The Astrophysical Journal, 616:1284–1300, 2004 December 1

AUTOMATED CLASSIFICATION OF *ROSAT* SOURCES USING HETEROGENEOUS MULTIWAVELENGTH SOURCE CATALOGS

T. A. McGlynn,[1] A. A. Suchkov,[2] E. L. Winter,[1,3] R. J. Hanisch,[2] R. L. White,[2] F. Ochsenbein,[4] S. Derriere,[4] W. Voges,[5] M. F. Corcoran,[1,6] S. A. Drake,[1,6] and M. Donahue[7]

Received 2004 February 19; accepted 2004 August 5

ABSTRACT

We describe an online system for automated classification of X-ray sources, ClassX, and we present preliminary results of classification of the three major catalogs of *ROSAT* sources, *ROSAT* All-Sky Survey (RASS) Bright Source Catalog, RASS Faint Source Catalog, and WGACAT, into six class categories: stars, white dwarfs, X-ray binaries, galaxies, active galactic nuclei, and clusters of galaxies. ClassX is based on a machine-learning technology. It represents a system of classifiers, each classifier consisting of a considerable number of oblique decision trees. These trees are built as the classifier is "trained" to recognize various classes of objects using a training sample of sources of known object types. Each source is characterized by a preselected set of parameters, or attributes; the same set is then used as the classifier conducts classification of sources of unknown identity. The ClassX pipeline features an automatic search for X-ray source counterparts among heterogeneous data sets in online data archives using Virtual Observatory protocols; it retrieves from those archives all the attributes required by the selected classifier and inputs them to the classifier. The user input to ClassX is typically a file with target coordinates, optionally complemented with target IDs. The output contains the class name, attributes, and class probabilities for all classified targets. We discuss ways to characterize and assess the classifier quality and performance, and we present the respective validation procedures. On the basis of both internal validation and external verification, we conclude that the ClassX classifiers yield reasonable and reliable classifications for *ROSAT* sources and have the potential to broaden class representation significantly for rare object types.

Subject headings: methods: statistical — surveys — X-rays: binaries — X-rays: general — X-rays: stars

1. INTRODUCTION

The classification of cosmic sources into physically distinct classes is a key element of research in all domains of astrophysics. Traditionally, this has involved painstaking manual analysis of detailed, homogeneous sets of observations. More recently, automated classifier tools have been used to help in the classification of objects from huge but still largely homogeneous surveys. Examples include analysis of the First (Odewahn 1995) and Second (Weir et al. 1995) Digital Sky Surveys and the Sloan Digital Sky Survey (SDSS; Stoughton et al. 2002). In this paper, we discuss how we can go beyond using single large surveys and combine information from multiple heterogeneous databases to classify astronomical sources. Using dynamic cross-correlations of electronically available data sets, the ClassX team has developed a series of classifiers that rapidly sort X-ray sources into classes. These facilities are now available to the community at the ClassX Web site.[8] Classification is distinct from correlation and identification with objects at other wavelengths. Our classification tools can use the nonexistence of counterparts at other wavelengths or use ensembles of potential counterparts to establish limits to parameters.

Our initial work has concentrated on the more than 100,000 unclassified sources detected by the *ROSAT* observatory[9] from 1990 to 1999. These high-energy sources are particularly rich in interesting objects: QSOs and other active galactic nuclei (AGNs), clusters of galaxies, young stars, and multiple systems containing white dwarf (WD), neutron star, or black hole companions. The *ROSAT* samples have been used in prior investigations (e.g., Rutledge et al. 2000; Zhang & Zhao 2003), but still only about 10% of the sources observed by *ROSAT* have a reliable classification. In most cases this identification rests on cross-correlation between the *ROSAT* object and tables of classified sources. In some cases detailed follow-up observations have been performed on a source-by-source basis. This is extraordinarily expensive in both telescope time and the time of astronomers analyzing these data. Direct comparison of *ROSAT* sources with massive optical catalogs (e.g., Rutledge et al. 2000; or the similar efforts for *XMM-Newton* data; cf. Watson et al. 2003, Yuan et al. 2003, and Lamar et al. 2003) enables the cross-identification of *ROSAT* sources, but unless the class of the counterpart is known, this does not determine the type of the source. However, flux information from multiple catalogs allows us to try to classify sources with more information than is available from the X-ray observations alone.

