Chapter 1

MINING ASTRONOMICAL DATABASES

Roberta M. Humphreys
Astronomy Department
University of Minnesota
roberta@aps.umn.edu

Juan Cabanela
Department of Physics, Astronomy and Engineering Science
St. Cloud State University
juan@lua.stcloudstate.edu

Jeffrey Kriessler
Efficient Channel Coding, Inc.
Cleveland, Ohio
jeffk@eccincorp.com

Abstract The development of software tools and techniques for the efficient access and analysis of large astronomical databases poses some unique challenges. We briefly describe some of the problems astronomical data and datasets present and give an example from our own efforts to automate the classification of galaxies, and then discuss where "clustering" algorithms may be applicable.

Keywords: databases galaxy classification data mining

Introduction

The number of space-based all-sky surveys ranging from gamma rays and X-rays to the far infrared and millimeter wavelengths plus the supporting digitization programs from the optical photographic sky surveys (POSS I and II and the UK-SRC) is rapidly increasing. When we add

in the new groundbased digital surveys, like 2MASS and DENIS in the near-infrared and the optical Sloan Digital Sky Survey (SDSS) a $\geq$40 TB-sized Internet-wide multi-wavelength astronomical dataset will soon exist. To derive the maximum scientific benefit from this vast resource of fundamental data and the recently proposed National Virtual Observatory will require efficient access, such as federated databases that stretch across several databases at physically different locations, and new software techniques and tools, often referred to as "data mining", for the analysis of large databases.

Astronomical databases, however, pose unique problems and challenges due not only to their very large size, but also to the variable quality of the data and the uncertainty of measurement over the entire electromagnetic spectrum, and to the nature of astronomical objects with their very wide dynamic range in apparent luminosity and in apparent size (angular diameter). Astronomical databases will not only possess an unprecedented number of objects, but the astronomical objects themselves may also have a large number of attributes leading to a very high dimensional dataset.

Many of the necessary techiques and software packages, including artificial intelligence techniques, like neural networks and decision trees have already been successfully applied to astronomical problems such as pattern recognition and object classification, while new clustering and data association algorithms that may have application to large astronomical databases are being developed by computer science groups. However, these new software packages are often developed and tested on idealized or "clean" datasets that lack the "real noise" and uncertainty of measurement encountered particularly in large astronomical databases. The APS Catalog of the POSS I is an excellent resource for perfecting and testing these data mining techniques.

1. THE APS PROJECT

The Automated Plate Scanner(APS) Catalog of the POSS I is an on-line database of fundamental data and parameters for over 100 million stars and galaxies derived from our digitized scans of glass copies of the blue and red plates of the original, first epoch Palomar Observatory Sky Survey (POSS I). It is large enough, 25 GB, to present a realistic challenge for testing "data mining" algorithms on a range of astrophysical applications. Its scientific usefulness and validity have been demonstrated by numerous studies by members of the APS group and by our users.

The Catalog contains coordinates, magnitudes, colors, and several other computed image parameters for all of the matched images on the blue and red plates. It provides information for the individual stars and galaxies down to fainter than 21st magnitude (in the blue). The calculation of accurate image parameters and the reliable separation of stellar and non-stellar images (galaxies) has long been a focus of our work with the APS data We were the first group to successfully apply AI techniques, specifically neural networks, to the image classification problem (see Odewahn *et al.* 1992. Odewahn *et al.* 1993, Odewahn 1995) Our neural network image classifier has been trained to the faint limit of the photograhpic plates and gives a success rate better than 90% to within one magnitude of the plate limit. It uses various image parameters with a back-propagation algorithm and two hidden layers to generate an output layer with two nodes, star or non–star(= galaxy). This "node_gal" value ranging from 0 to 1 also provides a confidence level of the classification and is cataloged with the image type.

The completed catalog of objects is available as an on-line database over the Internet (URL is http://aps.umn.edu). Querying is achieved with a custom-designed database management system called StarBase capable of handling millions of entries. StraBase was developed in collaboration with faculty and students of the University of Minnesota Computer Science Department. It uses specialized hashing on each image parameter derived for every catalog entry, including a two-dimensional hierarchical algorithm for positional search and retrieval. This level of optimization provides us with a DBMS that is faster and smaller than a commercial equivalent. A complementary image database is also available and includes all of the matched images in the object catalog as well as the unmatched images above the noise threshold on both the blue and red plates.

We have recently installed a federated database (FDBS) called Myriad (Lim *et al.* 1995) developed by Professor Jaideep Srivastava's group in our Computer Science Dept. The FDBS integrates the APS catalog and image database so that they appear as one easy-to-use resource. With a FDBS, the queries and transactions on the integrated database are performed as if it were a single database. The separate DBMS's are hidden from view by a flexible interface. It is important to emphasize that the FDBS permits horizontal access to the data, not just vertical. Queries can be made not only by sky position, but also by any parameter in either database.

