SKICAT: Sky Image Cataloging and Analysis Tool

Usama M. Fayyad
JPL, California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA 91109 U.S.A.
Fayyad@aig.jpl.nasa.gov

1 Introduction

In astronomy and space sciences, we currently face a data glut crisis. The problem of dealing with the huge volume of data accumulated from a variety of sources, of correlating the data and extracting and visualizing the important trends, is now fully recognized. This problem will become more acute very rapidly, with the advent of new telescopes, detectors, and space missions, with the data flux measured in terabytes. We face a critical need for information processing technology and methodology with which to manage this data avalanche in order to produce interesting scientific results quickly and efficiently. Developments in the fields of machine learning and AI can provide at least some solutions. Much of the future of scientific information processing lies in the implementation of these methods.

We present an application of supervised classification to the automation of the tasks of cataloging and analyzing objects in digitized sky images. The Sky Image Cataloging and Analysis Tool (SKICAT) was developed for use on the images resulting from the 2nd Palomar Observatory Sky Survey (POSS-II) conducted by the California Institute of Technology (Caltech). The photographic plates collected from the survey are digitized at the Space Telescope Science Institute. This process will result in about 3,000 digital images of 23,040 x 23,040 16-bit pixels each, totalling over 3 terabytes of data. When complete, the survey will cover the entire northern sky in three colors, detecting virtually every sky object down to a B magnitude of 22. This is at least one magnitude fainter than previous comprable photographic surveys. We estimate that there are on the order of 10^7 galaxies 10^9 stellar objects (including over 10^5 quasars) are detectable in this survey. This data set will be the most comprehensive large-scale imaging survey produced to date and will not be surpassed in scope until the completion of a fully digital all-sky survey.

The purpose of SKICAT is to enable and maximize the extraction of meaningful information from such a large database in timely manner. The system is built in a modular way, incorporating several existing algorithms and packages. There are three basic functional components to SKICAT, serving the purposes of sky object catalog construction, catalog management, and high-level statistical and scientific analysis.

2 Classifying Sky Objects

The first step in analyzing the results of a sky survey is to identify, measure, and catalog the detected objects in the image into their respective classes. Once the objects have been classified, further scientific analysis can proceed. For example, the resulting catalog may be used to test models of the formation of large-scale structure in the universe, probe Galactic structure from star counts, perform automatic identifications of radio or infrared sources, and so forth [4; 12; 11] Reducing the images to catalog entries is an overwhelming task which inherently requires an automated approach. The goal of our project is to automate this process, providing a consistent and uniform methodology for reducing the data sets. This will provide the means for objectively performing tasks that formerly required subjective and visually intensive manual analysis. Another goal of this work is to classify objects whose intensity (isophotal magnitude) is too faint for recognition by inspection, hence requiring an automated classification procedure. Faint objects constitute the majority of objects on any given plate. We target the classification of objects that are at least one magnitude fainter than objects classified in previous surveys using comparable photographic material. The goals of the video are to introduce the machine learning techniques we used, to give a general, high-level description of the application domain, and to report on the successful results which exceeded our initial goals. We aim to point out an instance where learning algorithms proved to a be useful and powerful tool in the automation of a significant scientific data analysis task.

Using decision tree and rule learning algorithms [5; 9] The SKICAT system classifies objects that are at least one magnitude fainter than objects cataloged in previous surveys. Wc have exceeded our initial accuracy target of 90% [8]. This level of accuracy is required for the data to be useful in testing or refuting theories on the formation of large structure in the universe and on other phenomena of interest to astronomers. The SKICAT tool is now being employed to both reduce and analyze the survey images as they arrive from the digitization instrument. We are also beginning to explore the application of SKICAT to the analysis of other surveys being planned by NASA and other institutions.

In order to produce a classifier that classifies faint objects correctly, the learning algorithm needs training data consisting offaint objects labeled with the appropriate class. The class label is therefore obtained by examining the CCD frames. Once trained on properly labeled objects, the learning algorithm produces a classifier that is capable of properly classifying objects based on the values of the attributes measured from the lower resolution plate image. Hence, in principle, the classifier will be able to classify objects in the photographic image that are simply too faint for an astronomer to classify by inspection. Using the class labels, the learning algorithms are basically being used to solve the more difficult problem of separating the classes in the multi-dimensional space defined by the set of attributes derived via image processing. This method is expected to allow us to classify objects that are at least one magnitude fainter than objects classified in photographic all-sky surveys to date.

By effectively defining robust features, we were able to obtain classifiers with an accuracy exceeding that of humans for faint objects [13] Since faint objects constitute the majority of objects on any plate, this results in a dramatic increase in the number of classified objects available for further scientific analysis. In effect, this shows that the pixels contained important information that the human visual system could not extract. Projection of the high-dimensional pixel space onto a powerful lower-dimensional feature space allowed us to transform the problem into one that is easily solvable by a supervised learning algorithm.By defining additional "normalized" image-independent attributes, we were able to obtain high accuracy classifiers within and across photographic plates.

