PHOTOMETRIC REDSHIFT AND THE LARGE SCALE DISTRIBUTION OF BROAD GALAXY MORPHOLOGIES

Nicholas Paul, Nick Virag, Lior Shamir

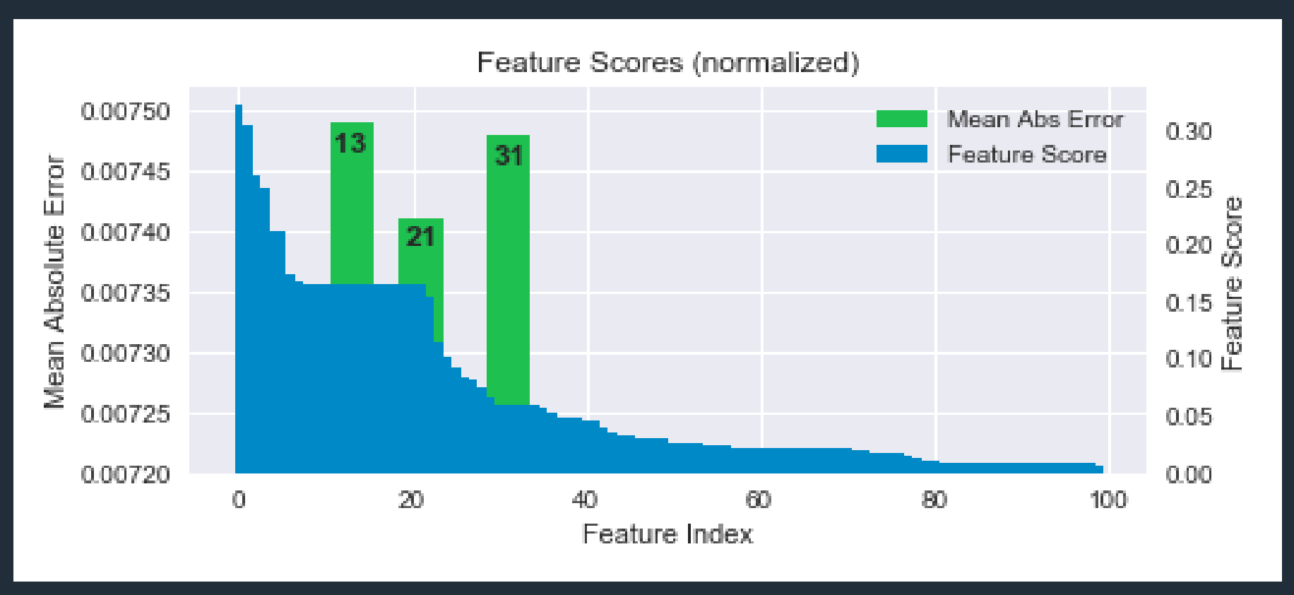
College of Arts and Sciences; Mathematics and Computer Science Department

Morphology carries important information about the physical characteristics of a galaxy. Here we used machine learning to produce a catalog of ~3,000,000 SDSS galaxies classified by their broad morphology into spiral and elliptical galaxies. Comparison of the catalog to Galaxy Zoo shows that the catalog contains a subset of $1.7*10^6$ galaxies classified with the same level of consistency as the debiased “superclean” sub-sample. In addition to the morphology, we also computed the photometric redshifts of the galaxies. Several pattern recognition algorithms and variable selection strategies were tested, and the best accuracy of mean absolute error of ~0.0062 was achieved by using random forest with a combination of manually and automatically selected variables. The catalog shows that for redshift lower than 0.085 galaxies that visually look spiral become more prevalent as the redshift gets higher. For redshift greater than 0.085 galaxies that visually look elliptical become more prevalent.

ALGORITHM

VARIABLE SELECTION

A dataset of 20,000 SDSS galaxies with spectroscopic redshift taken was used to train the algorithm. The photometric information was taken from the PhotoObjAll table of SDSS DR8. Several methods were used for variable selection, including hand-crafted variable selection and automatic statistical selection of variables.



Feature scores for selecting K features (blue) and MAE for using K features as the training set (green).

