Modelo de clasificación de cuerpos celestes

Alejandro Montoya Garcia, *Universidad de Antioquia*

***Index Terms*—Cuerpos celestes, Sloan Digital Sky Survey, SDSS, modelo de clasificación**

I. DESCRIPCIÓN DEL PROYECTO

the Sloan Digital Sky Survey"(SDSS)[1] es un dataset que ofrece datos públicos de observaciones espaciales. El problema planteado consiste en clasificar las diferentes observaciones por clase dónde cada observación puede pertenecer a uno de las siguientes clases: estrella, galaxia o cuásar. Para realizar esta clasificación se cuenta con 17 características de cada observación, esto podrá utilizarse en la astronomía para clasificar las más recientes observaciones de los diferentes telescopios, dado que es muy importante poder realizar estas clasificaciones de manera automática, porque se recopilan miles de datos de manera diaria y sería casi imposible revisar cada dato de manera Manual.

II. DESCRIPCIÓN DE LOS DATOS

El conjunto de datos original cuenta con 18 características en total, de entre estas se toman 17 como entrada y una como salida, las cuales son descritos en el cuadro I

Run,rerun, camcol y field son características que describen un campo dentro de una imagen tomada por el SDSS. Un campo es básicamente una parte de la imagen completa correspondiente a 2048 por 1489 píxeles.
Cada exposición espectroscópica emplea una placa de metal circular grande y delgada que coloca las fibras ópticas a través de orificios perforados en las ubicaciones de las imágenes en el plano focal del telescopio. Estas fibras luego alimentan los espectrógrafos. Cada placa tiene un número de serie único, que se denomina placa en vistas como SpecObj en el CAS. El espectrógrafo SDSS usa fibras ópticas para direccionar la luz a travez de un plano focal desde objetos individuales hasta el slithead. A cada objeto se le asigna un fiberID [2]

Nombre	Tipo de dato	Descripción
Entradas		
objid	float64	Identificador del objeto observado
ra	float64	Right ascension (abreviado RA) es la distancia angular medida hace el este a lo largo del ecuador celestial desde el sol en el equinoccio de marzo hasta el círculo horario del punto sobre la tierra en cuestión
dec	float64	declinación (abreviado dec), medida que junto con ra generan coordenadas astronómicas que especifican la dirección de un punto en la esfera celeste (tradicionalmente llamado en inglés los cielos o el cielo) en el sistema de coordenadas ecuatoriales.
u	float64	mejor ajuste de magnitud DeV/Exp para la banda de telescopio u
g	float64	mejor ajuste de magnitud DeV/Exp para la banda de telescopio g
r	float64	mejor ajuste de magnitud DeV/Exp para la banda de telescopio r
i	float64	mejor ajuste de magnitud DeV/Exp para la banda de telescopio i
z	float64	mejor ajuste de magnitud DeV/Exp para la banda de telescopio z
run	int64	número de corrida de la muestra
rerun	int64	número que especifica cómo fue procesada la imagen tomada
camcol	int64	columna de la cámara, va de 1 a 6 que identifica la línea de exploración dentro de la ejecución.
field	int64	número de campo, generalmente comienza en 11 (después de un tiempo de aceleración inicial) y puede llegar a 800 para recorridos particularmente largos.
specobjid	float64	identificador del objeto registrado según el CAS (concentration, asymmetry, smoothness)
redshift	float64	Resultado del proceso físico de cuando la luz u otra radiación electromagnética de un objeto incrementa su longitud de onda.
plate	int64	número del plato
mjd	int64	MJD(Modified Julian Date) de la observación, es usado para indicar la fecha en la que fue tomada la muestra
fiberid	int64	fiber ID
Salida		
Class	string	nombre del tipo de cuerpo celeste (Galaxia, Estrella o Quasar)

Cuadro I
DESCRIPCIÓN DE CARACTERÍSTICAS

III. ARTÍCULOS RELACIONADOS

Existen algunos trabajos relacionados en los que se buscan clasificar cuerpos celestes como los siguientes:

III-A. Resolving the celestial classification using fine k-NN classifier

Sangeeta et al.[3], buscan solucionar el problema de clasificación de imágenes relacionado a cuerpos celestes, haciendo énfasis en la identificación de planetas, para esto se hace uso del KNN classifier, más específicamente el fine KNN, sin embargo también usan otras variantes de KNN y de otras técnicas como Support Vector Machine y Decision trees, se entrena los modelos con un conjunto de imagenes de planetas pasados en forma de matriz de numeros como se muestra el cuadro II y da como resultado el identificador de planeta, finalmente la validación es realizada con Cross Validation con particiones de 10, 20, 30 y 40 folds, también se realizó una validación con Hold out en porcentajes de 10 %, 20 %, 30 % , 40 %, 50 % para testing.

