Julian Kates-Harbeck, CS229 Final Report

Classifying Galaxy Morphology using Machine Learning

Introduction:

The goal of this project is to classify galaxy morphologies. Generally, galaxy morphologies fall into one of two categories: "early types" (elliptical galaxies) and spiral galaxies, like the Milky Way. Additionally, a celestial object of interest may be categorized as neither, which usually corresponds to an image of a point source (e.g. a star) or to an artifact of the data. Although humans are generally successful at classifying astronomical images, the sheer amount of data that is currently produced is far too immense to be interpreted by humans. The Sloan Digital Sky Survey (SDSS) for example – the number one source for publicly available astronomical imaging data – has a growing catalog of over 230 million celestial objects. Successful classification of all of these objects could represent a major step forward in the field of observational astronomy, ultimately bringing us closer to the goal of better understanding the composition and origins of our universe. Several attempts have been made to classify the images from SDSS. One interesting project is the Galaxy Zoo Project (C. Lintott et al.): The idea of the project was to have humans – more specifically volunteers with some prior knowledge of the subject (e.g. science teachers) – interactively classify images of celestial objects online. Every given image was classified by several people, in order to produce a measure of reliability of the classification. Using these statistics, as well as several data reduction and cleansing methods, a data set of about 700,000 reliably classified galaxies was produced. A common object ID identifies the objects used in Galaxy Zoo with the images from SDSS. In this project, we use the classifications from Galaxy Zoo together with the corresponding observational data from SDSS as our data set.

This project consists of 2 separate approaches. In the first, we hand pick pre-computed features from SDSS for every galaxy and run classification using those features. In the second and far more challenging approach, we aim to classify galaxies using only the raw image data. As a training set, we use the *clean* images from Galaxy Zoo – those galaxies where at least 80% of observers agree on the morphology. We evaluate all classifiers using 10 fold cross validation.

I. Classification using Pre-computed Features

Data Selection:

The SDSS catalog, in addition to the raw image data, offers a set of preprocessed features for every object. The total number of such features is very large and many of them are not useful for morphological classification. Using the Weka package (M. Hall et al.), we analyze the classification power of different features. Using several attribute selection methods, we find that the feature set with the strongest classification power is very similar to the feature set used by Banerji *et al*. The features are described and the classification power is presented in table 1.

Evaluator vs. Feature	Meaning	Information Gain
Concentration	petroR90/petroR50 (ratio of radii containing 90% and 50% of the Petrosian flux	0.362
g-r color	Green minus red color (redshift removed)	0.304
deVAB_i	DeVaucouleurs fit parameter (infrared)	0.285
expAB_i	Exponential fit parameter (infrared)	0.284
mRrCc_i	Adaptive second moment	0.279
mE2_i	Adaptive shape measure	0.229
mE1_i	Adaptive shape measure	0.197
lnLexp_i	Exponential disk fit (log likelihood)	0.161
lnLdeV_i	DeVaucouleurs fit (log likelihood)	0.136
mCr4_i	Adaptive fourth moment	0.107
r-i color	Red minus infrared color (redshift removed)	0.101
lnLstar_i	Log likelihood of being well modeled as a star	0.021
texture_i	Texture parameter (infrared)	0.01

Table 1: Description of the different features and relative classification power (arb. scaling) .

Classification Results:

Using the optimized features, we employ several machine learning techniques on our data. Specifically, we use a simple Naïve Bayes classifier, a logistic regression classifier, an alternating decision tree, a random forest (using 10 trees and 4 random features per tree) and an SVM with a linear kernel. The results are presented in table 2. We can see that the classification performance reaches a **maximum** of **95.21%** with a **logistic regression** classifier. For most classification techniques, the performance lies between 93% and 95%. The only outlier is the Naïve Bayes classifier. This is an indication that the Naïve Bayes independence assumption is not well-justified. Indeed, for the SDSS feature set, several features often aim to describe the same characteristic: For example, both "Concentration" and "mE2_i"/"mE1_i" describe the width of the brightness distribution in the image.

II. Direct Image Classification

While we were able to achieve good results using the pre-compiled features from SDSS, a far more interesting problem is to classify the galaxies using only the raw image data. The goal is to train a classifier that can – just like a human – take a raw image of a galaxy and classify it as either spiral or elliptical.

