Automated Classification of Galaxy Images

2 authors:

Jorge de la Calleja
Universidad Politécnica de Puebla

43 PUBLICATIONS   **128** CITATIONS

Olac Luis Fuentes
University of Texas at El Paso

137 PUBLICATIONS   **1,094** CITATIONS

Some of the authors of this publication are also working on these related projects:

RedICA: Red temática CONACYT en Inteligencia Computacional Aplicada View project

SEXWISE: IBM Watson-powered Mobile App View project

Automated Classification of Galaxy Images

Jorge de la Calleja and Olac Fuentes

Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Enrique Erro 1,
Tonantzintla 72840, Puebla, México
jorge@ccc.inaoep.mx, fuentes@inaoep.mx

Abstract. In this paper we present an experimental study of the performance of three machine learning algorithms applied to the difficult problem of galaxy classification. We use the Naive Bayes classifier, the rule-induction algorithm C4.5 and a recently introduced classifier named random forest (RF). We first employ image processing to standardize the images, eliminating the effects of orientation and scale, then perform principal component analysis to reduce the dimensionality of the data, and finally, classify the galaxy images. Our experiments show that RF obtains the best results considering three, five and seven galaxy types.

1 Introduction

The morphology of galaxies is generally an important issue in the large scale study of the Universe. Galaxy classification is the first step towards a greater understanding of the origin and formation process of galaxies, and the evolution processes of the Universe [10]. Galaxy classification is important for two main reasons. First, to produce large catalogues for statistical and observational programs, and second for discovering underlying physics [7].

In recent years, with numerous digital sky surveys across a wide range of wavelengths, astronomy has become an immensely data-rich field. For example, the Sloan Digital Sky Survey [1] will produce more than 50,000,000 images of galaxies in the near future. This overload creates a need for techniques to automate the difficult problem of classification. Several methods have been used to solve this problem, such as neural networks [4, 6, 7, 9, 10, 13], oblique decision trees [11], ensembles of classifiers [2, 4], and instance-based methods [4].

We propose an approach to perform galaxy classification that first generates a representation that is independent of scale and orientation, then generates a more compact and manageable representation using principal component analysis, and finally classifies the galaxy images using machine learning algorithms. In previous work [4], we used locally-weighted regression and neural networks to perform galaxy classification, and now, we investigate the performance of three other learning algorithms: the Naive Bayes classifier, the rule-induction algorithm C4.5 and the random forest (RF) predictor. We also use ensembles of these algorithms to classify the images.

The paper is organized as follows: Section 2 gives a brief introduction of the Hubble tuning fork scheme for galaxy classification. In Section 3 we describe the

general architecture of the method, including the image analysis, data compression and learning stages. In Section 4 we show experimental results and finally in Section 5 conclusions and future work are presented.

2 The Hubble Tuning Fork Scheme

Galaxies are large systems of stars and clouds of gas and dust, all held together by gravity [1]. Galaxies have many different characteristics, but the easiest way to classify them is by their shape; Edwin Hubble devised a basic method for classifying them in this way [1]. In his classification scheme, there are three main types of galaxies: Spirals, Ellipticals, and Irregulars (Figure 1).
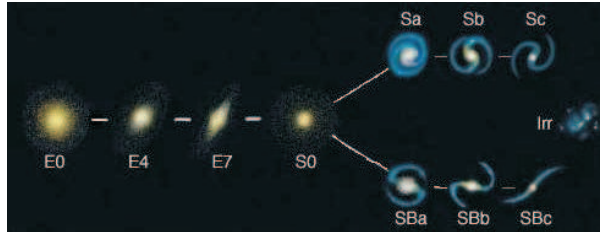

Fig. 1. The Hubble Tuning Fork Scheme.

Elliptical galaxies (E0, E4 and E7 in Figure 1) have the shape of an ellipsoid. Spiral galaxies are divided in ordinary and barred; ordinary spirals have an approximately spherical nucleus, while barred spirals have a elongated nucleus that looks like a bar. Spirals are classified as Sa, Sb, or Sc; barred spirals are labeled as SBa, SBb, or SBc. The subclassification (a, b or c) refers both to the size of the nucleus and the tightness of the spiral arms. An Sa galaxy has a bigger nucleus than an Sc galaxy, and the arms of the Sc are wrapped more loosely. S0 are spiral galaxies without any conspicuous structure in their disks. Irregular galaxies do not have an obvious elliptical or spiral shape.

