STAT 679 Final Project Progress Report

Xiaoyang Wang Ziang Zeng

1 Astronomical Challange

Our research focuses on classifying celestial objects into stars, galaxies or quasars using their spectral characteristics. With the advancement of astronomical technology, we can obtain a large amount of data, including images and spectral information, from telescopes and large-scale photometry.

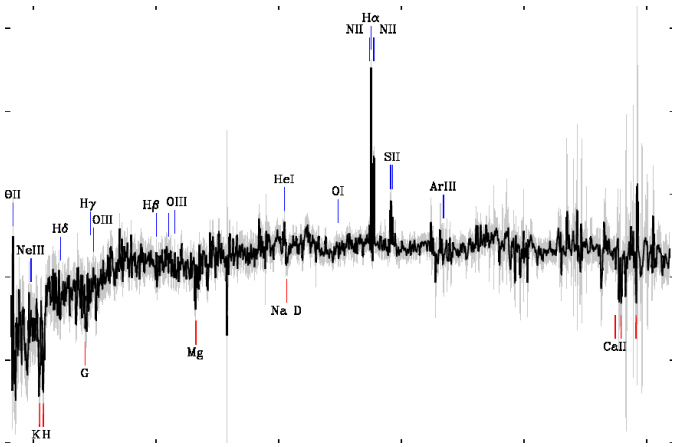
The central question of our research is: "How can we effectively use image and spectral data to accurately classify different types of stellar objects?"

2 Data

We plan to work with the astronomy dataset containing three types of data:

1. Image of the celestial objects:

2. Image of the spectra of the celestial objects.



3. Metadata of the celestial objects, whose variables are as follows:

vars	explanations
objid	Object Identifier
ra	Right Ascension angle (at J2000 epoch)
dec	Declination angle (at J2000 epoch)
u	Ultraviolet filter
g	Green filter
r	Red filter
i	Near Infrared filter
z	Infrared filter
run	Run Number
rerun	Rerun Number
camcol	Camera column
field	Field number
specobjid	Unique ID used for optical spectroscopic objects
class	Object class
redshift	Redshift value based on the increase in wavelength
plate	Plate
mjd	Modified Julian Date
fiberid	fiber ID
plateid	Plate ID

They can be found in this *link*. All the data is obtained from https://www.sdss.org.

3 EDA

First we will check and deal with missing values in the dataset. From figure 1 we can see that there are missing values for **i** and **z**. Then from figure 2, we can see that **i** and **z** are quite scattered. So, we use regression imputation to impute the missing values.

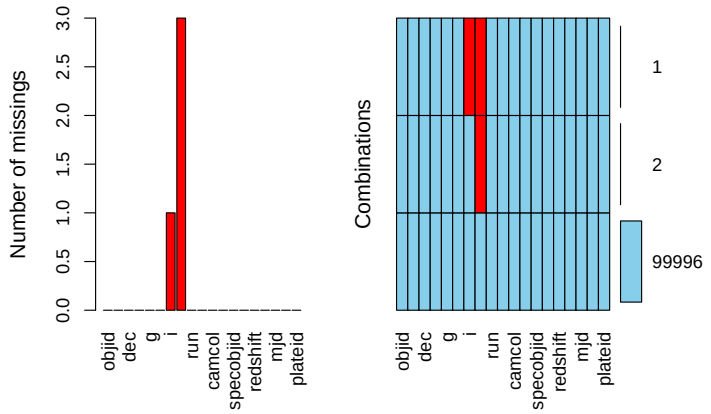


Figure 1: Missing Value Detection

Next, we give several visualizations to better understand the data. Figure 3 gives the distribution of classes, which are 33333 samples each, as selected from the website to make sure the data set is balanced.

Figure 4 gives the spread of stellar across the universe. The x-axis is right ascension angle (at J2000 epoch), and the y-axis is declination angle (at J2000 epoch). We can see some interesting patterns, but it is unclear for classification.

Figure 5 gives the correlationship between varibales. There are some variables that have strong correlationships.

