Galaxy Morphological Classification

Jordan Duprey and James Kolano

Abstract

To solve the issue of galaxy morphological classification according to a classification scheme modelled off of the Hubble Sequence, we implement a pipeline of various machine learning algorithms including Support Vector Machines (SVM's) and Gaussian Discriminant Analysis (GDA). We derived a set of features that we felt would best distinguish between the five classes of galaxies and used a forward search to find the optimal subset. Ultimately we found that the SVM performed significantly better than the GDA model at every step in the pipeline.

1 Introduction

In 1936, Edward Hubble developed a classification scheme to divide galaxies into distinct categories based upon morphological features. This system, which would eventually become known as the Hubble Sequence, consists of 4 major categories of galaxies: Elliptical, Lenticular, Barred Spiral, and Spiral. As technology has advanced and the sheer volume of data produced by telescopes have increased, a large number of galaxies have remained unclassified, creating a bottleneck in astronomical research. To address this issue, projects like Galaxy Zoo have attempted to crowdsource the classification process; however, this introduces a high degree of human error since the average participant in these projects does not have a background in astronomy. Moreover, some of the features depicted are virtually impossible to detect insofar as the image quality and angle at which the pictures can be taken are limited. To solve this problem, we have implemented various machine learning algorithms using a pipeline approach to include only the most relevant features at each step in the classification process. A few papers have been published on this topic, using various features and machine learning techniques. Our implementation looks to build off of these implementations and test a number of new parameters. Additionally, we are looking to test the efficacy of Gaussian Discriminant Analysis, which has to our knowledge not been applied to this problem.

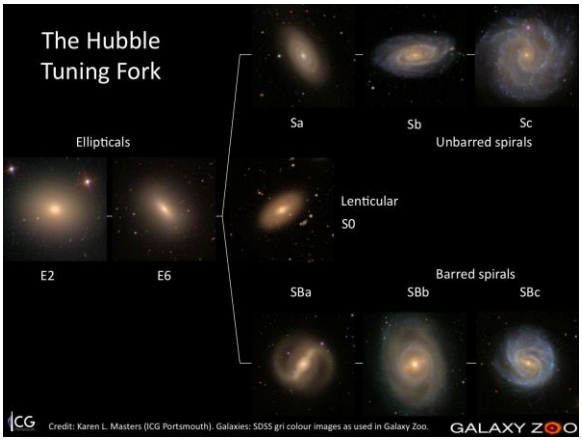

Figure 1. This depicts a few classic example galaxies in the four major categories of the Hubble Sequence

2 Dataset

The dataset consists of 334 images of galaxies along with their correct classification in the Hubble Classification Scheme. The images were found by searching Nasa's High Energy Astrophysics Science Archive Research Center (HEASARC) for galaxies that were given a New General Catalogue identification or Index Catalogue identification [1]. This qualification was included because in order to be given an identification in these catalogues, these galaxies were sure to have at least relatively clear

pictures and have a definite classification. The first 334 galaxies from the resulting list were selected to be our data. The images were obtained by following the links from HEASARC to Digitized Sky Survey images of each galaxy provided by Skyview. The correct classifications for these galaxies were then discovered by following the link to the Strasbourg Astronomical Data Center [2] [3].

3 Preprocessing and Features

Before we extracted any features from the images, we preprocessed the images to normalize our training and test data. First, we filtered out all the background noise, primarily additional stars, by and blacking out everything outside of the polygon defined by the contour with the largest arc length in the picture. Once the stars had been removed, the picture was centered on the cluster of pixels with the highest mean intensity in the middle fifth column and row of the picture, which correlated with the center of the galaxy. Lastly the color images were converted to greyscale and black and white.

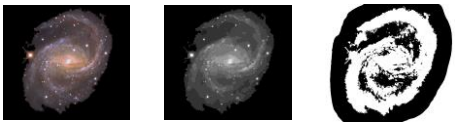

Figure 2. This depicts the chain of events that occur in the preprocessing stage.

For both of our models we relied on a set of 9 image features that took advantage of the three major morphological features used to distinguish galaxies in the Hubble Sequence: Bars, Rings, and Spiral Arms. There is one last thing to note before discussing the features; some of the pictures collected have been taken at such an angle that their morphological features are not discernable. This has been commonly adopted into a fifth category known as edge galaxies.

The first of our features was the number of times the brightness oscillated moving from the center of the galaxy to the edge totaled over each cardinal direction. We defined an oscillation as a change of at least .03 in mean brightness between two adjacent 10x10 pixel regions. We determined that should be useful for deciding between non-spiral (elliptical, lenticular, edge) and spiral galaxies since spiral arms will cause oscillations.