3 The Classification Method

The method that we developed for galaxy classification is divided in three stages: image analysis, data compression, and machine learning (see Figure 2). The method works as follows: It takes as input the galaxy images, which are then rotated, centered, and cropped in the image analysis stage. Next, using principal component analysis, the dimensionality of the data is reduced and we find a set of features. The projection of the images onto the principal components will be the input parameters for the machine learning stage. At the end, we will have the classification of the galaxies. The following three subsections describe each part in detail.
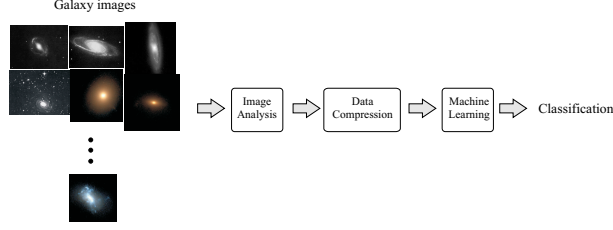
Fig. 2. The stages of the classification method.

3.1 Image Analysis

Galaxy images generally are of different sizes and color formats, and most of the time the galaxy contained in the image is not at the center. So, the aim of this stage is to create images invariant to color, position, orientation and size, all in a fully automatic manner. First, we find the galaxy contained in the image applying a threshold; that is, from the original image I, we generate a binary image B, such that

$$B(i,j) = \begin{cases} 1 \; if \; I(i,j) > threshold; \\ 0 \; otherwise. \end{cases} \tag{1}$$

Then we obtain $\bar{i}$ and $\bar{j}$, the center row and column of the galaxy in the image, given by

$$\bar{i} = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} iB(i,j) \tag{2}$$

$$\bar{j} = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} jB(i,j) \tag{3}$$

where m and n are the number of rows and columns, respectively, in the image. Then we obtain the covariance matrix of the points in the galaxy image

$$\mathbf{C} = \sum_{i=1}^{m} \sum_{j=1}^{n} B(i,j)[i - \bar{i}, j - \bar{j}]^T [i - \bar{i}, j - \bar{j}] \tag{4}$$

The galaxy's main axis is given by the first eigenvector (the eigenvector with the largest corresponding eigenvalue) of $\mathbf{C}$, the covariance matrix. We then rotate the image so that the main axis is horizontal. The angle is given by

$$\alpha = \arctan(\mathrm{p1}(1)/\mathrm{p1}(2)) \tag{5}$$

where $\mathrm{p1}(1)$ and $\mathrm{p1}(2)$ are the x and y values of the first principal component. Then we use an image warping technique to rotate the image (see Figure 3). After that, we crop the image, eliminating the columns that contain only background (black) pixels. Finally, we stretch and standardize the images to a size of 128x128 pixels. Figure 4 shows examples of the image processing stage for an elliptical galaxy, a spiral galaxy, and an irregular galaxy.
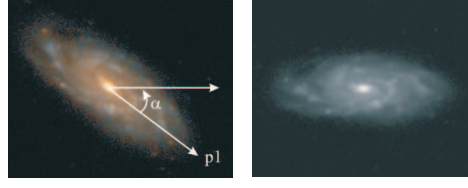
Fig. 3. Left: The first principal component (p1) is used to rotate the galaxy image. Right: Rotated galaxy.

3.2 Data Compression

Principal component analysis (PCA) is a statistical method that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components (PCs). PCA is generally used to reduce the dimensionality of a data set while retaining as much information as possible. Instead of using all the principal components of the covariance matrix, we may represent the data in terms of only a few basis vectors. We used 8, 13 and 25 PCs to perform the classification because they represent about 75%, 80% and 85% of the information, respectively, in the data set. More details about this technique can be found in [14].
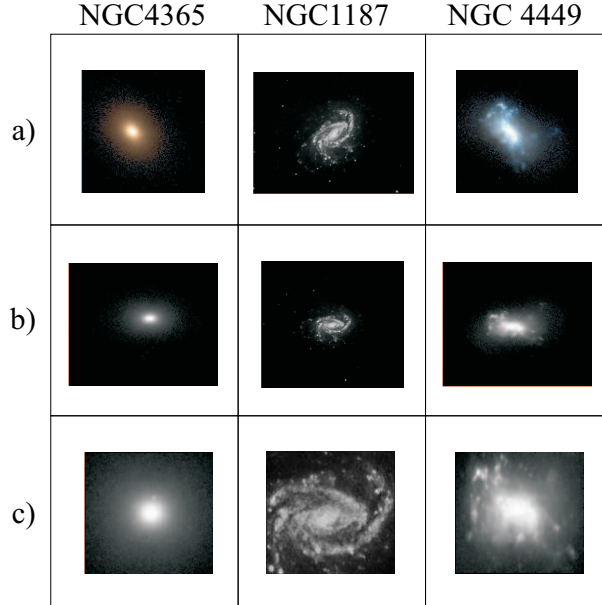

Fig. 4. Examples: a) Original images, b) Rotated images, and c) Cropped images.

