Volume Title
ASP Conference Series, Vol. **Volume Number**
Author

Galaxy Classification without Feature Extraction

Kai Lars Polsterer[1], Fabian Gieseke[2], and Oliver Kramer[2]

[1]*Astronomisches Institut, Department of Physics and Astronomy, Ruhr-University of Bochum, Germany*

[2]*Computational Intelligence, Computer Science Department, University of Oldenburg, Germany*

Abstract. The automatic classification of galaxies according to the different Hubble types is a widely studied problem in the field of astronomy. The complexity of this task led to projects like Galaxy Zoo which try to obtain labeled data based on visual inspection by humans. Many automatic classification frameworks are based on artificial neural networks (ANN) in combination with a feature extraction step in the pre-processing phase. These approaches rely on labeled catalogs for training the models. The small size of the typically used training sets, however, limits the generalization performance of the resulting models. In this work, we present a straightforward application of support vector machines (SVM) for this type of classification tasks. The conducted experiments indicate that using a sufficient number of labeled objects provided by the EFIGI catalog leads to high-quality models. In contrast to standard approaches no additional feature extraction is required.

1. Introduction

We analyzed the performance of the morphological classification process of galaxies and determined that a remarkable amount of computing power is required for pre-processing the data. Typically an automated classification of galaxies is realized via a multi-stage approach. In the first step the image is pre-processed, e.g., contrast enhanced or edge finding kernel-filters are applied. In a next step a small number of features is extracted from the image. Finally the generated features are used as input for classifiers like ANNs or decision trees (Wijesinghe et al. 2010; de la Calleja & Fuentes 2004). The classification approach we present in this work uses the raw image data without any feature pre-processing or extraction (see Fig. 1).

1.1. Image Data and Labels

The presented experiments are based on image data taken from the Sloan Digital Sky Survey (SDSS 2011). In Baillard et al. (2011) the EFIGI catalog of 4,458 nearby galaxies is presented. The Hubble type and morphological features of these galaxies have been determined by a group of human experts. This catalog was used to extract the required labels for the experiments. In a first step the image data for each galaxy was retrieved as a JPEG file. The resolution was adjusted to fit the whole galaxy and a 40×40 pixels^2 stamp was created.

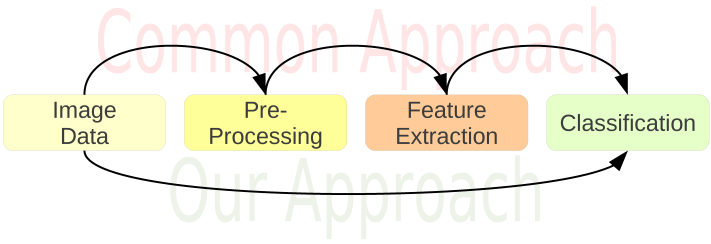

Figure 1. Comparison of Approaches: The standard multi-stage versus our raw feature approach.

1.2. The Classifiers: Support Vector Machines

The experiments rely on SVMs. Roughly speaking, the aim of a SVM is to find a hyperplane in a feature space which maximizes the margin between classes such that only a few training patterns lie within this margin (Hastie et al. 2009). The latter task can be formulated as an quadratic optimization problem, where the first term corresponds to maximizing the margin and the second term to the loss caused by patterns lying within the margin:

$$\begin{aligned} \underset{\vec{w}\in\mathcal{H}_0,\, \vec{\xi}\in\mathbb{R}^n,\, b\in\mathbb{R}}{\text{minimize}} \quad & \tfrac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n}\vec{\xi}_i \\ \text{s.t.} \quad & y_i(\langle\vec{w},\Phi(\vec{x}_i)\rangle + b) \geq 1-\xi_i, \\ & \text{and}\, \xi_i \geq 0 \end{aligned} \tag{1}$$

where $C > 0$ is a user-defined parameter. The function $\Phi : \mathbb{R}^d \to \mathcal{H}_0$ is induced by a kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with $k(\vec{x}_i, \vec{x}_j) = \langle\Phi(\vec{x}_i), \Phi(\vec{x}_j)\rangle$. A kernel function can be seen as a similarity measure for input patterns. The goal of the learning process is to find the optimal prediction function $f(\vec{x}) = \langle\vec{w}, \Phi(\vec{x})\rangle + b$. A common choice for the kernel function is the linear kernel

$$k(\vec{x}_i, \vec{x}_j) = \langle\vec{x}_i, \vec{x}_j\rangle \tag{2}$$

or the RBF kernel

$$k(\vec{x}_i, \vec{x}_j) = \exp\left(-\frac{\|\vec{x}_i - \vec{x}_j\|^2}{2\sigma^2}\right) \tag{3}$$

with σ as additional parameter.