PERFORMANCE EVALUATION

In order to evaluate the performance of each of the feature sets, we used the Mean Absolute Error, the Root Absolute Error, and the normalized Z value, Z_{norm}. The Z_{norm} statistic is used in the Brescia et al. [2014] paper and is defined by the following equation:

$$Z_{norm} = \frac{Z_{spec} - Z_{phot}}{1 + Z_{spec}}$$

Algorithm	Mean Abs. Error	Root Abs. Error	Z_{norm}
Simple Linear Regression	0.00621	0.01321	0.01273
MultiLayer Perceptron	0.00617	0.01313	0.01041
M5P	0.00611	0.01288	0.00562
ZeroR	0.00616	0.01301	0.06581
Decision Table	0.00616	0.01301	0.01586
Random Forest	0.00617	**0.01294**	**0.00222**

The simple linear regression algorithm predicts a multi-dimensional point by minimizing its squared error. MultiLayer Perceptron builds a multi-layer neural network of weighted perceptron nodes. The M5P algorithm implements M5 model trees and rules [Quinlan et al., 1992]. The rule-based ZeroR and Decision Table both use frequency tables to make a prediction. The tree-based Random Forest algorithm [Breiman, 2001] builds a classifier using a large number of random tree classifiers.

Training Set Size	Mean Abs. Error	Root Abs. Error	Z_{norm}
1000	0.00836	0.02027	0.00266
5000	0.00646	0.01446	0.00219
10000	0.00626	0.01342	0.00221
20000	0.00617	0.01294	0.00222

The accuracy of the prediction, measured using mean absolute error, root absolute error, and the Z_{norm} evaluation metric used in Brescia et a. [2014] increases as a function of the training set size. At a certain threshhold , ~15,000 samples, the increase in accuracy is no longer significant.

