Automated Analysis of a Large-Scale Sky Survey: The SKICAT System

Usama M. Fayyad
AI Group M/S 525-3660
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109
Fayyad@aig.jpl.nasa.gov

Nicholas Weir and **S. Djorgovski**
Palomar Observatory
M/S 105-24
California Institute of Technology
Pasadena, CA 91125
{weir,george}@deimos.caltech.edu

ABSTRACT

We describe the application of machine learning and state-of-the-art database management technology to the development of an automated tool for the reduction and analysis of a large astronomical data set. The 3 terabytes worth of images are expected to contain on the order of $5x10^7$ galaxies and $5x10^8$ stars. For the primary scientific analysis of these data, it is necessary to detect, measure, and classify every sky object. The size of the complete data set precludes manual reduction, requiring an automated approach. SKICAT integrates techniques for image processing, data classification, and database management. Once sky objects are detected, a set of basic features for each object are computed. The learning algorithms are trained to classify the detected objects and can classify objects too faint for visual classification with an accuracy level of about 94%. This increases the number of classified objects in the final catalog three-fold relative to the best results from digitized photographic sky surveys to date. The tasks of managing and matching the resulting hundreds of plate catalogs is accomplished using custom software and the Sybase relational DBMS. A full array of scientific analysis tools are provided for filtering, manipulating, plotting, and listing the data in the sky object database. We are currently experimenting with the use of machine discovery tools, such as the AUTOCLASS unsupervised classification program, on the data. SKICAT represents a system in which machine learning played a powerful and enabling role, and solved a difficult, scientifically significant problem. The primary benefits of our overall approach are increased data reduction throughput; consistency of classification; and the ability to easily access, analyze, and create new information from an otherwise unfathomable data set.

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 field of artificial intelligence (AI), machine learning, and related areas can provide at least some solutions. Much of the future of scientific information processing lies in the implementation of these methods.

In this paper we present an application of machine learning and data processing technology to the automation of the tasks of cataloging and analyzing objects in digitized sky images. The Sky Image Cataloguing and Analysis Tool (SKICAT) is being 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 being digitized at the Space Telescope Science Institute (STScI). This process will result in about 3,000 digital images of roughly $23,000^2$ pixels each, resulting in 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 21.5. This is at least one magnitude fainter than previous comprable photographic surveys. We estimate that at least $5x10^7$ galaxies $5x10^8$ stellar objects (including over 10^5

quasars) will be detected. This data set will be the most comprehensive large-scale imaging survey produced to date and will not be surpassed in quality or size until the completion of a fully digital all-sky survey.

The purpose of SKICAT is to facilitate the extraction of meaningful information from such a large database in an efficient and 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. In this paper we describe the implementation of these three components, with particular emphasis on the first. It is there where, to date, we have already realized the significant advantages of AI for a data reduction problem of this magnitude. However, the subsequent tasks of managing and updating the sky object database in the face of new and better data, not to mention the large-scale statistical analysis of the full data set, similarly cry out for the application of automated information processing and exploration technology. These aspects of SKICAT comprise a significant portion of our ongoing research.

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 [Weir92]. 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 this paper 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 therefore do not provide the details of either the learning algorithms or the technical aspects of the domain. We aim to point out an instance where learning algorithms proved to a be useful and powerful tool in the automation of scientific data analysis.

2. Background on Learning Algorithms

A familiar context in which machine learning classification techniques have been used is to overcome the "knowledge acquisition bottleneck" [Feig81] due to experts finding it difficult to express their knowledge in terms of concise situation-action rules. The growing number of large scientific databases provides another niche for machine learning applications. Problems include searching for and detecting patterns of interest, performing pre-processing necessary for subsequent analysis, as well as automating analysis subtasks. Sizes are now becoming too large for manual processing. Learning techniques can serve as effective tools for aiding in the analysis, reduction, and visualization of large scientific databases.

2.1. Induction of Decision Trees

A particularly efficient method for extracting rules from data is to generate a decision tree [Brei84, Quin86]. A decision tree consists of nodes that are tests on the attributes. The outgoing branches of a node correspond to all the possible outcomes of the test at the node. The examples at a node in the tree are thus partitioned along the branches and each child node gets its corresponding subset of examples. A well-known algorithm for generating decision trees is Quinlan's ID3 [Quin86] with extended versions called C4 [Quin90].

ID3 starts by placing all the training examples at the root node of the tree. An attribute is selected to partition the data. For each value of the attribute, a branch is created and the corresponding subset

of examples that have the attribute value specified by the branch are moved to the newly created child node. The algorithm is applied recursively to each child node until either all examples at a node are of one class, or all the examples at that node have the same values for all the attributes. Every leaf in the decision tree represents a classification rule.

