Morphological Galaxy Classification Using Machine Learning

Siddhartha Kasivajhula, Naren Raghavan, and Hemal Shah

Abstract

We compare the classification performance of three machine learning algorithms: Support Vector Machines (SVM), Random Forests (RF) [18], and Naïve Bayes (NB) as applied to morphological galaxy classification. Using both a set of morphic features derived from image analysis and the direct image pixel data compressed through PCA (Principal Component Analysis) into PCA features, we compare the performance of the different ML algorithms on each feature representation of a galaxy. Our experiments show that RF performed better than SVM and NB. Also, morphic features were more effective than our PCA features.

1 INTRODUCTION

Galaxies are gravitationally bound celestial entities composed of gas, dust, and billions of stars (and also Dark Matter as we now know, though this is irrelevant to our investigation). Galaxies form over billions of years, and their morphology – essentially their shape and general visual appearance – gives astronomers much information about their composition and their evolution. Galaxy classification is important because astrophysicists frequently make use of large catalogues of information to test existing theories against, or to form new conjectures to explain the physical processes governing galaxies, star-formation, and the nature of the universe. Currently, astronomers manually classify galaxies based on visual inspection of photographs. This method is slow, and is certainly not a worthy activity for an astronomer to be engaged in. This method is also prone to human error, and thus accounts for some inaccuracies and misclassifications.

Astronomy has recently seen an explosion of data, as programs like the Sloan Digital Sky Survey (SDSS) will generate nearly 50 million images of galaxies alone. Since access to this amount of data has only become possible in the past decade, computer aided celestial classification is a very young area, with much scope for machine learning and image processing application. Our goal is to apply machine learning algorithms to the repetitive task of galaxy classification on a massive data set. This will not only decrease classification error, but will also allow astronomers to pursue more stimulating tasks.

Other attempts have been made to apply neural networks [1], locally weighted regression using principal component analysis [2], and Naïve Bayes [3] classification techniques to this problem with varying success. Calleja and Fuentes [2] have achieved a 90% success rate for two classifications (spiral and elliptical) using locally weighted regression and 310 training examples. Goderya and Lolling [1] achieved 97% success on 171 training examples using neural networks, but only 57% success on test data. Other attempts at using neural networks have used features extracted from the images as well as raw pixel data as inputs to the neural networks.

We explore the effectiveness of various features extracted from galaxy image data, and the performance of different machine learning algorithms. The paper is structured in the following way: Section 2 introduces the Hubble tuning fork scheme for classifying galaxies. Section 3 discusses the system architecture in detail, including image preprocessing, feature extraction, and the machine learning techniques we used. Section 4 presents experimental results, and our conclusions are in Section 5.

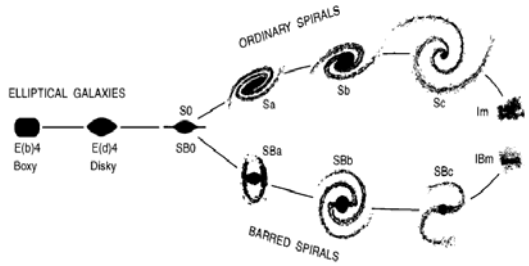
2 GALAXY CLASSIFICATION

Morphological galaxy classification is a system used by astronomers to classify galaxies based on their structure and appearance. The most

common classification scheme is the system devised by Sir Edwin Hubble in 1936. He proposed the following classifications:

Elliptical: E0, E3 , E5, E7
Spiral: S0, Sa, Sb, Sc
Barred spiral: SBa, SBb, SBc
Irregular: Im, IBm

This scheme is commonly referred to as the "Hubble Tuning Fork" and is traditionally depicted as shown in the figure below (the motivation behind this depiction is actually now known to be fallacious, but that is another story):



In classifying galaxies, we proceeded in the following manner: First, we used 3 classifications – Elliptical, Spiral, Irregular. Then, we used 7 classifications – E0, E7, Sa, Sc, SBa, SBc, I.