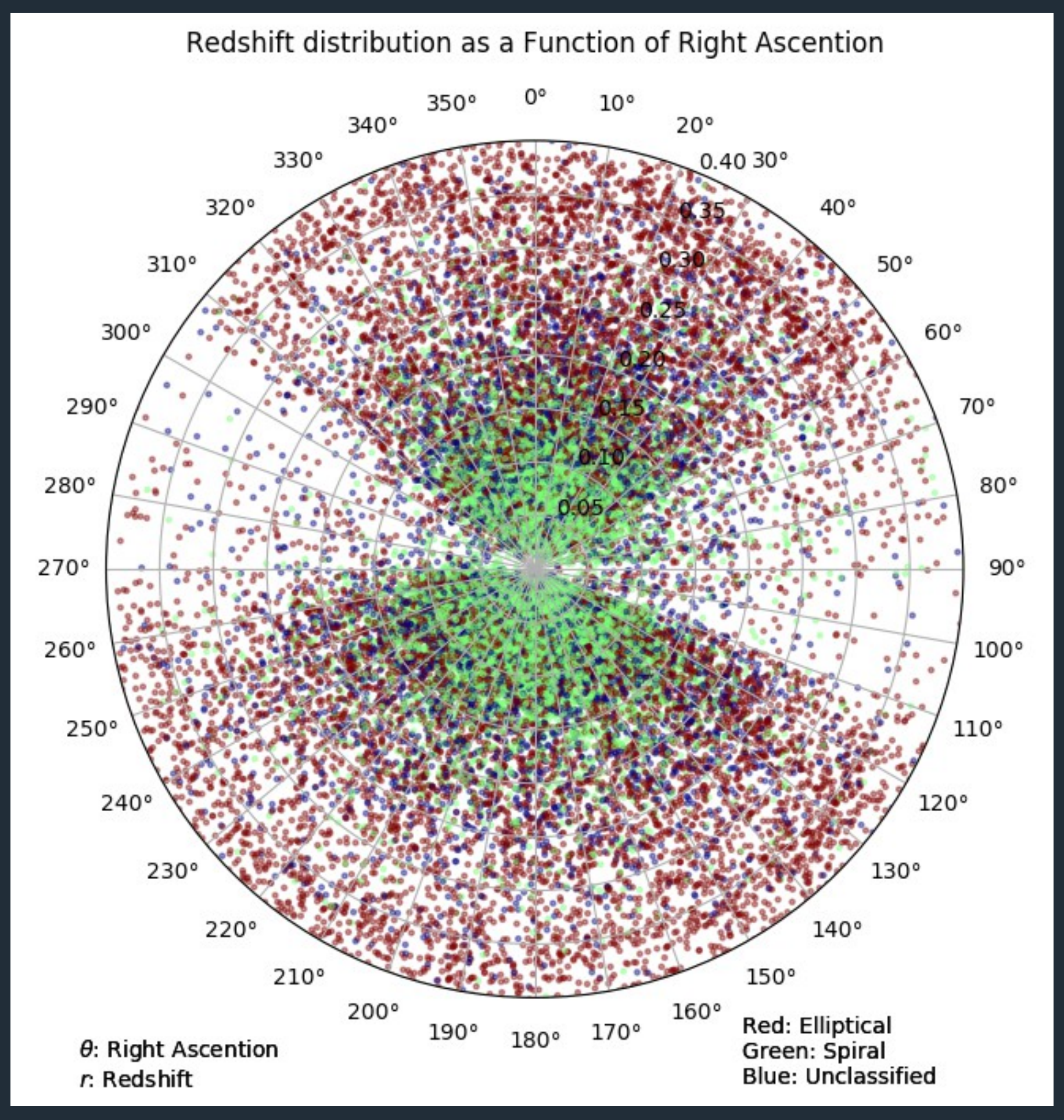
CATALOG

SOURCE

Data is sourced from the Sloan Digital Sky Survey (SDSS) DR 8. 2,912,341 galaxies are included, which have magnitude i<18, Petrosian radius larger than 5.5’’x, and are classified as spiral or elliptical.

SCOPE & SIZE

The catalog contains nearly 3 million samples collected from the Sloan Digital Sky Survey data release 8. It is the largest existing catalog of extragalactic objects classified by redshift values. The catalog also contains the certainty of each galaxy to belong in the broad morphological classes, elliptical and spiral [Kuminski and Shamir, 2016]. The previous largest existing catalog containing such information was created and published by Galaxy Zoo [Lintott et al. 2011]. The presented catalog contains roughly ~900,000 spiral galaxies and ~600,000 elliptical galaxies with consistency 98% with Galaxy Zoo.



There is a significant concentration of spiral (green) galaxies in the center of the plot. Moving outward from the plot’s center, spiral galaxy density increases and elliptical (red) galaxy density decreases. Unclassified (blue) galaxies are spread evenly throughout.

FEATURES

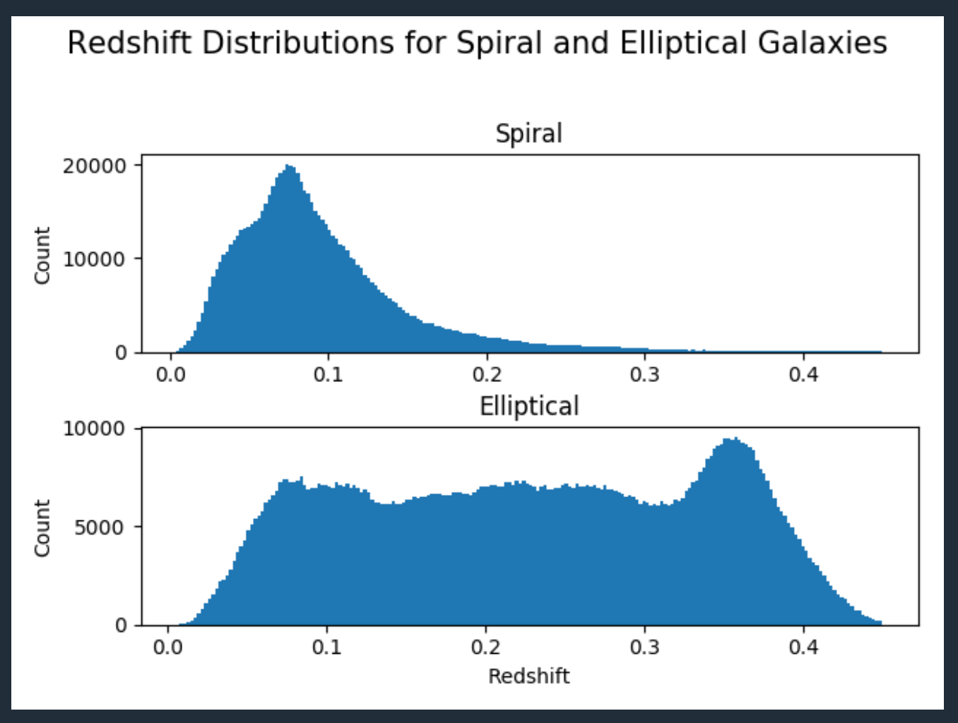
The following table lists the features contained in the final catalog. The features unique to our catalog are marked with an asterisk.

