Stellar Classification

Stars, Galaxies and Quasars

Aaron Gupta
SMCS
UPEI
Charlottetown, Canada
usgupta@upei.ca

***Abstract*—It is very important in Astrophysics to classify stellar objects. When astrophysicists use telescopes to spot a new object in the sky, they must try to place the object in one of the categories of known object types - such as Stars, Galaxies and Quasars. This paper attempts to conduct such a classification using Machine Learning to see if scientists can hand off this step to an algorithm and continue their research.**

***Index Terms*—astrophysics, classification, stellar objects**

I. Introduction

In this paper, K-Means Clustering and DBSCAN clustering were used to classify stellar objects based on the red and near-infrared photometric filter values and the redshift values. The red photometric filter values represent the frequency of light received from the stellar object in the red frequency, which is between 429-462THz. The near-infrared photometric filter values represent the frequency of light received from the stellar object in the near-infrared frequency, which is between 350-429THz. The redshift values represent the amount of shift in frequency of light, which relates to the distance of the object from the observatory. The further away the object, the more the redshift.

K-Means Clustering was run on the initial dataset since the number of clusters was known to be three - the first being Stars, the second being Galaxies and the third being Quasars. There was no scaling or normalization before K-Means, since K-Means does not require such preprocessing. To prepare for DBSCAN Clustering, a couple of preprocessing steps were taken. First, the data was scaled using SKLearn's standard scaler and normalized using SKLearn's normalize feature. Then, Principal Component Analysis with two components was run on the data to simplify it for DBSCAN. DBSCAN Clustering was then run on the dataset to obtain natural clusters.

The paper begins with a background on why the dataset was chosen and why the specific clustering algorithms were chosen. In section III, the distributions and relationships of features in the data are presented and analyzed. Section IV covers the results and analysis of K-Means Clustering and DBSCAN Clustering. Finally, section V has a conclusion and discussion of future research.

II. Background

When a telescope is pointed at a stellar object, the frequencies of bands of light are recorded by the computer. Light from different kinds of stars has different distributions of frequencies, each specific to the type, size, temperature, atmospheric composition and distance of the star. Similarly, information about the type, size and distance of galaxies and Quasars can be obtained. This information was present in the dataset chosen, hence the dataset was a good choice for the problem.

The important values to consider in my dataset were the red photometric filter, the near-infrared photometric filter and the redshift of each object. These values provide important information regarding the aforementioned properties of stellar objects, and as will be shown have distinct values over the different classes.

Since the problem in question is a labeled classification problem, K-Means and DBSCAN were the primary choices of algorithms. K-Means has proven to be quite effective when the number of clusters expected is known. DBSCAN provides the clusters which the data is naturally grouped into, which can coincide with the correct number of clusters needed.

III. The Data

A. Feature Distribution

The distribution of features as present were plotted in Fig. 1. The red and the near-infrared filter values show a similar distribution, whereas the redshift values follow their own. From this, not much was inferred.

B. Feature Relationships

The relationships between pairs of features were then plotted in Fig. 2.

- The red and the near-infrared filter values show close to a linear relationship, suggesting that these vary similarly across the dataset.
- The redshift and red filter values have a scattered relationship showing that the red filter values have a high variance when the redshift is low and a lower variance when the redshift is high.
- The redshift and near-infrared filter values show a similar relationship to the previous one, with the exception that the variance drops off quicker.

From these plots, it was inferred that objects closer to the observatory (which are usually stars) show high variance in properties and objects which are further away (which are usually galaxies) show low variance in properties. This coincides with the fact that galaxies are often quite similar to each other with respect to their images obtained on Earth.

C. Correlations between features

The correlation heatmap between features was plotted in Fig. 3. These correspond with the previous scatterplots and show the same relationships.

D. Feature-Label relationships

The distributions of each feature across the classes were then plotted in Figs. 4, 5, and 6.

- The red filter values have a lower mean for stars, a medium mean for galaxies and a higher mean for quasars. This means that stars have the lowest intensity of red light, followed by galaxies and quasars respectively.
- The near-infrared filter values follow a similar relationship, with the exception that the variance in means is lower - as noted in previous plots as well. This means that stars have the lowest intensity of near-infrared light, followed by galaxies and quasars respectively with smaller interquartile ranges.
- The redshift values show the highest difference between the classes, with stars having the lowest means (i.e the closest) and quasars having the highest means (i.e the furthest)

IV. K-Means and DBSCAN Clustering

The data was initially prepared by only keeping the necessary columns required for analysis, and no further preprocessing was done until DBSCAN.