Note that the critical decision in such a top-down decision tree generation algorithm is the choice of attribute at a node. Attribute selection in ID3 and C4 is based on minimizing an information entropy measure applied to the examples at a node. The measure favors attributes that result in partitioning the data into subsets that have low class entropy. A subset of data has low class entropy when the majority of examples in it belong to a single class. The algorithm basically chooses the attribute that provides the locally maximum degree of discrimination between classes. For a detailed discussion of the information entropy selection criterion see [Quin86, Fayy91, Fayy92].

2.2. THE GID3* AND O-BTREE ALGORITHMS

The criterion for choosing the attribute clearly determines whether a "good" or "bad" tree is generated by the algorithm[1]. Since making the optimal attribute choice is computationally infeasible, ID3 utilizes a heuristic criterion which favors the attribute that results in the partition having the least information entropy with respect to the classes. This is generally a good criterion and often results in relatively good choices. However, there are weaknesses inherent in the ID3 algorithm that are due mainly to the fact that it creates a branch for each value of the attribute chosen for branching. The overbranching problem in ID3 leads to several problems, since in general it may be the case that only a subset of values of an attribute are of relevance to the classification task while the rest of the values may not have any special predictive value for the classes. These extra branches are harmful in three ways [Fayy91]:

1. They result in rules that are overspecialized (conditioned on particular irrelevant attribute values).
2. They unnecessarily partition the data, thus reducing the number of examples at each node. Subsequent attribute choices will be based on an unjustifiably reduced subset of data.
3. They increase the likelihood of occurrence of the missing branches problem (see [Chen88,Fayy91] for more details).

The GID3* algorithm was designed mainly to overcome this problem. It utilizes a vector distance measure applied to the class vectors of an example partition, in conjunction with the entropy measure, to create for each attribute a *phantom attribute* that has only a subset of the original attribute's values. We generalized the ID3 algorithm so that it does not necessarily branch on each value of the chosen attribute. GID3* can branch on arbitrary individual values of an attribute and "lump" the rest of the values in a single *default branch*. Unlike the other branches of the tree which represent a single value, the default branch represents a subset of values of an attribute. Unnecessary subdivision of the data may thus be reduced. See [Fayy91] for more details and for empirical evidence of improvement.

The O-Btree algorithm [Fayy92b] was designed to overcome problems with the information entropy selection measure itself. O-Btree creates strictly binary trees and utilizes a measure from a family of measures (C-SEP) that detects class separation rather than class impurity. Information entropy is a member of the class of impurity measures. O-Btree employs an orthogonality measure rather than entropy for branching. For details on problems with entropy measures and empirical evaluation of O-Btree, the reader is referred to [Fayy91,Fayy92b].

Both O-Btree and GID3* differ from ID3 and C4 along one additional aspect: the discretization algorithm used at each node to discretize continuous-valued attributes. Whereas ID3 and C4 utilize a binary interval discretization algorithm, we utilize a generalized version of that algorithm which derives multiple intervals rather than strictly two. For details and empirical tests showing that this algorithm does indeed produce better trees see [Fayy91,Fayy93]. We have found that this ability improves performance considerably in several domains.

---

[1] See [Fayy90,Fayy91] for the details of what we formally mean by one decision tree being better than another.