Our approach differs from most previous efforts at multispectral classification in several basic ways. First it does not specifically constrain the information that is used to distinguish our output categories. Other authors have looked at the X-ray to optical ratios (Maccacaro et al. 1988) or X-ray/optical/radio

[1] NASA Goddard Space Flight Center, Greenbelt, MD 20771.
[2] Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218.
[3] Science Systems and Applications, Inc., Lanham, MD 20706.
[4] Centre de Données Astronomiques de Strasbourg, Observatoire Astronomique, UMR 7550, 11 rue de l'Université, F-67000 Strasbourg, France.
[5] Max-Planck-Institut für extraterrestrische Physik, Giessenbachstrasse, D-85740 Garching, Germany.
[6] Universities Space Research Association, 7501 Forbes Boulevard, Seabrook, MD 20706.
[7] Michigan State University, East Lansing, MI 48834.
[8] See http://heasarc.gsfc.nasa.gov/classx.
[9] See http://wave.xray.mpe.mpg.de/ROSAT.

correlations (Giommi et al. 1999). Here the authors manually define the regions in phase space that are to be assigned to specific classes. The algorithms below provide an automated method for *defining* a transformation between observational parameters and class rather than *imposing* one.

Second, the number of parameters that can be accommodated by the algorithms we use here is substantially greater than that in prior work. Rather than looking at ratios of two or three elements, even the simplest of our classifiers uses six independent quantities, and we have explored classifiers with a dozen or more independent features. Given the difficulties in simply visualizing these high-dimensionality phase spaces, extensions of earlier manual approaches are not feasible. This has important ramifications regarding the flexibility of the classifiers. For example, consider classification of Galactic versus extragalactic objects: Since Galactic absorption hardens the X-ray spectra and dims the optical brightness of extragalactic sources, a given hardness ratio or L_X/L_{opt} can imply different classification at different Galactic latitudes. Thus, while Maccacaro et al. (1988) can accommodate the effects of Galactic absorption on the hardness ratio, assuming a source to be extragalactic, our approach can, in principle, accommodate the effects of absorption automatically.

Unlike many of the previous works (Maccacaro et al 1988; Perlman et al. 1998; Giommi et al. 1999), our classifiers attempt to distinguish both Galactic and extragalactic classes of objects rather than focusing on selecting out one or two particular classes. Given this ambitious goal, these first results focus on fairly broad classes that contain the great majority of detected high-energy sources. Our experience suggests that to simultaneously provide fine-grained and wide-reaching classification will likely require a network of classifiers, starting with "broad band" classifiers of the type described below, that feed more specialized classifiers optimized for the various types. We defer further exploration of this topic to a later paper.

Finally, our approach lends itself to extending the sets of parameters that are used for classification. Once we have identified a training set of classified objects, we can use these objects to train classifiers using any interesting subset of observable parameters. For example, while a training set of AGNs may have been identified using correlations of X-ray, optical, and radio information, we can use this set to build classifiers based only on X-ray information. Thus, we can attempt to classify objects even if correlative optical or radio data are not available.

This paper concentrates on discussing the technique and evaluates its effectiveness in the classification of large samples in broad categories. Further work (e.g., Suchkov & Hanisch 2004b) will specialize this approach to analysis of specific classes.

With the recent and pending publication of several very large data sets covering much of the sky to considerable depth, we have begun to explore how well objects can be classified using data from these new large surveys. Thousands of previously classified sources are used to train classifiers, and these trained classifiers are then used to classify the 10^5 unclassified sources. In § 2 we discuss the sources of information we have used in our classifiers and how we dynamically extract information from the catalogs as needed, using capabilities that prototype generic Virtual Observatory[10] tools. Demonstrating the feasibility of this dynamic approach to extracting information was a major technical goal for this project.