The next feature we calculated was the absolute value of the brightness change moving from the center of the galaxy to the edge. This feature was selected to assist in separating barred spiral and spiral galaxies since thicker spiral arms will have a steeper intensity gradient. This complements the first feature by giving more weight to larger oscillations.

Our third feature, rotational symmetry, was chosen to distinguish non-spiral galaxies from spiral galaxies. To extract this feature, we rotated the galaxy 180 degrees about its center. We determined that this feature would be helpful since elliptical galaxies should have a very high degree of symmetry, while spirals would have a large number of pixels in their spiral arms that would not overlap. Furthermore, an ideal barred spiral would fall somewhere in the middle of this spectrum due to the symmetry of the bar that characterizes these galaxies.

The ellipticity of the brightest 90 percent area of the galaxy, was calculated by analyzing a new black and white image (derived from the greyscale image) with a threshold at 90% of the maximum brightness turned to black and white, fitting an ellipse around the remaining pixels. This defining feature of elliptical and lenticular galaxies segregated it from the spirals. Moreover, the barred center of barred spirals decides between barred spiral galaxies and non-

barred spiral galaxies since the center of barred spiral galaxies will have high ellipticity as opposed to the circular centers of non-spiral galaxies.

The next feature considered was the convexity of the bounding polygon of the galaxy, since non-spiral galaxies will be bounded by an ellipse whereas spiral galaxies will often have concavities in between their arms.

The mean brightness of the galaxy, within the bounding polygon was also added to the feature set to distinguish spiral galaxies from barred spiral galaxies, because barred spirals have overall lower mean brightness since they have longer arms that stretch further from the center of the galaxy.

Additionally, the ratio of white to black pixels inside the bounding rectangle of the galaxy in the black and white image was selected. We predicted that this feature would primarily serve as way to separate Spiral galaxies and edge galaxies from the rest since they will have lower ratios of white to black pixels.

The ratio of the perimeter of the bounding polygon of the galaxy to the bounding ellipse of the galaxy was also used. Spiral galaxies will have highest polygon perimeter since they are elliptical with many concavities, barred spiral galaxies will be next since they generally have a large concavity, and elliptical and lenticular galaxies will have the highest ratio of all.

The last feature we employed was the rBulge which is defined as the ratio of the shortest radius at which the brightness has dropped to 90% maximum brightness to the shortest radius to the edge of the galaxy [4]. Spiral galaxies and specifically barred spiral galaxies will drop in brightness quicker as gaps between arms allow the brightness to dip down abruptly.

After implementing forward search, we narrowed were able to determine which parameters were most useful at each step in the classification process for both models.

Model	SVM	GDA
Spiral vs Non-Spiral	0, 5, 6, 1, 2, 4, 3, 7, 8	8, 0, 7, 1, 4, 3, 6, 2, 5
Barred Spiral vs Spiral	6, 3, 4, 7, 1, 0, 5, 8, 2	0, 3, 7, 4, 6, 8, 1, 5, 2
Edge vs Elliptical / Lenticular	1, 6, 7, 0, 2, 4, 5, 8, 3	1, 7, 4, 6, 7, 0, 3, 2, 5
Elliptical vs Lenticular	2, 7, 0, 1, 3, 5, 8, 6, 4	2, 6, 8, 0, 1, 3, 5, 7, 4

Legend: 0 - Number of oscillations; 1 -Dark/Light pixel ratio, 2- Rotational symmetry, 3-Convexity, 4-Perimeter to area ratio, 5-rBulge, 6-Mean brightness, 7 – Ellipticity, 8- Oscillation Magnitude

Figure 3. The results of our forward search feature selection

4 Models

The models that we used were Gaussian Discriminate Analysis and Support Vector Machines. Gaussian Discriminate Analysis was selected as our experimental model in this project. If features are Gaussian distributed this the result of GDA would have a very low error and perhaps an undiscovered optimal model to use for machine learning galaxy classification.

Support Vector Machines were used as a reference model since we had seen many implementations that had succeeded with very low error [5]. The previous success of an SVM implementation with various sets of parameters indicated that the training examples were linearly separable.

Lastly, we decided to take a fragmented approach to classifying the galaxies to better examine the utility of each of the morphological features at each stage in the process.