3 SYSTEM ARCHITECTURE

The architecture of our system is divided into three main phases as shown below and is implemented in Matlab. In the **Image Preprocessing** phase, each galaxy is individually scaled, rotated, cropped, and centered to appear uniform for more accurate feature extraction. Then, in **Feature Extraction**, we measured 6 quantities which we call morphic features for each image. We also compressed our images using Principal Component Analysis (PCA) to derive PCA features [2]. Finally, in the **Classification** stage, we trained and predicted classifications with these features using Support Vector Machines (SVM), Random Forest (RF), and Naïve Bayes (NB) machine learning classifiers.

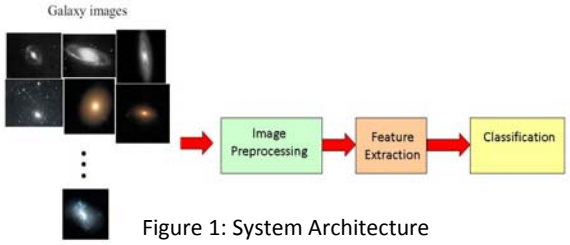


Figure 1: System Architecture

Image Preprocessing. We first applied a pre-determined threshold to pixel intensities to remove faint, extraneous noise. A binary image was then formed from the remaining pixels. We calculated the center of brightness using a simple centroiding technique. Using the (x, y) coordinate of each remaining pixel, we calculated the two principal components of the image using the SVD technique. The angle of the largest PC was used to rotate the main axis of the galaxy. We then projected the image pixels onto the PC vector basis and removed any pixels with locations outside of 3 standard deviations of the mean, effectively removing bright stars and other extraneous objects within the image. Afterwards, we scaled the images to a uniform size of 128 x 128.

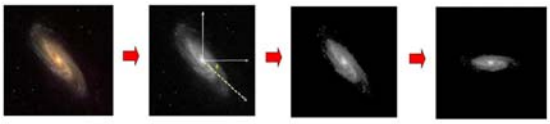


Figure 2: Image Preprocessing

Feature Extraction. The preprocessed images are then sent to the Feature Extraction phase where we calculated 6 morphic features from each image and generated PCA features. The morphic features are based on the perceived visual characteristics of the galaxy: elongation, form-factor, convexity, bounding-rectangle-to-fill-factor, bounding-rectangle-to-perimeter [1], and asymmetry index [15]. We used Canny edge detection with 5 standard deviations for the applied Gaussian filter and a threshold value of 0.5 to eliminate faint objects in the image. Then, we fitted an ellipse to the convex hull of the remaining pixels to calculate

most of our morphic features. Elongation is defined as (a – b) / (b + a), where a and b are the major and minor axes of our ellipse. Form-factor is the ratio of the area of the galaxy (number of pixels in the galaxy) to its perimeter (number of pixels in Canny edge detection). Convexity is the ratio of the galaxy perimeter to the perimeter of the minimum bounding rectangle. Bounding-rectangle-to-fill-factor (BFF) is the area of the bounding rectangle to the number of galaxy pixels within the rectangle. The asymmetry index is calculated by rotating the image 180 degrees and comparing its pixel intensities with those of the original, which has shown to be effective at differentiating spiral (high asymmetry) from elliptical (low asymmetry) galaxies.

The PCA features [2] were also calculated for each image. We calculated the principal component vector basis of the entire training set where each image was represented as a row vector. Then, we used the coefficients of each image projected (compressed) into this basis as a set of features [13], [17]. We compressed the images using PC vector bases of 8 and 24 elements, preserving 70% and 85% of the original image, respectively [2].

Classification. We then trained SVM, RF, and NB classifiers on training subsets and measured their classification accuracy on test subsets. We used libSVM [19], WEKA Random Forest [22], and Dr. Saeed Hashemi's Naive Bayes [20] classifier. We also explored using AdaBoost through the MATLABArsenal [21] package using each of the algorithms as a weak classifier.

4 EXPERIMENTAL RESULTS

Our data set is comprised of 119 images along with their classification labels, obtained from Zsolt Frei's galaxy catalog [23]. The catalog contains high-resolution images with minimal background noise.

After extracting features from the images, we evaluated the performance of the classification algorithms using various combinations of input features. We used 6 learning algorithms: SVM with RBF kernel, RF with 10 trees, NB, and their respective AdaBoosted versions. For each of these algorithms, we classified galaxies into 3 and 7 classes using only morphic features, only PCA features, and both morphic and PCA features combined. In addition, we used both 8 and 24 principal components.

