ISSN 1870-4069

Redes neuronales y árboles de decisión para la clasificación de objetos astronómicos

Rogelio Hernández Montes[1], Cynthia Alejandra Martínez Pinto[1], Silvana Guadalupe Navarro Jiménez[2]

[1] Tecnológico Nacional de México, ITCG, Jalisco, México
[2] Universidad de Guadalajara, Instituto de Astronomía y Meteorología, Jalisco, México
roger.hdez.m@hotmail.com, cynthia_amp@hotmail.com, silvananj@gmail.com

Resumen. El objetivo de esta investigación consiste en comprobar la factibilidad de realizar una clasificación de objetos astronómicos por medio de parámetros fotométricos. Esta identificación da como resultado objetos candidatos, que tienen la probabilidad de ser simbióticos. La búsqueda se realiza en las bases de datos de la misión GAIA, en su más reciente entrega de información (DR2). La clasificación se efectúa a través de inteligencia artificial. Para el estudio se utilizaron las técnicas de Redes Neuronales Artificiales (RNA), Random Forest (RF), Árboles de Decisión y Máquinas de Soporte Vectorial (SVM). Con estos algoritmos se realizó una clasificación binaria. Se desarrolló un software para la optimización de los hiperparámetros. Los resultados de la investigación mostraron que Random Forest tuvo un mejor comportamiento (98%) en la fase de comparación, y por su parte Árboles de decisión fue el que menor índice tuvo (90%). Además de tener un índice kappa superior, los bosques aleatorios presentan una menor complejidad de configuración.

Palabras clave: identificación automática, sistemas simbióticos, análisis de datos.

Neural Networks and Decision Trees for the Classification of Astronomical Objects

Abstract. The objective of this research was to verify the feasibility of classify astronomical objects using photometric parameters. This identification gives as result candidate objects, who have the probability of being symbiotic. The search was done in the databases of the GAIA mission data, in its most recent data release (DR2). The classification is made through artificial intelligence. For the study the techniques of Artificial Neural Networks (RNA), Random Forest (RF), Decision Trees and Support Vector Machines (SVM) were used. A binary classification was made with these algorithms. A software was developed for the optimization of hyperparameters. The results of the research showed that Random Forest had a better performance (98%) in the comparison phase, and Decision Trees was the one with the lowest rate (90%). In addition to having a higher kappa index, random forests have less configuration complexity.

Keywords: automatic identification, symbiotic systems, data analysis.

1. Introducción

En Astronomía se generan actualmente más de un Terabyte de datos por día [2]. La información generada es almacenada en bases de datos que son públicas en su mayoría y están localizadas en distintas partes del mundo. Tal cantidad de datos hace necesario el desarrollo de software especial, dedicado al análisis y clasificación de esta información. Entre los objetos de interés u objetos peculiares, se encuentran las estrellas simbióticas (SS), que son sistemas binarios, conformadas por una estrella de muy alta temperatura y una estrella fría, que generalmente son una enana blanca y una gigante roja, o una estrella variable tipo Mira, dichas estrellas se encuentran ligadas gravitacionalmente [1]. Un ejemplo ha sido observado en la simbiótica más cercana a la tierra: R Aqr [3].

Aunque la confirmación de su clasificación es a través del espectro, es posible identificar objetos que son candidatos a ser sistemas simbióticos utilizando observaciones fotométricas que, aunque no nos dan una clasificación precisa, nos pueden indicar la posibilidad de que se trate de uno de estos objetos y, por otra parte, permite distinguirlos de otro tipo de objetos con los cuales es frecuente confundir a los SS, por ejemplo las nebulosas planetarias (NPs), pues comparten algunas de sus características físicas, las cuales se ven reflejadas en sus características espectrales.

En astrofísica, la distancia a la que se encuentran los objetos es un parámetro fundamental, sin él no es posible conocer la luminosidad real de un objeto, su tamaño físico o determinar con precisión la distribución de los objetos en la galaxia y, con ello, la estructura de la misma. Así mismo, otros parámetros, como su densidad o magnitud absoluta, tampoco podrían ser determinados sin conocer su distancia. El proyecto GAIA fue pensado inicialmente para determinar el paralaje de más de 10^9 objetos en la vía láctea [4]. Dada la precisión de sus mediciones, es posible determinar distancias de hasta varios cientos de parsecs. Adicionalmente, GAIA está realizando mediciones fotométricas en tres bandas (G,BP y RP), determinaciones de movimientos propios y velocidades radiales de las estrellas y está obteniendo espectros de baja y alta resolución.

