Various project. 3.1. object classification classification is an

Various data miningalgorithms are used by astronomers in most of the applications in astronomy. However, studies and several projects have also been made by data mining expertsutilizing astronomical data because astronomy has produced many large datasetsthat are flexible to the approach along with other fields such as medicine andhigh energy physics. Examples of such projects are the SKICAT-Sky ImageCataloging and Analysis System for catalog production and catalog analysis fromdigitized sky surveys particularly the scans of the second Palomar ObservatorySky Survey; the JAR Tool- Jet Propulsion Laboratory Adaptive Recognition Toolused for recognition of volcanoes in the over 30, 000 images of Venus returnedby the Magellan mission; the subsequent and more general Diamond Eye and theLawrence Livermore National Laboratory Sapphire project. 3. 1.

Object classificationClassification is animportant preliminary step in the scientific process as it provides a methodfor organizing information in a way that can be used to make hypotheses and comparewith models. The two useful concepts in object classification are
the completeness andthe efficiency, also known as recall and precision. They aredefined in terms of true and false positives (TP and FP) and true and falsenegatives (TN and FN).

The completeness is the fraction of objects that aretruly of a given type that are classified as that type: and the efficiency is thefraction of objects classified as a given type that are truly of that type These two quantities areinteresting astrophysically because, while one wants both higher completenessand efficiency, there is generally a tradeoff involved. The importance of eachoften depends on the application, for example, an

investigation of rare objectsgenerally requires high completeness while allowing some contamination (lowerefficiency), but statistical clustering of cosmological objects requires highefficiency, even at the expense of completeness. 3. 1. 1. Star-GalaxySeparationDue to their smallphysical size in comparison to their distance from us, almost all stars areunresolved in photometric datasets, and thus appear as point sources. Galaxies, however, despite being further away, generally subtend a larger angle, and thusappear as extended sources.

However, other astrophysical objects such asquasars and supernovae, also appear as point sources. Thus, the separation ofphotometric catalogs into stars and galaxies, or more generally, stars, galaxies, and other objects, is an important problem. The sheer number ofgalaxies and stars in typical surveys (of order 108 or above)requires that such separation be automated. This problem is a wellstudied one and automated approaches were employed even before current datamining algorithms became popular, for example, during digitization by thescanning of photographic plates by machines such as the APM and DPOSS. Severaldata mining algorithms have been employed, including ANN, DT, mixture modeling, and SOM, with most algorithms achieving over 95% efficiency. Typically, this isdone using a set of measured morphological parameters that are derived from thesurvey photometry, with perhaps colors or other information, such as theseeing, as a prior. The advantage of this data mining approach is that all suchinformation about each object is easily incorporated.

3. 1. 2. GalaxyMorphologyGalaxies come in a rangeof different sizes and shapes, or more collectively, morphology. The mostwell-known system for

the morphological classification of galaxies is theHubble Sequence of elliptical, spiral, barred spiral, and irregular, along withvarious subclasses. This system correlates to many physical properties known tobe important in the formation and evolution of galaxies.

Because galaxy morphology is a complex phenomenon that correlatesto the underlying physics, but is not unique to any one given process, theHubble sequence has endured, despite it being rather subjective and based onvisible-light morphology originally derived from blue-biased photographicplates. The Hubble sequence has been extended in various ways, and for datamining purposes the T system has been extensively used. This system maps the categorical Hubble types E, S0, Sa, Sb, Sc, Sd, and Irronto the numerical values -5 to 10. One can, therefore, traina supervised algorithm to assign T types to images for which measuredparameters are available.

Such parameters can be purely morphological, orinclude other information such as color. A series of papers by Lahav andcollaborators do exactly this, by applying ANNs to predict the T type ofgalaxies at low redshift, and finding equal accuracy to human experts. ANNshave also been applied to higher redshift data to distinguish between normaland peculiar galaxies and the fundamentally topological and unsupervised SOMANN has been used to classify galaxies from Hubble Space Telescope images, where the initial distribution of classes is not known. Likewise, ANNs havebeen used to obtain morphological types from galaxy spectra. 3. 2.

Photometric redshiftsAn area of astrophysicsthat has greatly increased in popularity in the last few years is theestimation of redshifts from photometric data (photo-zs). This isbecause, although the distances are less accurate than those obtained withspectra, the sheer number of objects with photometric measurements can oftenmake up for the reduction in individual accuracy by suppressing the statisticalnoise of an ensemble calculation. The two common approachesto photo-zs are the template method and the empirical training setmethod. The template approach has many complicating issues, includingcalibration, zero-points, priors, multiwavelength performance (e. g., poor inthe mid-infrared), and difficulty handling missing or incomplete training data. We focus in this review on the empirical approach, as it is an implementationof supervised learning. 3.

2. 1. GalaxiesAt low redshifts, the calculation of photometric redshifts fornormal galaxies is quite straightforward due to the break in the typical galaxyspectrum at 4000A.

Thus, as a galaxy is redshifted with increasing distance, the color (measured as a difference in magnitudes) changes relatively smoothly. As a result, both template and empirical photo-z approaches obtainsimilar results, a root-mean-square deviation of ~ 0. 02 in redshift, which isclose to the best possible result given the intrinsic spread in the properties. This has been shown with ANNs SVM DT, kNN, empirical polynomialrelations, numerous template-based studies, and several other methods. Athigher redshifts, obtaining accurate results becomes more difficult because the4000A break is shifted redward of the optical, galaxies are fainter and thusspectral data are sparser, and galaxies intrinsically evolve over time.

Whilesupervised learning has been successfully used, beyond the spectral regime theobvious limitation arises that in order to reach the limiting magnitude of thephotometric portions of surveys, extrapolation would be required. In this regime, or where only small training sets are available, template-based results can beused, but without spectral information, the templates themselves are beingextrapolated. However, the extrapolation of the templates is being done in amore physically motivated manner. It is likely that the more general hybridapproach of using empirical data to iteratively improve the templates or thesemi-supervised method described in will ultimately provide a more elegantsolution. Another issue at higher redshift is that the available numbers ofobjects can become quite small (in the hundreds or fewer), thus reintroducingthe curse of dimensionality by a simple lack of objects compared to measuredwavebands.

The methods of dimension reduction can help to mitigate this effect.