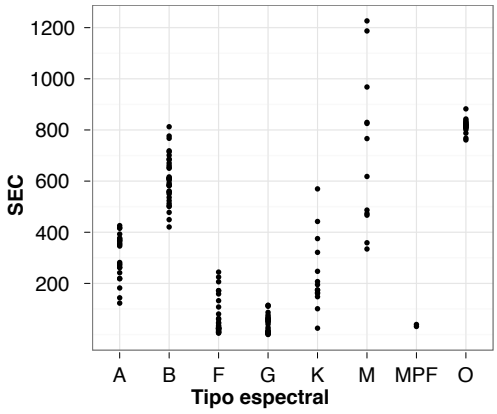
Este trabajo considera los momentos centrales 2, 3, 4, así como la suma de valores del espectro (η) creando de esta forma un nuevo conjunto de datos en el que estarían incluidas las variables η, C_{U-B}, C_{B-V} y μ_i, con $2 \leq i \leq 4$. Debido a la magnitud elevada de los resultados en las variables μ_i, $2 \leq i \leq 4$ se opta por trabajar con logaritmos, observando una dependencia lineal entre el color $U - B$ y $\log(\mu_4)$. La estimación de los coeficientes para un modelo de regresión lineal así como sus residuos aparecen en las Tablas 5a y 5b. En la Figura 4 se representa la relación entre $U - B$ y $\log(\mu_4)$.

4. Clasificación estelar mediante métodos basados en núcleos

Hemos visto en la sección anterior que existe una relación directa entre el color $U - B$ y $\log(\mu_4)$ del espectro de las estrellas pertenecientes al catálogo. Puesto que el color de una estrella está relacionado con su temperatura, y ésta con su tipo espectral, puede conjeturarse que los logaritmos de los momentos centrales



(a)



(b)

Figura 3: Comparación del espectro solar con algunos espectros presentados en la Figura 1. En (b) se muestra la suma de errores al cuadrado para cada uno de los objetos en el catálogo en comparación con el espectro solar según su tipo espectral.

	$U-B$		$B-V$		Int. conf. 95 %	
	$\overline{x}$	σ	$\overline{x}$	σ	$\overline{x}_{U-B}$	$\overline{x}_{B-V}$
O	-0.554	0.101	0.379	0.146	(-0.70, -0.40)	(0.19, 0.56)
A	0.125	0.0353	0.411	0.0921	(0.04, 0.20)	(0.28, 0.54)
B	-0.329	0.0521	0.424	0.0987	(-0.39, -0.23)	(0.31, 0.53)
F	0.172	0.0438	0.603	0.0361	(0.09, 0.25)	(0.53, 0.67)
G	0.613	0.131	0.913	0.0694	(0.47, 0.75)	(0.81, 1.01)
K	1.58	0.304	1.38	0.0961	(1.26, 1.90)	(1.20, 1.56)
M	1.69	0.613	1.71	0.148	(1.20, 2.19)	(1.46, 1.94)
MPF	-0.1.35	0.001	0.460	0.0002	(-0.45, 0.18)	(0.33, 0.58)

Tabla 2: Resumen de datos para los colores $U-B$ y $B-V$

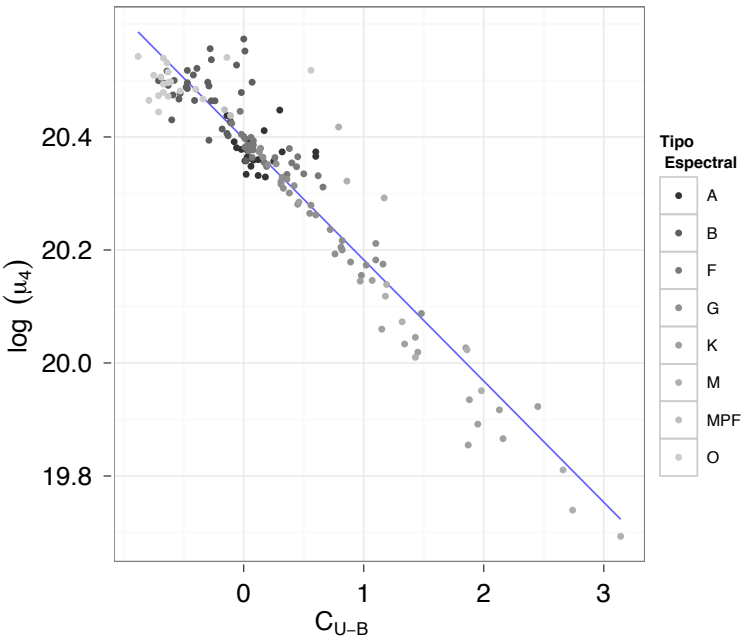


Figura 4: Ajuste de un modelo de regresión lineal entre el color $U-B$ y el logaritmo del momento central de orden 4.