Image Processing:

The imaging data retrieved from SDSS consists of 200 x 200 pixel RGB images. A typical collection of SDSS imaging data is shown in fig. 1. It quickly becomes obvious that before the raw image data can be used for classification, several issues need to be overcome. First, the image data contains a strong background noise level. The regions that appear "black" in the images actually have a consistent magnitude of about ¼ of the maximum magnitude of the image. Second, the imaging data, apart from the actual galaxies, contains many other bright secondary objects. These secondary objects may be artifacts of the camera, stars, quasars, or other galaxies. Third, the images are subject to random rotations. In our final classifier we do not want the rotation of the image to affect the classification of the galaxies. One of the most challenging aspects of this project was the development of a reliable image processing procedure that would remove the three issues listed above. The following is a summary of the developed image processing pipeline.

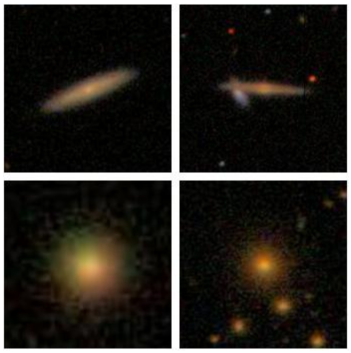

Figure 1. Typical Imaging data from SDSS: *Spiral* (top) and *Elliptical* (bottom) galaxies. *Single Objects* (left) and images with *Secondary Objects* (right).

First, the RGB-image is converted to a grayscale image ("flattened"), i.e. a single 200 by 200 matrix. In order to subtract the constant background, we find the minimum pixel value and subtract it from the matrix. We further zero all pixel values bellow a certain magnitude cutoff (we found that half the mean of the matrix was ideal). This further reduces random fluctuations in the ideally "black" regions of the image.

The second step – the removal of bright secondary objects – proved to be the most challenging. Due to their brightness, secondary objects are usually not removed by the above described procedure. The two key insights that lead to an efficient solution were the following: 1) The images queried from SDSS are almost exactly (within 1 pixel) centered on the objects of interest 2) Galaxies generally are almost perfectly invariant under 180° rotation. Thus, we can compare each pixel of our image matrix to the corresponding pixel rotated 180° about the center of the image. Taking advantage of 2), a "strong enough" difference in magnitude of these two pixels means that they are very likely part of secondary objects and can thus be set to zero. There are however a few subtleties involved in this algorithm. First, spiral galaxies sometimes

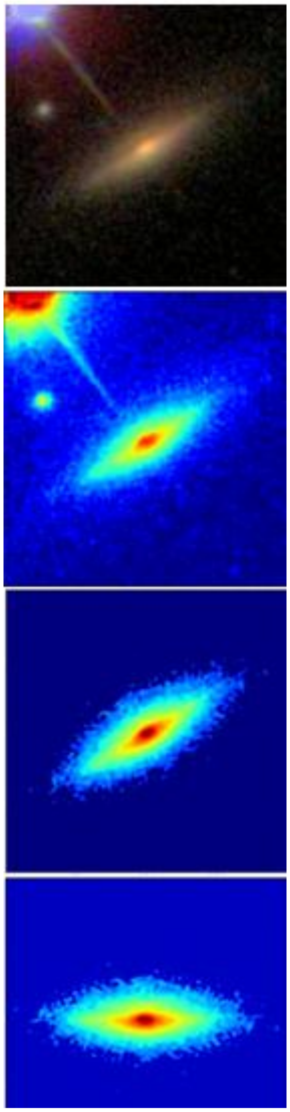

Figure 2. Image Processing Pipeline as illustrated with an example galaxy (greyscale images as heat maps): Raw image (top), flattened image (upper center), background and secondary objects removed (lower center), rotation normalization (bottom).

have structural elements (such as indications of galaxy arms) that are not exactly symmetric under 180° rotations. Second, the condition for “strong enough” must be optimized, since a too high threshold would miss objects, while a too low threshold might cause obstruction of the galaxy itself. We found that the optimal condition for “strong enough” is a relative difference in magnitude of over 0.5. Further, in order to guarantee that the main structure of the galaxy itself is unaffected, we run the algorithm only on pixels that are further than 30% of the total image radius away from the image center.