3 Conclusions

The implications of a tool like SKICAT for Astronomy may indeed be profound. One could reclassify any portion of the survey using alternative criteria better suited to a particular scientific goal (e.g. star catalogs vs. galaxy catalogs). This changes the notion of a sky catalog from the classical static entity "in print", to a dynamic, ever growing, ever improving, on-line database. The catalogs will also accommodate additional attribute entries, in the event other pixel-based measurements are deemed necessary. An important feature of the survey analysis system will be to facilitate such detailed interactions with the catalogs. The catalog generated by SKICAT will eventually contain about two billion entries representing hundreds of millions of sky objects. Unlike the traditional notion of a static printed catalog, we view our effort as targeting the development of a new generation of scientific analysis tools that render it possible to have a constantly evolving, improving, and growing catalog. Without the availability of these tools for the first survey (POSS-I) conducted over four decades ago, no objective and comprehensive analysis of the data was possible. In contrast, we are targeting a comprehensive sky catalog that will be available on-line for the use of the scientific community.

For future work, we are sudying the application of clustering algorithms (e.g. AutoClass [2]) to automate discovery of new classes in the large catalog databases created by SKICAT [13; 14; 3]

References

[1] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. 1984. *Classification and Regression Trees.* Monterey, CA: Wadsworth & Brooks.

[2] Cheeseman, P. and Stutz, J. 1995. Bayesian Classification (AutoClass): Theory and Results. In *Advances in Knowledge Discovery and Data Mining,* U. Fayyad et al (Eds.), Boston: AAAI/MIT Press.

[3] DeCarvalho, R., Djorgovski, S.G., Weir, N., Fayyad, U., Cherkauer, K., Roden, J., and Gray, A. 1995. Clustering Analysis Algorithms and Their Applications to Digital POSS-II Catalogs, in R. Hanisch, *et al.* (Eds.), *Astronomical Data. Analysis Software and Systems IV, A.S.P. Conf. Ser.* in press.

[4] Djorgovski, S.G., Weir, N., and Fayyad, U. M. 1994. Processing and Analysis of the Palomar - STScI Digital Sky Survey Using a Novel Software Technology. In D. Crabtree, R. Hanisch, and J. Barnes (Eds.), *Astronomical Data Analysis Software and Systems III, A.S.P. Conf. Ser.* 61, 195.

[5] Fayyad, U.M. and Irani, K.B. 1992a. The Attribute Selection Problem in Decision Tree Generation. In *Proc. of the Tenth National Conference on Artificial Intelligence AAAI-92,* pages 104-110, Cambridge, MA: MIT Press.

[6] Fayyad, U.M. and Irani, K.B. 1992b. On the Handling of Continuous-Valued Attributes in Decision Tree Generation. *Machine Learning,* vol.8, no.2.

[7] Fayyad, U.M. and Irani, K.B. 1993. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In *Proc. of the Thirteenth Inter. Joint Conf. on Artificial Intelligence,* Chambery, France: IJCAII.

[8] Fayyad, U.M., Weir, N., and Djorgovski, S.G. 1993. SKICAT: A Machine Learning System for the Automated Cataloging of Large-Scale Sky Surveys. In *Proc. of the Tenth International Conference on Machine Learning,* 1993.

[9] Fayyad, U.M. 1994. Branching on Attribute Values in Decision Tree Generation. In *Proc. of the Twelfth National Conference on Artificial Intelligence AAAI-94,* pages 601-606, Cambridge, MA, 1994. MIT Press.

[10] Quinlan, J. R. 1986. The induction of decision trees. *Machine Learning,* 1(1).

[11] Weir, N. 1994. *Automated Analysis of the Digitized Second Palomar Sky Survey: System Design, Implementation, and Initial Results.* Ph.D. Dissertation, California Institute of Technology.

[12] Weir, N., Djorgovski, S.G., Fayyad, U.M., Smith, J.D., and Roden, J. 1994. Cataloging the Northern Sky Using a New Generation of Software Technology. In *Astronomy From Wide-Field Imaging,* H. MacGillivray *et al.* (Eds.), Proceedings of the IAU Symp. #161, p. 205. Dordrecht: Kluwer.

[13] Weir, N., Fayyad, U.M., and Djorgovski, S.G. 1995. Automated Star/Galaxy Classification for Digitized POSS-II. *The Astronomical Journal,* in press.

[14] Weir, N.(Djorgovski, S.G., and Fayyad, U.M. 1995. Initial Galaxy Counts From Digitized POSS-II. *Astronomical Journal,* in press.