3.3 Machine Learning

Naive Bayes Classifier. The Naive Bayes classifier [8] is a probabilistic algorithm based on the assumption that the attribute values are conditionally independent given the target values. The Naive Bayes classifier applies to learning tasks where each instance x can be described as a tuple of attribute values $a_1, a_2, \ldots a_n$ and the target function $f(x)$ can take on any value from a finite set V. When a new instance x is presented, the Naive Bayes classifier assigns to it the most probable target value by applying the rule:

$$f(x) = argmax_{v_j \epsilon V} P(v_j) \Pi_i P(a_i \mid v_j) \tag{6}$$

To summarize, the learning task of the Naive Bayes is to build a hypothesis by estimating the different $P(v_i)$ and $P(a_i \mid v_j)$ terms based on their frequencies over the training data.

C4.5 This method operates by recursively splitting a training set based on feature values to produce a tree such that each example can end up in only one leaf. An initial feature is chosen as the root of the tree, and the examples are split among branches based on the feature value for each example. If the values are continuous, then each branch takes a certain range of values. Then a new feature is chosen, and the process is repeated for the remaining examples. Then the tree is converted to an equivalent rule set, which is pruned. For a deeper introduction of this method we refer the reader to [8] and [12].

Random Forest Predictor. A random forest (RF) is a classifier consisting of a collection of individual tree classifiers. Basically, random forest does the following:

1. Select *ntree*, the number of trees to grow, and *mtry*, a number no larger than the number of variables.
2. For $i=1$ to *ntree*:
3. Draw a bootstrap sample from the data. Call those not in the bootstrap sample the "out-of-bag" data.
4. Grow a "random" tree, where at each node, the best split is chosen among *mtry* randomly selected variables. The tree is grown to maximum size and not pruned back.
5. Use the tree to predict out-of-bag data.
6. In the end, use the predictions on out-of-bag data to form majority votes.
7. Prediction of test data is done by majority votes from predictions from the ensemble of trees.

Details about RF can be found in [3].

Ensemble Method. An ensemble consists of a set of classifiers whose individual decisions are combined in some way, normally by voting, to classify new examples. The ensemble method used here is bagging [5]. It was chosen because this method almost always improves the accuracies obtained by individual classifiers. The idea in this ensemble is to generate randomly n training sets with the examples from the original training set, and to use each of this subsets for creating a classifier. Each subset is obtained by sampling, with replacement, from the original training set, thus some of the examples will appear more than once, while others will not appear at all.

4 Experimental Results

We test our method with 292 galaxy images. Most of them were taken from the NGC catalog on the web page of the Astronomical Society of the Pacific[1], and their classification was taken from the interactive NGC online[2] catalog. For our purpose we consider three (E, S, Irr), five (E, S0, Sa+Sb, Sc+Sd, Irr) and seven (E, S0, Sa, Sb, Sc, Sd, Irr) galaxy types.

We used the Naive Bayes classifier, J48 (a particular C4.5 implementation) and the random forest classifier that are implemented in WEKA[3], and also the bagging ensemble method. We used 10-fold cross-validation for doing all the experiments. For C4.5 we used pruning and a confidence factor of 0.25. In the case of RF, 13 trees were used for creating the forest for all the experiments, however, we select different random features, i.e. five for the three-class case and two for five and seven classes.

Table 1 shows the accuracy for each of the individual classifiers, and for the ensembles, and we also show the standard deviation. The accuracies were obtained by averaging the results of 5 runs of 10-fold cross validation for each method. The columns *Ind*, *Ens* and *std* denote individual classifier, ensemble of classifiers and standard deviation, respectively.

Table 1. Accuracy for individual classifiers and ensembles.

