Highlights of Spanish Astrophysics VI, Proceedings of the IX Scientific Meeting of the Spanish Astronomical Society held on September 13 - 17, 2010, in Madrid, Spain. M. R. Zapatero Osorio et al. (eds.)

Revisiting the Hubble sequence in the SDSS DR7

M. Huertas-Company[1,2], J. A. L. Aguerri[3], M. Bernardi[4], S. Mei[1,2], and J. Sánchez Almeida[3]

[1] GEPI- Observatoire de Paris
[2] Université Paris 7 Denis-Diderot
[3] Instituto de Astrofísica de Canarias
[4] Department of Physics and Astronomy, University of Pennsylvania

Abstract

We present an automated morphological classification in 4 types (E, S0, Sab, Scd) of $\sim$ 700.000 galaxies from the SDSS DR7 spectroscopic sample based on support vector machines. The main new property of the classification is that we associate a probability to each galaxy of being in the four morphological classes instead of assigning a single class. The classification is therefore better adapted to nature where we expect a continuous transition between different morphological types. The algorithm is trained with a visual classification and then compared to several independent visual classifications including the Galaxy Zoo first-release catalog. We find a very good correlation between the automated classification and classical visual ones. The compiled catalog is intended for use in different applications and is therefore publicly available through a dedicated webpage

1 Introduction

Classification of objects is a key step in understanding and analyzing an astrophysical sample. In particular, morphology is a powerful tracer of the structure of a galaxy. Since Hubble's first classification of galaxies according to their shape [3], it has been shown that this phenomenological description hides important physical differences between galaxies and probably different evolutionary tracks. Elliptical galaxies appear with old stellar populations, high velocity dispersion, and small fraction of gas while spiral galaxies are more gas-rich, with younger stellar populations whose motion is rotation dominated. The main problem with morphology comes from estimation, since, even when done through visual inspection, there are several intrinsic problems that can hardly be overcome. What defines a given galaxy type? Is it just a shape and bulge fraction? or is it shape and stellar populations? or is it stellar dynamics? Almost eighty years after Hubble's definition, these questions remain

unanswered. It seems that, instead of being a closed definition, there is more like a continuous population of galaxies with some *canonical* objects, prototypes of elliptical, or spiral galaxies and then some galaxies that are more or less close to the definition. Consequently, it makes more sense to assign distances or probabilities of being in one of the canonical classes instead of having a binary definition that is not necessarily very close to reality. Lots of effort has been made to try to determine morphology in an automated and simple way by measuring some parameters, such as concentration, asymmetry, clumpiness, Gini index (e.g. [1]) however, all these methods deal with a finite number of classes and/or at some point require a degree of human intervention. In [4, 5] we presented a method based on support vector machines (galSVM). It was initially designed for high-redshift galaxies, and it has the advantage of dealing with an unlimited number of parameters and assigning probabilities instead of binary classes. In this paper, we revisit the Hubble sequence in the SDSS DR7 spectroscopic sample using this method and assign a probability to each galaxy of being in the following morphological classes: E, S0, Sab, Scd, instead of a closed class.

2 The sample

We used all the SDSS DR7 spectroscopic sample as the starting base. Then, the selection of objects was based on [9] who performed an unsupervised automated classification of all the SDSS spectra. Basically, we chose galaxies with redshift below 0.25, and with good photometric data and clean spectra, meaning objects not too close to the edges, not saturated, or not properly deblended. The final catalog contains 698420 objects for which we estimate the morphology as shown below. No additional selection criteria were added so that the catalog is not biased to any particular application.

3 Method

The classification method is based on support vector machines (SVM) implemented in the libSVM library [1]. SVM is a machine learning algorithm that tries to find the optimal boundary (not necessarily linear) between several clouds of points in an N-dimensional space. More information about the algorithm can be found in [4]. There are several interesting properties that make this algorithm attractive for galaxy classification. First, it can deal with an unlimited number of dimensions so that everything that is related to the classes one would like to separate can be included in the classification process. Second, it does not deliver a binary classification but a probability of belonging to a given class. This probability is related to the accuracy of the classification, the higher it is, the higher the success rate (and so the closer are the objects to the *canonical classes*), so that the accuracy of the classification can be studied in an objective way. This property is lacking in most of the existing classification schemes (specially in the visual techniques).

The SVM method needs a training sample, and all the behavior of the learning algorithm depends on how close this training sample is to the real sample one wants to classify.