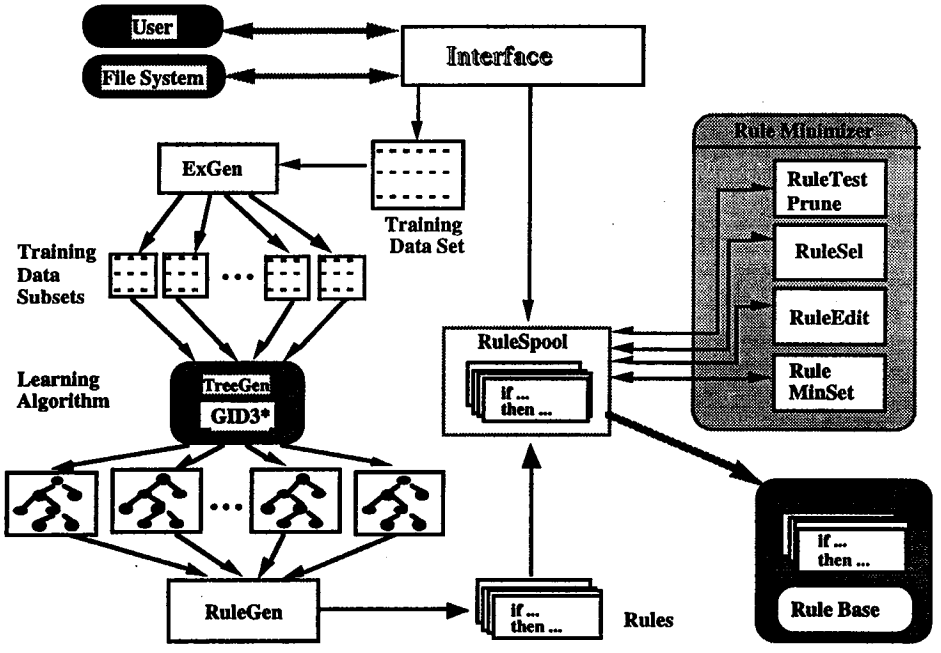

Figure 1. Architecture of the RULER Rule Induction System

2.3. THE RULER SYSTEM

There are limitations to decision tree generation algorithms that derive from the inherent fact that the classification rules they produce originate from a single tree. This fact was recognized by practitioners early on [Brei84,Quin87]. Tree pruning is used to overcome the fact that in any good tree there are always leaves that are overspecialized or predict the wrong class. The very reason which makes decision tree generation efficient (the fact that data is quickly partitioned into ever smaller subsets), is also the reason why overspecialization or incorrect classification occurs. It is our philosophy that once we have good, efficient, decision tree generators, they could be used to generate multiple trees, and only the best rules in each tree are kept. We initially developed the RIST system [Chen90] which later evolved into the RULER system to implement such a scheme. Figure 1 gives an overview of the RULER system.

RULER starts with a data set, and randomly divides it into a training subset and test subset. A decision tree is generated from the training set and its rules are tested on the corresponding test set. Using Fisher's exact test [Finn63] (the exact hyper geometric distribution) RULER evaluates each condition in a given rule's preconditions for relevance to the class predicted by the rule. It computes the probability that the condition is correlated with the class by chance [2]. If this probability is higher than a small threshold (say 0.01), the condition is deemed irrelevant and is pruned. In addition, RULER also measures the merit of the entire rule by applying the test to the entire precondition as a unit. This process serves as a filter which passes only robust, general, and correct rules.

By gathering a large number of rules through iterating on randomly subsampled training sets, RULER builds a large rule base of robust rules that collectively cover the entire original data set of examples. A greedy covering algorithm is then employed to select a minimal subset of rules that covers the examples. The set is minimal in the sense that no rule could be removed without losing complete coverage of the original training set. Using RULER, we can typically produce a robust set of rules that has fewer rules than any of the original decision trees used to create it. Furthermore, any learning algorithm that produces rules can be used as the rule generating

[2] The Chi-square test is actually an approximation to Fisher's exact test when the number of test examples is large. We use Fisher's exact test because it is robust for small and large data sets.

component. We use decision tree algorithms since they constitute a fast and efficient method for generating a set of rules. This allows us to iterate many times without requiring extensive amounts of time and computation.

3. CLASSIFYING SKY OBJECTS

SKICAT provides an integrated environment for the construction, classification, management, and analysis of catalogs from large-scale imaging surveys, in particular the digitized POSS-II. Due to the large amounts of data being collected, a manual approach to detecting and classifying sky objects in the images is infeasible: it would require on the order of tens of man years. Existing computational methods for classifying the images would preclude the identification of the majority of objects in each image since they are at levels too faint for traditional recognition algorithms or even manual inspection/analysis approaches. A principal goal of SKICAT is to provide an effective, objective, and examinable basis for classifying sky objects at levels beyond the limits of existing technology.

The photographic plates collected from the survey are being digitized at the Space Telescope Science Institute (STScI). This process will result in about 3,000 digital images of $23{,}040^2$ pixels each. A digitized plate is subdivided into a set of partially overlapping frames. Each frame represents a small part of the plate that is small enough to be manipulated and processed conveniently. Figure 2 depicts the overall architecture of the SKICAT plate catalog construction and classification process. Low-level image processing and object separation is performed by a modified version of the FOCAS image processing software developed at Bell Labs [Jarv81,Vald82]. In addition to defining the objects in each image, FOCAS also produces basic attributes describing each object. The paragraph below will explain the loop in the bottom left-hand corner in which machine learning is employed in the attribute measurement process. The image processing steps detect contiguous pixels in the image that are to be grouped as one object. Attributes are then measured based on this segmentation.

An extra step in attribute measurement involves selecting a subset of the objects in the frame, designating them as being "sure-thing" stars, and averaging their pixel values to define a point spread function (PSF) template. The purpose of this step is to facilitate the measurement of the *resolution attributes*: two parameters defining the distortion of the PSF best fitting each sky object. These attributes are particularly powerful and robust for classification purposes, as they have nearly identical distributions as a function of class for every portion of an image, as well as different images. To form the PSF template, the sure-thing stars must generally be selected by the astronomer. They represent the "archetypal" stars in that image. Once the stars are selected, the template is formed, and the resolution measurements are computed automatically. We refer to this problem as the *star selection subproblem.*

The total number of attributes measured for each object by SKICAT is 40. All steps are automated except for star selection and final sky object classification. The base-level attributes measured are generic quantities typically used in astronomical analyses [Vald82], including:

- isophotal, aperture, core, and asymptotic "total" magnitudes
- isophotal and "total" areas
- sky brightness and sigma (variance)
- peak, intensity weighted, and unweighted positions: xc, yc, icx, icy, cx, cy
- intensity weighted and unweighted image moments: ir1, ir2, ir3, ir4, r1, r2, ixx, iyy, ixy, xx, yy, xy
- ellipticity
- position angle (orientation)

Once all attributes, including the resolution attributes, for each object are measured, final classification is performed on the catalog. Our current goal is to classify objects into four major categories, following the original scheme in FOCAS: star (**s**), star with fuzz (**sf**), galaxy (**g**), and artifact (**long**). We may later refine the classification into more classes, however, classification

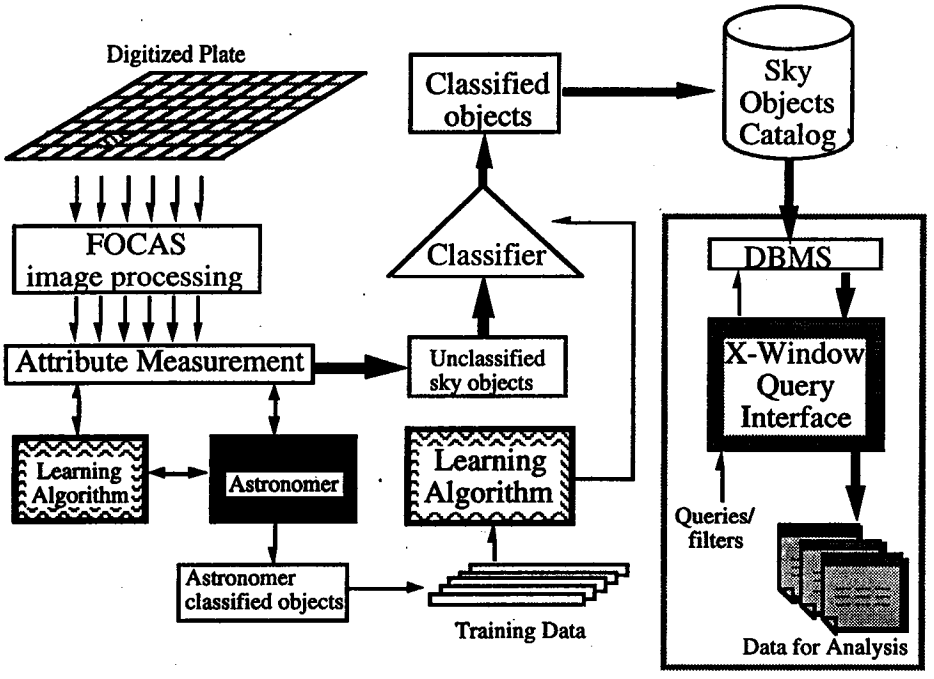

Figure 2. Architecture of the SKICAT Cataloguing and Classification Process

into one of these four classes represents adequate discrimination for primary astronomical analyses of the catalogs.

3.1. Classifying Faint Objects and the Use of CCD Images

In addition to the scanned photographic plates, we have access to CCD images that span several small regions in some of the frames. CCD images are obtained from a separate telescope. The main advantage of a CCD image is higher resolution and signal-to-noise ratio at fainter levels. Hence, many of the objects that are too faint to be classified by inspection on a photographic plate are easily classifiable in a CCD image. In addition to using these images for photometric calibration of the photographic plates, we make use of CCD images in two very important ways for the machine learning aspect:

1. CCD images enable us to obtain class labels for faint objects in the photographic plates.
2. CCD images provide us with the means to reliably evaluate the accuracy of the classifiers obtained from the decision tree learning algorithms.

Recall that the image processing package FOCAS provides the measurements for the base-level attributes (and the resolution attributes after star selection) for each object in the image. In order to produce a classifier that classifies faint objects correctly, the learning algorithm needs training data consisting of faint 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 provided by FOCAS. 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.

3.2. CLASSIFICATION RESULTS

Starting with digitized frames obtained from a single digitized plate, we performed initial tests to evaluate the accuracy of the classifiers produced by the machine learning algorithms ID3, GID3*, and O-BTree. The data consisted of objects collected from four different plates from regions for which we had CCD image coverage (since this is data for which true accurate classifications are available). The learning algorithms are trained on a data set from 3 plates and tested on data from the remaining plate for cross validation. This estimates our accuracy in classifying objects across plates. Note that the plates cover different regions of the sky and that CCD frames cover multiple minute portions of each plate. The training data consisted of 1,688 objects that were classified manually by one of the authors (NW) by examining the corresponding CCD frames. It is noteworthy that for the majority of these objects, the astronomer would not be able to reliably determine the classes by examining the corresponding survey (digitized photographic) images. All attributes used by the learning algorithms are derived from the survey images and not, of course, from the higher resolution CCD frames.