	Naive Bayes											
	3 classes				5 classes				7 classes			
PCs	*Ind*	*std*	*Ens*	*std*	*Ind*	*std*	*Ens*	*std*	*Ind*	*std*	*Ens*	*std*
8	83.23	0.7819	86.77	0.7893	46.02	1.3404	46.50	0.7819	41.70	1.7368	43.62	0.5725
13	80.68	0.7101	85.61	0.7283	44.53	1.1984	46.63	2.0185	37.46	2.4294	40.26	2.3314
25	75.88	0.5156	82.73	1.7241	40.33	1.1987	43.28	2.4780	34.58	1.2134	36.98	1.1596
	C4.5											
	3 classes				5 classes				7 classes			
PCs	*Ind*	*std*	*Ens*	*std*	*Ind*	*std*	*Ens*	*std*	*Ind*	*std*	*Ens*	*std*
8	88.35	0.6388	89.92	0.3080	43.01	2.2295	48.76	0.8947	40.54	1.7228	43.83	1.0561
13	87.39	0.2893	91.09	1.1865	46.77	1.9618	49.51	2.1580	38.35	3.7278	44.24	2.3525
25	86.84	1.0167	91.02	0.6582	45.81	4.2463	51.36	3.9718	36.02	2.0179	45.68	0.7880
	Random Forest											
	3 classes				5 classes				7 classes			
PCs	*Ind*	*std*	*Ens*	*std*	*Ind*	*std*	*Ens*	*std*	*Ind*	*std*	*Ens*	*std*
8	90.39	1.0208	91.22	0.3086	47.18	4.4414	50.47	0.5156	42.66	1.3413	46.16	1.4386
13	91.29	0.6677	91.64	0.1917	49.72	2.0179	51.77	0.5623	44.51	3.9993	47.12	1.4634
25	91.29	0.5180	91.64	0.5737	47.87	1.9263	54.72	2.3638	42.53	2.1143	48.62	1.7110

Analyzing the results, we can observe that RF obtained the best accuracy for all the galaxy classes, i.e. 91.64% accuracy for the three-class case, 54.72% accuracy for the five-class case, and 48.62% accuracy for the seven-class case; and the standard deviations were almost always the smallest. Only in the seven-class case, Naive Bayes obtained a smaller standard deviation than RF with

---

[1] www.apsky.org/ngc/ngc.html

[2] www.seds.org/∼spider/ngc/ngc.html

[3] WEKA is a software package that can be found at www.cs.waikato.ac.nz/ml/weka

0.5725, but its accuracy was of 43.62%. We can also note that in all cases ensembles obtained better results than individual classifier. Examining the results considering the number of PCs, we can say that 13 are enough to perform the classification, obtaining good results. This way we can reduce computation by using few attributes.

5   Conclusions

We presented a method that performs morphological galaxy classification in a fully automatic manner producing good results. The use of standardized images helps to improve the accuracy of the learning algorithms. We have shown experimentally that a small number of principal components is enough to classify the galaxies. Also, the ensemble permits to improve the classification accuracy.

Future work includes testing this method for other types of astronomical objects, such as nebulas and clusters, and extending the system to deal with wide-field images, containing multiple objects.

References

1. Ball, N. Morphological Classification of Galaxies Using Artificial Neural Networks. Master's thesis, University of Sussex, 2002
2. Bazell, D., Aha, D.W. Ensembles of Classifiers for Morphological Galaxy Classification. The Astrophysical Journal, 548:219-233, 2001
3. Breiman, L. Random Forests, Machine Learning, 45(1), 5-32, 2001
4. De la Calleja, J., Fuentes, O. Machine learning and image analysis for morphological galaxy classification, Monthly Notices of the Royal Astronomical Society, 349:87-93, 2004
5. Dietterich, T.G. Machine Learning Research: Four Current Directions. AI Magazine, 18(4):97-136, 1997
6. Goderya, S. N., Lolling, S.M.. Morphological Classification of Galaxies using Computer Vision and ANNs. Astrophysics and Space Science, 279(377), 2002
7. Lahav O. Artificial neural networks as a tool for galaxy classification, in Data Analysis in Astronomy, Erice, Italy, 1996
8. Mitchell, T. Machine Learning. McGraw Hill, 1997
9. Madgwick, D.S. Correlating galaxy morphologies and spectra in the 2dF Galaxy Redshift Survey. Monthly Notices of the Royal Astronomical Society, 338:197-207, 2003
10. Naim, A., Lahav O., Sodré, L. Jr., Storrie-Lombardi M.C. Automated morphological classification of APM galaxies by supervised artificial neural networks. Monthly Notices of the Royal Astronomical Society, 275(567), 1995
11. Owens, E.A., Griffiths, R.E., Ratnatunga K.U. Using Oblique Decision Trees for the Morphological Classification of Galaxies. Monthly Notices of the Royal Astronomical Society, 281(153), 1996
12. Quinlan, J.R. Induction of decision trees. Machine Learning, 1(1):81-106, 1986
13. Storrie-Lombardi, M.C., Lahav, O., Sodré, L., Storrie-Lombardi, L.J. Morphological Classification of Galaxies by Artificial Neural Networks. Monthly Notices of the Royal Astronomical Society, 259(8), 1992
14. Turk, M.A., Pentland, A.P. Face Recognition Using Eigenfaces, in Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, 586-591, 1991