[1] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

Figure 1: Examples of galaxies with their computed probability values

For morphological classification, the training sample is typically built using a visually classified subsample. We therefore used [2] classification as the training sample. In their paper, they provide a visual classification of 2253 SDSS galaxies brighter than $m_r = 16$ (compared to the full DR7 sample, which goes up to $m_r \sim 18$). Since our goal is to classify galaxies in 4 main classes (E, S0, Sab, Scd), we group them according to their morphological index T (Table 1 in [2]): E: $T < 1$, S0: $T = 1$, Sab: $2 < T < 4$, and Scd: $4 \leq T < 7$ before using them for training the algorithm. We included irregulars ($T = 6$) in the Scd class since there are not enough objects in the local universe (and in particular in the [2] catalog) to make a separate class for the training. SVM were originally thought to separate 2 classes. Some implementations were done to add multi-class separation but the accuracy is more difficult to assess. To avoid dealing with multi class problems, in this paper we proceeded in two steps. First we separated the sample in two main classes, i.e. early-type galaxies, which includes ellipticals and S0 galaxies, and late-type galaxies, which contain all the remaining morphological types from Sa to Scd/Im. Then we took the whole sample and classified it again using 2 different training sets that contain only early-type and late-type galaxies respectively. The probability computed in this second step can thus be seen as a conditional probability: "probability of being S0 or E given that it is an early-type galaxy" and "probability of being Sab or Scd given that it is a late-type galaxy". With this approach we were certain to have a broad classification in two types (which is enough for lots of science applications) with a high success rate, and then a more detailed one. Each galaxy in the catalog is therefore associated with 6 probability values, i.e. the probability of being in the two broad classes and the probability of being in the 4 subclasses. See Fig 1 for some examples.

4 Comparison with visual classifications

4.1 Comparison with Nair & Abraham (2010)

In a recent paper, Nair & Abraham [8] published a very detailed visual catalog of 14034 galaxies in the SDSS with $m_g < 16$. Galaxies in this sample are included in our classification, but most of them have not been used to build our training sample so they represent an ideal independent cross check. Since [8] classification is much more detailed than ours, we group their classes into 4 groups matching the 4 classes we have defined in this work. We consider elliptical galaxies objects with $TType = -5$, S0s, $TType = -2$, Sabs, $1 \leq TType \leq 3$, and finally Scd, $5 \leq TType \leq 10$ (see Table 1 of [8] for a definition of the $TType$ index used in their work). Globally, we observe a good correlation between the probability values and the visual class. For example, galaxies visually classified as ellipticals have on average a probability of ~ 0.8 of being ellipticals and ~ 0.2 of being S0. The two other probabilities are almost zero. Another interesting measurement is the fraction of *catastrophic* classifications, i.e. galaxies whose automated and visual classes are completely different. We define those cases as objects for which $P(E) > 0.8$ and $TType > 5$ or $P(\mathrm{Scd}) > 0.8$ and $TType = -5$, i.e. galaxies that are clearly elliptical (Scd) for our algorithm and visually classified as Sc or later (elliptical). There are only 2 objects verifying these conditions, and both are in the first case. They are indeed spiral galaxies, so the algorithm is wrong, but both have a large red bulge, which can probably account for the misclassification.

4.2 Galaxy Zoo

Recently, the Galaxy Zoo[2] team [7] has made publicly available the visual classification of the full DR7 performed through the aggregated efforts of hundreds of thousands of people over the course of many months. This work is an extraordinary effort (and probably the only way) to visually classify present and future extremely large surveys. The main drawback, however, is that it requires plenty of time (more than 2 years in this case) to collect all the information and put all the catalogs in place. It is therefore a very interesting question to see how our automated classification behaves compared to this visual classification. Our classification is indeed much faster and can be run several times with different parameters in just a few minutes, but it is not obvious whether we can reach an accuracy similar to the human brain. The classification made in the framework of the GalaxyZoo is less detailed than a *pure* visual classification, such as the one from [8] or [2]; i.e, they basically asked people if the galaxy is elliptical like (which should include S0s) or spiral like (with different subcategories like clockwise or anti-clockwise rotation), but without submorphological types. The confidence of the classification in the current release is measured by the fraction of votes received, since each galaxy is classified by several persons. A galaxy is then flagged as early-type or spiral-like if the fraction of votes in one of those categories is greater than 80%. We observe an extremely good correlation between both classifications even for faint galaxies not necessarily well represented in the training set. Galaxies flagged as ellipticals in the Galaxy

