Alex Millar and Roanna Rague

Final Project Proposal BMI Data Science MATH 4100

Optical Spectroscopic Object Classification

Motivating Questions

The definition of a star according to Oxford Lexico is: *a fixed luminous point in the night sky, which is a large, remote incandescent body*

With this definition in mind, the most fundamental building blocks of galaxies are stars. Stars help us understand the history and evolution of a galaxy, whether through age or distribution. The definition of 'Stellar Classification' is a scheme for assigning stars to types according to certain variables that are attributed to them, usually according to their temperature. (https://www.britannica.com/science/stellar-classification).

Among the astronomical objects that are discernible by optical/spectra data—primarily Solar Systems, Stars, Star constellations/clusters, Nebulae, Galaxies, Galaxy groups/clusters, and Black holes—those that are distinguishable as individual objects broadly fall into three categories: stars, galaxies, and active galactic nuclei (quasars). The classification/categorization of these astronomical objects are some of the most fundamental in astronomy.

Using the dataset: https://www.sdss.org/dr17/data_access/bulk and http://skyserver.sdss.org/dr17 which includes numerous data points, including right ascension angle, declination angle, ultraviolet filter, green filter, red filter, etc. With this dataset, we can construct a classification system to determine what type of stellar object we are working with. With this dataset, the data can be processed using dimensionality reduction, such as PCA/LDA, as well as cross-validated classification/clustering methods—such as SVM and K-means—to determine the proper classification of an astronomical object. Through our data analysis process, we will be able to answer the question, what astronomical object are we working with when presented with a set of astronomical data. Ultimately, being given certain data values and parameters our model/tool should predict the correct and accurate astronomical object class.

Hypothesis

What astronomical objects do we have based on the data we have been given?

1. If we can accurately determine what stellar object we are working with, what attributes truly determine what classifies the stellar object?
2. Are there other attributes that can cause false positives of our classification?
3. Can we be certain our data is accurate? If so, how do we prove this is the case?

4. Do any findings remain significant after including other independent variables?

Analysis Plan

- Acquire Optical and Spectra data to compile a data set, we will analyze three sources:
    - https://www.sdss.org/dr17/data_access/bulk
        - FITS files
    - http://skyserver.sdss.org/dr17
        - Web Scraping
    - https://skyserver.sdss.org/dr17/CrossMatchTools/ObjectCrossID
        - SQL Query
    - Steps to acquire the dataset:
        - Step 1: we will download and parse the relevant .fits files from the *sdss/dr17* website.
        - Step 2: if we are unable to properly parse the .fits files, we will then try to web-scrape optical/spectra data from the *sdss/dr17* website. Ideally, we will find HTML tables that we can find the necessary fields for, and then write them to one csv file.
        - Step 3: if we are unable to properly retrieve data from web scraping, we will run a SQL query on the *sdss/dr17/CrossMatchID* tool.
    - Once we have our data, we will reshape and clean data to validate we are working with a clean dataset.
    - Check shape
        - Reshape as needed
    - Check nulls/nans/0s/-9999s
        - Determine whether to impute or throw out
    - Check category distribution for Class Imbalance
        - Mitigate class imbalance as needed (during analysis)
- Explore data
    - Check for Univariate Outliers
        - Determine whether to throw out to keep or mitigate
    - Check for Multivariate Outliers
        - Determine whether to throw out to keep or mitigate
    - Explore Features
        - Visualize data
        - Identify
            - Patterns
            - Defining characteristics of classes
        - Determine criteria for Feature Selection (during analysis)
    - Select Classifier(s) most appropriate for data
- Analyze data
    - Dimensionality Reduction
        - Principle Component Analysis
        - Linear Discriminant Analysis

        - Clustering
            - K-Means
        - Classification
            - Methods
                - K-Nearest Neighbor
                - Random Forest
                - Support Vector Machine (SVM)
            - Cross-Validation
            - Train-Test-Split
            - Review/Visualize Results/Metrics
                - Accuracy/Precision
                - Confusion Matrix
                - Classification Report
                - ROC Curve/AUC
                - Class Prediction Error
            - Compare Models
            - Present our findings
- Explore ethical considerations

Project Timeline/Schedule

Delegation of tasks are assigned in our MS Teams Task board.

- Week 1 (March 14th-18th): Gather data -- acquire, parse and clean data, project proposal, validate we can use this dataset.
- Week 2 (March 21st-25th): Explore data – checking for outliers, exploring features, selecting classifiers
- Week 3 (March 28th – April 1st): Analyze data – Dimensionality reduction, clustering and classification. Model our data, do any post hoc tests, train-test our data
- Week 4 (April 5): Final testing, quality control and Project Milestone Report
- April 22 Final Project Report

References

1. SkyServer. N/A. **SkyServer**. Retrieved [March 3rd, 2022] from https://skyserver.sdss.org/dr17/
2. SkyServer. N/A. **Object CrossID**. Retrieved [March 18th, 2022] from https://skyserver.sdss.org/dr17/CrossMatchTools/ObjectCrossID.
3. SkyServer.N/A. **Bulk Data Downloads.** Retrieved[March 3rd, 2022]. https://www.sdss.org/dr17/data_access/bulk
4. Lexico.2022. **Lexico powered by Oxford**. Retrieved [March 18th, 2022]. https://www.lexico.com/.
5. Dana Bolles: NASA. N/A. **Stars**. Retrieved[March 3rd, 2022]. https://science.nasa.gov/astrophysics/focus-areas/how-do-stars-form-and-evolve.

6. Stellar Classifcation. May 14, 2013. **Stellar Classification**. Retrieved [March 3rd, 2022]. https://www.britannica.com/science/stellar-classification