Boletim da Sociedade Astronômica Brasileira, **30**, no. 1, 180-181


# Improving galaxy morphology with machine learning

P. H. Barchi[1], R. Sautter[1], F. G. da Costa[2], T. C. Moura[1], D. H. Stalder[1], Rosa, R.R.[1], e R.R. de Carvalho[1]

[1] National Institute for Space Research (INPE), São José dos Campos, SP, Brazil
e-mail: paulobarchi@gmail.com
[2] University of São Paulo (USP), São Carlos, SP, Brazil

**Abstract.** This work presents a new non-parametric approach to study galaxy morphology and application of machine learning experiments aiming to distinguish ellipticals (E) and spirals (Sp). We measure morphology with the following parameters: concentration (C), asymmetry (A), smoothness (S), entropy (H) and gradient pattern analysis parameter (GA). The dataset used for supervised learning experiments consists of 48,145 objects, with 44,760 galaxies labeled as Sp and 3,385 as E. The results are evaluated with metrics like precision ($P = TP/(TP + FP)$) and recall ($R = (TP/(TP + FN)$) for each galaxy class. Overall Accuracy ($OA = (TP + TN)/(TP + TN + FP + FN)$) is also presented for each experiment. In general, Decision Tress (DTs) have the best results and all supervised methods have over 97% of OA. The result of this ongoing research have potential to provide unbiased morphological classifications for hundreds of thousands of galaxies.

**Resumo.** Este trabalho apresenta avanços em morfologia de galáxias não-paramétrica e experimentos realizados com métodos de aprendizado de máquina sobre resultados de classificação de galáxias em elípticas (E) e espirais (Sp) com métricas morfológicas: concentração (C), métrica de assimetria (A), *smoothness* (S), entropia (H) e parâmetro de análise de padrão de gradiente (GA). O conjunto de dados utilizado para os experimentos com aprendizado supervisionado consiste de 48,145 objetos, com 44,760 galáxias rotuladas como Sp e 3,385 como E. Os resultados foram avaliados com precisão ($P = TP/(TP+FP)$), cobertura ($R = (TP/(TP+FN)$) para cada classe de galáxia considerada. Acurácia geral ($OA = (TP + TN)/(TP + TN + FP + FN)$) também é apresentada para cada experimento. No geral, as árvores de decisão apresentaram os melhores resultados e todos os métodos supervisionados atingiram mais de 97% de OA. O resultado desta pesquisa, ainda em desenvolvimento, tem potencial para fornecer classificação morfológica objetiva para centenas de milhares de galáxias.

**Keywords.** Galaxy Morphology – Computational Science – Machine Learning

## 1. Introduction

Astronomy has become an extremely data-rich enterprise with the advancement of new technologies in recent decades. New telescopes and instruments on board of satellites provide massive datasets. In view of their voluminous size, much of these data are never looked at, and therefore the potential extraction of information from these collected data is only partially accomplished, even though many answers of the contemporary science critically depend on the processing of such large amount of data (1; 2).

One of the key aspects of any extragalactic investigation is the definition of an unbiased sample that includes reliable morphological types. Galaxy morphological properties result from not only the internal formation and evolution processes but also from the interaction with the environment. Galaxies in groups or clusters may have diverse evolutionary paths compared to the isolated ones, which is clearly reflected in their morphology. Therefore, classification of galaxies into a meaningful taxonomy system is of paramount importance for galaxy formation and evolution studies.

Several attempts to objectively measure galaxy morphology have been tried. The most used system is based on Concentration, Asymmetry, Smoothness, Gini and M20 (CASGM), presented in (3; 4). The general rule for using a certain parameter to describe galaxy morphology is that it maximises the distinction between early and late type systems and minimize seeing effects. Such a parametrisation answers two immediate needs. First, to reproduce human classification by positioning the galaxies in the space of these parameters and second to establish a galaxy morphometry system that seeks structures in the quantitative morphology parameter space that may yield clues on the physical reasons for the formation and evolution of galaxies.

The main purpose of this investigation is to answer the question "How to morphologically classify galaxies using GalaxyZoo (12) classification, non-parametric features and Machine Learning methods?". The general schema of this work is represented in Figure 1.We present the first steps towards improving galaxy morphology with Machine Learning (ML). The dataset used for supervised learning experiments consists of 48,145 objects after preprocessing, with 44,760 galaxies labeled as Sp and 3,385 as E. The preprocessing removed 3,611 objects with missing data for one of the features, C. We used as features of the dataset the best morphological parameters from each type to classify galaxies: concentration (5), asymmetry (6), smoothness (S) (7; 8), entropy (9) and gradient pattern analysis parameter (10; 11).