Tipo	Estimación	Error estd.	Valor t	$\Pr(>\|t\|)$
O	$\beta_1 = 1.0395$	0.0233	44.54	0.0000
	$\beta_2 = 1.1909$	0.0368	32.39	0.0000
B	$\beta_1 = 0.7359$	0.0703	10.47	0.0000
	$\beta_2 = 0.9481$	0.1764	5.37	0.0000
A	$\beta_1 = 0.2446$	0.0440	5.57	0.0000
	$\beta_2 = 1.3359$	0.1978	6.75	0.0000
F	$\beta_1 = 0.4568$	0.0163	27.96	0.0000
	$\beta_2 = 0.8508$	0.0609	13.98	0.0000
G	$\beta_1 = 0.4813$	0.0244	19.70	0.0000
	$\beta_2 = 0.7052$	0.0345	20.45	0.0000
K	$\beta_1 = 0.5160$	0.0609	8.48	0.0000
	$\beta_2 = 0.5481$	0.0364	15.04	0.0000
M	$\beta_1 = 0.9411$	0.1143	8.23	0.0000
	$\beta_2 = 0.4511$	0.0617	7.31	0.0000

(a)

Tipo	Residuo	R^2		Est. F (p-valor)	Ajuste
		Multiple	Ajustado		
O	0.049	0.984	0.983	105 (<2e-16)	Muy bueno
B	0.231	0.474	0.458	28.9 (6.69e-06)	Muy Malo
A	0.174	0.685	0.685	45.6 (1.11e-06)	Malo
F	0.067	0.879	0.879	195 (7e-14)	Bueno
G	0.065	0.941	0.939	418 (<2e-16)	Muy bueno
K	0.072	0.950	0.945	226 (3.77e-09)	Muy bueno
M	0.160	0.842	0.827	53.4 (2.58e-05)	Bueno

(b)

Tabla 3: Estimación de parámetros y resumen de residuos para el modelo de regresión lineal $C_{B-V} = \beta_1 + \beta_2 C_{U-B}$

Id. Objeto	Objetos	η	μ_2	μ_3	μ_4
1	HD242908	1.46e+02	1.62e+04	2.10e+06	8.32e+08
2	HD215835	1.47e+02	1.56e+04	2.04e+06	7.98e+08
3	HD12993	1.49e+02	1.54e+04	2.00e+06	7.79e+08
4	HD35619	1.54e+02	1.61e+04	2.08e+06	8.25e+08
5	HD44811	1.49e+02	1.60e+04	2.03e+06	8.07e+08
⋮	⋮	⋮	⋮	⋮	⋮

Tabla 4: Momentos estadísticos para los cinco primeros objetos del catálogo JHC.

Estimación		Error Std.	Valor t	Pr(> |t|)
(Ord. Origen)	85,72	2,24	38,2	$< 2e-16$
$\log(\mu_4)$	$-4,20$	0,11	$-38,1$	$< 2e-16$

(a)

Mín	1Q	Mediana	3Q	Máx
-0.5395	-0.1311	-0.0336	0.0737	1.0403

(b)

Tabla 5: Modelo de regresión lineal para las variables $U-B$ y $\log(\mu_4)$. En (b) se presenta un resumen de los residuos para el ajuste del modelo.

Núcleo	Función	Error
Gausiano	$\kappa(x,x') = \exp(-\sigma\|x-x'\|^2)$	0.14907
Polinómico	$\kappa(x,x') = (\text{escala}\cdot\langle x,x'\rangle)^{\text{grado}}$	0.18634
Lineal	$\kappa(x,x') = \langle x,x'\rangle$	0.18634
Tangente hiperbólica	$\kappa(x,x') = \tanh(\text{escala}\cdot\langle x,x'\rangle)$	0.50932
Laplace	$\kappa(x,x') = \exp(\sigma\|x-x'\|)$	0.09316
Bessel	$\kappa(x,x') = \frac{\text{Bessel}^n_{\nu+1}(\sigma\|x-x'\|)}{(\|x-x'\|)^{-n(\nu+1)}}$	0.19876
RBF ANOVA	$\kappa(x,x') = \left(\sum_{k=1}^n \exp(-\sigma(x^k-x'^k)^2)\right)^d$	0.16149

Tabla 6: Distintas funciones núcleo utilizadas en el experimento.

de un espectro son capaces de clasificar estrellas en clases de tipo espectral.

