Automated Classification of Galaxy Zoo Images

CS229 Final Report

Michael J. Broxton - broxton@stanford.edu

1. Introduction

The Sloan Digital Sky Survey (SDSS) in an ongoing effort to collect an extensive image and spectral catalogue of deep sky objects. Billed as "the most ambitious and influential survey in the history of astronomy," the SDSS contains rich information that has had a significant impact on our understanding of the history, structure, and evolution of our universe [York 2000]. Unfortunately the enormous size of the SDSS data catalogue makes it unwieldy and difficult to analyze by hand. One viable approach for processing SDSS data on a large scale was demonstrated in 2008 by the Galaxy Zoo 1 project [Lintott 2008], in which Internet users were asked to categorize millions of images of galaxies into three morphological classes: elliptical, spiral, or 'other.[1]' However, after one year of collecting user input, this effort ultimately classified only 1 million out of 230 million objects in the survey catalog. With more observations being added on a daily basis, there is an ever growing need for robust, automated analysis & classification techniques.

The goal of this project is to demonstrate that machine learning algorithms can be trained to produce results that consistently agree with galaxy classifications produced by human click-workers. Using classification statistics released by the Galaxy Zoo project for nearly 900,000 objects [Lintott 2010] as a training & validation data set, we show that modern machine learning architectures can be trained to perform well on the SDSS galaxy classification task. We directly compare galaxy morphology classifications produced by our algorithms to those produced by Galaxy Zoo users. With this as our error metric, we test various machine learning approaches using two different feature modalities that are readily available from the Sloan Digital Sky Survey: (1) *hand-engineered features* that are good proxies for galaxy morphology (e.g. color, luminosity, spectral features, or structural parameters); and (2) 423x423 pixel, 3-channel *color images* of the galaxies. Note that human classifiers on the Galaxy Zoo website based their classifications solely on color imagery (i.e. feature set #2). It is therefore a primary objective of this project to determine whether a machine classifier can produce comparable results using the image feature set alone. In Section 2 we describe these feature modalities in more detail, then in Section 3 we present two machine learning architecture for processing color galaxy images. Results are presented in Section 4, and we discuss our conclusions in Section 5.

green - red	green minus red color (red shift removed)
red - infrared	red minus infrared color (red shift removed)
deVAB_i (de Vaucouleurs fit axial ratio)	axial ratio for a fit to a galaxy bulge model
expAB_i (Exponential fit axial ratio)	axial ratio for a fit to a galaxy disk model
lnLexp_i (Disk fit log liklihood)	liklihood of being well modeled as a galaxy with a disk-like feature
lnLdeV_i (Star log likelihood)	liklihood of being well modeled as a galaxy with a central bulge feature
lnLstar_i (Star log likelihood)	liklihood of being well modeled as a point source
petroR90 i/petroR50 i (concentration)	ratios of radii containing 90 and 50 percent of the Petrosian flux
mRrCc_i	second moment of object intensity in the CCD (robust to noise)
aE_i	adaptive ellipticity - (based on mE1 and mE2)
mCr4_i	adaptive fourth moment
texture_i	ratio of range of fluctuations in surface brightness to full dynamic range of object

Table 1: These 12 hand-tuned features capture the color, morphology, and photometric properties of galaxies. They are available for download for all SDSS observations, and were chosen here because they were used in previous galaxy classification studies (specifically [Banerji 2010]).

2. Feature Modalities

It has been shown in various studies [Odewhan 1994 , Lahav 1995, Ball 2006, Elting 2008, Banerji 2010] that promising automated galaxy classification rates can be achieved by leveraging the wealth of morphological and photometric statistics available in

[1] The 'other' category in the Galaxy Zoo data set is used for hybrid objects like galaxy mergers or partially occluded galaxies; or objects that were incorrectly classified as galaxies by SDSS automated data pipeline, such as Stars and Quasars.

the SDSS database. For consistency with past studies, we chose the same twelve features used in [Banerji 2010] (listed in Table 1). These capture a range of properties including the galaxy's color, degree of ellipticity, and surface brightness profile. Classification using hand-tuned features serves as a baseline case for the purposes of this study.