Table 1. Summary of results using all attributes.

ID3		GID3*		O-Btree		RULER	
#rules	accuracy	#rules	accuracy	#rules	accuracy	#rules	accuracy
73	75.6%	58	90.1%	54	91.2%	45	94.2%

Using all the attributes, including the two resolution attributes derived after star selection, the classification results are shown in Table 1.

The results for RULER above are shown with O-Btree as the decision tree generation component and were obtained by cycling through tree generation and rule merging 10 times. Results Using GID3* as the tree generating component for RULER are similar. Using ID3 in the inner loop, however, the results were not as good: the accuracy in this case was only around 85%.

When the same experiments were conducted without using the *resolution scale* and *resolution fraction* attributes the results were significantly worse. The error rates jumped above 20% for O-BTree, above 25% for GID3*, and above 30% for ID3. The respective sizes of the trees grew significantly as well.

The initial results may be summarized as follows:

1. Algorithms GID3* and O-BTree produced significantly better trees than ID3.
2. Classification accuracy results of better than 90% were obtained when using two user-defined attributes: *resolution fraction* and *resolution scale*.
3. Classification results were not as reliable and stable if we exclude the two resolution attributes.

We took this as evidence that the resolution attributes are very important for the classification task. Hence, we turned to the task of automating the star selection subproblem. Furthermore, the results point out that the GID3* and O-BTree learning algorithms are more appropriate than ID3 for the final classification task. As expected, the use of RULER resulted in improvement in performance.

3.3. THE STAR SELECTION SUBPROBLEM

Based on the initial results of the previous section, it was determined that using the resolution attributes is necessary, since without them the error rates were significantly worse. We do not have the option of leaving star selection as a manual step in the process, since it is a time consuming task and will easily become the bottleneck in the system. We decided to use a machine learning approach to solve the star selection subproblem.

The star selection subproblem is a binary classification problem. Given a set of objects in an image, the goal is to classify them as sure-thing stars and non-sure-thing stars. Unlike the overall

classification problem, the star selection problem turned out to be a much easier classification problem, as suitable "sure-thing" stars are much brighter and easier to distinguish than the average sky object. The data objects from all three plates described above were classified manually by one of the authors (NW.) into *sure-stars*, *non-sure-stars*, and *unknowns*. The goal of the learning subproblem is to construct classifiers for selecting out sure-stars from any collection of sky objects. The results of applying the learning algorithms to the data sets described above, using only the basic set of attributes derived by FOCAS, gave the results shown in Table 2.

Table 2. Sure-star selection results.

ID3		GID3*		O-Btree	
#rules	accuracy	#rules	accuracy	#rules	accuracy
41	95%	35	97.3%	29	98.7%

In this case, using RULER with O-Btree did not change the results significantly. Note that a 98.7% accuracy rate on this subproblem is more than sufficient to indicate that this subproblem is essentially completely solved. Consequently, this allows us to automate all the steps in the plate processing and obtain an overall classification rate of better than 94% as shown in Table 1. One note about this learning subproblem: the results reflect the accuracy in selecting sure-thing stars and not the classification error rate. In other words, we only care about the performance in terms of sure-thing stars selected correctly. Sure-stars classified as galaxies or unknowns does not concern us since all we need is a subset of good stars with which to form an unbiased PSF template. Since this is not the main classification task, we only present the relevant performance aspects to avoid confusion.

In order to achieve stable classification accuracy results on classifying data from different plates, we had to spend some effort in defining normalized attributes, other than resolution scale and fraction, that are less sensitive to plate-to-plate variation. It was determined that the base-level attributes such as area, background-sky-levels, and average intensity are image-dependent as well as object-dependent. It was also determined that a new set of user-defined attributes needed to be formulated. These attributes were to be computed automatically from the data, and are defined such that their values would be normalized across images and plates. The technique we use to derive such attributes is to derive non-linear curves in two dimensions defined by two of the base-level attributes and then define a new attribute to be the distance of each object in the 2-D plane to that curve. These quantities are ones that astronomers use, and many of them have physical interpretations.

It is beyond the scope of this paper to give the detailed definitions of these new attributes. As expected, defining the new "normalized" attributes raised our performance on both intra- and inter-plate classification to acceptable levels varying between 92% and 98% accuracy with an average of 94%. Note that without these derived attributes the cross-plate classification accuracy drops to 60%-80% levels when classifying data from different plates. Encoding of these attributes represents an implicit imparting of more domain knowledge to the learning algorithm.

3.4. Comparison with Neural Nets

In order to compare against other learning algorithms, and to preclude the possibility that a decision tree based approach is imposing *a priori* limitations on the achievable classification levels, we tested several neural network algorithms for comparison. The results indicate that neural network algorithms achieve similar, and sometimes worse, performance than the decision trees. The neural net learning algorithms tested were:

1. traditional backpropagation,
2. conjugate gradient optimization, and
3. variable metric optimization.