Nombre	Tipo de dato	Descripción
Entrada		
image	int64[255][255]	matriz que contiene las intensidades de color de una imagen de un planeta
Salida		
class	string	identificador del planeta

Cuadro II
RESOLVING THE CELESTIAL CLASSIFICATION USING FINE K-NN CLASSIFIER, CARACTERISTICAS

III-B. k-Nearest Neighbors for automated classification of celestial objects

Similar al artículo anterior Li L et al.[4], Realizan una implementación de la técnica de KNN para la clasificación de datos de rayos X de cuerpos celestes, en este caso buscan clasificar galaxias activa(AGN), estrellas y galaxias normales, las características usadas se muestran en el cuadro III, en este caso realizan una implementación tradicional de KNN variando la cantidad de vecinos de 2 a 17 y para la validación se utiliza Holt out con división de la data en 50

Nombre	Tipo de dato	Descripción
Entrada		
optical index	float64	Bandas de muestras opticas de rayos X
cr	float64	
ourcecount-rate in the broad energy band	float64	
hardness ratio 1	float64	
hardness ratio 2	float64	
source extent	float64	
likelihood of source extent	float64	
infrared index J-H	float64	
infrared index H-Ks	float64	
Salida		
class	string	galaxia, estrella o AGN

Cuadro III
K-NEAREST NEIGHBORS FOR AUTOMATED CLASSIFICATION OF CELESTIAL OBJECTS, CARACTERÍSTICAS

III-C. Development of accurate classification of heavenly bodies using novel machine learning

techniques

Wierzbiński. M et al [5], haciendo uso del conjunto de datos SDSS(el mismo del presente trabajo) abordan el mismo problema de clasificación que proponemos, clasificar las muestras de un telescopio en 3 grupos distintos, estrellas, galaxias y quasar, para esto utilizan varias técnicas como decision tree, Ada boost, KNN, SVM, logistic regression, etc.., en cada uno entrenando 2 veces, la primera con los valores predeterminados para cada caso y la segunda haciendo uso de algoritmos genéticos para determinar los parámetros óptimos para el modelo, en principio las características seleccionadas son las mismas descritas en el cuadro I y posteriormente utilizan PCA para reducir el numero de características dejando solo las 3 variables producidas por el algoritmo, ra, dcc y redshift; como estrategia de validación se utiliza cross-validation con 5 folds.

III-D. Study of Star/Galaxy Classification Based on the XGBoost Algorithm

Chao, L.tran[6], También utilizan el conjunto de datos SDSS, esta vez para la clasificación de estrellas y galaxias, la técnica empleada es la de XGBoost, pero de igual modo utilizan otras técnicas de aprendizaje alternas de entre las que destacan adaboost y gradient boosting decision tree(GBDT), el conjunto de datos es mismo representado en el cuadro I, igualmente se utiliza el como metodología de validación el cross validation, en este caso con 10 folds.

III-E. Resultados

Los resultados con el accuracy score de algunos de los modelos entrenados en los artículos anteriores se encuentran ponderados en el cuadroIV

IV. EXPERIMENTOS

La base de datos posee 10000 registros distribuidos en las clases como se señala en el cuadro V donde se puede apreciar que la clase QSO solo es el 8.5 % de los registros totales, mientras que el resto se distribuye de forma casi equitativa entre GALAXY y STAR, por lo cual se puede considerar que esta base de datos se encuentra desbalanceada, razón por la cual se toma como metodología de validación stratified kfolds donde se asegura mantener la distribución de datos en cada partición, con el propósito de obtener los mejores resultados en cada modelo se tendrán variaciones de la cantidad de particiones para la validación.

