Análisis de la creación de un dataset de estrellas y su clasificación con diferentes métodos usando un clasificador supervisado

Héctor Abraham Galván García
Juan Pablo Murillo Macías

***Abstract*—El propósito de esta práctica es comparar diferentes métodos de clasificación y analizar la eficacia de cada uno de estos, tomando los datos de un dataset propio para así determinar las posibles causas de confusión y las mejores combinaciones de características en el clasificador supervisado.**

Palabras clave: Reconocimiento, Patrón, Entrenamiento, Clasificación, Training (data) set, Clasificador Supervisado, Distancia euclidiana, Matriz de confusión, Ruido (discrepancia), Eficacia, K_{nn}, Mínima distancia.

I. Introducción

El reconocimiento de patrones es natural en todo ser vivo, y cada uno de ellos lo hace de manera diferente. Por ejemplo, las personas en su mayoría se reconocen por vista al comparar sus características físicas con las que ya conocen de otros individuos, pero una persona ciega tiene que usar otro tipo de *reconocimiento*, como el tono o volumen de la voz. Todas las características que se reconocen de un objeto o individuo y/o lo describen son llamadas *patrones*.

En muchos de los casos de reconocimiento de patrones el problema es definir las características de los objetos que los identifican por clases, ya que no todas son importantes para el posterior proceso de clasificación en cada una de esas clases. Para determinar qué grupo de características describen mejor al objeto en cuestión se hace una *selección de características*. Una vez se tienen las características, se le llama *extracción de características* a la medición de los distintos objetos.

Ahora bien, la extracción de características puede ser difícil ya que previamente se tienen que elegir **aquellas que mejor puedan clasificar los objetos en sus diferentes clases y no las más fáciles de extraer**. Una vez extraídas empieza el entrenamiento, que sirve para identificar cuáles características generan más *discrepancia* a la hora de la clasificación. Un clasificador cuenta con la etapa de entrenamiento y la etapa de clasificación.

Ya identificadas y medidas se puede formar un set de entrenamiento, que contiene las características elegidas que servirán para el correcto entrenamiento y así mejorar la toma de decisiones del *clasificador*. Estas decisiones se basan en la delimitación de las medidas de los objetos. Cuando el programador o analista define esos límites, el clasificador es *supervisado*.

II. Creación del training set

En este caso, para comprobar el funcionamiento de distintos clasificadores y su eficacia (porcentaje de clasificación correcta), se tomaron estrellas como objeto a clasificar. Como no hay bases de datos completas y abiertas al público de telescopios o centros de investigación de astronomía, se recopiló la información de las estrellas de dos sitios principales: Wikipedia y Stellarium. Elegir estrellas da la ventaja de que existen diversas clasificaciones estelares. Se usará el sistema de clasificación Harvard, que responde a los siguientes rangos:

Clase	Temperatura[1] (Kelvin)	Color convencional	Color aparente[2 3 4]	Masa[1] (Masa solar)	Radio[1] (Radio solar)	Luminosidad[1] (bolométrica)	Líneas de hidrógeno	Fracción de la Secuencia principal[5]	Líneas de absorción
O	≥ 33.000 K	azul	azul	≥ 16 $M_\odot$	≥ 6,6 $R_\odot$	≥ 30.000 $L_\odot$	Débil-Media	~0.00003%	Nitrógeno, carbono, helio y oxígeno
B	10.000–33.000 K	azul a blanco azulado	azul a blanco azulado	2,1–16 $M_\odot$	1,8–6,6 $R_\odot$	25–30.000 $L_\odot$	Medio	0,13%	Helio, hidrógeno
A	7.500–10.000 K	blanco	blanco a blanco azulado	1,4–2,1 $M_\odot$	1,4–1,8 $R_\odot$	5–25 $L_\odot$	Fuerte	0,6%	Helio, hidrógeno
F	6.000–7.500 K	blanco amarillento	blanco	1,04–1,4 $M_\odot$	1,15–1,4 $R_\odot$	1,5–5 $L_\odot$	Medio	3%	Metales: hierro, titanio, calcio, estroncio y magnesio
G	5.200–6.000 K	amarillo	blanco amarillento	0,8–1,04 $M_\odot$	0,96–1,15 $R_\odot$	0,6–1,5 $L_\odot$	Débil	7,6%	Calcio, helio, hidrógeno y metales
K	3.700–5.200 K	naranja	anaranjado	0,45–0,8 $M_\odot$	0,7–0,96 $R_\odot$	0,08–0,6 $L_\odot$	Muy débil	12,1%	Metales y óxido de titanio
M	≤ 3.700 K	rojo	rojo anaranjado	≤ 0,45 $M_\odot$	≤ 0,7 $R_\odot$	≤ 0,08 $L_\odot$	Muy débil	76,45%	Metales y óxido de titanio

Fig. 1. Clasificación estelar Harvard.