Our novel contribution is to show that modern machine learning architectures can classify galaxies using exactly the same 'input' as was available to human classifiers; namely color galaxy images alone. In fact, we will show that image features consistently outperform hand-engineered morphological and photometric features. In order to achieve this result we found it necessary to first compute the Fourier transform of the imagery, take the magnitude of the result (i.e. discarding phase information), and then 'cropping' this power spectrum image to retain only the low frequency components. This *spectral signature* still contains most of the salient information from the original image, but is invariant to phase and robust to the high frequency image noise that is prevalent in Galaxy Zoo images. This approach was inspired to a certain extent by models from neuroscience of complex visual receptive cells that respond to the frequencies and orientations of imagery, but exhibit a high degree of phase invariance. Such complex cells can be well modeled as acting on the power spectra of visual stimuli [David 2005]. In our case, the intuitive justification for this pre-processing step is that the frequency content and orientation of galaxy images contain salient features that are useful for detecting spiral structure and elliptical eccentricity respectively; whereas the phase is less important since the positions in an image of disks, bulges, spiral arms, and other features do not encode much useful information for classification purposes.

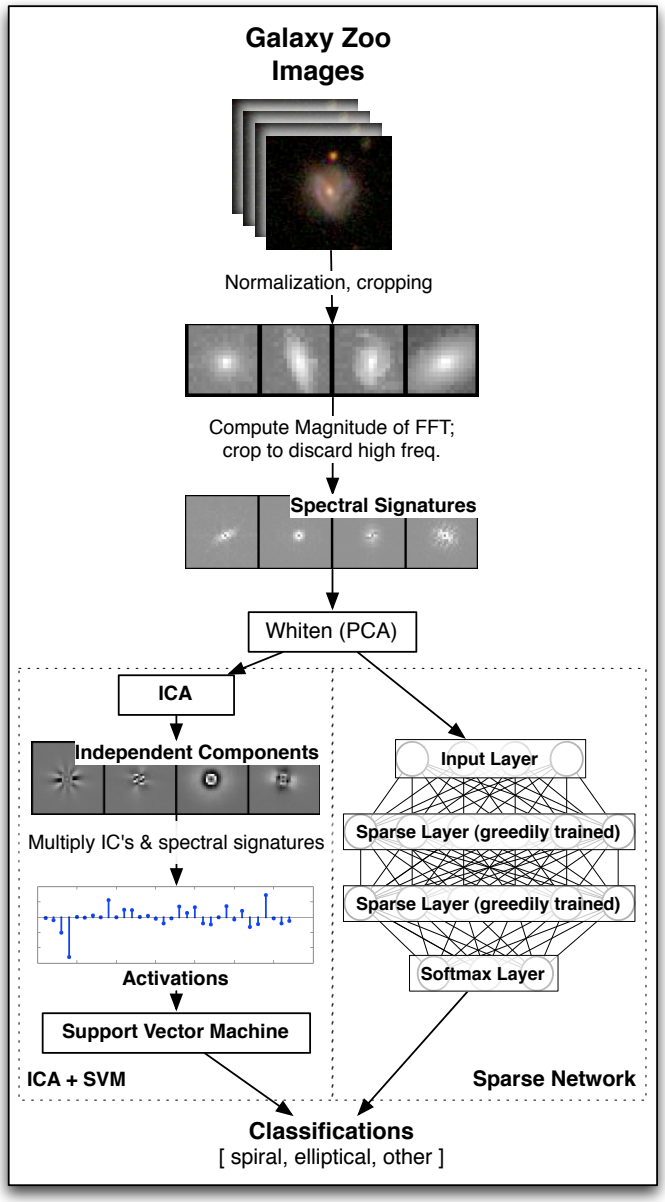
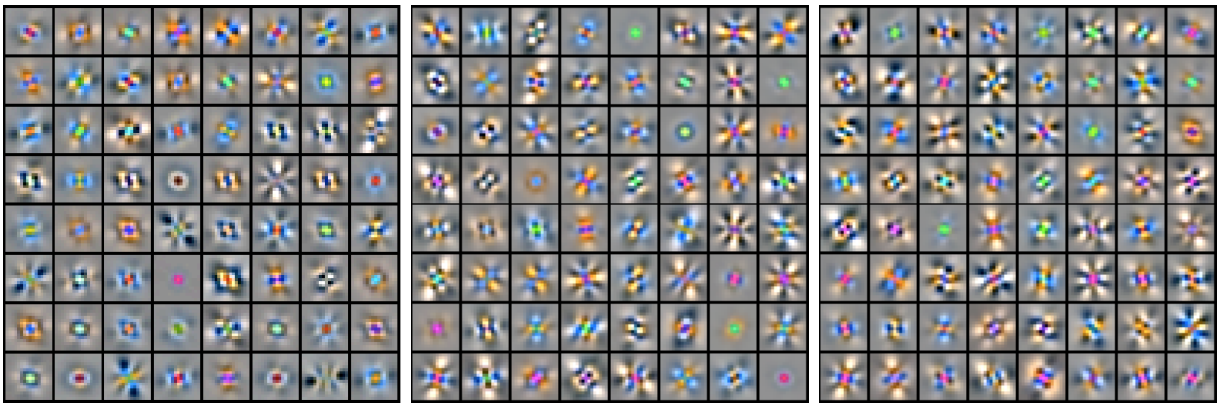