También se realiza un proceso de identificación y remoción de datos atípicos sobre los registros de las clases GALAXY y STAR, para identificar qué columnas poseían datos atípicos se hace usa del diagrama de caja con el cual se puede visualizar que las características ra, u, g, r, i, field, specobjid, redshift, plate y fiber id, tras lo cual se remueven descartando en total 1022 registros dejando 8978 en el conjunto para entrenar los modelos quedando al final con la distribución mostrada en el cuadro VI.

Como medida de desempeño se utilizará Matthews correlation coefficient(MCC) y Balanced Accuracy score (BAS), estas medidas fueron seleccionadas a causa de que la distribución de datos está desbalanceada, el MCC será la medida principal, esta consiste en tratar la clase verdadera y la clase predicha como variables y se calcula la correlación entre ellas, a mayor correlación mejores predicciones [7]; por otra lado el Balanced Accuracy es definido como el promedio de los recall por clase, además se va tener en cuenta el tiempo de ejecución medido en segundos.

V. MODELOS

Para este proyecto se plantea el entrenamiento de 5 tipos de modelos cada uno con diferentes configuraciones de parámetros y además de probar con diferentes distribuciones del método de validación que en este caso será stratified k folds, tanto la cantidad de folds como los diferentes parámetros de los modelos serán variados en búsqueda del mejor resultado en cada uno, los modelos y los resultados:

Articulo	Tecnica	accuracy			
		cross-validation (5 folds)	cross-validation (20 folds)	holt-out (10 %)	holt-out(50 %)
Resolving the celestial classification using fine k-NN classifier	Fine KNN		84,4	100	87,3
	Medium gaussian SMV		87,8	88,9	79,4
	complex tree		82,2	100	81
	Bagged trees		91,1	100	85,7
k-Nearest Neighbors for automated classification of celestial objects	KNN (K = 10)				97,73
Development of accurate classification of heavenly bodies using novel machine learning techniques	Votting	99,16			
	Random forest	99,11			
	svm	99,07			
Study of Star/Galaxy Classification Based on the XGBoost Algorithm	XGBoost	79,48			
	GBDT	77,64			
	Adaboost	77,56			

Cuadro IV
ACCURACY DE MODELOS PRESENTES EN ARTÍCULOS

Class	Registros	%
GALAXY	4998	49.98
STAR	4152	41.52
QSO	850	8.5

Cuadro V
DISTRIBUCIÓN DE REGISTROS POR CLASE

Class	Registros	%
GALAXY	4704	52.39
STAR	3424	38.14
QSO	850	9.47

Cuadro VI
DISTRIBUCIÓN DE REGISTROS POR CLASE SIN OUTLIERS

V-A. Análisis discriminante cuadrático (DCA)

Para este modelo se varía el parámetro priors entre 1 y 0.001, los 5 mejores resultados se pueden ver en el cuadro VII

train time	**bas**	**bas std**	**mcc**	**mcc std**	**priors**	**folds**
0.0125	0.5102	0.0912	0.4643	0.2417	1.000	5.0
0.0125	0.5102	0.0912	0.4643	0.2417	0.100	5.0
0.0125	0.5102	0.0912	0.4643	0.2417	0.010	5.0
0.0062	0.5102	0.0912	0.4643	0.2417	0.001	5.0
0.0104	0.4535	0.1089	0.3519	0.2414	1.000	6.0

Cuadro VII
RESULTADOS DCA

En este modelo especifico se puede ver que en realidad variar el valor de priors no afecta el rendimiento del modelos, los cambios mas notorios solo se notan al cambiar la cantidad de folds en la metodología de validación, por lo cual se puede tomar cualquiera de los valores definidos como mejor para el parámetro prior.

V-B. K-nearest neighborhood(KNN)

En este caso se varía el parámetro de n-neighborhoods con distintos valores entre 3 y 10 , los mejores resultados se encuentran en el cuadro VIII

train time	**bas**	**bas std**	**mcc**	**mcc std**	**n neighbors**	**folds**
0.0031	0.884	0.004317	0.8179	0.0126	5.0	5.0
0.0031	0.879	0.007478	0.8145	0.0176	7.0	5.0
0.0000	0.877	0.009204	0.8123	0.0220	7.0	6.0
0.0078	0.8788	0.010507	0.8116	0.0246	5.0	6.0
0.0000	0.8801	0.011827	0.8083	0.0216	3.0	6.0

Cuadro VIII
RESULTADOS KNN

De nueva cuenta el mayor impacto en el rendimiento es dado por los folds del metodo de validación, aun asi se puede notar que el usar 5 neighbors con 5 folds presenta los mejores resultados, aunque por muy poco, tras lo que vale la pena notar que hay resultados cuyo tiempo de entrenamiento fue tan reducido que no registro tiempo dentro de la magnitud dada, por lo cual en ultimas se toma como mejor parametro los 7 neighbors bajo la validación con 6 folds.