In order to maximize the information provided by our data set, we implemented 10-fold cross validation with the hold-out images randomly selected from the overall training set (without replacement) at each iteration. Since the training set differs at each iteration, we recalculated the PC vector basis each time. We ran cross validation 3 times and reported the average of these runs as our overall accuracy and standard deviation for each algorithm to normalize resulting random variations due to the data distribution in the folds. We calculated the mean accuracy and standard deviations for both the training and testing sets.

Table 1 below shows the mean accuracies and standard deviations we obtained for each of the classifiers. *Ind* denotes individual classifiers, *std* denotes standard deviation, *M* denotes using only morphic features, *nPC* denotes using only *n* principal components, and *M+nPC* denotes using both morphic features and *n* principal components.

From our results, RF with only morphic features performed best. SVM performed better than NB in most cases. Also, classification using morphic features alone was more effective than using PCA features, and the optimal number of PCA features was found to be between 8 and 24.

SVM								
	3 class				7 class			
Features	Ind Mean	Ind std	AdaBoost Mean	AdaBoost std	Ind Mean	Ind std	AdaBoost Mean	AdaBoost std
M	80.41	16.17	80.41	16.17	20.12	16.06	7.56	8.68
8PC	80.99	14.68	81.54	13.89	24.81	11.67	7.56	7.94
24PC	79.57	17.45	72.62	14.31	28.36	16.27	10.62	10.6
M+8PC	78.18	19.89	80.4	16.43	17.78	15.73	7.56	8.31
M+24PC	80.49	14.41	80.49	14.41	26.88	12.7	7.5	8.66
Random Forest								
	3 class				7 class			
Features	Ind Mean	Ind std	AdaBoost Mean	AdaBoost std	Ind Mean	Ind std	AdaBoost Mean	AdaBoost std
M	85.72	12.6	81.02	13.72	27.47	10.7	14.35	15.41
8PC	81.11	11.9	79.66	12.79	17.65	11.47	9.91	9.86
24PC	77.93	13.25	79.85	15.9	22.9	12.63	8.95	9.03
M+8PC	83.06	10.5	78.8	14.13	25.83	12.7	11.57	11.4
M+24PC	80.8	11.99	80.49	14.41	30.03	13.57	10.83	12.79
Naïve Bayes								
	3 class				7 class			
Features	Ind Mean	Ind std	AdaBoost Mean	AdaBoost std	Ind Mean	Ind std	AdaBoost Mean	AdaBoost std
M	79.91	16.62	76.23	15.74	7.56	8.68	7.59	8.45
8PC	73.09	18.42	70.82	19.14	23.55	14.77	22.92	14.32
24PC	72.25	17.71	72.56	16.46	23.79	13.86	23.41	12.98
M+8PC	80.4	16.43	75.15	18.63	7.56	8.31	7.67	8.27
M+24PC	80.49	14.41	79.86	17.82	7.5	0.79	7.61	0.67

Table 1: Classification Test Accuracy

5 CONCLUSIONS AND FUTURE WORK

We believe that RF performed better than SVM in general because our features were not linearly separable even when projected to higher dimensions with the RBF kernel. RF, with overfitted decision trees, seemed more robust with such nonlinear data. Our AdaBoost results are intriguing because we expected that AdaBoosting would have improved classification accuracy.

We believe that morphic features were more effective than PCA features because they are less susceptible to morphological variations of galaxies in the same category. PCA features depend on raw pixel data, which is more susceptible to these same morphological variations. This is consistent with the results, which showed that too many PCs in our PCA feature set (preserving more raw data) degraded classification performance.

The accuracy of our calculated morphic features was affected by several factors. The cloud of stars and dust surrounding the galactic core resulted in spurious edges, making it difficult to determine the perimeter of the galaxy. The bounding ellipse may be distorted by bright background stars that impact the calculation of the convex hull, faint spiral arms that are missing due to thresholding, and edge-on galaxy images.