Feature	Description
dr8id	SDSS DR 8 object id
ra	Right ascension
dec	Declination
elliptical*	Probability that the object is an elliptical galaxy
spiral*	Probability that the object is a spiral galaxy
zphot*	Photometric redshift
zerr*	Estimated photometric redshift error

MORPHOLOGIES

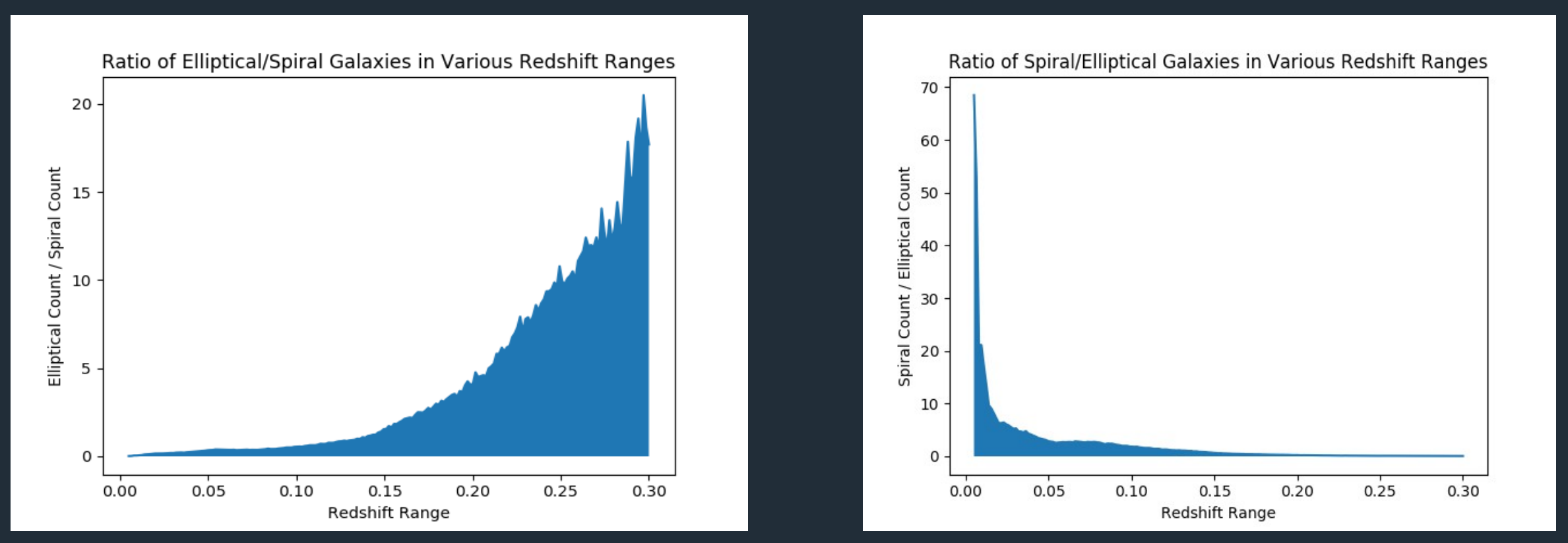
DISTRIBUTION

For spiral galaxies, redshift values peak at approximately 0.08 and quickly drop off after 0.1. Elliptical galaxies peak at approximately 0.6, plateau until approximately 0.33 and peak again at 0.35. Spiral galaxies tend to have a mush smaller and significantly more concentrated redshift range while elliptical galaxies are spread out evenly from 0.02 to 0.43.