Although we have had considerable success with our neural network-based object classifier for our research applications, for many astrophys-

ical problems the actual morphological type of the galaxy is very important, especially for studies of galaxy formation and evolution and large-scale structure in the universe. Working with members of our Computer Science Department, we have recently had some success applying data mining and pattern recognition codes to identifying the most useful parameters for automating the classification of the galaxy images by their morphological types.

2. THE MORPHOLOGICAL CLASSIFICATION OF GALAXIES

Ever since the discovery of galaxies, it has been known that these assemblies of stars, dust and gas have different morphological shapes. In 1936, Edwin Hubble established a system to classify galaxies into three fundamental types. Elliptical galaxies had an elliptical shape with no other discernible structure. Spiral galaxies had an elliptical nucleus surrounded by a flattened disk of stars and dust containing a spiral pattern of brighter stars. The irregular galaxies as their name suggests, were irregularly shaped and did not fit into the other two categories. As more galaxies were observed, it became apparent that the galaxy types formed a continuous sequence starting from nearly spherical galaxies toward more flattened ellipticals, through the lenticulars, galaxies with a large nucleus and small disk with no spiral structure, to the true spirals starting with tightly wound spiral arms and proceeding to less tightly wound arms and concluding with the irregulars. In other words, Hubble arranged galaxies in order of increasing complexity. Although many subdivisions and refinements have been made within the Hubble classification system, we are primarily concerned here with identifying the four basic types of galaxies: ellipticals, lenticulars, spirals, and irregulars.

The classification of galaxies is typically performed by visual inspection of photographic plates. This is by no means an easy task, requiring a great deal of practice and time on the part of the classifier. Large catalogs of galaxies containing a few tens of thousands of galaxies [e.g. the Third Reference Catalogue of Bright Galaxies (de Vaucouleur *et al.* 1991)] take years to compile. With today's large all–sky surveys, generating millions of galaxy images, human classification is no longer a viable option. Furthermore, although the types are well defined, human classifications tend to be subjective and it is difficult for independent researchers to reproduce results. Often it is very difficult to distinguish between adjacent types. For instance, a lenticular galaxy viewed face-on, looking down at the disk, looks very similar to an elliptical and all morphological catalogs have fewer face-on lenticulars than edge-on,

where the presence of a disk is more readily discerned. Studies also show that morphological catalogs of galaxies produced by even the best human classifiers disagree with other classifiers between 10% and 20% of the time. Therefore, in order to produce the large, objective, reproducible, morphological catalogs necessary for galaxy formation and evolution studies computer generated classifications are required.

There have been a few recent attempts to create an automated classification system, generally using artificial neural networks (Odewahn 1995, Naim *et. al.* 1995). While limited success has been achieved, these computer-based classifiers have yet to produce large, unbiased morphological galaxy catalogs. The reason for this seems to be that while it has been possible to train a neural network to correctly classify a well defined, hand picked, set of galaxies, when applied to the large random samples of galaxy images upon which any classifier must ultimately be applied, they fail to give results that can equal human classifications.

In our attempts to solve this problem we have visually classified some 1500 galaxies images obtained from the APS database in the region of the north galactic pole. Although this training set was chosen based solely on the brightness and size of galaxy images on 9 photographic plates, it is important to note that galaxies which were hard to classify (less than 1 %) were removed from this sample. The first problem is to identify a set of parameters which can separate the galaxies by their types. This has turned out to be quite challenging. The human eye can easily recognize complicated patterns in images such as spiral arms which tend to be spotty, blochy affairs that are difficult for automated techniquesi. Often it is necessary to rely on secondary effects such as color (spiral galaxies tend to be bluer than ellipticals) which are not specifically part of the classification system as originally conceived. If a picture is worth a thousand words, with a little imagination a galaxy image can be described by hundreds of parameters, all of which may have some relation to the morphological type. Currently, we calculate over five hundred such parameters for each galaxy in the APS database. Unfortunately, we cannot simply present all of these parameters to a neural network and let the training algorithm determine which are the most important. We would merely end up with a network that has memorized the training sample perfectly, but performs poorly on samples not seen during training. In order to have a reasonable chance of spanning a five hundred dimensional parameter space we would require a training sample of many millions: the thing we are trying to avoid.

With drastic increases in training set size ruled out for practical reasons, another option is to limit the number of parameters presented to

a neural network. The question is, which parameters to choose? If one or two parameters yielded adequate separation, we could merely plot all the parameters in turn and see which provided the greatest distance between the clumps defining the various types. Unfortunately, this is not the case. While several parameters show trends with galaxy type, no combination of two or three parameters is capable of solving the problem.