El problema de la extracción de algún tipo de estructura a partir de los datos puede plantearse a través de los *métodos basados en núcleos*. En este tipo de procedimientos se utilizan representaciones implícitas del espacio de datos dentro de un espacio de características mediante una función núcleo κ, es decir, dada la proyección $\Phi : X \to H$, del espacio de datos al espacio de características, se devuelve el producto interno $\kappa(x,y) = \langle\Phi(x),\Phi(y)\rangle$. Una vez se tiene la proyección en el espacio de características se aplica un algoritmo de aprendizaje.

Una propiedad interesante de los sistemas basados en núcleos es que, una vez seleccionada una función κ válida, puede trabajarse en espacios de cualquier dimensión sin comprometer el coste computacional (la representación de las características nunca se lleva a efecto). De hecho no es necesario saber qué características están siendo utilizadas. En la Tabla 6 se especifican los núcleos utilizados [3].

Se han probado diferentes combinaciones de las variables $\log(\eta)$ y $\log(\mu_i)$, $2 \leq i \leq 4$ para diferentes funciones núcleo, optando finalmente por las dos primeras

[3]El parámetro *escala* en los núcleos polinómico y tangente hiperbólica normaliza los patrones sin la necesidad de normalizar los datos obtenidos de las componentes principales.

	Comp.1	Comp.2	Comp.3
Desv. estándar	1.252	1.174	0.228
Prop. varianza	0.522	0.4596	0.017
Prop. acumulada	0.522	0.982	1.000

(a) Importancia de las componentes principales

	Comp.1	Comp.2	Comp.3
$\log(\eta)$	0.432	0.708	0.559
$\log(\mu_2)$	0.790		-0.612
$\log(\mu_4)$	0.434	-0.707	0.559

(b) Cargas factoriales

Tabla 7: Resumen del ACP para las variables $\log(\eta)$, $\log(\mu_2)$, $\log(\mu_4)$.

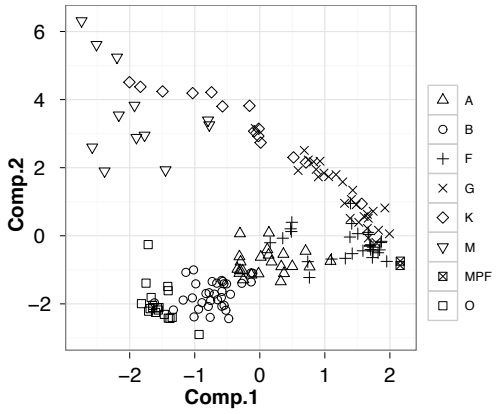
componentes de un análisis de componentes principales (ACP) de las variables $\log(\eta), \log(\mu_2)$ y $\log(\mu_4)$, ver Tabla 7. La última columna de la Tabla 6 recoge el error producido por la clasificación utilizando las dos primeras componentes principales. En la Figura 5 se muestra el resultado de la clasificación de las dos primeras componentes principales para el núcleo Gausiano. Puede observarse en la imagen la disposición de los datos forma un clúster abierto, añadiendo una dificultad más a la tarea de clasificación. Las matrices de confusión para dichos núcleos aparecen en la Tabla 8.

5. Clasificación mediante redes de Kohonen

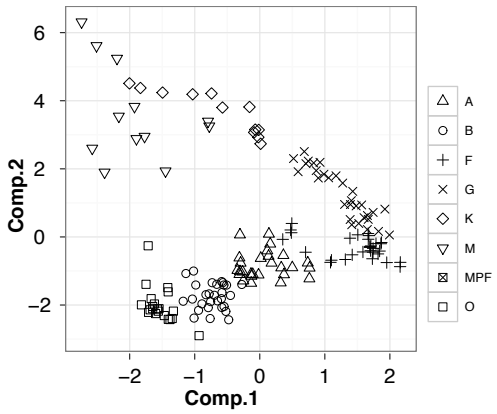
Los métodos adaptativos son una manera eficaz de analizar las características intrínsecas de las variables estadísticas registradas en un catálogo estelar. Concretamente, y mediante las redes de Kohonen (supervisadas o no) pueden establecerse mapas discretos a partir de datos multidimensionales, de forma que las localizaciones espaciales (coordenadas) de los nodos de la red son indicativas de las características de las variables en estudio.

