WORCAP 2017
XVII Workshop de Computação Aplicada

Data Mining and Advances in Non-Parametric Galaxy Morphology

Paulo Henrique Barchi; Rubens Andreas Sautter; Tatiana Coelho Moura; Diego Herbin Stalder; Reinaldo Roberto Rosa; Reinaldo Ramos de Carvalho

paulobarchi@gmail.com

The volume of digital data of stars, galaxies, and the universe has multiplied in recent decades due to the rapid development of new technologies as new satellites, telescopes, and other observatory instruments. The process of scientific discovery is increasingly dependent on the ability to analyze massive amounts of complex data from scientific instruments and simulations. Such analysis has become the bottleneck of the scientific process. By studying global properties of early-type galaxies (ETGs), researchers have been able to constrain models of galaxy formation and evolution. This work presents advances in non-parametric galaxy morphology and machine learning experiments performed over results of galaxy classification into elliptical (E) and spiral (S) with morphological parameters: concentration (CN), asymmetry metrics (A3), smoothness metrics (S3), entropy (H) and gradient pattern analysis parameter (GA). Except concentration, all parameters performed a image segmentation preprocessing. The dataset used for supervised learning experiments consists of 48145 objects after preprocessing, with 44760 galaxies labeled as S and 3385 as E. The preprocessing removed 3611 objects with missing data for one of the features: CN. The supervised methods used are Decision Tree (DT) and Support Vector Machine (SVM). These are preliminary results from an ongoing research about morphological parameters to classify galaxies – a full publication about it will be released soon. The target of our dataset (considered as true label) is the classification from Galaxy Zoo 1 project. The experiments were conducted to explore different method parametrization, if it is applicable. For the unsupervised learning experiments with K-Means and Agglomerative Clustering, we used a balanced dataset with 1962 objects. The results are evaluated with precision (P = TP/(TP+FP)) and recall (R = (TP/(TP+FN)) for each galaxy class: spiral (S) and elliptical (E). F-score (F1 = 2 * (P*R)/(P+R)), Overall Accuracy (OA = (TP+TN)/(TP+TN+FP+FN)) and Kappa index (K) are also presented for each experiment. 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 (K) of 0,9, although the interval 0,8 <= K <= 1 is considered of excellent concordance. The recall was also affected by this unbalance. However, all supervised methods have over 97% of OA.

Data Mining. Galaxy Morphology. Machine Learning. Computational Astrophysics.