Unlike backpropagation, the latter two training algorithms work in batch mode and use standard numerical optimization techniques in changing the network weights [Hert91]. They compute the weight adjustments simultaneously using matrix operations based on the total error of the network

on the entire training set. Their main advantage over traditional backpropagation is the significant speed-up in training time.

The results can be summarized as follows: The performance of the neural networks was fairly unstable and produced accuracy levels varying between 30% (no convergence) and 95%. The most common range of accuracy on average was between 76% and 84%. Note that we had to perform multiple trials, each time varying:

1. the number of internal nodes in the hidden layer,
2. the initial network weight settings, and
3. the learning rate constant for backpropagation.

Upon examining the results of the empirical evaluation, we concluded that the neural net approach did not offer any clear advantages over the decision tree based learning algorithm. Although neural networks, with extensive training and several training restarts with different initial weights to avoid local minima, could match the performance of the decision tree classifier, the decision tree approach still holds several major advantages. The most important is that the tree is easy for domain experts to understand. In addition, unlike neural network learning algorithms, the decision tree learning algorithms GID3* and O-BTree do not require the specification of parameters such as the size of the neural net, the number of hidden layers, and random trials with different initial weight settings. Also, the required training time is orders of magnitude faster than the training time required for a neural network approach.

The stability of the performance of the decision tree algorithms, and the fact that a decision tree (or classification rule) is a lot easier to interpret and understand than a neural network, provided strong reasons for favouring the decision tree approach. We, therefore, decided to adopt the decision tree approach for our problem.

3.5. Verification & Reliability Estimates

As mentioned earlier, in addition to using the CCD frames to derive training data for the machine learning algorithms, we also use them to verify and estimate the performance of our classification technique. This is done by testing on data sets that are drawn independently from the training data. An additional source of internal consistency checks comes from the fact that the plates, and the frames within each plate are partially overlapping. Hence, objects inside the overlapping regions will be classified in more than one context. By measuring the rate of conflicting classifications, we can obtain further estimates of the statistical confidence in the accuracy of our classifier. For the purposes of the final catalog production, a method is being designed for resolving conflicts on objects within regions of overlap. We have not yet collected reportable results on this aspect of the problem.

In order to demonstrate the difficulty and significance of the classification results presented so far, consider the example shown in Figure 3. This figure shows four image patches each centered about a faint sky object that was classified by SKICAT. These images were obtained from a plate that was not provided to SKICAT in the training cycle and the objects are part of a region in the sky containing the Abell 1551 cluster of galaxies near the North Galactic Pole. SKICAT classified the top two objects as stars and the bottom two as galaxies. According to astronomers (at least two of the authors), the objects shown in Figure 3 are too faint for reliable classification. As a matter of fact, an astronomer visually inspecting these images would be hard pressed to decide whether the object in the lower right hand corner is a star or galaxy. The object in the upper right hand corner appears as a galaxy based on visual inspection. Upon retrieving the corresponding higher resolution CCD images of these objects, it was clear that the SKICAT classification was indeed correct. Note that SKICAT produced the prediction based on the lower resolution survey images (shown in the figure). This example illustrates how the SKICAT classifier can correctly classify the majority of faint objects which even the astronomers cannot classify. Indeed, the results