En el caso del catálogo JHC se han seleccionado aleatoriamente el 75 % del total de espectros en el catálogo para el entrenamiento de redes de Kohonen con diferentes topologías. El 25 % restante se utilizarán para validar la red. La Tabla 9a muestra la distancia media de los datos a los nodos asignados, y el número de estrella clasificadas correctamente en el conjunto de validación (40 estrellas) para redes no supervisadas.

A la vista de los resultados, puede observarse que: (i) la clasificación correcta de los espectros no supera el 90 % de los datos. Este resultado se repite en las topologías 4×4, 6×6, 8×8, 9×9. (ii) La topología de red de 4×4 es mejor, en términos de gasto computacional, que las demás topologías, aún teniendo una distancia media al nodo más cercano mayor. La matriz de confusión atendiendo al tipo espectral se muestra en la Tabla 10. Utilizando esta topología para el



(a)



(b)

Figura 5: Proyección sobre las dos primeras componentes principales de las variables $\log(\eta), \log(\mu_2)$ y $\log(\mu_4)$ en (a). Clasificación por la aplicación del núcleo de base radial Gausiano en (b).

	A	B	F	G	K	M	MPF	O
A	20	0	3	0	0	0	0	0
B	2	30	0	0	0	0	0	2
F	3	0	24	2	0	0	0	0
G	0	0	4	23	1	0	0	0
K	0	0	0	3	11	0	0	0
M	0	0	0	0	1	10	0	1
MPF	0	0	2	0	0	0	0	0
O	0	0	0	0	0	0	0	19

(a) Núcleo Gausiano.

	A	B	F	G	K	M	MPF	O
A	20	0	3	0	0	0	0	0
B	1	31	0	0	0	0	0	2
F	2	0	26	1	0	0	0	0
G	0	0	4	24	0	0	0	0
K	0	0	0	2	12	0	0	0
M	0	0	0	0	0	12	0	0
MPF	0	0	0	0	0	0	2	0
O	0	0	0	0	0	0	0	19

(b) Núcleo Laplace.

Tabla 8: Matrices de confusión para las clasificaciones realizadas por diferentes núcleos κ.

Topología	Dist. media	Clasif. correcta
2×2	31.3	4
3×3	9.65	28
4×4	6.28	36
5×5	3.23	34
6×6	2.19	36
7×7	1.66	33
8×8	1.25	36
9×9	1.09	36
10×10	0.73	35

(a) Redes no supervisadas.

Topología	Dist. media	Clasif. correcta
2×2	60.09	26
3×3	15.95	38
4×4	10.47	37
5×5	7.81	35
6×6	3.22	35
7×7	2.02	34
8×8	1.59	35
9×9	1.34	35
10×10	0.85	32

(b) Redes supervisadas.

Tabla 9: Resumen de las clasificaciones para diferentes topologías de redes de Kohonen.

	O	B	A	F	G	K	M
O	4	0	0	0	0	0	0
B	0	5	0	0	0	0	0
A	0	0	7	2	0	0	0
F	0	0	0	10	0	0	0
G	0	0	0	0	8	0	0
K	0	0	0	0	0	0	1
M	0	0	0	0	0	1	2

Tabla 10: Matriz de confusión para la red 4×4.

espectro solar obtenemos el primer nodo como ganador, con una distancia media a dicho nodo de 1.585.

En el caso de utilizar redes supervisadas, será necesario definir previamente una variable dependiente con el que realizar el aprendizaje, en nuestro caso utilizaremos la variable *tipo espectral*. Se establecen los mismos conjuntos de datos de entrenamiento y test que en la sección anterior. Recordamos que las distancias de los datos originales X (espectros) y la información adicional Y (tipo espectral), se calculan de forma independiente en sus respectivos espacios. Ambas se escalan de forma que la máxima distancia sea igual a 1. Además la distancia global se pondera de la siguiente manera:

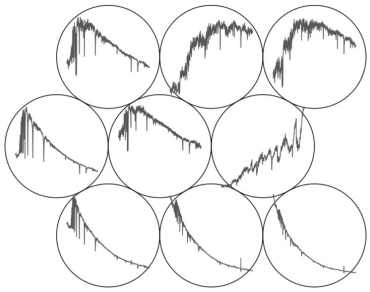
$$D(o,u) = \alpha D_x(o,u) + (1-\alpha) D_y(o,u), \tag{5.1}$$

donde $D(o,u)$ indica la distancia combinada de un objeto o a un nodo u de la red, siendo D_x y D_y las distancias en cada uno de los espacios considerados X e Y. Dependiendo de la variable utilizada, continua o discreta, se utilizarán diferentes distancias, euclídea o tanimoto, respectivamente.

