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, 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.

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, 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

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 parametros optimos para el modelo, 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, est 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), 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 cuadroII

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.

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 II
ACCURACY DE MODELOS PRESENTES EN ARTÍCULOS