Universidad de Guadalajara
Centro Universitario de Ciencias Exactas e Ingenierías

Practicas profesionales

Clasificacioń de imagenes astronomicas con machine learning

Alumno: Eduardo Tonathiu Piña Medina

Guía de uso del repositorio github

Índice

1. Introducción

La clasificación de galaxias es una tarea importante dentro de la astronomía; sin embargo, dada la gran cantidad de imágenes que existen, se deben buscar diferentes métodos que permitan clasificar dichos datos de una manera más rápida y eficiente.

Un enfoque es la implementación de algoritmos de *machine learning*, tanto algoritmos no supervisados como supervisados, siendo la diferencia que los no supervisados no requieren de un etiquetado en las imágenes de las galaxias.

Los datos de las imágenes fueron tomados del proyecto Galaxy Zoo 2 [2, 1].

Tanto los datos como el código pueden ser encontrados en el repositorio Classify galaxies with Unsupervised and Supervised learning.

2. Generación del Dataset

Para obtener los datos tanto de entrenamiento como de prueba, se creó un código encargado de obtener 10,000 imágenes para cada conjunto de datos.

Lo que hace el código 1_Dataset_generation es tomar el archivo gz2 filename_mapping y gz2 hart16 para encontrar las galaxias que están etiquetadas y pueden ser utilizadas. De ahí, se toman las 20,000 imágenes aleatoriamente y se descargan los *.jpg*s.

2.1. Preprocesamiento de imagen

El código 1_Image_Preprocessing se encarga de convertir las imágenes *.jpg* a formato *.h5*, pues este formato es más fácil de manejar. Además, se encarga de comprimir los archivos *.jpg* y *.h5* en archivos *.tar.gz* para reducir el espacio de almacenamiento.

2.2. Reducción de imagen

Las imágenes ya en formato *.h5* requieren mucho almacenamiento, haciendo imposible manejar toda la información en un entorno de Colab. Para que la información sea manejable, se hace un "recorte" en la imagen para centrar más las galaxias, manejando imágenes de 174×174 píxeles en RGB (aprox. 10^9 píxeles).

Lo que se hace en 2_Reduce_Dimension primeramente es obtener la información de las imágenes convertidas a *.h5*; esto se guarda en los archivos **log_data**, donde se almacena el ID de la galaxia, número de galaxia y clasificación de la galaxia.

Hay un inconveniente con la clasificación dada, y es que hay muchas clases (más de 800); por ello se hizo una simplificación de las clases, mostrada en el apéndice A.

Una vez simplificada la información de las clases, es necesario simplificar la información de las imágenes para que sea de un tamaño manejable. Para ello, además del "recorte", se entrenó un modelo IPCA para reducir la imagen de dimensión (174,174,3) a (75,). Se eligió este tamaño al considerar una imagen de dimensión (5,5,3)[1].

Una vez reducida, es requerido, para un buen funcionamiento de los algoritmos, que los valores no sean muy dispersos, motivo por el cual se implementa un modelo Scaler.

Tras reducir y escalar los datos de entrenamiento y prueba, estos fueron guardados con el formato *.pkl* y se encuentran disponibles para su uso. De igual manera, los modelos IPCA y Scaler están disponibles.

3. Aprendizaje no supervisado

Se aplicaron diferentes algoritmos no supervisados con la intención de observar el desempeño de cada uno. Los algoritmos empleados fueron:

- K-means
- Agrupamiento jerarquizado
- HDBSCAN
- Mapas auto-organizados
- Modelos de mezcla gaussiana

La forma en la que se explora el desempeño de los algoritmos sigue prácticamente el mismo enfoque. El primero de ellos son histogramas de clase para

[1] Un área de mejora puede ser recortar aún más las galaxias o usar una reducción menos drástica a 147 elementos por imagen (7,7,3)

cada *cluster*[2] para observar si alguna clase predomina en algún *cluster*. Una evaluación más global son los histogramas del clasificador, donde se muestra la frecuencia de cada *cluster*. Finalmente, se observan los centroides, lo que representa el punto central de cada *cluster*, la imagen promedio de todas las galaxias de un *cluster*.

3.1. K-menans

K-means es ejecutado en 1_Apply_K-Means. En este código es el único en el que se muestran cómo se ven las imágenes reducidas de las galaxias. Se usaron 12 *clusters* para hacerlo coincidir con las clases de GZ2. Los resultados están disponibles aquí.

3.2. Agrupamiento jerarquizado

El agrupamiento jerarquizado es ejecutado en 2_Apply_Hierarchical Clustering, haciendo uso de *linkage*. Un diferenciador de este método es el dendrograma, que permite saber la distancia que existe entre los diferentes *clusters* encontrados. Los resultados están disponibles aquí.

3.3. HDBSCAN

El HDBSCAN trabaja siguiendo una idea similar a la del agrupamiento jerarquizado. Es ejecutado en 3_Apply_HDBSCAN. Una característica de este algoritmo es que no permite controlar explícitamente el número de *clusters*, además de crear un *cluster* para manejar los datos que son considerados ruido. Este modelo tiene un gráfico similar al dendrograma, llamado árbol de mínima expansión, que indica también la similitud y cercanía entre los *clusters* encontrados. Los resultados están disponibles aquí.