Section 3 describes the actual classification tools and the training process we have used. We have used a supervised classification technique: oblique decision trees (Murthy et al. 1994). We discuss the reasons for this choice and the applicability of our approach to other supervised and unsupervised classification algorithms.

Section 4 discusses the ways in which we test our classifiers for accuracy. Internal validation looks at the performance of the classifier with respect to the sources we used to train it, and to the general characteristics of our newly classified sources. Can the classifier recover the classes of the data used to train it? External verification uses data sets independent of those used to train the classifier and compares how well the classifier predicts these results. Substantial numbers of our sources (several thousand) have been classified by other surveys, notably the SDSS. Comparing our results with these external data sets is a powerful test of our classifiers especially when the external data set is sufficiently deep.

Section 5 gives results for classification of the major *ROSAT* samples. We present the classification probabilities for each source in our original samples. Since we are classifying nearly 200,000 sources, only excerpts are included here, but the full tables are available for download from the ClassX Web site. Section 6 summarizes the status of the classifiers and describes our plans to extend our results to other X-ray data sets such as *XMM-Newton* and to integrate our classifiers in the growing Virtual Observatory.

2. DATA SOURCES AND DATA COLLECTION

2.1. *Data Sets*

2.1.1. *WGACAT*

The White-Giommi-Angelini Catalog (WGACAT)[11] was created by reprocessing the data from the pointed phase observations of the *ROSAT* PSPC. The result was a catalog of 88,579 sources with X-ray count rates in three energy bands and a variety of supporting data. About 20% of the sources in this sample have classifications derived from cross-correlations with other catalogs. The cross-correlation catalogs are described by White et al. (2000). The cross-correlations were performed from the less specific (i.e., giving only limited information about the type of the counterparts) to more specific catalogs, and the latter matches were used for the classification. The X-ray positions and fluxes from WGACAT were supplemented with the source extent information derived from the *ROSAT* PSPC (ROSPSPC) catalog.[12] The ROSPSPC includes only high-quality source detections in the standard image processing.

The pointed phase of *ROSAT* observations lasted nearly 8 years, and during that time the observations provided coverage of about 15% of the sky. Many regions were observed more than once, and objects in those regions may have multiple entries in WGACAT and the ROSPSPC catalog. When objects shared a common WGACAT ID, only a single value was included in our sample. The catalog contains a quality flag, and the data with higher quality were retained preferentially. In cases with equal quality flags, the entry nearest the center of the

[10] See http://www.ivoa.net and http://us-vo.org.

[11] See http://wgacat.gsfc.nasa.gov, maintained by N. E. White, P. Giommi, & L. Angelini.

[12] See http://heasarc.gsfc.nasa.gov/W3Browse/rosat/rospspc.html.

TABLE 1
Basic Classes and the Number of Class Objects in the WGACAT and RASS BSC Samples

Class	WGACAT All	WGACAT Unique[a]	RASS	Origin of X-Ray Emission
Star	6027	4678	4694	Corona or shocked stellar wind
WD	152	98	78	Hot atmosphere
XRB[b]	494	271	192	Accretion disk of a neutron star or black hole
AGN[c]	4589	3031	726	Central accretion disk, XRBs, galactic wind
Galaxy	1614	1305	1015	XRBs, hot corona, galactic wind
Cluster (of galaxies)	1717	1508	210	Hot intracluster gas
Unclassified	73986	65872	...	
Total	88579	76763	6915	

[a] In the case of multiple entries for a source, only the entry with the highest quality flag is used.
[b] Including cataclysmic variables.
[c] Including quasars, radio galaxies, and BL Lac galaxies.

field of view of the observation was retained. This resulted in a WGACAT sample of 76,763 sources, 18% of which had existing classifications.