La liberación de los datos más reciente (25 de abril de 2018) denominada “data release 2” (DR2), contiene una gran base de datos con miles de objetos observados con información sobre paralajes, posiciones precisas, fotometría en G, BP y RP, determinación del error de cada uno de estos parámetros, además de determinaciones muy precisas de los movimientos propios de una fracción importante de los objetos y de sus velocidades radiales. Con esta información es posible determinar estrellas simbióticas candidatas que cubren las características fotométricas de las estrellas del estudio y presentarlas a la comunidad científica para su posterior confirmación a través de sus respectivos espectros.

Para hacer la clasificación de estas estrellas se utilizan dos técnicas de aprendizaje automático, las redes neuronales y los árboles de decisión. El algoritmo de las redes es preciso y requiere ser alimentado con valores muy puntuales, para lograr obtener la clasificación deseada, mientras que el algoritmo de los árboles es menos riguroso en su entrada de datos, pero la clasificación es aleatoria que en las redes neuronales. Sin embargo los resultados son alentadores para el manejo de una clasificación binaria, mostrando una ventaja sobre las redes neuronales.

2. Metodología

2.1. Adquisición y tratamiento de los datos

Para la recuperación de los datos de las estrellas simbióticas se recurrió al catálogo de Belczyński, et al. (2000), donde recopilan un total de 188 objetos bajo esta clasificación [5]. De este catálogo, se extrajo el nombre y se procedió a hacer una búsqueda cruzada entre las bases de datos de GAIA y 2MASS, utilizando SIMBAD, la cual es una base de datos astrofísicos que es actualizada diariamente. Esta base de datos reúne la información publicada de los objetos e incluye la identificación cruzada entre las diversas base de datos de los grandes proyectos de "survey". De la información obtenida, se recuperaron los valores de las mediciones fotométricas en distintos filtros (G, Bp, Rp, B, V, J, H, K, estos últimos de 2MASS). Los parámetros recupera-dos de GAIA fueron: *source_id, ra, ra_error, dec, dec_error, parallax, parallax_error, parallax_over_error, phot_g_mean_flux, phot_g_mean_mag, phot_bp_mean_flux, phot_bp_mean_mag, phot_rp_mean_flux, phot_rp_mean_mag, bp_rp, bp_g, g_rp, radial_velocity, radial_velocity_error, phot_variable_flag, teff_val, a_g_val*. Estos valores se describen a continuación en la tabla 1.

Tabla 1. Descripción de los campos de la base de datos de GAIA.

Campo recuperado	Descripción
source_id	Identificador único perteneciente a la base de datos de la misión GAIA. En este se encuentra codificado la posición del objeto.
RA	Valor perteneciente a la misión GAIA. Ascensión recta del objeto, el valor se encuentra expresado en formato ICRS
ra_error	Valor perteneciente a la misión GAIA. Error estándar de la ascensión directa.
Dec	Valor perteneciente a la misión GAIA. Declinación del objeto, el valor se encuentra expresado en formato ICRS
dec_error	Valor perteneciente a la misión GAIA. Error estándar de la declinación del objeto.
Parallax	Valor perteneciente a la misión GAIA. Paralaje estelar del objeto
parallax_error	Valor perteneciente a la misión GAIA. Error estándar del paralaje.
parallax_over_error	Valor perteneciente a la misión GAIA. Valor obtenido de la división del paralaje sobre su error.
phot_g_mean_flux	Valor perteneciente a la misión GAIA. Flujo medio integrado en la banda G.
phot_g_mean_mag	Valor perteneciente a la misión GAIA. Magnitud media en la banda G calculada en base a la magnitud de Vega.
phot_bp_mean_flux	Valor perteneciente a la misión GAIA. Flujo medio integrado en la banda BP.