Dentro de esta clasificación estelar, las clases de las estrellas son, de la más grande a la más chica respectivamente: O, B, A, F, G, K y M. Aún así hay subclasificaciones para estrellas de estas clases, como B9V de la estrella Arkab, pero para fines de una clasificación más atomizada sólo se tomaron las letras iniciales. Ahora bien, los patrones reconocibles de una estrella en Harvard son varios, pero debido a las limitaciones para la extracción de características, se decidió seleccionar sólo las siguientes cinco: Temperatura (en grados Kelvin), magnitud, masa, luminosidad y radio, siendo éstos últimos tres en escalas solares. Otro factor que influyó en esta elección es que para esas características los rangos están bien definidos en cuestión de números, lo que facilitará la tarea del clasificador.

Para la extracción de características, se tomó la clase y de la magnitud de la estrella del sitio web Stellarium, y los demás valores fueron obtenidos del Wikipedia. Se estructuró un aglomerado de datos en una hoja de cálculo para evitar confusiones a la hora de extracción de la siguiente manera:

NOMBRE	CÓDIGO HD	TEMP(K)	MAGNITUD	MASA(S)	LUMINOSIDAD(S)	RADIO(S)	CLASE	LINK
Achernar	10144	14510	0.46	8	1076	10	B	https://es.wikipedia.org/wiki/Achernar
Adhafera	89025	7030	3.41	3	207	10	F	https://es.wikipedia.org/wiki/Aldhafera
Alchi2a	105452	7000	4	1.2	4	1.67	F	https://es.wikipedia.org/wiki/AlchiBa

Fig. 2. Primera estructuración del training set.

Los datos de las estrellas se obtuvieron primero por nombre y después se ordenaron por clase para que el clasificador, a la hora de leer el archivo, muestre la *matriz de confusión* ordenada por clase. Al final, se logró extraer las características de 70 estrellas con las siguientes cantidades de cada una:

CLASE	CANTIDAD
O	1
B	17
A	12
F	7
G	13
K	15
M	5
TOTAL	70

Fig. 3. Cantidades de estrellas por clases.

Por último, todos los datos y sólo los datos se copiaron a otra hoja de cálculo, poniendo la clase al final y finalmente guardando el archivo como delimitado por comas (csv).

III. Clasificador de mínima distancia

Cuando se conoce el número de clases y los patrones son geométricamente separables, se puede usar una función de decisión para clasificar. Si los patrones forman grupos, se suele usar un clasificador que use la distancia como parámetro de decisión. En este caso, el *clasificador de mínima distancia* usa la fórmula de la distancia euclidiana para clasificar los patrones.

$$d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Fig. 4. Fórmula para calcular la distancia euclidiana.

Para calcular la eficacia del clasificador se programó una matriz de confusión $n * n$ donde n es el número de clases, que compara si la clase original del dataset es igual o no a la resultante después del entrenamiento y de la clasificación. La matriz identidad (diagonal principal) indica la eficacia de clasificación de cada clase en el orden en el que lee los archivos, y los elementos que estén en la misma fila de la clase pero fuera de la diagonal principal representan fallos de clasificación.

Se obtuvieron los siguientes resultados de clasificación tomando sólo una de las características y luego todas:

Característica	**Eficacia**
Radio	35
Luminosidad	31
Masa	33
Magnitud	28
Temperatura	94
Todas	42

Fig. 5. Eficacia de mínima distancia.

Se resalta en color rojo la característica que por sí sola, genera una peor clasificación, y en verde la que clasifica mejor. Además de éstas primeras seis pruebas preliminares, se realizaron pruebas de las otras 25 posibles combinaciones de características, arrojando que se clasifica mejor siempre que se toma la temperatura, a excepción de cuando se junta con la luminosidad. A su vez, la luminosidad es la que al encontrase como característica activa, genera más discrepancia (ruido) en la clasificación, incluso más que la magnitud. No hubo eficacia mayor al 94%.