The problem of finding clusters in large dimensional spaces however, falls within the sphere of data mining. Working with the data mining group at the University of Minnesota, we applied the program Mineset to the task. This program allows quick evaluation and ranking of the parameters, as well as creating a decision tree classifier. Using the 10 best parameters we have been able to achieve a classifier with an 85% accuracy treating the Ellipticals and the lenticulars as a single class. While this is still short of our goal of creating a classifier as good as the human classifiers, it is a step in the right direction. In order to improve our classifier we continue to seek parameters that can provide better separation between the morphological types. However, it is possible that our 500 parameters already have enough information to correctly classify galaxies and by limiting the number to only ten we are ignoring useful information. At the same time, examination of the misclassified galaxies often reveals an anomaly in the image which confuses the computer classifier. Examples include forground stars or faint background galaxies within the galaxy image, or the presence of dust lanes in an otherwise structureless Elliptical galaxy. While a human classifier routinely discounts these deviations, automated classifiers see only the parameters presented to them. A possible solution is to train a large number of neural networks, each of which is presented a small number of the parameters. The final classification is then taken to be a weighted average of all the classifiers output. This procedure allows a more robust classification to be performed as one or two deviant parameters can be out-weighed by the vast majority of normal parameters for that galaxy type.

3. TESTING CLUSTERING ALGORITHMS ON AN ASTRONOMICAL DATABSE

To a computer scientist, "clustering" is a discovery process (Stonebraker *et al.* 1993, Chen *et al.* 1996) that groups objects such that the similarity between objects in the same group is maximized and the similarity between objects in different groups is minimized (Jain and Dubes 1988, Kaufman and Rousseeuw 1990, Chen *et al.* 1996). In astron-

omy, we often group objects into "populations" with distinct properties. There is obviously a large overlap between what an astronomer would call a "population" and what a computer scientist would call a "cluster." In this section we will concentrate on explaining how new clustering algorithms can be applied to existing (and future) astronomical databases to allow for automated identification of astronomical populations.

There are many different types of astronomical populations. If the properties distinguishing the populations are spatial (positions on the sky or in space), the populations identified may be real physical "clusters" of objects. For example, it has been well established that galaxies tend to lie near other galaxies, in "galaxy clusters." Spatial clusters are one very common form of population in astronomy. However, in addition to these spatial clusters, populations with similar physical and image parameters may exist both within the spatial clusters and independent of them. An example of a non-spatial population is the famous Hertzsprung-Russell diagram, a plot of color versus absolute brightness for stars. Stars tend to lie in relatively restricted regions of this parameter space. The stars in this diagram can therefore be classified into separate populations, including Main Sequence stars, Giants, White Dwarfs, etc. These stellar populations have other properties, outside of color and absolute brightness, that distinguish them, indicating they are indeed physically distinct types of stars. If we treat spatial, physical, and image parameters as part of a multivariate description of each object in an astronomical database, we can see that astronomical populations are just what a computer scientist would call a "cluster."

Traditionally the identification of astronomical populations has been done "by eye" through examination of parameter space plots. Recently, astronomers have started using computers to identify some populations in astronomical databases, usually by looking for populations they expect to exist. The promise of applying clustering algorithms to astronomical databases lies in the application of precisely defined criteria in identifying populations, criteria that are not subject to psychological or physiological biases. Developing clustering algorithms for astronomical datasets poses a number of challenges due to both the characteristics of the data discussed as well as the types of the desired clusters. The clusters may be of variable sizes and densities, and of arbitrary shapes. We need to develop clustering algorithms that are capable of accommodating different clustering objectives.

For spatial clustering, a new class of hierarchical agglomerative clustering algorithms are useful in identifying clusters with varying, non-uniform densities and arbitrary shapes. These algorithms use a dynamic-

modeling approach to measure the similarity between two clusters; thus allowing them to automatically adjust to the characteristics of each cluster. Two clusters are merged only when the distinctiveness of parameter values between the clusters is comparable to the internal scatter of the parameter values within each cluster. These algorithms forego the definition of a user-specified model (and the biases possible in such models), instead automatically adjusting the parameters distinguishing clusters from one another. Therefore, these algorithms are suitable for identification of spatial clusters of stars and galaxies, since they can identify clusters with homogeneous internal parameters even if the clusters vary in density, shape, or size.

An example of such a clustering algorithm is Chameleon (Karypis *et al.* 1999). Chameleon operates on a sparse graph where nodes represent objects and weighted edges represent the similarities between objects. This sparse graph representation allows Chameleon to scale to the large data sets becoming common in astronomy. Initial investigations with Chameleon show that it is able to correctly identify clusters of varying size, orientation, shape. In addition to this strength, Chameleon is tolerant of noise and outliers, something common to all astronomical datasets. It is our goal to apply Chameleon to a subset of the APS catalog, a catalog of over 200,000 galaxies in the region surrounding the North Galactic Pole, called the MAPS-NGP. Using Chameleon on the MAPS-NGP, we hope to identify not only previously recognized galaxy clusters (Abell 1958) but new ones as well. The challenge when applying such an algorithm to astronomical datasets is that every object will usually have a great variety of parameters describing it, thus essentially, instead of a sparse graph, we have a dense graph.