[2] http://galaxyzoo.org/

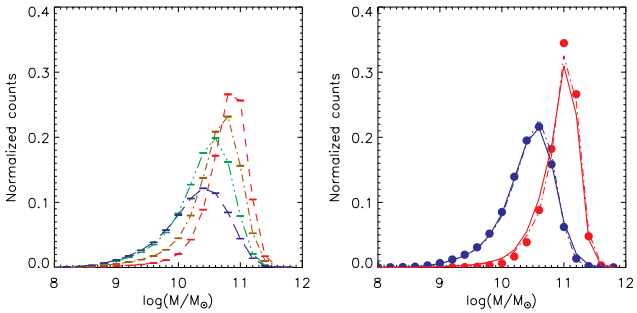

Figure 2: Observed distribution of masses for different morphological types computed using different estimators described in the text (see text for details). In the left panel the whole sample is shown using the probability weighting. Red short dashed line: ellipticals; yellow dashed dotted line: S0s; green dashed three dotted line: Sabs; blue long dashed line: Scds. In the right panel, we show galaxies flagged as SPIRAL and ELLIPTICAL in the galaxy zoo. Red solid lines are galaxies flagged as ellipticals in Galaxy Zoo (FLAG ELLIPTICAL = 1), red dashed line is the distribution obtained using probability weighting and red dots are galaxies with $p(\mathrm{E}) > 0.5$. Blue solid lines are galaxies flagged as spirals (FLAG SPIRAL = 1) in the Galaxy Zoo, blue dashed line is the distribution obtained using probability weighting and blue dots are galaxies with $p(\mathrm{Sab}) + P(\mathrm{Scd}) > 0.5$.

Zoo catalog have a median probability of 0.92 of being elliptical or S0 and the same for galaxies classified as spirals. This means that *robust* classifications in Galaxy Zoo are also very sure classifications in our catalog.

5 How to use the catalog?

The most important new point of the classification presented in this work is the measurement of probabilities. Therefore, a morphological class is not defined as a closed box, but there is more like a continuous transition from one class to another. How can this new property can be used for selecting a particular population and studying its properties? If one wants to perform luminosity or mass functions for a given morphological type, the optimal way (in terms of optimal estimation) is to make use of the probability measure as a weight for the galaxy counts. All the galaxies contribute then to the mass function of a given morphological type weighted by its probability. As a result, a galaxy that is 95% Sd and 0.5% E will still contribute to the mass function of elliptical galaxies with a weight of 0.005.

Another approach is to make probability cuts. This way, we decide that galaxies belong to a given class by applying a probability threshold. This approach (even if not optimal) should be closer to the classical approach from visual classifications in which galaxies only contribute in one given class. The threshold to apply depends on the application. As illustration, we show in Fig. 2 the mass distributions obtained for the full sample with both

estimators as well as a comparison with the Galaxy Zoo classification.

6 Summary and conclusions

We have presented an automated morphological classification of the SDSS DR7 spectroscopic sample. The algorithm used is based on SVM, and the most interesting and new property is that it associates a probability value to each galaxy instead of a single class. This way, the transition between one class and another is continuous, which should be a better approximation to nature and to visual classifications. The results obtained are in good agreement with existing visual classifications and are robust even at the faint end of the sample. The probability measurements can be used as a weighting factor for computing statistical quantities, such as luminosity or mass functions, or as a selection criterion to be sure that a *cleaned* sample of galaxies is selected. The classification is intended for use in many different applications and is therefore freely available at `http://gepicom04.obspm.fr/sdss_morphology/Morphology_2010.html` and soon from the CasJobs database.

References

[1] Conselice, C. J., Bershady, M. A. &, Jangren, A. 2000, ApJ, 529, 886

[2] Fukugita, M., Nakamura, O., Okamura, S., Yasuda, N., Barentine, J. C., Brinkmann, J., Gunn, J. E., Harvanek, M., Ichikawa, T., Lupton, R. H., Schneider, D. P., Strauss, M. A., & York, D. G. 2007, AJ, 134, 579

[3] Hubble, E. P. 1926, ApJ, 64, 321H

[4] Huertas-Company, M., Rouan, D., Tasca, L., Soucail, G., & Le Fèvre, O. 2008, A&A, 478, 971

[5] Huertas-Company, M., Tasca, L., Rouan, D., Pelat, D., Kneib, J. P., Le Fèvre, O., Capak, P., Kartaltepe, J., Koekemoer, A., McCracken, H. J., Salvato, M., Sanders, D. B., & Willott, C. 2009, A&A, 497, 743

[6] Huertas-Company, M., Aguerri, J. A. L., Bernardi, M., Mei, S., & Sánchez Almeida, J. 2011, A&A, 525, 157

[7] Lintott, C., Schawinski, K., Bamford, S., Slosar, A., Land, K., Thomas, D., Edmondson, E., Masters, K., Nichol, R., Raddick, J., Szalay, A., Andreescu, D., Murray, P., & Vandenberg, J, 2011, MNRAS, 410, 166

[8] Nair, P. B., & Abraham, R. G. 2010, ApJS, 186, 427

[9] Sánchez Almeida, J., Aguerri, J. A. L., Muñoz-Tuñón, C., & de Vicente, A. 2010, ApJ, 714, 487