3.4. Mapas auto-organizados

Los mapas auto-organizados son ejecutados en 4_Apply_Self_Organizing Maps. De igual manera, se siguieron empleando 12 *clusters*. Los resultados están disponibles aquí.

[2]Las clasificaciones de GZ2 son llamadas clases, mientras que la clasificación empleada por los algoritmos no supervisados se llamará *cluster*.

3.5. Modelos de mezcla gaussiana

Este modelo trabaja usando funciones gaussianas para cada *cluster*. El código es implementado en 5_Apply_Gaussian_Mixture_Models. Los resultados están disponibles aquí.

4. Aprendizaje supervisado

El aprendizaje no supervisado presentó diferentes dificultades. La principal es el desbalance tan grande que existe en los datos empleados. Para tratar de mejorar los resultados, se optó por un enfoque de clasificación supervisada.

4.1. CNN no balanceado

Se entrenó con el conjunto de datos desbalanceado 0_CNN para tenerlo como punto de partida. En términos generales, se tuvo un desempeño malo.

4.2. CNN balanceado

Para lograr el conjunto de datos balanceado, se tuvo que combinar los dos conjuntos de datos (entrenamiento y prueba) y seleccionar aleatoriamente 400 imágenes por cada clase. Sin embargo, existieron algunas clases con menos de 400 elementos. Para completar la cantidad requerida, se crearon sintéticamente imágenes similares, rotando mediante la rotación de la imagen[3]. Se optó por este enfoque debido a que los métodos comunes de aumento de datos no generaban bien las imágenes sintéticas.

La CNN es entrenada en 1_CNN_with_dataset_balanced[4]. Los modelos entrenados pueden ser encontrados aquí.

4.3. Pruebas de CNN

Las pruebas a las que fueron sometidos los modelos de CNN fueron las métricas estándar como la exactitud, pérdida, precisión, *recall* y *F1-score*,

[3]Un punto de mejora es cómo se crean las imágenes sintéticas

[4]Otro punto de mejora es implementar *k-fold cross validation* para mejorar los resultados.

así como el cálculo de las curvas ROC para cada clase y la matriz de confusión para cada modelo. En 2_Test_CNN se prueban los modelos de CNN, mientras que los resultados de ambos modelos están disponibles aquí.

Referencias

[1] Ross E Hart, Steven P Bamford, Kyle W Willett, Karen L Masters, Carolin Cardamone, Chris J Lintott, Robert J Mackay, Robert C Nichol, Christopher K Rosslowe, Brooke D Simmons, et al. Galaxy zoo: comparing the demographics of spiral arm number and a new method for correcting redshift bias. Monthly Notices of the Royal Astronomical Society, 461(4):3663–3682, 2016.

[2] Kyle W Willett, Chris J Lintott, Steven P Bamford, Karen L Masters, Brooke D Simmons, Kevin RV Casteels, Edward M Edmondson, Lucy F Fortson, Sugata Kaviraj, William C Keel, et al. Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. Monthly Notices of the Royal Astronomical Society, 435(4):2835–2860, 2013.

A. Abreviaciones

Tabla 1: Abbreviations used in Galaxy Zoo 2 and their meanings.

Abbreviation	Meaning
E	Elliptical galaxy
Er	Completely round elliptical galaxy
Ei	Elliptical galaxy with intermediate shape
Ec	Elongated elliptical galaxy (cigar shape)
S	Galaxy with disk characteristics
Ser	Edge-on disk with round bulge
Seb	Edge-on disk with square bulge
Sen	Edge-on disk without bulge
SB	Disk with central bar
Sd	Disk without prominent bulge
Sc	Disk with barely noticeable bulge

Abbreviation	Meaning
Sb	Disk with obvious bulge
Sa	Disk with dominant bulge
A	Object that is a star or an artifact
1t	Spiral structure with 1 tight arm
2t	Spiral structure with 2 tight arms
3t	Spiral structure with 3 tight arms
4t	Spiral structure with 4 tight arms
+t	Spiral structure with more than 4 tight arms
?t	Spiral structure with an indeterminate number of tight arms
1m	Spiral structure with 1 medium arm
2m	Spiral structure with 2 medium arms
3m	Spiral structure with 3 medium arms
4m	Spiral structure with 4 medium arms
+m	Spiral structure with more than 4 medium arms
?m	Spiral structure with an indeterminate number of medium arms
1l	Spiral structure with 1 loose arm
2l	Spiral structure with 2 loose arms
3l	Spiral structure with 3 loose arms
4l	Spiral structure with 4 loose arms
+l	Spiral structure with more than 4 loose arms
?l	Spiral structure with an indeterminate number of loose arms
(r)	Ring
(l)	Lens/arc
(d)	Disturbed galaxy
(i)	Irregular galaxy
(o)	Other unusual feature
(m)	Galaxy merger
(u)	Dust lane