2. Classification Experiments and Results

Based on the available data we conducted several morphological classification experiments. The SVMs were trained with one half of the galaxy sample and tested on the other half. For all experiments a linear kernel was used. To tune the involved parameters we resort to 5-folds cross validation performed on the training set.

Experiment 1: Discriminating elliptical/lenticular from spiral galaxies. This experiment shows a performance of $\approx$ 84% correctly classified galaxies. Towards the intermediate Hubble type S0 the amount of miss-classifications rises (see Tab. 1).

Experiment 2: Discriminating elliptical from spiral galaxies. A classification accuracy of $\approx$ 88% is reached. The EFIGI catalog is highly unbalanced concerning the amount of pure elliptical galaxies: Therefore a balanced sub-set was used for this experiment. With the unbalanced data-set $\approx$ 95% accuracy is reached. This is comparable to results of classifications with feature extraction.

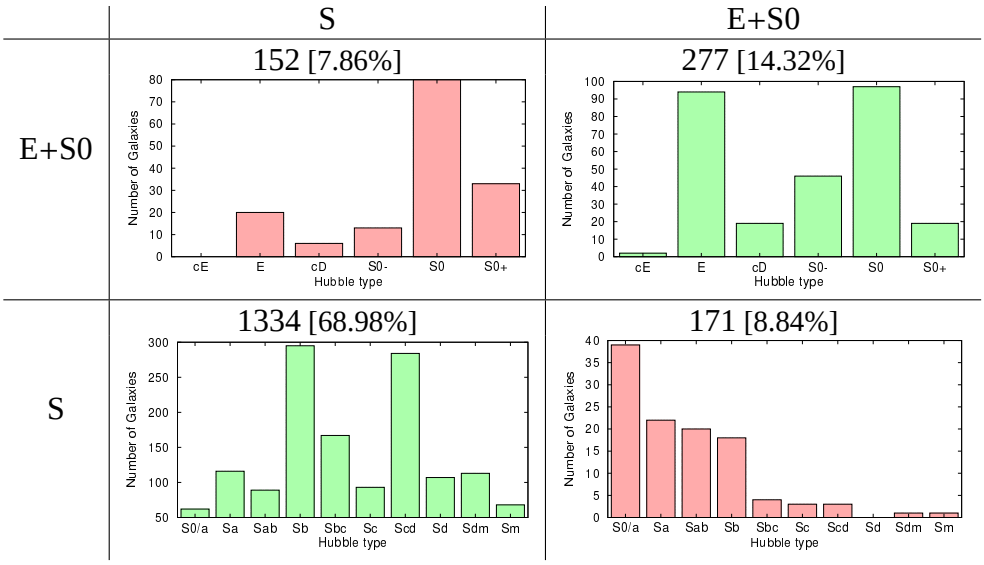

Table 1. Confusion matrix of experiment 1: True labels (left side), predictions (top).

Experiment 3: Detecting a bar in a galaxy. The discrimination between barred and un-barred galaxies fails when using raw features. Just $\approx$ 60% of the galaxies can be classified correctly (see Tab. 2).

	Bar	No Bar
No Bar	501 [38.87%]	37 [2.87%]
Bar	733 [56.87%]	18 [1.39%]

Table 2. Confusion matrix of experiment 3: True labels (left side), predictions (top).

3. Self-Organizing Maps

By using an ANN a discrete and low-dimensional map is created which represents the input objects in their high-dimensional feature space. The resulting map is called a self-organizing map (SOM). It reflects the similarity of the input objects in high dimensions. With all EFIGI galaxies such a SOM was trained. This map clearly separates spiral from elliptical galaxies (see Fig. 2).

4. Conclusions

Both, the classification experiments and the SOM show that simple classification tasks can be solved with raw features. Complex tasks like detecting a bar seem to require other sophisticated features extraction schemes. Our experimental evaluation indicates that distant regions in the SOM correspont to classification tasks which can be addressed easily without feature extraction. Therefore automated dimension reduction methods like a SOM are excelent pre-evaluation frameworks.

