Astronomical Data Analysis Software and Systems IV
ASP Conference Series, Vol. 77, 1995
R. A. Shaw, H. E. Payne, and J. J. E. Hayes, eds.

Clustering Analysis Algorithms and Their Applications to Digital POSS-II Catalogs

R. R. de Carvalho[1], S. G. Djorgovski, and N. Weir

California Institute of Technology, MS 105-24, Pasadena, CA 91125

U. Fayyad, K. Cherkauer[2], J. Roden, and A. Gray

Jet Propulsion Laboratory, MS 525-3600, Pasadena, CA 91109

Abstract. We report on the preliminary results of experiments using a Bayesian cluster method to cluster objects present in photographic images of the POSS-II. Our goal is to explore the power of unsupervised learning techniques to classify objects meaningfully, and perhaps to discover previously unrecognized object categories in digital sky surveys. Our primary finding is that the program we used, *AutoClass*, was able to form several sensible categories from a few simple attributes of the object images, separating the data into four recognizable and astronomically meaningful classes: stars, galaxies with bright central cores, galaxies without bright cores, and stars with a visible "fuzz" around them. Also, in an independent experiment we found out that the two types of galaxies have distinct color distributions (the more concentrated class being redder, as indeed expected if they are predominantly early Hubble types), although no color information was given to *AutoClass*. This illustrates the power of unsupervised classification techniques to discriminate between astronomically distinct types of objects on the basis of data alone. We believe that the application of such algorithms to large-scale astronomical sky surveys can aid in cataloging the detected objects, and may even have the potential to discover new categories of objects.

1. Introduction

The last two decades have witnessed the cataloging of the northern and southern hemispheres through the use of high-quality photographic plates combined with CCD frames (van Altena 1993). These digital sky surveys amount to ∼5–6 TB worth of data, resulting in catalogs of many millions—or even billions—of objects. This richness of information requires new, efficient tools to explore the resulting data spaces (Weir et al. 1993a).

---

[1]On Leave of Absence from Observatório Nacional/Cnpq, Rio de Janeiro, CEP 20921, Brazil

[2]University of Wisconsin, Madison, WI 53706

A crucial point in constructing scientifically useful object catalogs is the star/galaxy separation. Various supervised classification schemes can be used to produce consistent results in this task (Valdes 1982; Beard et al. 1990; Odewahn et al. 1992; Weir et al. 1995). However, a more difficult problem is systematically and objectively to provide at least rough morphological types for the galaxies detected, without visual inspection of the plates or scans—which is impractical for obvious reasons. We have thus started to explore new clustering analysis and unsupervised classification techniques for this task. Our goal is to try to separate astronomically meaningful morphological types on the basis of the data themselves, rather than some preconceived scheme.

Thus we investigate the possibility of finding natural (data-based) partitions of the attribute spaces which show high correlations between the plate-measured attribute space, and the CCD-based attribute space, or a high degree of separation between expected classes such as stars versus galaxies, spirals versus ellipticals, or galaxies of different concentrations. These partitions of the data may be used for investigations of unusual regions of the attribute space, and may even lead to a discovery of the previously unknown objects or classes of objects.

2. Data and Methodology

We use the data from the digitized version of the Second Palomar Observatory Sky Survey (POSS-II). For brief descriptions of the survey, see (Djorgovski et al. 1994; Reid and Djorgovski 1993; Weir et al. 1993b; Weir et al. 1994; Weir 1995). We have used data from 3 fields from POSS-II, numbers 380 (J-Band), 442 (J-Band), and 679 (J and F Bands).

The following attributes were used for the analysis: (1) resolution scale, (2) resolution fraction (these two are described in Valdes 1982), (3) ellipticity, (4) normalized core magnitude, (5) normalized area, (6) first intensity moment, and (7) the S parameter introduced by Collins et al. (1989). We have used only objects classified as galaxies and stars by using the Decision Tree technique (Weir et al. 1995). It is important to emphasize that we are not intentionally using legitimate attributes like colors, mean surface brightness and concentration index, which are available in our catalogs, because at this point they can help us understand the association between the classes which come out from the experiment and the large scale distribution of galaxies. Also, the classification is not given to the algorithm but is only used to judge its performance.

AutoClass (Cheeseman et al. 1988) is an unsupervised learning algorithm that fits user-specified probability distribution models to a set of examples represented as feature vectors. Classes are represented probabilistically as particular parameterizations of the models. In these experiments, we used multi-dimensional Gaussian models. *AutoClass* uses Bayesian techniques to estimate the parameter values of each class. It also tries to find the most probable number of classes by comparing the likelihoods of the fits for different numbers of classes. Objects are then assigned probabilistic memberships in the output classes.