Figure 1: Our two machine learning architecture for galaxy classification learn sparse features on the Fourier power spectrum of galaxy images. In this study we consider two architectures that first learn useful features from the imagery in an unsupervised setting, and then we conduct supervised training using class labels from the Galaxy Zoo data set.

Figure 1 shows these pre-processing steps in more detail. Images are first normalized and then cropped to a 200x200 pixel window centered on the galaxy. The square root of the Fourier power spectrum is then computed by taking the magnitude of the FFT for each galaxy image. The resulting 200x200 pixel frequency space image is further cropped to the center 32x32 pixel region; in essence throwing out high frequency components of the FFT that are dominated by image noise. Finally, the data is whitened using PCA (retaining 99% of variance). We refer to the final whitened spectral image as the *spectral signature* of the galaxy.



(a) (b) (c)

Figure 2: Three sparse feature dictionaries learned from the Fourier power spectrum of galaxy zoo images. The features produced using ICA (a) and the Sparse Autoencoder (b) are similar; both contain filters that select for an assortment of frequencies (round, ring-like features) and orientations (asymmetric, oriented features). Some sparse autoencoder features change after fine-tuning (c), while others remain largely the same. Note that color seems to play an important role in feature morphologies. Different color channels often select for different frequency characteristics within the same filter.

3. Learning Architectures

We have developed two separate systems that learn sparse features on the Fourier power spectrum of Galaxy Zoo images. The *ICA+SVM* system (left half of Figure 1) uses Independent Component Analysis (ICA) to learn M independent components that exhibit a high degree of selectivity for frequencies (via round, ring-like features) and orientations (via asymmetric, oriented features). Figure 2(a) shows typical independent components learned from spectral signatures. We then compute *activations* for a given galaxy image by multiplying spectral signatures by each independent component. This produces M activations per galaxy (we selected $M = 64$ in this study). Finally, we train a SVM using a linear kernel on these activations to match known classifications in our training data set.

Our second approach employs a neural network with hidden layers that are greedily trained as Sparse Autoencoders [Ng 2010]. During pre-training, this *Sparse Network* system learns a set of features like those depicted in Figure 2(b). Features are subsequently refined during the supervised fine-tuning phase (see Figure 2(c)) at the same time as the network trains its Softmax output layer to identify galaxies. Sparse weights are held constant during the first 50 iterations of fine-tuning, and then allowed to change during the remaining 200-300 iterations or until the algorithm converges. In this study we tested networks with one or two sparse hidden layers; each containing 100 sigmoidal units trained with a lifetime sparsity constraint as well as a regularization term that helped to prevent over-fitting.

Note that the ICA+SVM approach produces ‘hard’ classifications with only one label per galaxy, whereas the Sparse Network (by virtue of its softmax output layer) produces ‘soft’ classifications with three separate probabilities that indicate the likelihood that a galaxy is a member of each class. We will show in Section 4 that ‘soft’ labels are useful for rejecting low-confidence matches that would otherwise appear as false positives in the final results.

4. Results

We consider three subsets of the Galaxy Zoo data for testing and validation of these algorithms. The *clean* subset contains galaxies for which 80% or more of human classifiers agree on a given galaxy’s class. We also consider an *uncertain* subset for which only 50% or more of human classifiers agree. Finally, we test performance on the *full* galaxy zoo data set (rejecting only those entries with incomplete metadata from the SDSS) in which there is sometimes substantial human disagreement. The *uncertain* and *full* subsets contains images that are far more challenging for humans to categorize consistently, and it shows how the performance of our algorithm degrades as data becomes increasingly ambiguous. All