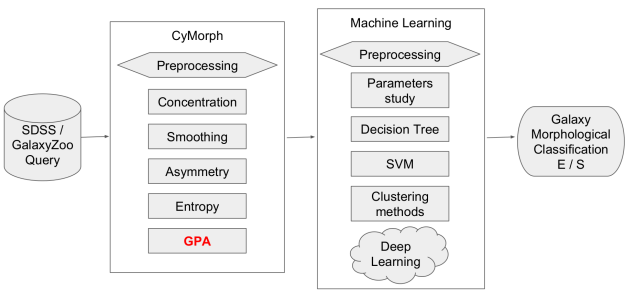


**Figure 1.** General schema proposed to morphologically classify galaxies into early and late-type.

## 2. Non-parametric Galaxy Morphology

Among the possible configurations of metrics, we focus on the five parameters that, in principle, better describe the morphology of a galaxy: 1) Concentration (C) - is given by the ratio of the circular radii containing 65% and 25% of the Petrosian flux of the galaxy, respectively, $C = log(R_{65\%}/R_{25\%})$; 2) Asymmetry (A) - is measured by the correlation between an image and its $\pi$-rotated variant. We adopted the prescription introduced by (13), where $A = 1 - r(Im, Im^{\pi})$. In this equation, $Im$ represents the original image, $Im^{\pi}$ is the rotated (by $\pi$ radians) corresponding image, and $r$ indicates the Pearson correlation coefficient; 3) Smoothness (S) - is measured by the correlation between an image and its smoothed variant. We estimated smoothness as $S = 1 - r(Im, Im^{F})$, where $Im$ is the original image and $Im^{F}$ is its smoothed version; 4) Entropy (H) - is the Shannon entropy, namely the average amount of information resulting from a stochastic process - the clumpiest an image is the larger is its entropy; and 5) GPA - technique to separate early from late-type galaxies by the second moment of gradient from images All these parameters measured, except GPA, are already described in more detail in (13) and references therein. All these quantities are measured within a package named CyMorph, write in Cython (a C-extension for Python).

## 3. Data Mining in Galaxy Morphology

CyMorph presents a consistent non-parametric morphology system, which can achieve better classification results if a data mining process is employed to gather the best information from the group of metrics with Machine Learning (ML) methods. Basically, ML can be divided into Supervised and Unsupervised Learning. Supervised Learning (SL) is a learning process guided by some form of supervision to build a classification model. For this, the dataset must be divided into train, validating and model testing set. Unsupervised Learning (UL), differs from SL because has no supervision, i.e., there is no model to guide the learning process.

Different configurations for supervised (Support Vector Machine – SVM, and Decision Tree – DT) and unsupervised learning methods (K-means and Agglomerative Hierarchical Clustering – AHC) were tested.

Support Vector Machines (SVM) constructs the optimal hyperplane that will divide the target classes. An optimal hyperplane is the one that maximizes the separation margins between the classes, providing a unique solution for the problem (15).

Decision Tree (DT) is a supervised machine learning method to classification and regression. The goal here is to create a model which predicts the classification by learning simple decision rules inferred from the dataset (16).

One of the more general-purpose clustering methods (non-supervised machine learning), K-means finds clusters of similar sizes, flat geometry, not many clusters, and accepts specification of clusters (17).

## 4. Concluding Remarks

Morphology is a key ingredient in the process of selecting a sample of galaxies for studying the physical mechanisms responsible for shaping the galaxies as we observe today. Also, considering that the following decades will be dominated by photometric (image) rather than spectroscopic data (e.g. LSST, Pan-STARRS, etc.), it is critical to have robust measurements that capture the essential morphological information and avoid redundancy.

In general, DTs have the best results, considering CN as the most important feature to separate galaxies into spiral and elliptical (responsible attribute for the first decision in all DTs). The Grid Search applied in the supervised methods optimized the OA. Due to the unbalance in the dataset (44760 galaxies labeled as S and 3385 as E), none experiment reached Kappa index ($\kappa$) of 0,9, although the interval $0,8 \leq \kappa \leq 1$ is considered of excellent concordance. The recall was also affected by this unbalance. However, all supervised methods have over 97% of OA.

The research from (14) presented here shows satisfactory preliminary results for separating early from late-type galaxies with non-parametric morphology features improved by data mining. The result of this ongoing research have potential to provide unbiased morphological classification for hundreds of thousands of galaxies.

## References


Feigelson, E. D., & Babu, J. (Eds.). 2006, Statistical challenges in astronomy (Springer Science & Business Media)

Zaïane, O. R. 1999, Introduction to data mining (University of Alberta)

Conselice, C.J. and Bershady, M.A. and Jangren, A. 2000, The Astrophysical Journal, 886-910, 529, doi:10.1086/308300

Lotz, J. M. and Primack, J. and Madau, P. 2004, The Astronomical Journal, 1, 128, 163

Kent, S.M. 1985, The Astrophysical Journal. Supplement Series, 59, 115

Abraham, R.G., Bergh, S.V.D., Glazebrook, K., Ellis, R.S., Santiago, B.X., Surma, P. and Griffiths, R.E. 1996, The Astrophysical journal. Supplement series. Chicago. 107, 1, 1.

Lotz, J.M., Primack, J. and Madau, P. 2004, The Astronomical Journal, 128, 1, 163

Hamming, R.W., 1998. Digital Filters (3rd ed.). Courier Dover Publications.

Bishop, C., 2007. Pattern Recognition and Machine Learning (Springer, New York)

Rosa, R.R., Sharma, A.S. & Valdivia, J.A. 1999, International Journal of Modern Physics C, 10, 1, 147

Baroni, M.P.M.A., Rosa, R.R., da Silva, A.F., Pepe, I., Roman, L.S., Ramos, F.M., Ahuja, R., Persson, C. & Veje, E. 2006, Microelectronics Journal, 37, 4, 290

Willett, K. W., Lintott, C. J., Bamford, S. P., Masters, K. L., Simmons, B. D., Casteels, K. R., ... & Melvin, T. 2013, Monthly Notices of the Royal Astronomical Society, 435, 4, 2835

Ferrari, F., de Carvalho, R. R., & Trevisan, M. 2015, The Astrophysical Journal, 814, 1, 55

Barchi, P. H. and Sautter, R. & da Costa, F. G. & Moura, T. C. & Stalder, D. H. & Rosa, R. R. & de Carvalho, R. R. 2016, Journal of Computational Interdisciplinary Sciences, 7, 3, 114–120.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. 1998, IEEE Intelligent Systems and their Applications, 13, 4, 18.

Quinlan, J. R. 1986, Machine learning, 1, 1, 81

Hartigan, J. A., & Wong, M. A. 1979, Journal of the Royal Statistical Society. Series C (Applied Statistics), 28, 1, 100