X-ray source extent measurements were not included in the WGACAT. We obtained the required data by correlating the WGACAT sources against the ROSPSPC catalog, using a correlation radius of 30″, and selecting the closest candidate in the case of multiple matches. X-ray extent information was obtained only for sources within 19′ of the pointing axis or 34,633 (45%) of the 76,763 distinct WGACAT sources; for the sources without such information the extent parameter was set to 0. The ROSPSPC catalog and WGACAT were derived from the same set of observations (the pointed phase observations of the *ROSAT* PSPC instrument).

Setting the size to 0 where the extent is not known biases the classifier against classes in which sources have a real extent, notably clusters of galaxies. We have found that omitting extent information entirely makes it very difficult to distinguish such classes. This approach is the most effective way to use the information we have. The presence of a ROSPSPC catalog counterpart is noted in the tables in § 5.

The distribution of the previously classified sources in both the full WGACAT sample and our subset of it is shown in Table 1. While many objects had more specific classifications (e.g., specific spectral types for stars or Hubble types for galaxies), the classes chosen represent distinct physical origins for the X-ray emission. Understanding classification in these broad categories is a necessary prerequisite to attempting more detailed classifications. These classes in Table 1 represent categories in which there were sufficient entries to train the classifier. There were some categories—supernova remnants, nebulae, open star clusters—for which there were only a handful of classified sources. These were eliminated from our training set.

Figure 1 gives the photon count rate distribution for the WGACAT sample for the classified and unclassified sources. While brighter sources are more likely to be classified, there are many classified sources down to the faint end of the observed brightness distribution. The classified sources sample the entire flux space of the WGACAT. Figures 2 and 3 give the overall sky coverage of the WGACAT sources for both the entire sample and the classified sources. Although the WGACAT source distribution is highly nonuniform, the distributions of the (known) classified WGACAT objects are similar to those of the entire catalog. The nonuniform distribution reflects the concentration of *ROSAT*'s pointed observations on "interesting" targets and regions. About 15% of the sky was covered in the PSPC pointed-phase observations.

2.1.2. *ROSAT All-Sky Survey*

The *ROSAT* All-Sky Survey (RASS) catalogs (Voges et al. 1999, 2000) contain X-ray sources detected during the survey phase of the *ROSAT* mission with the PSPC instrument. The entire sky was surveyed with exposures highest toward the ecliptic poles. While the survey covers the entire sky, it is generally less deep than pointed observations in the areas of overlap. Overall, 124,735 objects were detected: 18,806 of these were published in the RASS Bright Source Catalog (BSC) and 105,924 in the Faint Source Catalog (FSC).

Figures 4 and 5 give the photon count rate distribution of the RASS classified and unclassified sources. Since the classified sources were restricted to the BSC, the sampling of faint objects is quite poor. The sky distribution of objects detected in the RASS is shown in Figures 6 and 7. There is a marked increase in density toward the ecliptic poles and there are a few regions not observed in the all-sky survey, but the sky coverage is, by design, much more uniform than during the pointed phase.

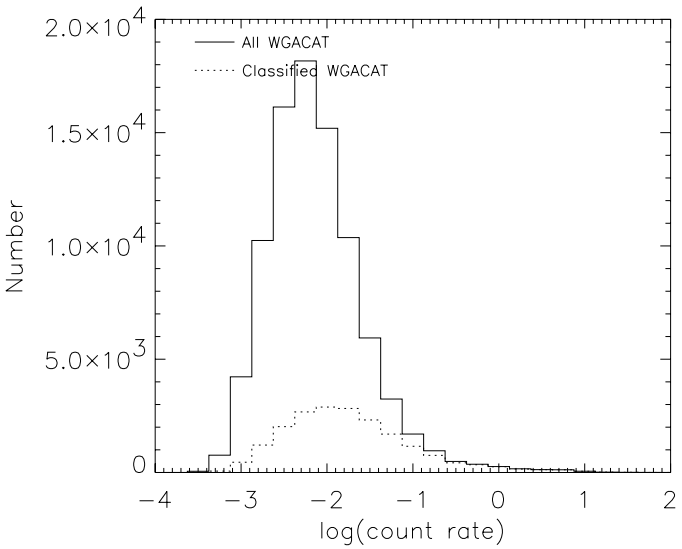

Fig. 1.—Photon count rate distribution for all WGACAT sources (*solid line*) and classified WGACAT sources (*dashed line*).