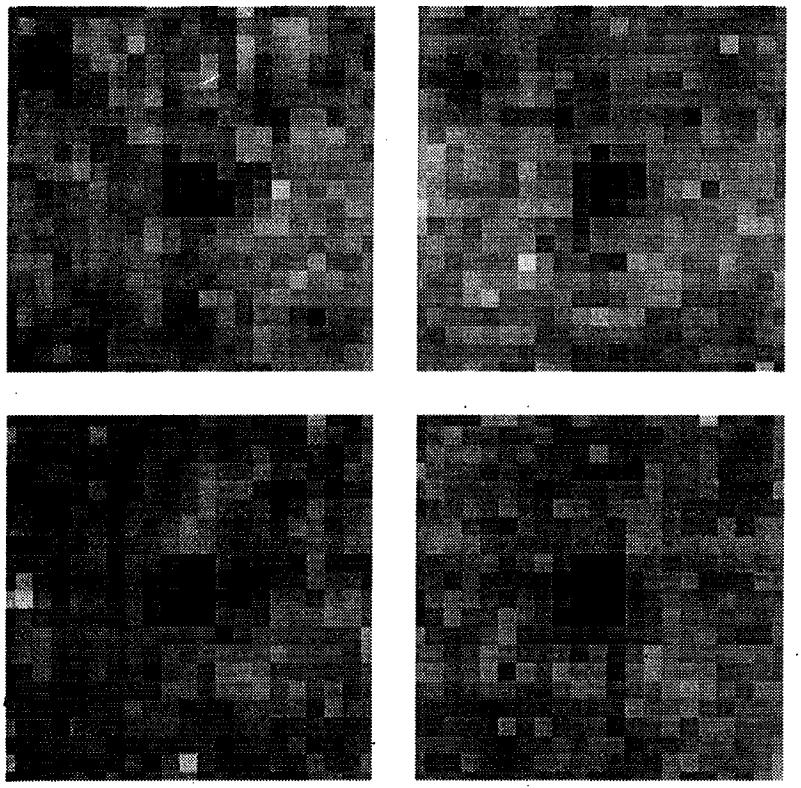

Figure 3. An illustrative example: four faint sky objects.

indicate that SKICAT has a better than 90% success rate a full magnitude below the comparable limit in previous automated Schmidt plate surveys.

4. CATALOG MANAGEMENT

The current version of SKICAT uses the Sybase commercial database package for catalog, storage, modification, and management. Each of the plate and CCD catalogs, produced as described in Section 3, must be registered in the SKICAT system tables, where a complete description and history of every catalog loaded to date is maintained. Catalog revisions, e.g., from deriving new and improved plate astrometric coordinates, photometric corrections, or even improved classifications, are also logged. The system is designed to manage a database of image catalogs constantly growing *and* improving with time.

One of the most difficult, yet critical, aspects of the data management process is the matching of identical sky objects detected in multiple, independent images. The most important science to be derived from the POSS-II depends upon uniformly integrating object measurements from a large number of overlapping plates. The advantages are two-fold: (1) permitting the objective analysis of a much larger portion of the sky than covered by a single 6.5^{o} x 6.5^{o} photographic plate, and (2), providing cross-spectral information through matching catalogs of the same sky field in different colors. It is the large solid angular coverage of all-sky surveys which set them most apart in observational phase space; this property facilitates certain types of science which are possible with no other form of data. In addition, cross-correlation of sources detected at different wavelengths is particularly fruitful in maximizing the scientific return from virtually any type of astronomical observation, especially from major surveys. A direct comparison of emission from astronomical objects in different parts of the electromagnetic spectrum can lead to astrophysical insights and better understanding of their nature. Matching and cross-identifications of large numbers of sources, in an objective and uniform way, is thus an increasingly more important data processing challenge.

We have implemented a SKICAT utility for matching any number of catalogs, object by object, in a consistent, though user-definable, fashion. With a modest amount of programming effort, the system can even be made to accommodate astronomical catalogs from sources other than plate scans or CCDs, i.e., from vastly different spectral regimes. The resulting matched catalog contains independent entries for every measurement of every object present in the constituent catalogs. The matched catalog may be queried using a sophisticated filtering and output mechanism to generate a so-called object catalog, containing just a single entry per matched object. Such queries may generate either additional Sybase objects tables or ASCII files, suitable for input to any number of plotting and analysis packages[Weir92].

A particularly promising aspect of SKICAT is the facility, as new data are added in, to query plate and CCD overlap regions in the matched catalog and dynamically update the constituent catalogs (their photometry, astrometry, classifications, etc.) in light of these results. While this process is accomplished by applying the appropriate SKICAT tools manually at this point, it is clear that an automated approach to maintaining and improving database uniformity in this way would be rewarding. This goal, of creating a "living," growing data set, instead of a data archive fixed for all time, has been an overriding one from the very start of the development of SKICAT.

5. Conclusions and Future Work

In this paper, we gave a brief overview of the machine learning techniques we used for automating the sky object classification problem. SKICAT classifies objects that are at least one magnitude fainter than objects cataloged in previous surveys. This project represents a step towards the development of an objective, reliable automated sky object classification method.

The initial results of our effort to automate sky object classification in order to automatically reduce the images produced by POSS-II to sky catalogs are indeed very encouraging. We have exceeded our initial accuracy target of 90%. 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 addition to using machine learning techniques to automate classification, we used them to aid in the attribute measurement process. Since measurement of the resolution attributes requires interaction with the user in selecting sure-thing stars for template fitting, we used the same machine learning approach to automate the star selection process. By defining additional "normalized" image-independent attributes, we were able to obtain high accuracy classifiers for star selection within and across photographic plates. This in turn allows us to automate the computation of the powerful resolution attributes for each object in an image.

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 a 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 4 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.

As part of our plans for the future we plan to begin investigation of the applicability of unsupervised learning (clustering) techniques such as AUTOCLASS [Chee88] to the problem of discovering clusters or groupings of interesting objects. The initial goals will be to answer the following two questions:

1. Are the classes of sky objects used currently by astronomers justified by the data: do they naturally arise in the data?
2. Are there other classes of objects that astronomers were not aware of because of the difficulty of dealing with high dimensional spaces defined by the various attributes?

