Astronomical Data Analysis Software and Systems XII
ASP Conference Series, Vol. 295, 2003
H. E. Payne, R. I. Jedrzejewski, and R. N. Hook, eds.

Automated Object Classification with ClassX

A. A. Suchkov, R. J. Hanisch, R. L. White, M. Postman, & M. E. Donahue

Space Telescope Science Institute

T. A. McGlynn, L. Angelini, M.F. Corcoran, S.A. Drake, W.D. Pence, N. White, & E.L. Winter

Goddard Space Flight Center

F. Genova, F. Ochsenbein, P. Fernique, & S. Derriere

Centre de Données astronomiques de Strasbourg

Abstract. ClassX is a project aimed at creating an automated system to classify X-ray sources and is envisaged as a prototype of the Virtual Observatory. As a system, ClassX creates a pipeline by integrating a network of classifiers with an engine that searches and retrieves multi-wavelength counterparts for a given target from the worldwide data storage media. At the start of the project we identified a number of issues that needed to be addressed to make the implementation of such a system possible. The most fundamental are: (a) classification methods and algorithms, (b) selection and definition of classes (object types), and (c) identification of source counterparts across multi-wavelength data. Their relevance to the project objectives will be seen in the results below as we discuss ClassX classifiers.

1. Classifiers

We apply machine learning methods to generate classifiers from 'training' data sets, each set being a particular sample of objects with pre-assigned class names that have measured X-ray fluxes and, wherever possible, data from other wavelength bands. In this paper, a classifier is represented by a set of oblique decision trees (DT) induced by a DT generation system OC1. An X-ray source is input into a classifier as a set of X-ray fluxes and possibly data from the optical, infrared, radio, etc. The discussion below includes some results obtained with classifiers trained on the data from the ROSAT WGA, GSC2, and 2MASS catalogs.

1.1. Classifier Metrics

In order to quantify the quality and efficiency of classifiers, we have introduced a variety of metrics. They include the classifier's preference, P_{ij}, which is the probability that a class i object will be classified as class j (Figure 1); its affinity,

© Copyright 2003 Astronomical Society of the Pacific. All rights reserved.

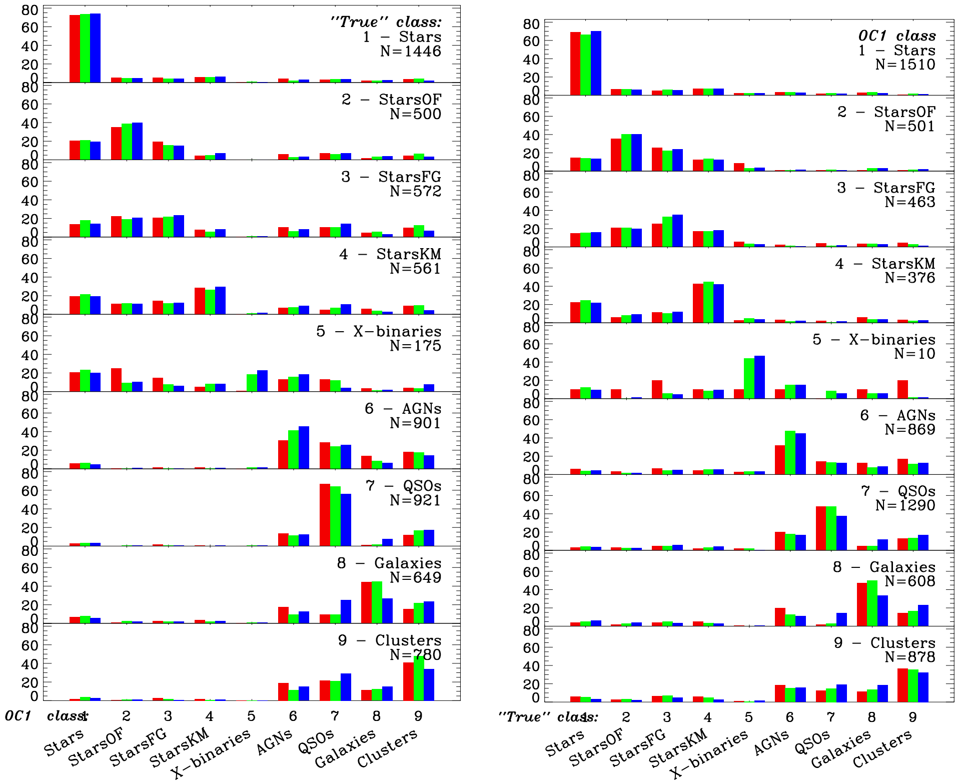

Figure 1. Preference, P_{ij}, (left) and affinity, A_{ij}, (right) for three classifiers trained using (a) X-ray magnitudes, (b) GSC2 magnitudes, and (c) both GSC2 and X-ray magnitudes along with coordinates and GSC2 "extended" vs. "point" source parameter. Notice that the OC1 classifiers separate stellar objects from non-stellar ones quite reliably. At the same time, a confusion between different types of stars or, say, QSO and AGN should be expected because of original misclassification and significant overlap of the respective object types in the parameter space.