The Gaussians used to model the classes can range from noncovariant (i.e., axis-aligned) to fully covariant based on prior knowledge the user may have of the attributes and problem. In our experiments here, we used only models that

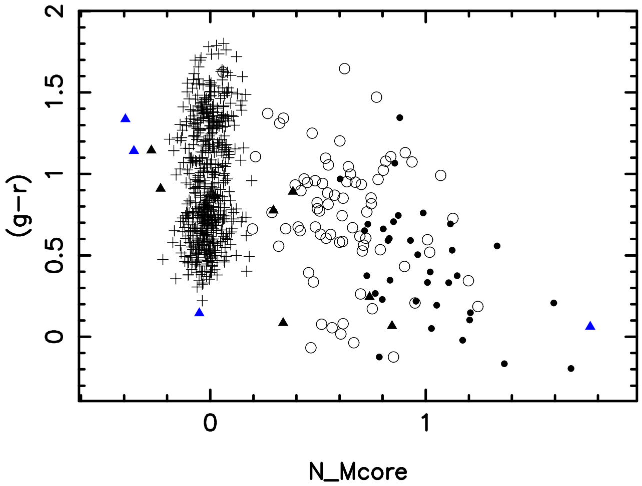

Figure 1. The $(g-r)$ color versus normalized core magnitude, for the four types of objects found by *AutoClass*: galaxies without a bright core (open circles), galaxies with a bright core (solid circles), stars (crosses), and stars with fuzz (solid triangles).

had no covariance. We ran some simple tests using synthetic data to verify that *AutoClass's* behavior was reasonable in each of these cases.

3. Discussion

In our first experiment we used data from the fields 380 and 442. *AutoClass* was able to find four natural classes of objects in the data space. These four classes were, by visual inspection, identified with stars, galaxies with a bright core, galaxies without a bright core, and stars with fuzz around them. Thus, the object classes found by *AutoClass* are astronomically meaningful, even though the program itself does not know about stars, galaxies, and such! These results were obtained using data in a given bin of magnitude ($17^{\mathrm{m}} < r < 18^{\mathrm{m}}$), although the same trends were found for a bin one magnitude fainter. The results are robust and repeatable from field to field.

By inspecting the so-called confusion matrix, we found that each cluster identified by *AutoClass* corresponds to the type of the objects, as classified by the Decision Tree (a supervised classification approach). The Decision Tree was trained to recognize only two classes of objects, stars and galaxies, and no attempt was made to make any morphological distinctions among the galaxies.

Another experiment was done using another field in two colors (442 J and F), both in order to check the previous finding, and also to explore a little more deeply the meaning of these classes. Again, *AutoClass* found the same four significant classes in the data space, which confirms the robustness of the method.

Figure 1 displays the $(g - r)$ color versus the normalized core magnitude (one of the attributes used in the experiment). As can be seen, the two morphologically distinct classes of galaxies, represented by solid and open circles, populate different regions of the data space, and have systematically different colors, *even though AutoClass was not given the color information.* In this figure we display stars as crosses and stars with fuzz around as solid triangles.

The confusion matrix for such experiments indicates that stars and galaxies, as classified by Decision Tree, are well separated in different classes. Galaxies are distributed in two classes, representing redder and bluer systems, respectively—presumably the early and late Hubble types, respectively.

We are now exploring our database from POSS-II in a systematic way using such techniques to map the large scale structure (clustering in the physical space) in an unbiased fashion. One project is to objectively define and discover clusters and groups of galaxies, which can then be used for a variety of follow-up studies.

A full paper will be presented in near future describing in detail the application of *AutoClass* to POSS-II and similar data.

References

Beard, S. M., MacGillivray, H. T., & Thanisch, P. F., 1990. MNRAS, 247, 311

Cheeseman, P., et al. 1988, in Proc. Fifth Machine Learning Workshop, ed. J. Laird (San Mateo, Calif., M. Kauffmann), p. 54

Collins, C. A., Heydon-Dumbleton, N. H., & MacGillivray, H. T. 1989. MNRAS, 236, 7p

Djorgovski, S., Weir, N., & Fayyad, U. 1994, in Astronomical Data Analysis Software and Systems III, ASP Conf. Ser., Vol. 61, eds. D. R. Crabtree, R. J. Hanisch, & J. Barnes (San Francisco, ASP), p. 195djorgovskis

Odewahn, S. C., Stockwell, E. B., Pennington, R. L., Humphreys, R. M., & Zumach, W. A. 1992. AJ, 103, 318

Reid, I. N., & Djorgovski, S. 1993, in Sky Surveys: Protostars to Protogalaxies, ASP Conf. Ser., Vol. 43, ed. B. T. Soifer (San Francisco, ASP), p. 125

Valdes, F. 1982, in Instrumentation in Astronomy, IV, ed. D. L. Crawford, SPIE Proc., 331, 465

van Altena, W. F. 1993, in Astronomy from Wide-field Imaging, IAU Symp. 161, ed. H. T. MacGillivray et al. (Dordrecht, Kluwer), p. 193

Weir, N., Djorgovski, S., Fayyad, U., Smith, J. D., & Roden, J. 1993a. in Astronomy from Wide-field Imaging, IAU Symp. 161, ed. H. T. MacGillivray et al. (Dordrecht, Kluwer), p. 205

Weir, N., Fayyad, U., Djorgovski, S., Roden, J., & Rouquette, N. 1993b, in Astronomical Data Analysis Software and Systems II, ASP Conf. Ser., Vol. 52, eds. R. J. Hanisch, R. J. V. Brissenden, & J. Barnes (San Francisco, ASP), p. 39

Weir, N., Djorgovski, S., Fayyad, U., Smith, J. D., & Roden, J. 1994, in Astronomy From Wide-Field Imaging, IAU Symp. 161, ed. H. T. MacGillivray et al. (Dordrecht, Kluwer), p. 205

Weir, N. 1995, Ph.D. Thesis, California Institute of Technology.

Weir, N., Djorgovski, S., & Fayyad, U. 1995. AJ, in press