The longer term goal is to evaluate the utility of unsupervised learning techniques as an aid for the types of analyses astronomers conduct after objects have been classified into known classes. Typically, astronomers examine the various distributions of different types of objects to test existing astrophysical models. Armed with prior knowledge about properties of interesting clusters of sky objects, a clustering system can search through catalog entries and point out potentially interesting object clusters to astronomers. This will help astronomers catch important patterns in the data that may otherwise go unnoticed due to the sheer size of the data volumes.

ACKNOWLEDGMENTS

We thank the Sky Survey team for their expertise and effort in acquiring the plate material. The POSS-II is funded by grants from Eastman Kodak Co., The National Geographic Society, The Samuel Oschin Foundation, NSF Grants AST 84-08225 and AST 87-19465, and NASA Grants NGL 05002140 and NAGW 1710. We thank Joe Roden of JPL for help on evaluating the performance of the learning algorithms. This work was supported by a NSF graduate fellowship (NW), Caltech President's Fund, NASA contract NAS5-31348 (SD& NW), and the NSF PYI Award AST-9157412 (SD).
The work described in this paper was carried out in part by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

REFERENCES

[Brei84] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). Classification and Regression Trees. Monterey, CA: Wadsworth & Brooks.

[Chee88] Cheeseman, P. et al (1988) "Bayesian Classification." *Proc. of the 7th Nat. Conf. on Artificial Intelligence AAAI-88*, pp. 607-6611, Saint Paul, MN.

[Chen88] Cheng, J., Fayyad, U.M., Irani, K.B. and Qian, Z. (1999) "Improved Decision Trees: A gegneralized version of ID3" *Proc. of the 5th Int. Conf. on Machine Learning.* pp. 100-107. Morgan Kauffman.

[Chen90] Cheng, J., Fayyad, U.M., Irani, K.B. and Qian, Z. (1990) "Applications of machine learning techniques in semiconductor manufacturing." *Proceedings of the SPIE Conference on Applications of Artificial Intelligence VIII.* pp. 956-965, Orlando, Fl.

[Feig81] Feigenbaum, E.A. (1981) "Expert systems in the 1980s." In Bond, A. (Ed.) *State of The Art Report on Machine Intelligence.* Maidenhead: Pergamon-Infotech.

[Fayy90] Fayyad, U.M. and Irani, K.B. (1990). "What should be minimized in a decision tree?" *Proceedings of Eighth National Conference on Artificial Intelligence AAAI-90*, Boston, MA.

[Fayy91] Fayyad, U.M. (1991). *On the Induction of Decision Trees for Multiple Concpt Learning*. PhD Dissertation, EECS Dept. The University of Michigan.

[Fayy92a] Fayyad, U.M. and Irani, K.B. (1992) "On the handling of continuous-valued attributes in decision tree generation." *Machine Learning,* vol.8, no.2.

[Fayy92b] Fayyad, U.M. and Irani, K.B. (1992) "The attribute selection problem in decision tree generation." Proceeedings of the Tenth National Conference on Artificial Intelligence, AAAI-92. San Jose, CA.

[Fayy93] Fayyad, U.M. and Irani, K.B. "Multi-interval discretization of continuous-valued attributes for classification learning." *Proc. of the 13h International Joint Conference on Arificial Intelligence IJCAI-93*. Chambery, France: Morgan Kauffman.

[Finn63] Finney, D.J. , Latscha, R., Bennett, B.M., and Hsu, P. (1963). *Tables for Testing Significance in a 2x2 Contingency Table*. Cambridge: Cambridge University Press.

[Hert91] J. Hertz, A. Krogh, and R.G. Palmer (1991) *Introduction to the Theory of Neural Computation.* Reading, MA: Addison-Wesley.

[Jarv81] Jarvis, J. and Tyson, A. (1981) *Astronomical Journal* 86:41.

[Quin86] Quinlan, J.R. (1986) "The induction of decision trees." *Machine Learning* vol. 1, no. 1.

[Quin90] Quinlan, J.R. (1990) "Probabilistic decision trees." *Machine Learning: An Artificial Intelligence Approach vol. III.* Y. Kodratoff & R. Michalski (eds.) San Mateo, CA: Morgan Kaufmann.

[Reid91] Reid, I.N. et al (1991) "The scond Palomar sky survey" *Publications of the Astronomical Society of the Pacific*, vol. 103, no.665.

[Vald82] Valdes (1982) *Instrumentation in Astronomy IV*, SPIE vol. 331, no. 465.

[Weir92] Weir, N. Djorgovski, S.G., Fayyad, U. et al (1992) "SKICAT: A system for the scientific analysis of the Palomar-STScI Digital Sky Survey." *Proc. of Astronomy from Large databases II*, p. 509, Munich, Germany: European Southern Observatory.