	"Clean" Subset (80% or better Human Agreement)					"Uncertain" Subset (50% or better Human Agreement)					Full Dataset				
Classifier (Feature Set)	*Elliptical*	*Spiral*	*Other*	***Overall***	***% Reject***	*Elliptical*	*Spiral*	*Other*	***Overall***	***% Reject***	*Elliptical*	*Spiral*	*Other*	***Overall***	***% Reject***
Prior Art [Banerji et. al. 2010] (Hand-tuned Features)	*97%*	*97%*	*86%*	**--**	--	--	--	--	**--**	--	*91%*	*92%*	*95%*	**--**	--
Softmax Regression (Hand-tuned Features)	*97.3%*	*95.3%*	*37.5%*	**94.8%**	*3.2%*	*90.5%*	*85.1%*	*45.2%*	**85.6%**	*6.6%*	*90.2%*	*83.8%*	*43.2%*	**84.6%**	*8.4%*
Support Vector Machine (Hand-tuned Features)	*97.3%*	*97.9%*	*0*	**97.0%**	--	*88.9%*	*90.1%*	*0%*	**86.6%**	--	*87.7%*	*88.4%*	*0%*	**84.4%**	--
ICA + Support Vector Machine (Power Spectral Features)	*97.3%*	*98.7%*	*5.6%*	**97.6%**	--	*90.2%*	*90.2%*	*6.1%*	**87.4%**	--	*86.5%*	*87.6%*	*0%*	**83.4%**	--
Sparse Network (1 hidden layer) (Power Spectral Features)	*98.4%*	*98.4%*	*64.7%*	**98.3%**	*0.6%*	*92.9%*	*93.9%*	*68.6%*	**91.6%**	*9.0%*	*90.6%*	*93.7%*	*58.8%*	**89.4%**	*13%*
Sparse Network (2 hidden layers) (Power Spectral Features)	*98.8%*	*98.2%*	*70.6%*	**98.5%**	*0.5%*	*93.7%*	*93.9%*	*68.5%*	**92.0%**	*9.1%*	*91.0%*	*93.5%*	*59.7%*	**89.5%**	*13%*
Sparse Network (2 hidden layers) (Hand-tuned + Spectral Features)	*99.1%*	*98.7%*	*73.3%*	**98.8%**	*0.3%*	*95.0%*	*93.2%*	*59.4%*	**91.8%**	*6.5%*	*92.8%*	*94.1%*	*50.0%*	**91.2%**	*14%*

Table 2: Results from classification tests for different feature sets and classifiers. The best performance is obtained using a combination of spectral signatures and hand-tuned features, however this only improved slightly over performance with spectral signatures only. Spectral features consistently out perform hand-tuned features in all of our tests. Note that these are results of individual tests; we have not yet had time to conduct thorough cross-validation. Some data was not available, and some tests with 'hard' classification outputs did not allow for us to reject inconclusive entries. These entries are marked with '--'.

tests were conducted using 10,000 randomly selected galaxies from one of the three data subsets. Images were split into 7,500 records for training and 2,500 records for testing.

Table 2 shows results for various combinations of learning algorithms and feature sets. Results are broken out by class, with an overall classification rate shown in bold face. In the first row we have included results from the [Banerji 2010] study, which they obtained by training a three layer neural network on hand-tuned features. In order to draw an accurate comparison to this prior work, we have adopted a similar strategy to the one employed by Banerji et. al. for interpreting classification results. This first requires rejecting a certain number of 'inconclusive' classifications made by the machine algorithm for galaxies that did not receive more than a 50% probability of belonging to any one of the three classes. Culling out weak classifications in this manner decreases the number of false positives in all classes, and improves classification rates (i.e. the fraction of true positives out of all classifications considered for a given class) by several percent. With the 50% threshold, we typically rejected between 1% and 10% of the total galaxies being classified. The "% Reject" column in Table 2 shows the exact percentage of testing examples rejected in each test.

Note that the 50% threshold was chosen arbitrarily, and should be considered a parameter that can be tuned to minimize false positive rates at the expense of increasing the number of false negatives. However, we did not have time to study this tradeoff closely, but in future work one could find the optimal cutoff that perfectly balances false positives and false negatives for any given machine learning algorithm.