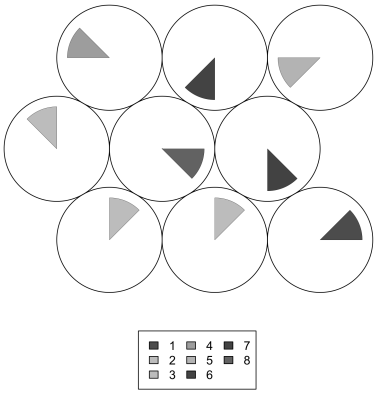
Fijaremos $\alpha = 0{,}2$, es decir, la variable *tipo espectral* tiene una mayor importancia que la variable *espectro* en la clasificación. En la Tabla 9b se recogen los resultados para distintas topologías de red supervisada. Se observa una buena clasificación del conjunto test para la topología 3×3 (95 % correctos), sin embargo la distancia media al vector de codificación es de 15,95. Esto significa que los 9 nodos de esta red son incapaces de separar cada uno de los tipos espectrales que se dan en el catálogo. La topología 4×4 tiene una distancia media al vector de codificación de 10,47 y 9 nodos para clasificar el tipo espectral. En las Figuras 6 y 7 se recogen algunas representaciones de los resultados para topologías 3×3 y 4×4, respectivamente.

6. Conclusiones

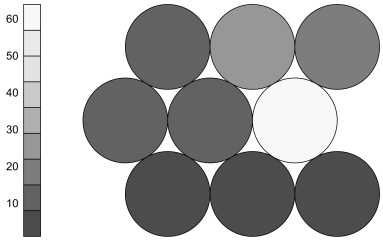
Este trabajo aborda la aplicación de diferentes técnicas de clasificación y reducción de la dimensión en el catálogo estelar JHC. Sin embargo, en una fase



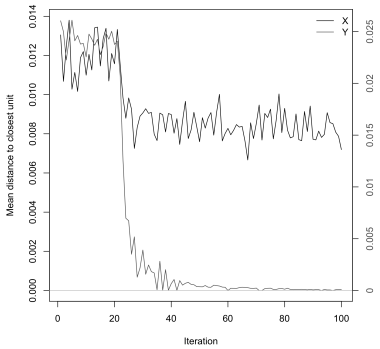
(a) Vectores de codificación.



(b) Asignación del tipo espectral (variable dependiente).

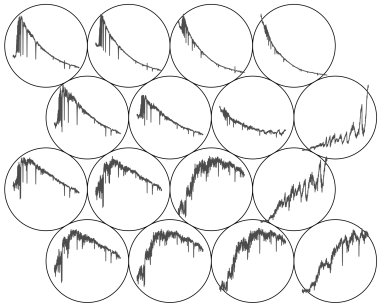


(c) Distancia media de los objetos asignados a un nodo.

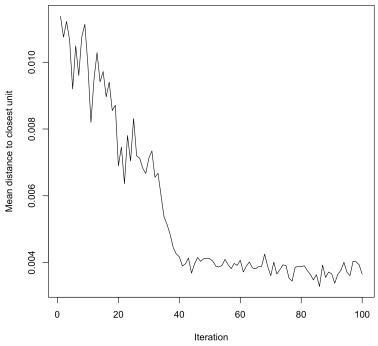


(d) Distancia media al vector más cercano durante el entrenamiento.

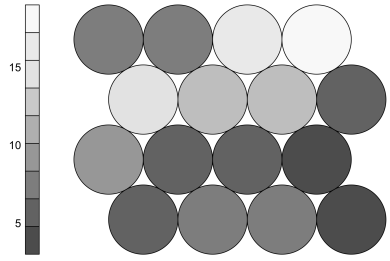
Figura 6: Visualización de la red de Kohonen 3×3 supervisada.