IV. Clasificador de vecinos cercanos

Cuando existen varios grupos, se usa un *clasificador de vecinos cercanos* o K_{nn} donde K es el número de vecinos. Para el proceso de entrenamiento recibe un conjunto de patrones incluyendo las clases involucradas y usa el cálculo de la mínima distancia. Para el proceso de clasificación es de acuerdo al siguiente orden: 1.- Se recibe un patrón/muestra a clasificar. 2. Se calculan las distancias entre el patrón anterior con respecto a las instancias de entrenamiento. 3- Opcionalmente y para fines del clasificador ya programado, se ordenan las instancias. 4.- En base a las distancias, se verifica la clase que cumple con el número K de vecinos más cercanos.

Se empleó la matriz de confusión del clasificador de mínima distancia, obteniendo los siguientes resultados:

Característica	**Eficacia**
Radio	58
Luminosidad	53
Masa	33
Magnitud	32
Temperatura	82
Todas	71

Fig. 6. Eficacia de vecinos cercanos.

Al igual que en el clasificador anterior, se remarcó con verde la mejor característica para clasificar y con rojo la peor. A su vez, se repitió el caso de que la luminosidad causa la mayor discrepancia cuando se utiliza. Además, nótese que la eficacia más alta (82% de temperatura) no supera a la de mínima distancia, debido a que la única estrella de clase O se clasifica mal. Se probaron todas las combinaciones posibles y no hubo eficacia mayor al 82%. Un número K de vecinos de 2 a 5 da la misma eficacia, y a partir de 6 la eficacia baja. Alternativamente se empleó un dataset de prueba al que se agregó una estrella extra de clase O, lo cuál subió la eficacia de esa clase al 50% y la general al 89%.

V. Conclusiones

Es muy difícil que un clasificador supervisado tenga un 100% de eficacia, aunque se puede presentar el caso. Para este dataset, la mayor eficacia fue del 94% y se presentaron errores mostrados en la siguiente matriz de confusión para mínima distancia y vecinos cercanos:

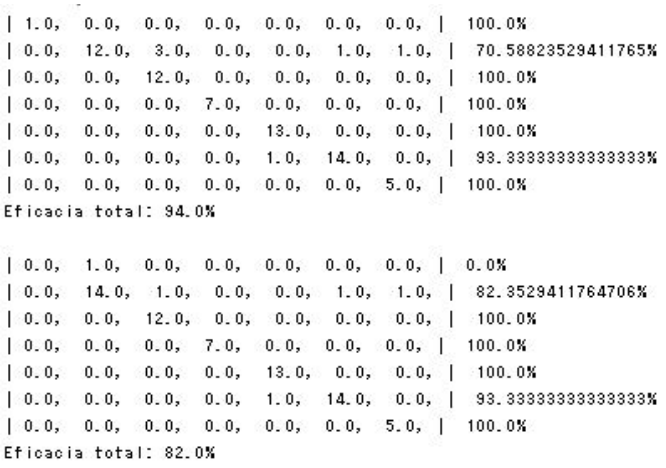

Fig. 7. Mayor eficacia con temperatura.

Primeramente, se pueden identificar tres de los cinco errores de mínima distancia en la clase B con certeza , debido a que su temperatura es muy parecida a estrellas de clase A, K ó M. El error en la clase K fue una estrella cuya temperatura está en el rango de las de clase G y por ende, las clasificó así. También, cabe aclarar que la magnitud por sí sola es la peor para clasificar porque son cantidades muy pequeñas y varían muy poco, por lo que el clasificador se confunde. Lo mismo pasa con la luminosidad pero con cantidades muy grandes.

Segundamente, los errores de vecinos cercanos que presenta la clase O es un caso particular y curioso al mismo tiempo. Debido a que el dataset sólo tiene una estrella de esa clase, en el proceso de entrenamiento sólo entrena con ese patrón, y por eso lo clasifica mal. Este argumento se respalda en el hecho de que al agregar una estrella extra de clase O, la eficacia subiera al 50%. No se probó con más estrellas de clase O para ratificar esta teoría porque no se encontró información completa de más estrellas de esa clase. Es por esto que la efectividad de vecinos cercanos es inferior a la de mínima distancia, aunque si se tuvieran más muestras para entrenar es muy seguro que la efectividad fuera superior.

VI. Bibliografía

R. Duda, P. Hart and D. Stork, Pattern classification, 2nd ed. New York: Wiley-Interscience, 2000.

M. Friedman and A. Kandel, Introduction to Pattern recognition. New Jersey: World Scientific, 2005.