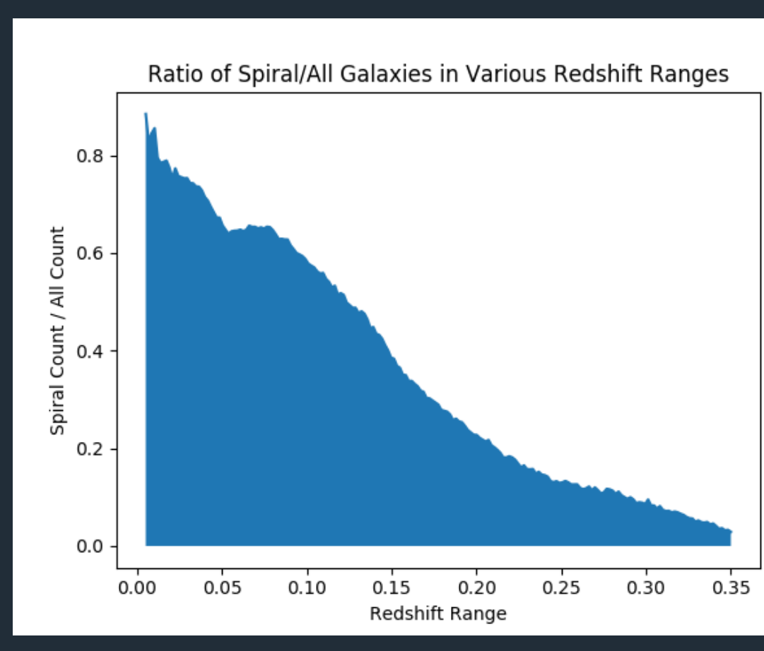


Frequency distribution of spiral and elliptical galaxies.

MORPHOLOGY RATIOS



Frequency distribution of spiral and elliptical galaxies.



Ratio of spiral galaxies over all galaxies. There is a small plateau occurring from approximately 0.05 to 0.07.

SUMMARY

Using our automatic and manually selected feature set in conjunction with the random forest machine learning algorithm, we have created a solution that works for much broader redshift ranges than previously tested by sources such as Collister et al. [2007]’s range of $0.4 < z < 0.7$, with a Z_{norm} value of 2^{-3}, compared to 3×10^{-5} achieved by Brescia et al. [2014]. These photometric redshift results provide a less expensive, less technically daunting alternative to measuring redshift stereoscopically. Photometric redshift estimation is often sufficient for many applications involving analysis of large populations of celestial objects [Oyaizu et al., 2008], a task that would be much more difficult if attempted through pure stereoscopic means.

References:
M Brescia, S Cavuoti, G Longo, and V De Stefano. A catalogue of photometric redshifts for the sdss-dr9 galaxies. Astronomy & Astrophysics, 568:A126, 2014.
Leo Breiman. Random forests. Machine Learning, 45(1):5-32, 2001.
John R Quinlan et al. Learning with continuous classes. In 5th Australian joint conference on artifcial intelligence, volume 92, pages 343-348. Singapore, 1992.
Evan Kuminski and Lior Shamir. A computer-generated visual morphology catalog of 3,000,000 sdss galaxies. The Astrophysical Journal Supplement Series, 223(2):20, 2016.
Chris Lintott, et al. Galaxy zoo 1: data release of morphological classifcations for nearly 900 000 galaxies. Monthly Notices of the Royal Astronomical Society, 410(1):166-178, 2011.
Hiroaki Oyaizu, et al. A galaxy photometric redshift catalog for the sloan digital sky survey release 6. The Astrophysical Journal, 674(2):768, 2008.
Adrian Collister, et al. Megaz-lrg: a photometric redshift catalog of one million sdss luminous red galaxies. Monthly Notices of the Royal Astronomical Society, 375(1):68-76, 2007.

IIS-1546079