Finally, we wish to process the images such that rotation of the original image will not affect the classification. The key insight here is that galaxies share a common ellipsoid structure in the images, i.e. they all have two orthogonal axes of symmetry. By calculating the second moments in x and y of the image as well as the correlation of x and y, we can construct the covariance matrix of the galaxy image. This allows us to deduce the major variance carrying axis (Teague). We can now use the angle of this axis and rotate every image onto its major axis.

In summary, we now have a clean, centered image with consistent definitions of the x and y axes, without secondary objects and with all pixels other than those of the main object set to zero. A summary of the image processing pipeline is given in figure 2.

Image Analysis:

Using the clean image data, we train both a “blind” classifier that processes the image without any insight into the data as well as a more intelligent classifier that uses scientific knowledge of galaxies in order to extract features that might be especially indicative of morphology.

In our blind approach, we compute the 2D Fourier Transform of the image, discarding the high frequency terms (which carry mostly noise). Specifically, taking the central 10 by 10 Fourier components showed the best performance. When performing classification using these 100 Fourier coefficients directly, we achieve a classification accuracy of 88.23% with an ROC area of 0.931 using an SVM with a linear kernel. After further dimensionality reduction using ICA and retaining only the coefficients of the 20 most prevalent independent components (20 yielded the best results) as features, we achieved **91.52% classification accuracy** with an **ROC area of 0.878** using an **SVM with a Gaussian kernel**,

Our knowledge based classifier takes advantage of three main distinguishing characteristics between elliptical and spiral galaxies. First, the typical color spectrum of elliptical galaxies is different from that of spiral galaxies, since most elliptical galaxies are very old (they are often also called “early type” galaxies) and thus are subject to a stronger redshift than the usually “ younger” spiral galaxies. We use the three RGB matrices of the raw unflattened image – corresponding to the intensities from the green, red and infrared color channels of the original photometric SDSS image, together with a bitmask obtained in the image processing step separating the galaxy from the background, to extract the color composition of the galaxy. We then compute the fractions of the R, G and B pixels of the total image magnitude, thus obtaining the fractions of green, red and infrared light of the total brightness of the galaxy, respectively.

Second, we take advantage of the fact that spiral galaxies, unless they are viewed directly from their axis of rotation, appear as very eccentric ellipsoids, while elliptical galaxies show very uniform circular symmetry with low eccentricity. We use up to 4th order normalized central moments in x and y as well as generalized TSR (translation, rotation, scaling) invariant moments (Hu; Flusser) of the images to capture this information. Since our images are centered and rotated onto their major axis, the x and y moments won't be affected by rotation of the galaxies.

Classifier	Accuracy	ROC Area
Pre-computed features		
Logistic Regression (PCF)	**95.21%**	0.986
Random Forrest	94.86%	0.982
SVM (SMO, linear kernel)	94.12%	0.926
Decision Tree	93.76%	0.979
Naïve Bayes	91.43%	0.977
Blind image classification		
Image Analysis + ICA + SVM	**91.52%**	0.878
Knowledge-based image classification		
Image Analysis + feature selection + SVM (poly. Kernel 3)	**95.89%**	0.943
Image Analysis + feature selection + Random Forest (60 trees)	95.69%	0.968
Image Analysis + feature selection + Logistic Regression	95.18%	0.989

Table 2: This table summarizes classification performance from the different parts of the project. The best performing classifiers are printed in bold face.

Third, we make use of the difference in radial intensity distributions between elliptical and spiral galaxies. Generally, elliptical galaxies experience an exponentially decreasing intensity profile with radius, while spiral galaxies are better modeled using a deVacouleurs profile: $I_{spiral}(r) \propto e^{-kr^{\frac{1}{4}}}$, $I_{elliptical}(r) \propto e^{-kr}$ (for some constant k) (Sérsic). By taking advantage of the symmetry of the galaxies and their rotational normalization onto the major axis, we can capture most of the information by using the 1D density profiles along the major and minor axes of the image instead of the full 2D radial profile, thus simplifying the computation substantially. Specifically, we compute the distances in x and in y from the center that capture 50% and 90% of the total magnitude of the 1D intensity profile along the major and minor axes, respectively. We then take ratios of these 4 values, including cross terms. These features capture information about the radial density distribution well.