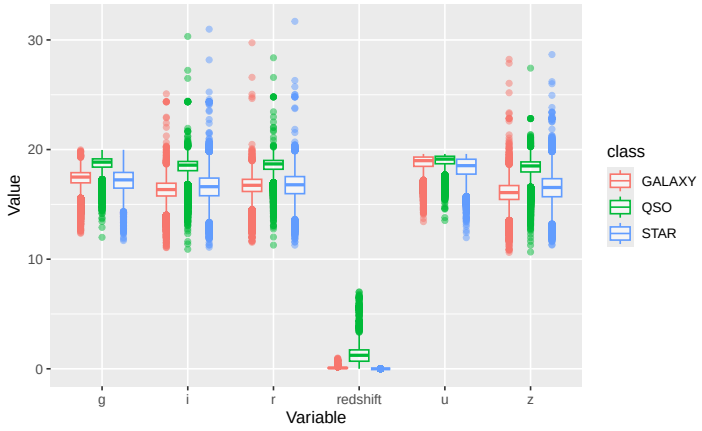


Figure 2: Boxplot

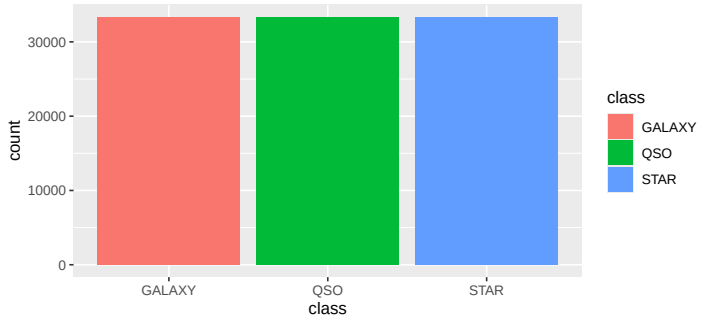


Figure 3: Distribution of Classes

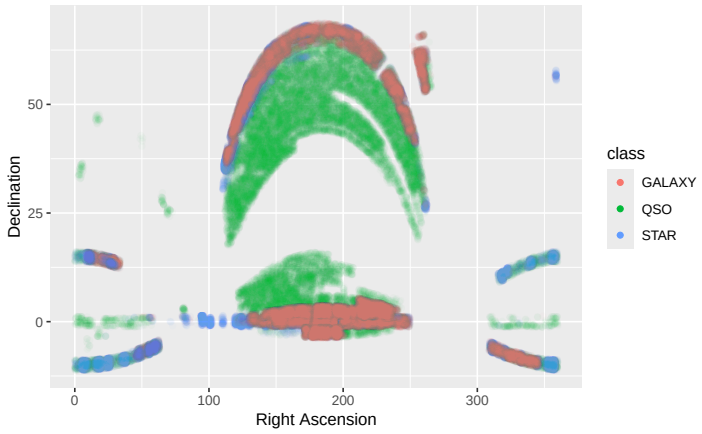


Figure 4: Spread of Stellars

4 Methods

We plan to employ both Deep Neural Networks (DNNs) and statistical models for object classification within the Sloan Digital Sky Survey (SDSS). For image classification, we have selected several DNN candidates based on their proven performance: ResNet, which supports very deep networks through its residual structure; VGG, which utilizes small-sized convolutional kernels (3x3) and pooling layers; and ResNeSt, which enhances the ResNet architecture by introducing a split-attention mechanism.

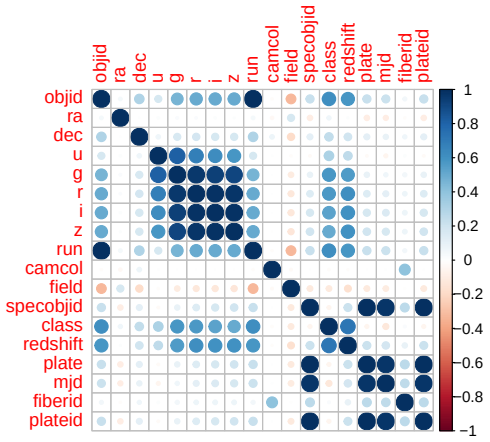


Figure 5: Correlation

Regarding statistical models, we intend to use either classification trees or logistic regression models, incorporating ensemble learning methods such as bagging or boosting to improve performance. Our ultimate goal is to develop a hybrid model that combines image classifiers with meta-classifiers, aiming for superior accuracy.

5 Results

6 Future Work

The next steps are clear. We will first finish the rest models: DNN for spectrum and classification model for metadata. Then, we will combine those three models with voting classifier to gain a better classification result.