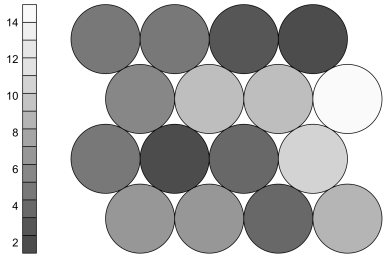
(a) Vectores de codificación.



(b) Distancia media al vector más cercano durante el entrenamiento.



(c) Número de objetos representados en cada nodo.



(d) Distancia media de los objetos asignados a un nodo.

Figura 7: Distintas representaciones de una red de Kohonen 4×4.

previa a la aplicación se han planteado dos dificultades añadidas: (i) de tipo estadístico, ya que la reducción de la dimensión hecha a partir de los logaritmos de los momentos centrales generaba un clúster abierto. (ii) de tipo computacional, ya que a la hora de trabajar con este tipo de base de datos, por lo general, no hay uniformidad en los formatos de los distintos catálogos, y los objetos, así como las variables estadísticas medidas son muy heterogéneas.

Se ha constatado que existe una relación directa entre el color $U - B$ y logaritmo del momento central de orden 4 del espectro en las estrellas de éste catálogo. La función núcleo de Laplace es la que provoca menor error en la clasificación de los objetos estelares, realizando una clasificación satisfactoria del 90.68 % del total de los espectros. Por otro lado, las redes de Kohonen tienen un comportamiento eficiente en términos de clasificación adaptativa, y otorgan buenos resultados para topologías 4×4 tanto para redes supervisadas como no supervisadas.

A. Enlaces de interés

Se indican a continuación varios enlaces a páginas web en las que el lector podrá profundizar en alguno de los conceptos astronómicos que aparecen en el texto. También se ofrece la dirección del paquete FITSio para manipulación de archivos en formato FITS con el programa estadístico `R`.

- Color de una estrella: `http://cas.sdss.org/dr6/en/proj/advanced/color/`
- Tipo espectral de una estrella: `http://en.wikipedia.org/wiki/Stellar_classification`
- Acceso al catálogo JHC: `ftp://ftp.stsci.edu/cdbs/grid/jacobi`
- FITSio: `http://cran.r-project.org/web/packages/FITSio/index.html`

Referencias

[1] Berihuete A. (2010). *Modelos astroestadísticos de reducción de la dimensión aplicados al análisis de señales espectrales originadas en la superficie solar*. Tesis Doctoral, Universidad de Cádiz.

[2] Babu G.J. y Feigelson E.D. (1993). *Statistical Challenges in Modern Astronomy*, Springer-Verlag

[3] Babu G.J. y Feigelson E.D. (1996). *Astrostatistics*, Chapman y Hall.

[4] Jacoby G.H., Hunter D.A. y Christian C.A. (1984). A library of stellar spectra, *ApJS*, **56**, 257-281.

[5] Jiménez A. (2003). *Una visión unificada de las redes neuronales y la estadística multivariante*. Tesis Doctoral, Universidad de Cádiz.

[6] Jiménez A., Berihuete A. y Gutiérrez J.M. (2008). Los observatorios virtuales astronómicos como nuevo campo de aplicación de la minería de datos. *XXX Congreso Nacional de Estadística e I.O.*

[7] Sarro L.M. y Berihuete A. (2008). Feature selection in SUMER spatial spectra using wavelet decomposition and ICA. *AIP*, Doi: 10.1063/1.3059067.

[8] Starck J.L, Llebaria A. y Loredo T. (2008). Astrostatistics. *Stat. Methodol.*, **5**, 289-396.

Acerca de los autores

A. Berihuete es profesor sustituto interino en el Departamento de Estadística e I.O. de la Universidad de Cádiz. Doctor en Estadística por dicha Universidad y miembro del grupo de investigación ESTIO. En la actualidad colabora con el Spanish Virtual Observatory en el desarrollo de técnicas estadísticas para su aplicación en conjuntos de datos de tipo astronómico.

A. Jiménez es técnico superior estadístico del Centro Integrado de Tecnologías de la Información de la Universidad de Cádiz. Doctor en Estadística por dicha Universidad y miembro del grupo de investigación ESTIO.

F. Álvarez es catedrático de escuela universitaria en la Universidad de Cádiz. Doctor en Estadística por dicha Universidad y miembro del grupo de investigación ESTIO.

J.M. Gutiérrez es profesor emérito de la Universidad de Cádiz. Miembro del grupo de investigación ESTIO.