Campo recuperado	Descripción
phot_bp_mean_mag	Valor perteneciente a la misión GAIA. Magnitud media en la banda BP.
phot_rp_mean_flux	Valor perteneciente a la misión GAIA. Flujo medio integrado en la banda RP .
phot_rp_mean_mag	Valor perteneciente a la misión GAIA. Magnitud media en la banda de RP .
bp_rp	Valor perteneciente a la misión GAIA. Color resultante de la diferencia de las magnitudes de las bandas BP y RP
bp_g	Valor perteneciente a la misión GAIA. Color resultante de la diferencia de las magnitudes de las bandas BP y G
g_rp	Valor perteneciente a la misión GAIA. Color resultante de la diferencia de las magnitudes de las bandas G y RP
radial_velocity	Valor perteneciente a la misión GAIA. Velocidad radial espectroscópica en el marco de referencia baricéntrico solar
radial_velocity_error	Error de la velocidad radial.
phot_variable_flag	Valor perteneciente a la misión GAIA. Bandera de variabilidad fotométrica. Los posibles valores para este campo son *variable* que indica que el objeto se identificó y procesó como variable, *constant* lo cual hace referencia a que no se encontró variación, y *not_available* que es usado cuando no ha sido procesado y/o exportado al catálogo.
teff_val	Valor perteneciente a la misión GAIA. Temperatura estelar efectiva.
a_g_val	Valor perteneciente a la misión GAIA. Valor de extinción en la banda G.
B	Valor recuperado de SIMBAD. Magnitud en la banda B.
V	Valor recuperado de SIMBAD. Magnitud en la banda V.
J	Valor de 2MASS, recuperado con SIMBAD. Magnitud en la banda J.
H	Valor de 2MASS, recuperado con SIMBAD. Magnitud en la banda H.
K	Valor de 2MASS, recuperado de SIMBAD. Magnitud en la banda K.

Del catálogo sólo se recuperaron 102 objetos ya que algunos de ellos presentan irregularidades en sus campos, entre ellas se encuentran estrellas con parallax de 0 o falta de mediciones en las bandas G, BP o RP y por lo tanto la falta en los colores bp_rp, bp_g y g_rp. Además de la recuperación de las estrellas simbióticas, también se procedió a descargar 20,000 estrellas la secuencia principal descrita en el artículo Gaia Data Release 2-Observational Hertzsprung-Russell diagrams [6] donde se construye y describe el diagrama de Hertzprung-Russell. También se incluyeron otros conjuntos de objetos evolucionados como lo son las nebulosas planetarias.

2.2. Unión y corrección de los datos

Los objetos fueron recuperados en formato .vot en el caso de la base de datos de GAIA, y en formato Web para SIMBAD. A causa de la generación de distintos archivos y en distinto formato, fue menester la unión de todos los valores en un sólo formato, además de realizar correcciones y la generación de otros campos necesarios para la clasificación. Se agregó una columna que contiene la magnitud absoluta de cada uno de los objetos de estudio, debido a que GAIA sólo contiene una magnitud observada. Es importante realizar el cálculo de la magnitud absoluta para poder conocer la verdadera luminosidad de un objeto sin que esta se vea afectada por la distancia real del objeto y así tener todos los objetos con la magnitud aparente que tendrían si estuvieran a una distancia de 10 parsecs.

La magnitud absoluta en G se determinó utilizando la ecuación 1, con la que se realiza también la corrección por extinción, y por su parte la ecuación 2 se utilizó para convertir el paralaje a distancia:

$$Gabs=5+G-5\log(d)+Ag, \quad (1)$$

$$d=(1/parallax), \quad (2)$$

donde:
Gabs: Magnitud absoluta en la banda G; G: Magnitud observada en la banda G; Ag: Valor de extinción en G; d: distancia expresada en parsecs; parallax: valor de parallax expresado en segundos de arco

2.3. Normalización de los datos

Los distintos conjuntos de datos se etiquetaron de acuerdo a un criterio binario: si el objeto es una estrella simbiótica o no. Posterior al etiquetado se realizó una fusión de los datos generados para cada tipo de objeto con el fin de crear un único set de datos. Sobre el conjunto de datos resultante se procedió a realizar una normalización. La cual se realizó de acuerdo a la ecuación 3. Los valores utilizados para los distintos parámetros del vector de características fueron recuperados a través de consultas sobre las bases de datos GAIA (DR2) y 2MASS. El proceso consistió en encontrar la combinación de valores frontera para cada parámetro de los objetos. Los valores se muestran a continuación en la tabla 2. La normalización generada por la ecuación es necesaria para poder comparar los valores entre ellos y así definir el rango de valores máximos y mínimos que podrá adoptar cada parámetro. Para la investigación se seleccionó como valor máximo, identificado en la ecuación como d2, el valor de 1 y para el valor mínimo, identificado como d1, el valor de 0. Por lo tanto, todos los parámetros del vector de características estarán comprendidos en el rango [0,1]:

$$y=((x-xmin)(d2-d1))/(xmax-xmin)+d1, \quad (3)$$

donde:
y: valor de x normalizado,
x:Valor a normalizar,
xmin: valor mínimo posible para x,
xmax: valor máximo posible para x,
d1: valor mínimo que podrá tomar y,

d2: valor máximo que podrá tomar y.

Tabla 2. Valores frontera usados para la normalización de los parámetros utilizados.

Parámetro	Valor mínimo	Valor máximo
bp_rp	-5.489193	9.800095
g_rp	-4.157661	3.650905
bp_g	-2.073900	7.795892
g_abs_mag	-39.51653	26.91888
teff_val	3229	9803

2.4. Selección y personalización de los algoritmos

Separación del conjunto de datos. Con el conjunto de datos normalizados se procedió a hacer una separación del 16.5% del total, con el fin de ser usados como validación para los 4 algoritmos que se programaron, 16.5% para las pruebas, y el 66.6% para el entrenamiento. El conjunto de prueba se usó para verificar el nivel de aprendizaje de los algoritmos, mientras que el set de validación se utilizó para comprobar que el algoritmo fuera capaz de generalizar la clasificación y detectar de esta manera una especialización sobre los datos de entrenamiento y prueba.

Los algoritmos seleccionados para la clasificación binaria fueron: redes neuronales, "random forest", árboles de decisión y máquinas de soporte vectorial. La selección de estos algoritmos se llevó a cabo después de una revisión del estado del arte con referencia a clasificación estelar. Las redes neuronales y "random forest" han sido utilizados y recomendados en investigaciones previas [7][8] cuyo carácter es similar tanto en campo de investigación como en similitud de los datos con los que se cuenta. Con el fin de poder comparar estos algoritmos se han añadido los algoritmos de árboles de decisión y máquinas de soporte vectorial.

De la información recuperada de ambas base de datos, se procedió a seleccionar los campos con los que trabajarían los algoritmos seleccionados. El criterio de selección utilizado, fue tomar aquellos datos que ayudan a una mejor separación de las clases a identificar. Con los parámetros elegidos, el vector de características final se compone con los campos que se muestran a continuación en la tabla 3.

Tabla 3. Descripción de los campos del vector de características usado para los algoritmos de clasificación.

Parámetro	Valor máximo
Gabs	Magnitud absoluta en la banda G
teff_val	Temperatura efectiva
bp_rp	Color resultante de la diferencia de las magnitudes de la banda Bp y Rp
b_rp	Color resultante de la diferencia de las magnitudes de la banda G y Rp
bp_g	Color resultante de la diferencia de las magnitudes de la banda Bp y G

j_h	Color resultante de la diferencia de las magnitudes de la banda J y H
h_k	Color resultante de la diferencia de las magnitudes de la banda H y K
b_v	Color resultante de la diferencia de las magnitudes de la banda B y V
v_k	Color resultante de la diferencia de las magnitudes de la banda V y K

2.5. Redes neuronales

El primer acercamiento para la clasificación de las estrellas simbióticas se realizó a través de las redes neuronales supervisadas. El modelo de red neuronal que se usó fue el perceptrón multicapa. El algoritmo de entrenamiento seleccionado fue el de retro-propagación con gradiente descendente y una función sigmoidal para activación.
Con el fin de determinar la mejor topología para la red neuronal se automatizó el proceso de creación, entrenamiento y prueba del algoritmo mediante un programa en Java. Con el proceso automatizado se procedió a codificar una búsqueda aleatoria sobre espacios continuos de distintos parámetros de la red neuronal. Dichos parámetros se muestran a continuación en la tabla 4.

Tabla 4. Descripción de los parámetros utilizados para la creación automática de redes neuronales.