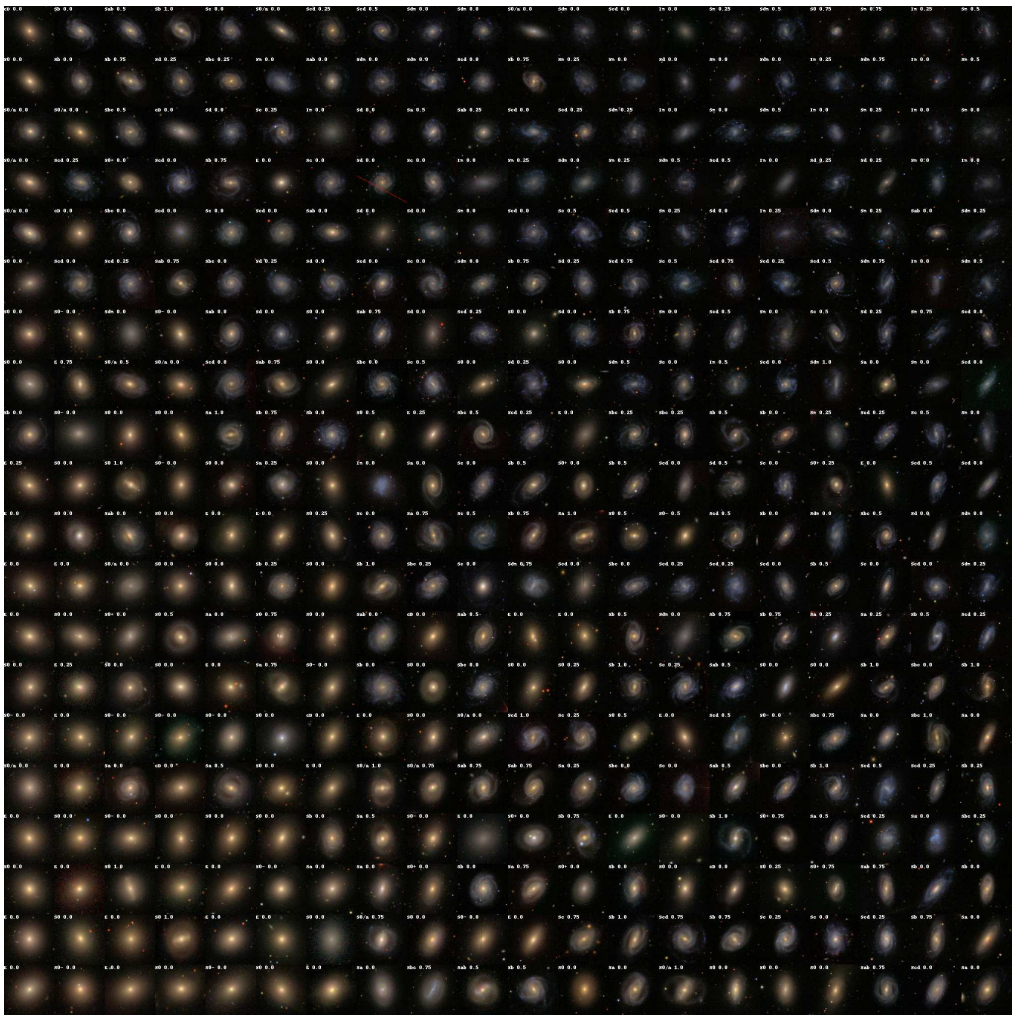

Figure 2. Self-Organizing Map of the EFIGI Galaxies.

Acknowledgments. This work is based on data of the SDSS project.

References

Baillard, A., Bertin, E., de Lapparent, V., Fouqué, P., Arnouts, S., Mellier, Y., Pelló, R., Leborgne, J.-F., Prugniel, P., Makarov, D., Makarova, L., McCracken, H. J., Bijaoui, A., & Tasca, L. 2011, A&A, 532, A74. 1103.5734

de la Calleja, J., & Fuentes, O. 2004, MNRAS, 349, 87

Hastie, T., Tibshirani, R., & Friedman, J. 2009, The elements of statistical learning: data mining, inference and prediction (Springer), 2nd ed.

SDSS 2011, Sloan digital sky survey. URL http://www.sdss.org

Wijesinghe, D. B., Hopkins, A. M., Kelly, B. C., Welikala, N., & Connolly, A. J. 2010, MNRAS, 404, 2077. 1001.5322