Galaxy Morphological Classification with Big Data: An Analysis of Large Spectral and Photometric Datasets*

Extended Abstract

Son Pham†
CU Boulder
Boulder, CO
son.pham-2@colorado.edu

ABSTRACT

CCS CONCEPTS

• **Mathematics of computing** → *Mathematical analysis*; • **Information systems** → *Data management systems*; • **Computing methodologies** → *Machine learning*; *Distributed computing methodologies*; • **Applied computing** → *Astronomy*.

KEYWORDS

ACM proceedings

ACM Reference Format:
Son Pham. 2022. Galaxy Morphological Classification with Big Data: An Analysis of Large Spectral and Photometric Datasets: Extended Abstract. In *Proceedings of CSCI 6502 - Big Data Analytics (Spring '22)*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.475/123_4

*Produces the permission block, and copyright information
†Note

Spring '22, March 2022, Boulder , CO, USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

1 INTRODUCTION

Big data analytics is transforming the way scientists and astronomers study the universe. With each large scale telescope on the ground or in space, dozens of gigabytes of data is being generated per day - an amount that will take an increasingly amount of time to explore and process. For example, the Hubble Space Telescope generates about 120G of scientific data every week [2]. Sky surveys conducted by telescopes such as the Sloan Digital Sky Survey (SDSS) produces data releases each year that can be as high as 100's of TBs. SDSS provides a wide range of data types, such as optical spectra, infrared spectra, and imaging.
Recent discoveries, such as the presence of over 100 black holes in the center of our Milky Way, was realized using data from decades ago generated by the Chandra satellite. Scientific and technological advancement in combination led the way to this capability. With the increase in computational performance, astronomers now have the capabilities to explore these large datasets without the need to invest in large ground-based optical telescopes or work in a research lab. As big data analytics continue to grow, these discoveries will become more common as scientists continue to collect and process more of the data that is available. As technological advancements grow, they will have more readily available tools and platforms for their analysis. With the cummilation telescopes and technologies present today, the entire electromagnetic spectrum can be observed within a patch of sky of interest.
SDSS datasets, such provide many different data types, provides the opportunity to explore galaxies, stars, and

Parameter	Description	Unit
Total unique area covered	14,555	sq. deg
Total area of imaging (including overlaps)	31,637	sq. deg
Individual image field size	1361x2048 (0.0337)	pixels (sq. deg)
catalog objects	1,231,051,050	(-)
unique detections	932,891,133	(-)
Median PSF FWHM, r-band	1.3	arcsec
Pixel scale	0.396	arcsec

quasars too distant to view for amateur astronomers and conduct many different types of analysis on them. The analytics of interest to the authors include the use of optical imaging and spectral data. Legacy imaging generated from prior SDSS programs, if used with machine learning classification techniques, can provide automatic classification of morphological properties of these galaxies. Quatities measured from these images and spectra readings can also provide information such as magnitudes, redshifts, and object classifications. In particular, redshift can be a good indicator of galaxy morphology.

This paper will go over in detail the approach to perform morphological classification of galaxies using these available information from various sources. The main source of interest for the author is SDSS's data, which includes 100's of terabytes of data covering more than one-third of the entire celestial sphere [1].

2 LITERATURE SURVEY

2.1 External Studies

2.2 Studies Based on This Dataset

3 DATASETS

SDSS Imaging

4 APPROACH

4.1 Data Cleaning and Pre-Processing

4.2 Data Integration

4.3 Statistical Analysis

One of the major issues in working with large datasets is that the standard analytical evaluation of statistical formulas requires an entire batch of data samples to be stored in memory for computation. Having 100's of thousands of samples can quickly exceed the memory limit of an average laptop or computer. This is especially the case when running analysis on embedded systems where memory is a driving constraint of performance. Stream processing allows a user to run calculations as the data comes in and releases memory that held samples of data from previous use.

Welford's method for statistical analysis allows the use of stream processing to compute some statistical parameters [3]. This method gives accurate estimates of the mean and variance without having to store all the data in memory. The standard process for computing standard deviation is to compute the mean of the data in one pass, then calculate the square deviation of values from the mean in the second pass. In crude methods of numerically computing deviation and means, one can compute the same standard deviation in one pass. Equation 1 shows this method by accumulating the sums of x_i and x_i^2. This subtraction can result in loss of accuracy if the square of the mean is large while the variance is small.

$$\sigma = \sqrt{(n \sum_{1 \le i \le n} x_i^2 - (\sum_{1 \le i \le n} x_i)^2)/n(n-1)} \tag{1}$$

Welford's method simply keeps a running sum of the data, number of samples, and deviation from data collected so far, and the user can view the sample variance and mean at anytime during the computational process to view their progression. Equation 2 show the recurrence formula from Welford which takes into account only the current and previous sample values. Equation 3 shows the computation of the mean, and Equation 4 for the standard deviation.

$$M_{2,n} = M_{2,n-1} + (x_n - \bar{x}_{n-1})(x_n - \bar{x}_n) \tag{2}$$

$$\sigma_n^2 = \frac{M_{2,n}}{n} \tag{3}$$

$$s_n^2 = \frac{M_{2,n}}{n-1} \tag{4}$$

4.4 Imaging

4.5 Spectroscopy

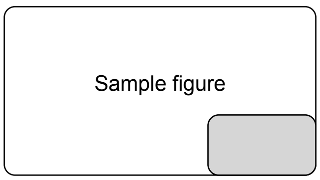

Figure 1: Sample figure

4.5.1 *Subsubsection.*

Paragraph. Nulla

5 EVALUATION

6 RESULTS

7 APPLICATIONS

8 CONCLUSION

REFERENCES

[1] Sloan Digital Sky Survey. 2022. Data Release 17 Scope. (2022). https://www.sdss.org/dr17/scope/

[2] Nola Taylor Tillman. 2022. Hubble Space Telescope: Pictures, facts and history. (2022).

[3] B. P. Welford. 1962. Note on a Method for Calculating Corrected Sums of Squares and Products. *J. ACM* 4, 3 (Aug. 1962). https://www.jstor.org/stable/1266577