A. K-Means

A K-Means clustering algorithm was set up with 3 clusters, 10000 maximum iterations and a random state of 42. This algorithm was then trained over the dataset and the obtained labels were plotted in Fig. 7.

The inertia obtained was 145612.22. The result is that K-Means is perhaps too simple an algorithm for this dataset.

B. DBSCAN

The data was then scaled, normalized and analyzed for 2 principal components. The resulting distribution is presented in Fig. 8. Further, the relationships between the components and classes were plotted in Figs. 9 and 10. These relationships were not subject to further analysis.

A DBSCAN clustering algorithm was set up with minimum distance 0.0375, 3 minimum samples and euclidean distance as the metric. This algorithm was then trained over the dataset and the obtained lables were plotted in Fig. 11.

There were clearly more than 3 clusters obtained. The result is that the data is not naturally grouped into 3 classes for the parameters chosen.

V. Conclusion and Further Research

The results in this paper show that satisfactory clustering was not obtained in either algorithm. This presents one of two possible conclusions: firstly, the features in the data are not the ones needed to correctly classify stellar objects in this manner; and secondly, the algorithms used or their parameters were not appropriate for the data.

Further research should include trying out different algorithms and tuning parameters appropriately and finding another dataset with the correct features for classification.

References

[1] fedesoriano. (January 2022). Stellar Classification Dataset - SDSS17. Retrieved January 15th, 2022 from https://www.kaggle.com/fedesoriano/stellar-classification-dataset-sdss17.

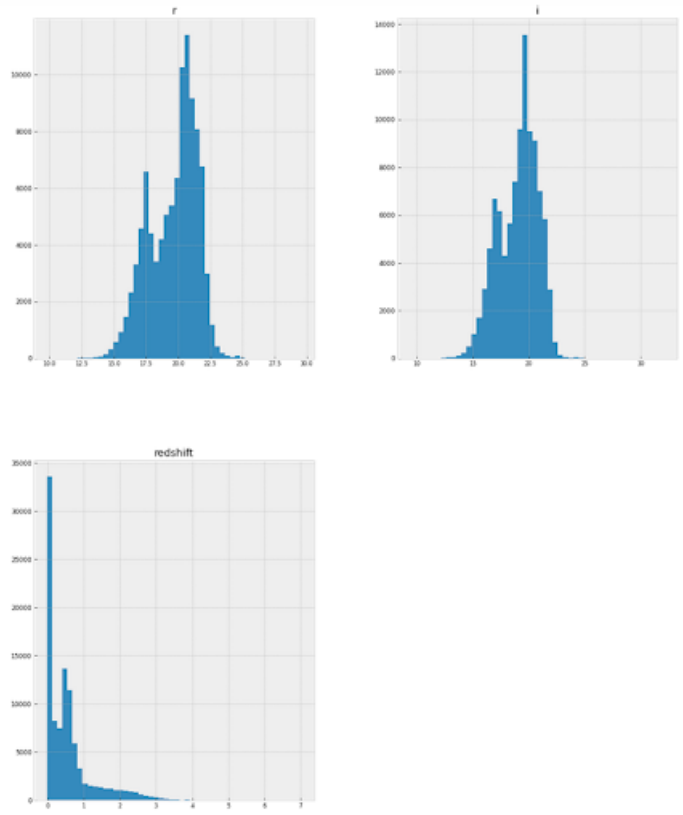

Fig. 1. Distribution of features.