Another approach to the problem of identifying astronomical populations is unsupervised clustering. Unsupervised clustering can be used to discover structure within a large dataset as well as grouping similar objects together independent of any user-defined classes. Once such structuring of the data is preformed, further exploration of the data set is made easier, since analyses that would normally have been applied to the entire data set can now be applied to hierarchically structured subsets. One unsupervised clustering algorithm is the Principal Direction Divisive Partitioning (PDDP) algorithm, which has been demonstrated to be a fast method of constructing a hierarchical clustering tree top-down (Boley 1998, Chen *et al.* 1996). PDDP is fast and scalable to very large datasets, producing data structures that identify the attributes distinguishing one cluster from another, in a completely unsupervised way. Initial application of the PDDP algorithm to the MAPS-NGP

should provide a way of identifying distinct populations of objects, with information on just what makes them distinct.

We have found that the problems with existing data mining techniques when applied to astronomical datasets are largely tied to the vast size of modern astronomical datasets plus the density of records. Often, records for an individual astronomical object will contain dozens of fields (each with a value). And in modern, multi-wavelength astronomy, cross-identification of the same object in separate databases is a common technique (see, Cabanela and Dickey 2001 for an example of cross-identified radio and optical data used to identify a subset population of low surface brightness galaxies from the general galaxy population). Data mining techniques will be required to handle hundreds of fields associated with each record, with the values of items in those fields retrieved through integrated and efficient access to multiple, distributed, databases around the world. When data mining techniques meet this challenge, the process of astronomical discovery should accelerate tremendously, and the astronomer's efforts will be able to focus on interpreting the relationships for astronomical populations (maybe we'll use the term "cluster" by then) instead of searching for them.

References

Abell, G.O. 1958, "The Distribution of Rich Clusters of Galaxies", Astrophys. J. Suppl., 3, 211.

Boley, D., 1998, "Principal Direction Divisive Partitioning", Data Mining and Knowledge Discovery, 2, 325.

Cabanela, J.E. & Dickey, J.M. 2001, "Galaxies on the Blue Edge", Astron. J., to be published.

Chen, M.S., Han, J., & Yu, P.S., 1996, "Data Mining: An Overview from a Database Perspective", IEEE Transactions on Knowledge and Data Engineering, 8(6), 866.

de Vaucouleurs, G., de Vaucouleurs, A., Corwin, Jr., H. G., Buta, R. J., Paturel, G., & Fouque, P. 1991, Third Reference Catalogue of Bright Galaxies (Springer-Verlag New York Inc., New York).

Han, S., Boley, D., Gini, M., Gross, R., Hastings, K., Karypis, G., Kumar, V., Mobasher, B., & Moore, J., 1998, "A Web Agent for Document Categorization and Exploration," Autonomous Agents '98 Conf.

Jain, A.K., & Dubes, R.C., 1988, "Algorithms for Clustering Data," Prentice Hall.

Karypis, G., Aggarwal, R., Kumer, V., & Shekhar, S., 1999, "Multilevel Hypergraph Partitioning: Applications in VLSI Domain," IEEE Transactions of VLSI Systems, 7(1), 69.

Kaufman, L., & Rousseeuw, P.J., 1990, Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons.

Lim, e.-P., Hwang, S.-Y., Srivastava, J., Clements, D., & Ganesh, M., 1995, Software Practice and Experience, 25, 533.

Naim, A., Lahav, O., Sodre, L. & Storrie-Lombardi, M. C., 1995, "Automated Morphological Classification of APM Galaxies by Supervised Artificial Neural Networks", MNRAS, 275, 567.

Odewahn, S. C., Stockwell, E. B., Pennington, R. L., Humphreys, R. M. & Zumach, W., 1992, "Automated Star–Galaxy Discrimination with Neural Networks", Astron. J., 103, 318.

Odewahn, S. C., Humphreys, R. M., Aldering, G., & Thurmes, P. M., 1993, ""Star–Galaxy Separation with a Neural Network II. Multiple Schmidt Plates", Pub. Astron. Soc. Pacific, 105, 1354.

Odewahn, S. C., 1995 "Automated Classification of Astronomical Images", Pub. Astron. Soc. Pacific, 107, 770.

Stonebraker, M., Agrawal, R., Dayal, U., Neuhold, E.J., Reuter, A., 1993, "DBMS Research at a Crossroads: The Vienna Update," Proc. Of the 19th VLDB Conference.