Following Banerji's example, we did in fact study the trade off between false positives and false negatives *for each class in isolation*. For example, the '*other'* category rarely had the highest probability of the three classes (probably due to there being relatively few training examples of this class in the data set), so our algorithms superficially appeared to do poorly at classifying objects in *other* category. However, we found that *other* objects could be more reliably identified if we chose a low threshold for determining whether an object might be in the *other* category. We found that *P(other) > 19%* was the optimal threshold that evenly traded off false positives and false negatives in this class for the training data. Optimal thresholds for the elliptical and spiral probabilities were found to be around 43% and 45% respectively. It should be noted, however, that setting thresholds in this manner allows an object to be placed into more than one category (or even into none at all). This may or may not be desirable depending on the application, but we chose to analyze our results in this fashion here so that our comparison to Banerji's study could be made as accurately as possible.

5. Conclusions & Discussion

Techniques using spectral signatures outperformed hand-tuned features in all tests, although slightly better results were obtained by training a sparse network using a concatenated vector containing of both of these feature types. The sparse network techniques outperformed others, with additional hidden layers yielding a modest increases in classification performance. Overall, we demonstrate a 2% improvement over Banerji et. al.'s classification rates using our sparse network approach. This is significant, especially when considering that we achieved this performance using the same color images that served as 'input' for the human classifiers, and demonstrated that one need not rely on hand-tuned features to achieve good classification performance.

Of the two sparse learning techniques presented here, the Sparse Network seems to be the better performing and more flexible of the two. Additional hidden layers can be added to increase performance somewhat, although we did notice that additional layers do increase the likelihood of over-fitting to the training data. We also found that the Sparse Network was better at learning an over-complete basis set that the ICA+SVM method. Finally, the 'soft' classifications provided by the softmax output layer of the Sparse Network provide an intuitive and easily justified way to filter out 'inconclusive' classifications. This guarantees better classification rates if you can afford to throw out some galaxies that are difficult to classify.

In the end, we have demonstrated a modest but meaningful improvement over existing techniques, with classification rates of 99% for 'clean' galaxies that are easy for humans to classify. For the full data set we have demonstrated classification rates for elliptical and spiral galaxies of 93-94%; and this number could be increased even higher by raising the threshold and rejecting even more 'inconclusive' galaxies. Considering that human agreement with the full data set is sometimes less than 50% and often less than 80%, this may be approaching the best performance achievable using Galaxy Zoo data.

Bibliography

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

N. M. Ball, et. al. Robust Machine Learning Applied to Astronomical Data Sets. I. Star-galaxy Classificiation of the Sloan Digital Sky Survey DR3 using Decision Trees. The Astrophysical Journal. Vol. 650. pp. 497-509. (2006)

S. V. David and J. L. Gallant. Predicting Neuronal Responses during Natural Vision. Network: Computations in Neural Systems. Vol. 16(2/3). pp. 239-260. (2005)

C. Elting, C. A. L. Bailer-Jones, K. W. Smith. Photometric Classification of Stars, Galaxies and Quasars in the Sloan Digital Sky Survey DR6 Using Support Vector Machines. American Institute of Physics Conference Series. Vol. 1082. pp. 9-14 (2008)

O. Lahav, et. al. Galaxies, Human Eyes and Artifical Neural Networks. Science. pp. 859-862. (1995)

Y. Lecun, et. al. Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE. Vol. 86. No. 11. pp. 2278-2324 (1998)

C. Lintott, et. al. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society. Vol. 389. pp. 1179–1180 (2008)

C. Lintott, et. al. Galaxy Zoo 1 : Data Release of Morphological Classifications for nearly 900,000 galaxies. Monthly Notices of the Royal Astronomical Society. pp. 1–14 (2010)

A. Ng. CS294A Lecture Notes: Sparse Autoencoder. Retrieved Nov. 17th, 2010 from http://www.stanford.edu/class/archive/cs/cs294a/cs294a.1104/sparseAutoencoder.pdf

S.C. Odewahn, M.L. Nielsen. Star-Galaxy Separation using Neural Networks. Vistas in Astronomy. Vol. 38, pp. 281-286. (1994)

M. A. Nieto-Santisteban, A. S. Szalay, and J. Gray. ImgCutout, an Engine of Instantaneous Astronomical Discovery. ADASS XIII ASP Conference Series, Vol. XXX (2004)

Sloan Digital Sky Survey. Skyserver retrieved Nov. 17th, 2010 from http://cas.sdss.org/astro/en/tools/search/sql.asp

D. J. York, et. al The Sloan Digital Sky Survey: Technical Summary. The Astronomical Journal, Volume 120, Issue 3, pp. 1579-1587. (2000)