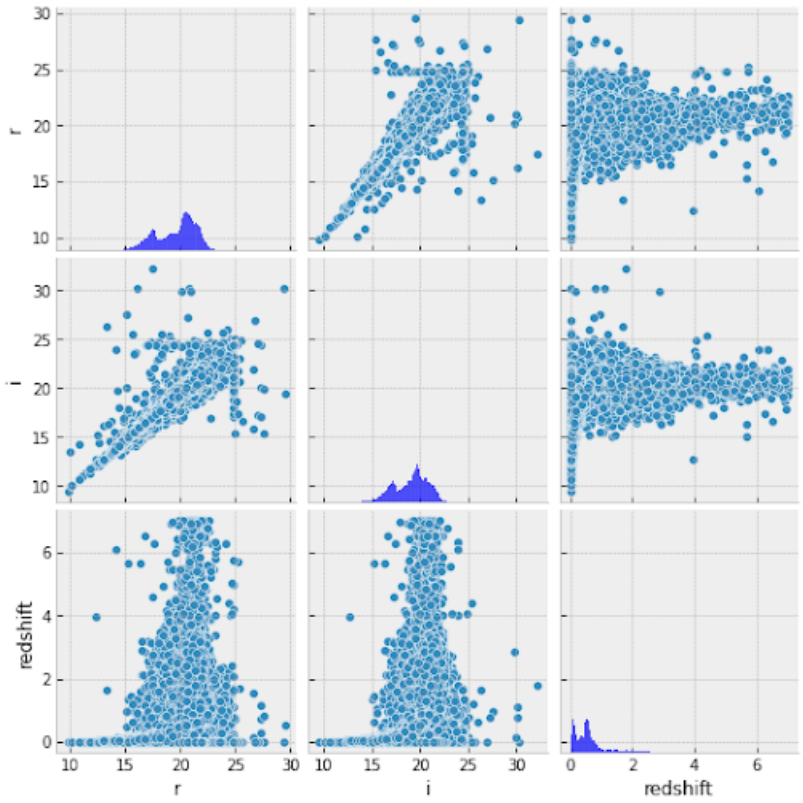

Fig. 2. Relationships between features.

	r	i	redshift
r	1.000000	0.962868	0.433241
i	0.962868	1.000000	0.492383
redshift	0.433241	0.492383	1.000000

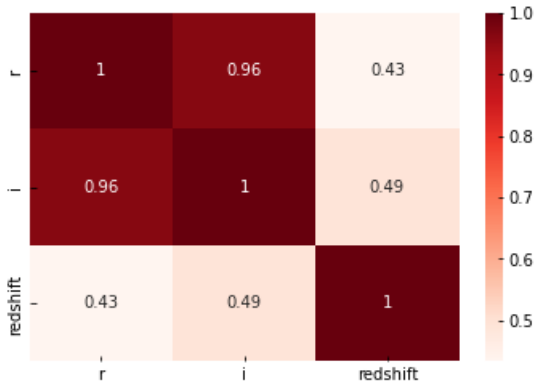

Fig. 3. Correlations between features.

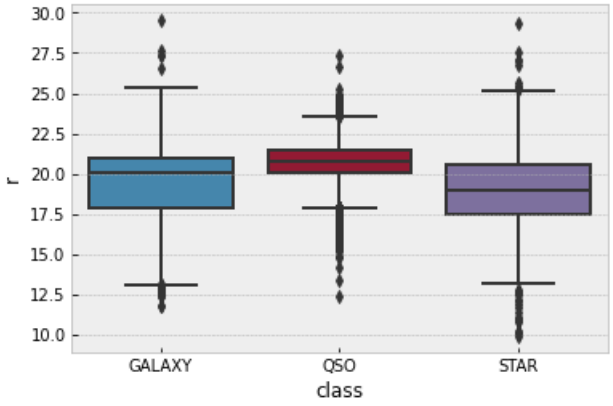

Fig. 4. Red Photometric Filter Values Across Classes.

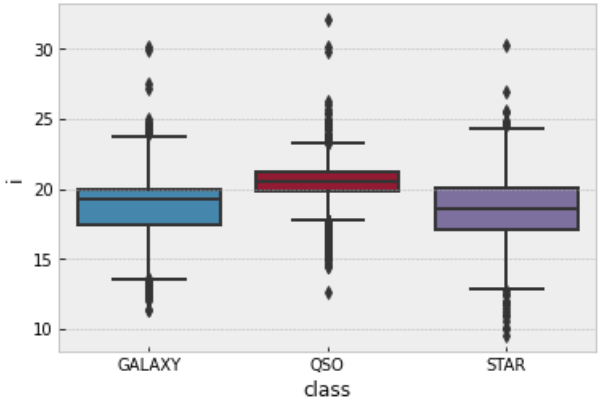

Fig. 5. Near-infrared Photometric Filter Values Across Classes.

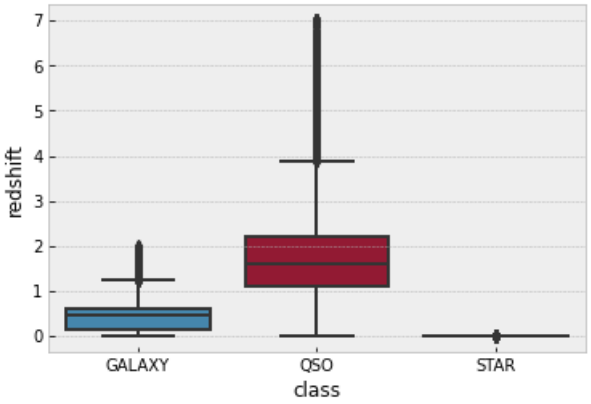

Fig. 6. Redshift Values Across Classes.