Future work to improve classification accuracy includes incorporating more training data and integrating photometric features measured in different spectra [1] with morphic features. Implementing the bar-to-bulge ratio and the integrated curvature morphic features [5] may also boost the accuracy on 7-category classification. Additionally, we can classify galaxies in stages: first weakly classify a large number of galaxies into 3 categories, and then further classify these into 7 categories using this a priori knowledge.

6 REFERENCES

1. Goderya, S.N., Lolling, S.M., “Morphological Classification of Galaxies Using Computer Vision and Artificial Neural Networks: A Computational Scheme.” Astrophysics and Space Science. Vol. 279, no. 4, pp. 377 – 387.
2. Calleja, J., Fuentes, O., “Machine Learning and Image Analysis for Morphological

Galaxy Classification." Monthly Notices of the Royal Astronomical Society, Vol. 24, 2004, pp. 87-93.
3. Calleja, J., Fuentes, O., Automated Classification of Galaxy Images. Lecture Notes in Computer Science, Vol. 3215, 2004, pp. 411-418.
4. Bazell, D., Aha, D., "Ensembles of Classifiers for Morphological Galaxy Classification". 2001.
5. Au, K., Genovese, C. Connolly, A., "Inferring Galaxy Morphology Through Texture Analysis". 2006.
6. Calleja, J., Fuentes, O., "Automated Classification of Galaxy Images". 2003?
7. Goderya, S., Andreasen, J.., Philip, N.S., "Advances in Automated Algorithms for Morphological Classification of Galaxies Based on Shape Features." 2004.
8. Lekshmi, S., Revathy, K., Nayar, S.R.P. "Galaxy Classification Using Fractal Signature." 2003.
9. Ofer Lahav, "Galaxy Classification by Human Eye and by Automated Algorithms". 2003.
10. Zhang, Y., Zhao, Y., "Classification in Multidimensional Parameter Space: Methods and Examples." 2003.
11. Odewahn, S., "Galaxy Classification Using Artificial Neural Networks."
12. Odewahn, S.C., "Automated Galaxy Classification in Large Sky Surveys." 1999.
13. Turk, M., Pentland, A., "Face Recognition Using Eigenfaces." 1991.
14. N. M. Ball, J. Loveday, et. all. "Galaxy Types in the Sloan Digital Sky Survey Using Supervised Artificial Neural Networks." Mon. Not. R. Astron. Soc. 348, p.1038-1046. 2004.
15. Abraham, R. G., Tanvir, N. R., "Galaxy morphology to I=25 mag in the *Hubble Deep Field*."
16. Bazell, D., "Feature Relevance in Morphological Galaxy Classification." Mon. Not. R. Astron. Soc. 316, 519-528. 2000.
17. Turk, M., Pentland, A. "Eigenfaces for Recognition", Journal of Cognitive Neuroscience, vol. 3, no. 1, pp.71-86, 1991.
18. Brieman, L. "Random Forests", Machine Learning, 45(1), 5-32, 2001.
19. Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
20. Heshami, Saeed. Naïve Bayes classifier. Homepage at http://torch.cs.dal.ca/~saeed
21. Yan, Rong. MATLABArsenal. Software available at http://finalfantasyxi.inf.cs.cmu.edu/MATLABArsenal/MATLABArsenal.htm
22. Ian H. Witten and Eibe Frank (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.
23. Frei, Z., and Gunn, J. available at http://www.astro.princeton.edu/~frei/catalog.htm

NOTE: We consulted with several astronomers, many of whom are directly involved in the Sloan Digital Sky Survey; we have chosen to do morphological galaxy classification based on the advice that we received from them. They identified this as one of the most cumbersome areas in celestial classification, and the one that has proven the most difficult to automate. Following is a list of some of the astronomers who are advising us in this project:

David Weinberg, Scientific Spokesperson for the Sloan Digital Sky Survey.
Dr. Cecilia Barnbaum, Professor of Astronomy, Valdosta State University.
Dr. Kenneth Rumstay, Professor of Astronomy, Valdosta State University.
Robert Brunner, Asst. Professor of Astronomy, University of Illinois at Urbana-Champaign (who has employed machine learning algorithms for celestial classification before, and is intimately familiar with the research in this area).
Chris Lintott, Presenter for the BBC's Sky at Night TV program, and maintainer of galaxyzoo.com.