V-C. Gradient Boosting Tree(GBT)

Con gradient Boosting tree se varían los parámetros de n samples split y n estimators (cantidad de árboles), las 5 mejores combinaciones se visualiza el cuadro IX

train time	**bas**	**bas std**	**mcc**	**mcc std**	**trees**	**split sample**	**folds**
0.6437	0.9769	0.0041	0.9844	0.0028	10.0	5.0	5.0
0.6281	0.9769	0.0041	0.9844	0.0028	10.0	2.0	5.0
0.6588	0.9769	0.0055	0.9844	0.0038	10.0	2.0	6.0
0.6588	0.9768	0.0055	0.9842	0.0038	10.0	5.0	6.0
0.6614	0.9768	0.0055	0.9842	0.0038	10.0	3.0	6.0

Cuadro IX
RESULTADOS GBT

En general con 10 arboles se ve que se tienen los mejores resultados con muy poca diferencia entre los diferentes valores de split sample por lo cual la desición ultima se toma con el tiempo, dando como mejor modelo aquel con 10 arboles con 2 split samples bajo la validación de 5 folds

V-D. *Support vector machines(SVM)*

Con las SVM se varían 3 parámetros, el kernel entre linear y rbf, el gamma com 0.1 y 0.01, y el C entre 0.001 y 10, las 5 mejores combinaciones se encuentran en el cuadro X

train time	bas	bas std	mcc	mcc std	kernel	gamma	C	folds
0.7469	0.9853	0.0021	0.9825	0.0042	linear	0.01	10.0	5.0
0.7656	0.9853	0.0021	0.9825	0.0042	linear	0.10	10.0	5.0
0.3219	0.9792	0.0024	0.9720	0.0033	linear	0.01	1.0	5.0
0.3250	0.9792	0.0024	0.9720	0.0033	linear	0.10	1.0	5.0
0.3562	0.9716	0.004	0.9599	0.0057	rbf	0.01	10.0	5.0

Cuadro X
RESULTADOS SVM

en este caso los mejores resultados se tienen con el kernel linear, gamma = 0.01 y c=10, sin embargo al tener en cuenta el tiempo de ejecución se puede notar que con c = 1 este es reducido a menos de la mitad perdiendo al rededor de 0.1 en el resultado por lo cual en ultimas se tiene como mejores parámetros el kernel linear, c = 1, gamma = 0.01 bajo la validación con 5 folds.

V-E. *Artificial neural network (ANN)*

En ANN se plantea una red con una sola capa oculta, en este caso se varía los parámetros de epochs entre 3 y 5, y el las neuronas en la capa oculta entre 3 y 8, los 5 mejores resultados están en el cuadro XI

train time	bas	bas std	mcc	mcc std	neurons	epochs	folds
89.7461	0.96	0.0009	0.9486	0.0066	8.0	5.0	4.0
91.4844	0.9465	0.0082	0.9347	0.0112	5.0	5.0	4.0
58.8359	0.9515	0.0045	0.9247	0.0058	8.0	3.0	4.0
91.4609	0.9455	0.0090	0.9234	0.0178	3.0	5.0	4.0
54.9531	0.9423	0.0064	0.9182	0.0065	5.0	3.0	4.0

Cuadro XI
RESULTADOS ANN

Al compara en conjunto el tiempo de entrenamiento el resultado de la medidas de desempeño el modelo generado con 8 neuronas en la capa oculta y 3 epochs

VI. SELECCIÓN DE CARACTERÍSTICAS

VI-A. *Análisis de correlación*

Ahora se realiza un análisis de correlación entre las variables del cual se pueden ver los resultados en la figura 1 De esta se puede extraer que hay una alta correlación (definida como mayor a 0.75) entre diferentes variables, para ser mas especifico entre:

- g - u
- r - g
- i - g
- i - r
- z - g
- z - r
- z - i
- run -dec
- plate - specobjid
- mjd - specobjid

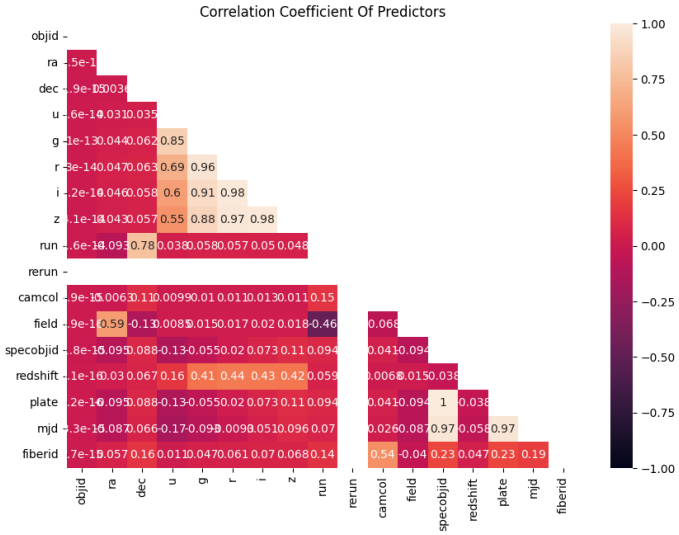

Figura 1. correlación de características

- mjd - plate

De este modo las características g, r, i, z, run, dec, plate, mjd y specobjid podrían ser consideradas para ser removidas (aclarando que no todo el conjunto)

VI-B. *Búsqueda secuencial descendente*

Utilizando de base los resultados del análisis de correlación se realiza una selección de característica utilizando el factor de inflación de la varianza (vif), el cual cuantifica que tan fuertemente esta relacionado una variable con el resto del conjunto, como criterio de remoción de los modelos, de esta forma se remueve la característica que presente el mayor valor de vif y se calcula nuevamente, este proceso se puede ver en los cuadros XII, XIII y XIV.

Variable	VIF
specobjid	1.360869e+09
plate	1.326817e+09
z	9.921808e-01
i	9.776206e-01
r	9.619112e-01
g	9.447992e-01
u	9.353706e-01
mjd	1.421353e-03

Cuadro XII
VALOR DE VIF PASO 1

Variable	VIF
r	110.441123
g	70.075669
z	47.243776
i	40.092877
u	10.587015
mjd	1.249808

Cuadro XIII
VALOR DE VIF PASO 2 Y 3

Finalizado el proceso se puede determinar como características candidatas a ser descartadas a g, r, i, plate y specobjid, con esto se alcanza un porcentaje de reducción de características de 29.4 %.

Variable	VIF
u	1.547196
z	1.516996
mjd	1.086585

Cuadro XIV
VALOR DE VIF PASO 3, 4 Y 5

REFERENCIAS

[1] "Sloan digital sky survey dr14 — kaggle."
[2] "Understanding sdss imaging data - sdss-iii."
[3] S. Yadav, A. Kaur, and N. S. Bhauryal, "Resolving the celestial classification using fine k-nn classifier," *2016 4th International Conference on Parallel, Distributed and Grid Computing, PDGC 2016*, pp. 714–719, 2016.
[4] L. Li, Y. Zhang, and Y. Zhao, "K-nearest neighbors for automated classification of celestial objects," *Science in China, Series G: Physics, Mechanics and Astronomy*, vol. 51, pp. 916–922, 7 2008.
[5] M. Wierzbiński, P. Pławiak, M. Hammad, and U. R. Acharya, "Development of accurate classification of heavenly bodies using novel machine learning techniques," *Soft Computing*, vol. 25, pp. 7213–7228, 5 2021.
[6] L. Chao, Z. Wen-hui, and L. Ji-ming, "Study of star/galaxy classification based on the xgboost algorithm," *Chinese Astronomy and Astrophysics*, vol. 43, pp. 539–548, 10 2019.
[7] "Matthews correlation coefficient is the best classification metric you've never heard of — by boaz shmueli — towards data science."