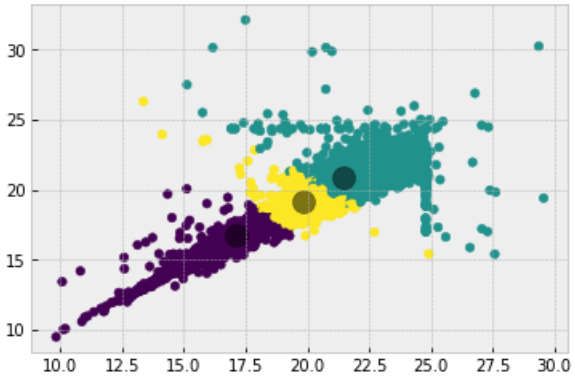

Fig. 7. K-Means Clusters.

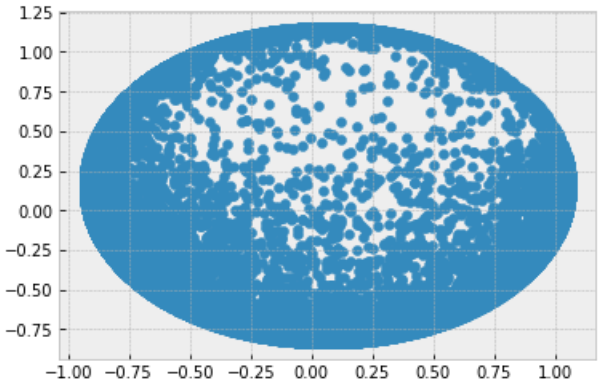

Fig. 8. Principal Component Analysis distribution.

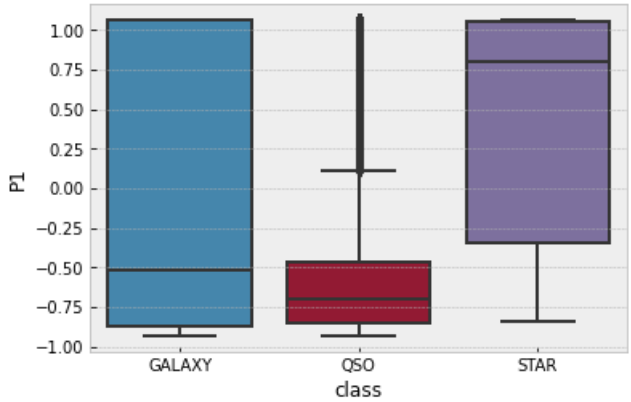

Fig. 9. Principal Component 1 Values Across Classes.

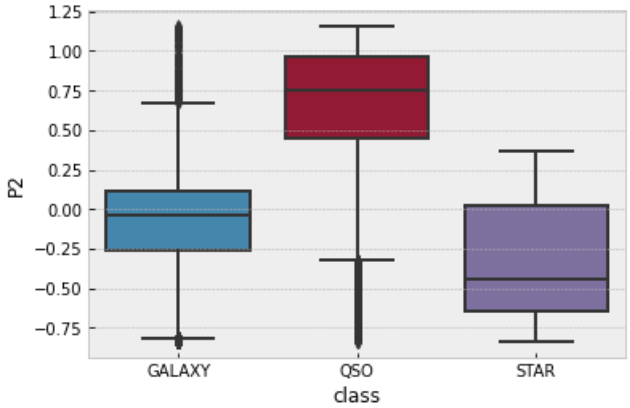

Fig. 10. Principal Component 2 Values Across Classes.

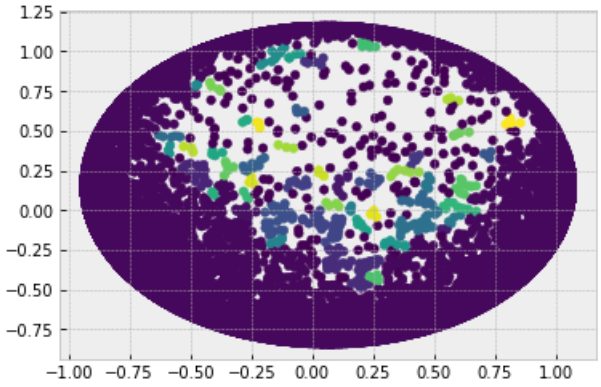

Fig. 11. DBSCAN Clusters.