A_{ij}, which is the probability that an object classified as class i does in fact belong to class j (Figure 1), and the power, S_i, which is the ratio of the probability that an object classified as class i is indeed class i to the probability that a randomly selected object belongs to class i. Additional useful characteristics are completeness, $C_i = P_{ii}$, and reliability, $R_i = A_{ii}$.

1.2. Classifier Networks

Using different sets of training parameters (attributes), we get different classifiers for the same list of class names (e.g., Figure 1). We integrate them into a network, in which each classifier makes its own class assignment and is optimized for handling different tasks or different object types. We envision that, having a set of X-ray sources, a user would generally select a certain classifier to make, for instance, the most complete list of candidate QSOs, but a different classifier would be used to make a most reliable list of such candidates. Additional classifiers would be selected to make similar lists for other object types. Figure 1 suggests that one would prefer the xray_gsc2 classifier to pick up cluster candidates, while AGNs call for the xray_only classifier.

2. Training Set Deficiencies

A classifier is adversely impacted by source misclassification, counterpart misidentification, data bias, etc. As the training data improve, so do the classifiers. In Figure 2, about 50% of class “Stars” sources (stars without spectral classification) come from the LMC/SMC region. This introduces a coordinate bias that affects classifiers generated from those data. Certain metrics of a classifier can be improved if stars from the LMC/SMC region are dropped from the respective training sets.

3. Validating Pre-assigned Classes with ClassX

An X-ray source in a training set may have an inappropriate class name or incorrect optical or other counterpart. Candidates with these deficiencies can be identified as a classifier is applied to the training data. In Figure 2, an OC1 classifier is seen to noticeably enhance the contrast between “extended” and “point” sources for StarsKM, QSOs, and Clusters, suggesting that the sources contributing to that enhancement were probably misclassified in the training set. They can further be examined and then reclassified if warranted, which would improve the training set itself.

4. Counterpart Search Strategies with ClassX

Classifiers trained using optical counterparts proved to be much better if a counterpart is selected as a brightest objects within 30 arcsec as opposed, for instance, to a nearest object within 30 arcsec or a brightest object within 60 arcsec. Thus, classifier validation in ClassX offers a way to find the best strategies to search for multi-wavelength counterparts.

5. Class Ambiguity in ClassX

A class is rarely a clear-cut notion. One person’s QSO is another’s AGN or galaxy. The overlap of object properties that often results in confusion in the object type assignement is an essential issue for any classification system. With ClassX, one can isolate sources with a greater degree of class name ambiguity and look into why their classification in the training set differs from the OC1 classification (see Figure 2).

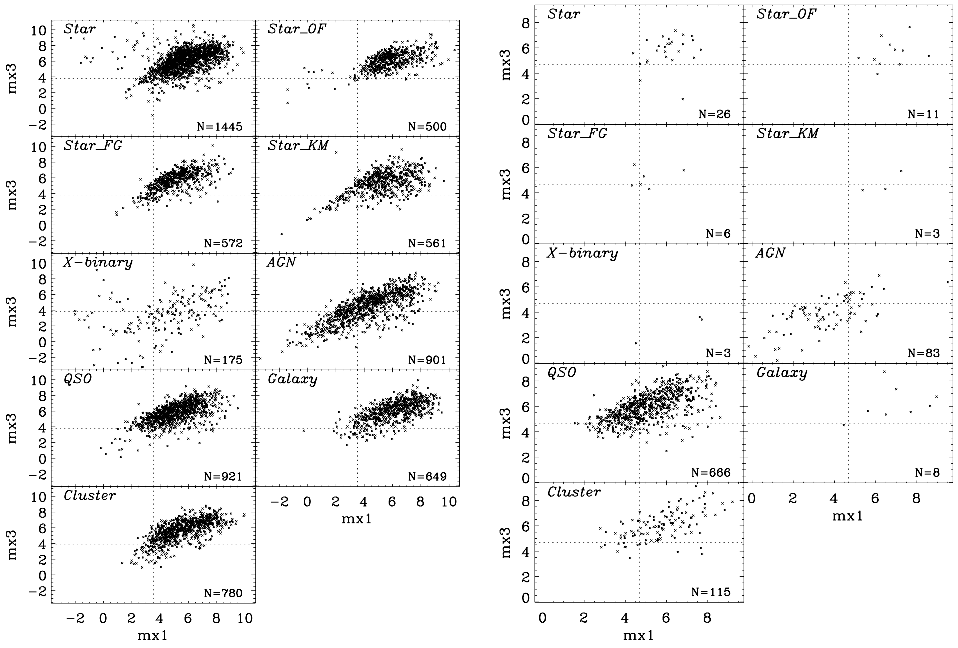

Figure 2. X-ray soft versus hard magnitudes for classes from the WGA catalog (left) and the OC1 classifier (right). On the right, only the WGA class QSO is shown. Most of the brightest and the faintest WGA QSOs have been classified by the OC1 classifier as AGN and Cluster, respectively, which partly reflects the class name ambiguity for these sources.

6. ClassX Outputs

A network classifier outputs the class name and the probability that the source belongs to the assigned class. It also outputs the probabilities that the source belongs, in fact, to other classes in the class name list. This is useful, for instance, for assessing how close the source association is with various classes in parameter space.