Parámetro de la Red Neuronal	Selección
Factor de aprendizaje	Búsqueda aleatoria en el espacio continuo [.3. .9]
Numero de Capas	Búsqueda Aleatoria en el espacio discreto 1-3
Numero de neuronas por capa	Búsqueda Aleatoria en el espacio discreto 3-255

El algoritmo de automatización se diseñó para trabajar limitado por tiempo. Dado un tiempo específico, crea la mayor cantidad de sujetos (redes neuronales) y se evalúan. Una vez finalizado el tiempo especificado sólo despliega la información correspondiente al sujeto con las mejores características registradas.

El proceso de evaluación que se realiza sobre cada sujeto es llevado a cabo separando el set de datos. La separación se llevó como se describe en la tabla 5.

Tabla 5. Descripción de la división y uso del conjunto de datos.

Fracción del set de datos total (porcentaje)	Uso	Separación
2/3 (66%)	Entrenamiento	Realizada de forma aleatoria al comienzo del entrenamiento de cada red neuronal
1/6 (16.6%)	Pruebas	Realizada de forma aleatoria al comienzo del entrenamiento de cada red neuronal
1/6 (16.6%)	Validación	Realizada y separada antes del proceso de creación de redes neuronales.

2.6. Random forest

El algoritmo Random forest consta de una gran cantidad de árboles de decisión, que a diferencia de las redes neuronales, cuenta con una menor cantidad de parámetros para su construcción y entrenamiento [10]. Los valores que deben ser cambiados son el número de clasificadores que serán generados, el número de variables a tomar, y un parámetro opcional es limitar la profundidad máxima que puede tener cada árbol generado.

La principal característica del algoritmo random forest es que aunque sea un algoritmo de aprendizaje supervisado como las redes neuronales, no es posible controlar la topología de los estimadores generados por el algoritmo. Otra de las características notables del algoritmo es que para la creación de cada uno de los árboles estimadores, los parámetros que usará no son controlados, sino que son escogidos aleatoriamente. Este algoritmo también se distingue por el hecho de que el set de datos que requiere, puede o no estar normalizado.

Para el caso de estudio se realizaron pruebas variando el número de clasificadores entre [1,3000]. Para todas las pruebas realizadas se estableció como máximo número de variables a tomar por cada estimador la longitud del vector de características. Por lo que la configuración para este algoritmo quedo de la manera que se describe en la tabla 6.

Tabla 6. Descripción de los parámetros utilizados en la generación de estimadores para el algoritmo random forest.

Parámetro	Selección
Profundidad máxima	Sin límite
Árboles	[1, 3000]
Mínimo de muestras por hoja	1
Máximo de características a tomar	Número de características
Número máximo de hojas	Sin límite

2.7. Árboles de decisión

Los arboles de decisión son un algoritmo que genera un árbol como un diagrama de flujo, donde cada nodo interno representa un valor del vector de características, las ramas representan decisiones, y los nodos hojas representan un resultado de la clasificación. Este es un algoritmo de clasificación de aprendizaje supervisado, no paramétrico, que no requiere supuestos distribucionales, permite modelar relaciones no lineales, y no es sensible a la ausencia de datos [12]. La forma básica de funcionamiento es la creación de particiones recursivas de acuerdo con reglas de asignación, partición y parada.

Entre las principales ventajas de este algoritmo es que crea un modelo de caja blanca, por lo que es fácil de comprender el resultado. Y requiere poca preparación del set de datos, ya que no requiere la eliminación de datos faltantes, o normalización.

2.8. Máquinas de soporte vectorial

Las máquinas de soporte vectorial son un algoritmo de aprendizaje supervisado. Este algoritmo puede manejar fácilmente variables continuas y paramétricas. El funcionamiento básico del algoritmo es la creación de un hiperplano en un espacio multidimensional para la separación de las clases. Las máquinas de soporte vectorial según lo expuesto por Auria [13] tienen la ventaja de presentar un buen rendimiento cuando los datos son no linealmente separables. Además de poder manejar una alta dimensionalidad de datos.

3. Resultados

Con el fin de poder comparar la eficacia de los algoritmos utilizados para clasificar el set de datos de validación, se tomó en cuenta la precisión mostrada en la matriz de confusión.