5 Results

SVM

Classifying	Spiral (barred spiral, spiral) vs Non-Spiral (elliptical, lenticular, edge)	Barred spiral vs Spiral	Edge vs Elliptical or Lenticular	Elliptical vs Lenticular
Parameters Used	Number of oscillations, rBulge, mean brightness	Mean brightness	Dark/Light pixel ratio, mean brightness, ellipticity	Rotational Symmetry
Train Error	6.62%	23.5%	21.54%	30%
Test Error	12.5%	11.11%	14.28%	20%

GDA

Classifying	Spiral (barred spiral, spiral) vs Non-Spiral (elliptical, lenticular, edge)	Barred spiral vs Spiral	Edge vs Elliptical / Lenticular	Elliptical vs Lenticular
Parameters Used	Oscillation magnitude	Number of oscillations, convexity, ellipticity	Dark/Light pixel ratio, ellipticity	Rotational Symmetry, mean brightness, oscillation magnitude
Train Error	42.5%	49.7%	16.9%	46.6%
Test Error	31.25%	44.44%	20.0%	20.0%

6 Discussion

In order to prevent overfitting and the use of parameters that were not relevant for a specific classification type, we used forward search feature selection to find the best features to use for each classification. This usually resulted in 1-3 features. Of these, the mean brightness of the galaxy and the ellipticity of the center of the galaxy were the two parameters that seemed to be the most useful.

Our results clearly show that Support Vector Machines are superior to Gaussian Discriminant Analysis for galaxy classification with our parameters. Gaussian Discriminant Analysis had very little success, usually obtaining training error in the range of 40 percent. This shows that the parameters cannot be well modeled as Gaussian with the morphological features that were used. Even after adjusting the amount of features as well as the size of k in our kFold Cross Validation from 10 to both 5 and 20, the GDA failed to find a fit that was comparable to that of the SVM. Additionally, the fit of the Support Vector Machine shows that the features we selected here were also linearly separable.

7 Future

Normalizing and preprocessing our images more would be an important step forward in our work, yet a major undertaking. A common obstacle in the algorithm came from the fact that galaxies that are not directly perpendicular to the Earth are seen at an inclination. Thus, the images contain galaxies at various inclinations. Circular spiral galaxies seen at an inclination will appear elliptical. Since a number of our parameters depend upon analyzing the ellipticity of the galaxy, the height and width of the galaxy, and the density of pixels in an area and these parameters assume that the galaxy is seen directly perpendicular, they are thrown off by these inclined galaxies. In order to solve this problem, the perpendicular view of these inclined galaxies would need to be projected from their images. This is a complicated procedure since it is difficult to tell if a galaxy is inclined, simply naturally very elliptic, or a barred spiral galaxy with only two short arms.

Another problem that often arose was that the brightness and the size of galaxies in the images were not consistent. Some of the elliptical galaxies took up the entire image without including any of the dark space around the galaxy, while other images had an expanse of space around the galaxy. Our features that worked with the borders of the galaxy were hard to implement when the borders of the galaxy were sometimes not found to be included in the picture. Preprocessing that works to normalize the maximum and minimum brightness in the images and the size of the galaxy within the image would serve to remedy this problem.

Lastly, despite having tested the applications of nine different parameters in separating the types of galaxies, there are a number of others that could perhaps be useful if we were to implement them. Some of these include spectral signature and light color (implemented by N. Ball in Robust Machine Learning Applied to Astronomical Data Sets) [6], and photometric, texture, and spectral data available from Galaxy Zoo (implemented A. Gauci et al. in Machine Learning for Galaxy Morphology Classification) [7].

8 References

[1] NASA's HEASARC. Retrieved Nov. 20, 2014 from http://heasarc.gsfc.nasa.gov/db-perl/W3Browse/w3table.pl?tablehead=name%3Dmcg&Action=More+Options

[2] Skyview. Retrieved Nov. 20, 2014 from http://skyview.gsfc.nasa.gov/current/cgi/titlepage.pl

[3] Strasbourg Astronomical Data Center. Retrieved Nov. 20, 2014 from http://cds.u-strasbg.fr/

[4] D. Bazell, Feature Relevance in Morphological Galaxy Classification, Royal Astronomical Society, pp. 519-528, Feb. 2000.

[5] Zhang Y and Zhao Y, Classification in Multidimensional Parameter Space, Publications of the Astronomical Society of the Pacific, vol. 115, no. 810, pp. 1006-1018, Aug. 2003.

[6] N. Ball et al, Robust Machine Learning Applied to Astronomical Data Sets. I. Star Galaxy Classification of the Sloan Digital Sky Survey DR3 Using Decision Trees, The Astronomical Journal, vol. 650, no. 1, October 2006.

[7] A. Gauci et al, Machine Learning for Galaxy Morphology Classification, Royal Astronomical Society, pp. 1-8, June 2010.