Data Mining for Astronomy

Given the size of the universe and the ease with which massive swaths of it can be imaged at high resolution, it is no wonder that huge amounts of data are becoming the standard input for astronomical research. Over the years, many surveys of the sky have been completed at varying scales and over a wide spectrum of electromagnetic frequencies. Images from these surveys have been stored in a large number of huge public databases. While in the past astronomers have often been able to find and study objects of interest with only simple scans of the data, as scientists explore ever more subtle and remote objects in space, it is becoming necessary to rely on automated techniques to sift through the mountains of pixels and find the interesting bits. Data mining has proven very effective at helping astronomers identify which of the innumerable of specks of light are the objects they wish to look at.

Chandrika Kamath and her colleagues used such techniques to find double-bent galaxies in the Faint Images of the Radio Sky at Twenty-cm (FIRST) survey [4]. Astronomers are interested in finding double-bent galaxies because these indicate the presence of galaxy clusters. The FIRST survey is a radio frequency record of the sky that, at the time of Kamath et al.’s paper, consisted of 32,000 two million pixel pictures. Before Kamath came along, astronomers had been manually scanning these images (which mostly consist of noise) for blobs that look like double-bent galaxies. This process was both highly time consuming and subjective. Data mining addressed both these issues. Computers can crunch through the data much faster than humans and objective and quantifiable statistical rules and algorithms eliminate most of the previous subjectivity. Kamath et al. used an iterative process of progressively paring down the data to find the radio signals that were likely sent by double-bent galaxies:

First, the raw image data was preprocessed and condensed into a catalog of significant image ‘blobs’. Each blob was modeled in the catalog as an elliptical Gaussian. Nearby Gaussians were then grouped as being part of the same radio source, and the list of candidate radio sources were pruned down based on a few relatively safe heuristics.

The heart of the data mining was performed on this condensed data set. Experts (FIRST astronomers) were consulted to determine which set of features in the data (e.g. relative positions and orientations of a radio source’s Gaussians, overall shape of a radio source, etc) would be useful in discriminating between the various types of galaxies and identifying double-bent morphologies. Then, this feature set was used as the input to an ensemble of decision tree classifiers. Each of these classifiers was trained on a random sample of galaxies pre-classified by the FIRST experts. Once trained the ensemble of trees voted on the proper classification of novel radio sources.

It is worth noting that while this mining process was largely objective, an element of human subjectivity does enter in the choice of features to use as parameters. Indeed, Kamath et al. point out that the classification results are much more susceptible to variations in the choice of feature set than in the more mechanical choice of which classifier to use.
Kirk Borne and Allison Chang took a somewhat different approach to automated classification [1]. To aid the search for extra-solar planets, they used a clustering algorithm to identify various types of stellar systems. More specifically, they ran a K-Means clustering algorithm (with K = 4) on 1600 stars from the USNO-B catalog for nearby stars. In the same feature space, the authors then plotted the 129 stars that were currently listed in the extrasolar catalog – a catalog of stars known to have extra-solar planets. The authors used this comparison to determine which of the clusters was most likely to contain the most stars that had extra-solar planets orbiting around them. Finally, the authors listed the 1600 stars based on their proximity in feature space to the chosen cluster, claiming this list reflects ranks the likelihood of each of the star systems containing extra-solar planets.

While Borne and Chang were primarily creating artificial clusters simply as a tool for the classification of a single type of object (stars with extra-solar planets), their approach hints at how data mining might be used to reveal more complex statistical information. For example, Borne and Chang did not find clear distinctions between the majority of their clusters, but perhaps in a different feature space, distinct clusters could be found. In such a case, the data might reveal fundamental and previously unknown distinctions between various types of stellar systems. This would suggest new taxonomies of stellar systems, and might lead to the discovery of underlying hidden variables that cause the separation of the feature clusters.

Clustering is also excellent at revealing anomalous data points. Borne and Chang’s clusters were all contiguous except one, which contained a single anomalous star. While the authors did not explore this anomaly, in general it is useful to have outliers identified since outliers often merit special study to find out what makes them different. Indeed, elsewhere researchers have set loose unsupervised clustering algorithms to just generally find rare celestial objects, betting that such objects when found will be interesting to study [3].

In order to find such statistical subtleties, it is often necessary to explore the full range of possible feature spaces. Unfortunately, the many astronomical databases that currently exist are, for the most part, incompatible with each other. This means that radio image surveys, for example, cannot be easily comparable with infrared image surveys. Since each particular database has its own shortcomings and limitations, it would be very useful to be able mine data across multiple databases. Efforts are being made along these lines: The US National Virtual Observatory (NVO) is a meta-database that combines data from disparate surveys into a common format. This database has great potential as a data mining resource, but combining all surveyed astronomical data is not easy. Cross-correlating between multiple surveys is more difficult than simply matching up spatial coordinates. If two images of the same region of space have differing resolutions, two objects that appear distinct in the first image may be merged in the second. Or, if the two images are sensitive to differing electromagnetic frequencies, a strong signal in the first image may not even show up in the second.

In spite of these difficulties, the NVO has managed to accumulate a huge amount of data that has already begun to be mined [2]. While mining is already being actively applied to celestial object classification, in the future, data crunching might also be used to reveal more subtle patterns and relationships in the vast catalogs of the sky.
References


Ye, Nong. *The Handbook of Data Mining*. 