Los resultados de evaluar los distintos algoritmos pueden considerarse, en forma general, como muy positivos, ya que las estrellas simbióticas son un tipo de objeto difícil de identificar. Además que de la manera tradicional de hacer la identificación sobre objetos peculiares es hacerlo sobre los espectros de los objetos. Sin embargo en la investigación realizada se optó por valores fotométricos debido a la falta de disponibilidad de espectrometría en GAIA. Los algoritmos lograron separar las estrellas no simbióticas de las estrellas simbióticas. Sin embargo los algoritmos clasificaron estrellas simbióticas como no simbióticas. El algoritmo random forest mostró una mejor clasificación sobre redes neuronales.

A continuación en la tabla 7 se despliega la matriz de confusión del algoritmo random forest. Los resultados de las distintas técnicas fue una completa separación de las estrellas no simbióticas de las simbióticas, por lo que la matriz no refleja falsos positivos. Sin embargo las técnicas presentan falsos negativos, clasificaron estrellas simbióticas como estrellas no simbióticas.

Tabla 7. Matriz de confusión correspondiente al algoritmo random forest.

0	1	
147	0	0=0
1	36	1=1

En la tabla 8 se muestra la matriz de confusión obtenida del algoritmo de automatización sobre redes neuronales.

Tabla 8. Matriz de confusión correspondiente al algoritmo sobre redes neuronales.

0	1	
147	0	0=0
2	35	1=1

En la tabla 9 se muestra la matriz de confusión resultante del algoritmo de árboles de decisión.

Tabla 9. Matriz de confusión correspondiente al algoritmo de árboles de decisión.

0	1	
139	0	0=0
6	39	1=1

En la tabla 10 se despliega la matriz de confusión obtenida a partir del algoritmo de máquinas de soporte vectorial.

Tabla 10. Matriz de confusión correspondiente al algoritmo de máquinas de soporte vectorial

0	1	
139	0	0=0
3	42	1=1

Los resultados mostrados para los algoritmos, representa al mejor sujeto generado durante las pruebas. Para las redes neuronales, se utilizó el algoritmo de automatización durante 72 horas. Los resultados obtenidos fue la generación de 1128 modelos, por lo que se tiene que en promedio se han generado 15.6 modelos de redes neuronales por hora, lo que da un tiempo promedio de entrenamiento de 3.49 segundos en promedio para cada modelo. Por otra parte, para la generación de clasificadores utilizando el algoritmo random forest, se procedió a la creación de 1128 sujetos, y el tiempo que se llevó la ejecución fue de 10 horas.

El criterio para la evaluación de cada sujeto fue la precisión obtenida al generar la matriz de confusión. La evaluación fue realizada dentro de la automatización y únicamente se tenía control sobre el mejor sujeto. Por este motivo, al momento de que se generaba un sujeto superior, se sustituía el mejor sujeto por el nuevo.

Con motivo de comparar mejor la matriz de confusión los distintos algoritmos se calculó el índice kappa para cada algoritmo en base al mejor sujeto presentado. La comparación se muestra a continuación en la tabla 11. De acuerdo a la escala propuesta por Landis y Koch [9] el algoritmo random forest presenta un grado de concordancia casi perfecto.

Tabla 11. Valores de los índices kappa para los algoritmos.

	Redes neuronales	Random forest	Árboles de decisión	Máquinas de soporte vectorial
Índice Kappa	0.9654	0.9829	.9075	.9548

Con el fin de ilustrar el funcionamiento de random forest (RF) se extrajo uno de los árboles estimadores perteneciente al mejor sujeto generado. Dicho árbol es desplegado en la figura 1. Al observar el árbol se aprecia que la forma de clasificación que utilizó es similar a la creación de un diagrama color-magnitud, proceso usado en astrofísica para la caracterización. La magnitud usada fue *Gabs* y el color que usa el *g_rp*.

La primera separación que realiza el árbol es sobre el color, acto seguido confirma las regiones a través de las magnitudes. Sólo en el par de nodos hoja más profundos se realizó una segunda comprobación por color. El presente árbol es uno de los distintos

árboles generados, por lo que su configuración puede no ser correcta para todos los casos presentes en el set de validación.