The combined features from the color study, the distribution moments and the density profiles are then fed to several different classifiers, as summarized in table 2. The best performance is achieved using an **SVM** with a **3rd degree polynomial kernel**: we achieve a maximum **classification accuracy** of **95.89%** with an **ROC area** of **0.986**, exceeding the benchmark performance in section I using the pre-computed features.

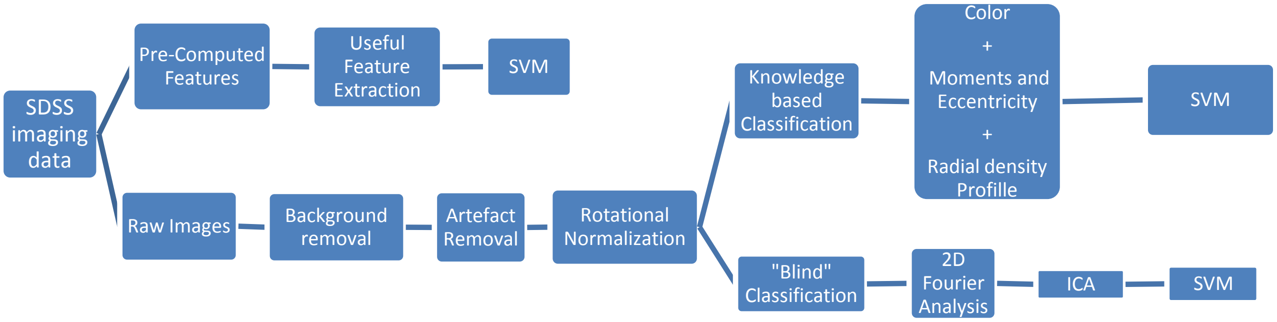

Figure 3. Summary of the learning systems used in this project. Classification is performed using pre-computed features as well as raw image data. For the raw image data, we use both knowledge-based and "blind" features.

Conclusion and Future Work

We have presented a system consisting of an image processing algorithm in conjunction with a knowledge based classifier that can with high accuracy classify galaxies using only the same raw image data that was available to humans in the Galaxy Zoo study. Comparing the blind and insightful classifiers, we have also shown that using physical insight into the details of a problem can lead to significant improvement of classification performance. The summary of our learning systems is given in figure 3, and the summary of classification performance is given in table 2

There are two major areas for possible improvement of our system. First, due to the computational complexity of the problem of image cleansing and feature extraction and limitations of our computational processing and memory resources, we were only able to create a training set of 1000 galaxies for testing and training; a larger training set will very likely result in even better classification performance. Second, one physical difference between elliptical and spiral galaxies that we have found difficult to capture in our feature set is the very fine spiral arm structure: In some cases, a spiral galaxy viewed from above (close to or along the axis of rotation) can show (usually very slight) indications of spiral arm structure, while an elliptical galaxy is always azimuthally homogeneous. A more advanced feature set including this information would most likely increase the classification accuracy even further.

Acknowledgements:

I would like to thank Dr. Mark Allen from the Stanford Physics Department for his helpful input and suggestions for this project.

References:

M. Banerji et al. "Galaxy Zoo: reproducing galaxy morphologies via machine learning." Monthly Notices of the Royal Astronomical Society. Vol. 406 2010: pp. 342-353.

Flusser, Jan. "On the independence of rotation moment invariants." Pattern Recognition 2000: 1405-1410.

M. Hall et al. "The WEKA Data Mining Software: An Update." SIGKDD Explorations, Volume 11, Issue 1 2009.

Hu, M. "Visual pattern recognition by moment invariants." IRE Transactions on Information Theory, 8 2 Feb. 1962.

C. Lintott et al. "Galaxy Zoo 1: Data Release of Morphological Classifications for nearly 900,000 galaxies." Monthly notices of the Royal Astronomical Society 2010.

—. "Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey." Monthly Notices of the Royal Astronomical Society 2008.

Sérsic, J. L. "Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy." Boletin de la Asociacion Argentina de Astronomia 1963: p. 41.

Teague, M. R. "Image analysis via the general theory of moments." Journal of the Optical Society of America August 1980: 920-930.