El índice Gini permite seleccionar dos elemetos de una población, estos deben ser de la misma clase, donde la probabilidad de que esto suceda es uno, cuando la población es "pura". Utiliza dos variables categóricas: "Success" o "Failure" y entre más grande sea el índice Gini, mayor es la homogeneidad de los datos. El cálculo de los subnodos usa la ecuación 4 que consiste en suma de los cuadrados de probabilidad para success y failure [11]:

$$(p^2 + q^2), \tag{4}$$

donde p: es success y q es failure.

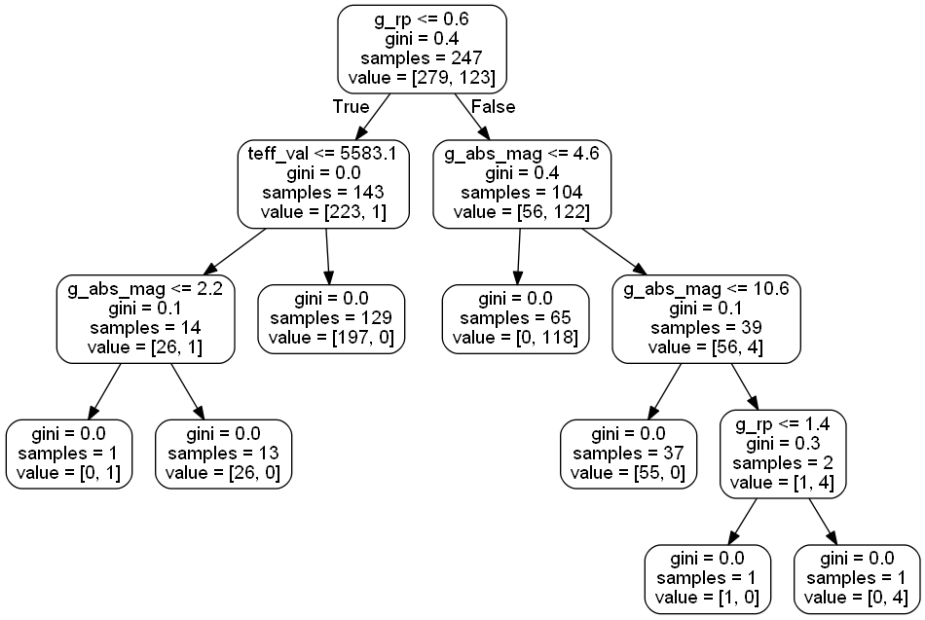

Fig. 1. Árbol de decisión generado aleatoriamente por random forest (RF).

4. Conclusiones

Después de analizar el comportamiento de los distintos algoritmos, se aprecia que random forest (RF) representa una mejor opción al momento de tener que seleccionar un algoritmo para una clasificación binaria de 20,136 objetos. El algoritmo random forest presenta un resultado casi similar al de las redes neuronales, sin embargo es importante entender que el algoritmo es completamente aleatorio. Por esta razón mejorar los resultados de este algoritmo realizando un cambio en sus parámetros no garantiza obtener mejores resultados.

El algoritmo random forest crea un número interno de estimadores y para evaluar el modelo generado se realiza una "votación" entre todos los estimadores, donde la

clasificación generada por el algoritmo es la clase con más votaciones. Por esta razón aunque se recomiende el uso de random forest para problemas de clasificación binaria, es muy importante que se analice el comportamiento de los modelos al aumentar el número de clases disponibles. Dado que se incluirán más tipos de objetos, será necesario evaluar la conveniencia de separar cada clase con un RF independiente.

También se pudo observar que el algoritmo de los bosques aleatorios tiene el mejor índice kappa en comparación con los demás métodos evaluados, incluso con los árboles de decisión, que aunque manejan una estructura de datos similar a los Random Forest, el algoritmo para la clasificación demuestra ser mucho más eficiente, gracias a la cantidad de árboles que genera el RF, donde cada árbol vota por una clase y el resultado es la clase con mayor número de votos en todo el bosque [11]. Por su parte las máquinas de soporte vectorial, tuvieron un buen desempeño, pero como pasa con las redes neuronales, su desempeño es el resultado de una buena parametrización, por lo que se tendrá una gran cantidad de posibles configuraciones. Y como el mismo caso que RNA, encontrar los parámetros ideales conlleva un gran trabajo.

Las fortalezas que se encontraron de algoritmo de random forest sobre las redes neuronales, son la facilidad de implementación y generación de modelos, así mismo el tiempo invertido para la obtención de resultados similares juega un papel importante para optar por los bosques aleatorios.

5. Recomendaciones y trabajo futuro

Aun no se ha concluido la presente investigación. El trabajo que falta por realizar es seguir generando modelos tanto de redes neuronales como random forest con el fin de alcanzar un mayor grado de confiabilidad en la clasificación de estrellas simbióticas. Una vez alcanzado este objetivo, se añadirán más tipos de objetos al conjunto de datos con el fin de ver el comportamiento de los clasificadores.

Como recomendación al momento de trabajar con el algoritmo de redes neuronales para clasificaciones binarias, se usaba una función de error basada en el error cuadrático medio, con lo cual se obtenía un índice kappa alrededor de .7 al momento de variar sólo esta función se mejoró considerablemente la precisión de la red, alcanzado un índice kappa de .96. La función de error utilizada fue la cross entropy la cual presenta mejores resultados para problemas de clasificación binaria. A lo largo de esta investigación se encontró que la mejor forma de mejorar los resultados de los algoritmos random forest y árboles de decisión, además variar el número de árboles estimadores en el caso del primer algoritmo, es experimentar con la profundidad máxima de los árboles. Un punto importante es explorar las posibilidades donde la profundidad es menor al tamaño del vector de características, ya que de otra forma los algoritmos tendrán un buen rendimiento durante la fase de entrenamiento, pero esto no asegura un buen resultado con un set de datos que el algoritmo no ha visto, ya que a más profundidad los árboles tienden a especializarse en los datos y forzar las clasificaciones.

Referencias

1. Belczyński, K., Mikołajewska, J., Munari, U., Ivison, R. J., Friedjung, M.: A catalogue of symbiotic stars. Astronomy and Astrophysics Supplement Series, 146(3), 407–435 (2000)

2. Hernández Cervantes, L., Santillán González, A., González-Ponce, A.: Observatorios Virtuales Astrofísicos. Revista Digital Universitaria, Vol. 10, No. 10 (2009)
3. H. M. Schmid, A. Bazzon, J. Milli, et al.: SPHERE/ZIMPOL observations of the symbiotic system R Aquarii-I. Imaging of the stellar binary and the innermost jet clouds. A&A Vol. 602, p. A53 (2017)
4. Gaia Collaboration, T. Prusti, J. H. J. de Bruijne, A. G. A. Brown, A. Vallenari, C. Babusiaux, C. A. L. Bailer-Jones, U. Bastian, M. Biermann, D. W. Evans, et al.: The gaia mission. A&A 595, p A1 (2016)
5. K. Belczyński, J. Mikołajewska, U. Munari, R. J. Ivison, M. Friedjung: A catalogue of symbiotic stars, Astron. Astrophys. Suppl. Ser. 146 (3) 407–435 (2000)
6. Gaia Collaboration, C. Babusiaux, F. van Leeuwen, M. A. Barstow, C. Jordi, A. Vallenari, D. Bossini, A. Bressan, T. Cantat-Gaudin, M. van Leeuwen, et al.: Gaia Data Release 2-Observational Hertzsprung-Russell diagrams. Astronomy & Astrophysics 616, A10 (2018)
7. Kheirdastan, S., Bazarghan, M.: SDSS-DR12 bulk stellar spectral classification: Artificial neural networks approach. Astrophysics and Space Science, 361(9), 1–8. doi:10.1007/s10509-016-2880-3 (2016)
8. Colas F., Brazdil, P.: Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks. In: Bramer M. (eds) Artificial Intelligence in Theory and Practice. IFIP AI 2006. IFIP International Federation for Information Processing, vol. 217, Springer, Boston, MA (2006)
9. Landis, J. Richard, and Gary G. Koch.: The measurement of observer agreement for categorical data. Biometrics 159–174 (1977)
10. Keller, C. A., Evans, M. J.: Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10. Geoscientific Model Development, 12(3), 1209–1225 (2019)
11. Árboles de decisión y Random forest. https://bookdown.org/content/2031/, last accessed 2019/04/01.
12. Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J.: Classification and regression trees. Belmont, CA: Wadsworth. International Group, 432 (1984)
13. Auria, L., Moro, R. A.: Support vector machines (SVM) as a technique for solvency analysis. (2008)