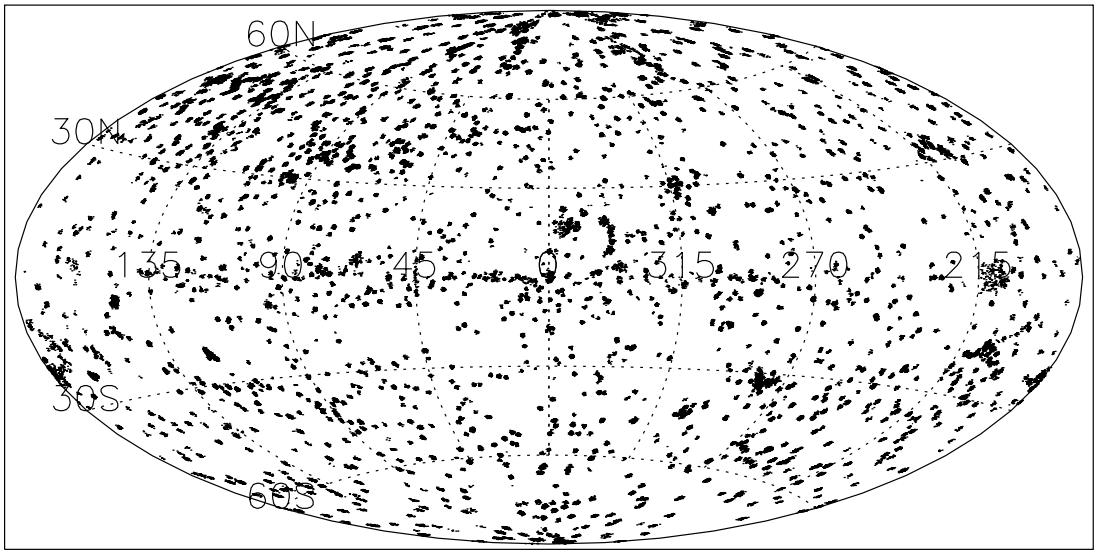

FIG. 2.—Galactic distribution of all WGACAT sources.

2.2. *The ClassX Pipeline*

The ClassX processing pipeline gathers the data used for classification. A generic pipeline that can gather data from many catalogs in many wavebands was constructed for ClassX, and we have looked at many different sources of information. However, in this paper only X-ray, optical, and radio data were used. The catalogs used and information extracted are shown in Table 2. The correlative data from each band are gathered separately, filtered, and then combined to form a single package of data for use by the classifier itself. The classifiers X and XOR are described further in § 3.5. The ClassX pipeline makes extensive use of the standard representation of tabular and catalog data developed in the Virtual Observatory initiative, VOTable (Ochsenbein et al. 2000a).[13]

[13] Also see http://www.ivoa.net/twiki/bin/view/IVOA/IvoaVOTable for updates to the VOTable standard.

Optical counterparts of X-ray sources are found using a search radius of 30″; this gives a reasonable completeness level while keeping the number of chance coincidences manageable. The correlations were done using the VizieR (Ochsenbein et al. 2000b) system.

If no counterpart was found, the object was dropped from consideration for use by classifiers needing information from that waveband. If a single counterpart was found, then the data from that counterpart were used. When multiple counterparts were found, a rule for resolving the ambiguity was needed. Both nearest and brightest counterparts were tried. Using the brightest counterpart was found to provide somewhat more accurate results and was used here; however, a function combining the two would likely be better still.

For radio data, only the existence or nonexistence of the radio counterpart was used in the classifier. The combination of the NRAO VLA Sky Survey (NVSS; Condon et al. 1998) and SUMSS catalogs gave us radio coverage over approximately 92% of the sky. Since the determination of the coverage

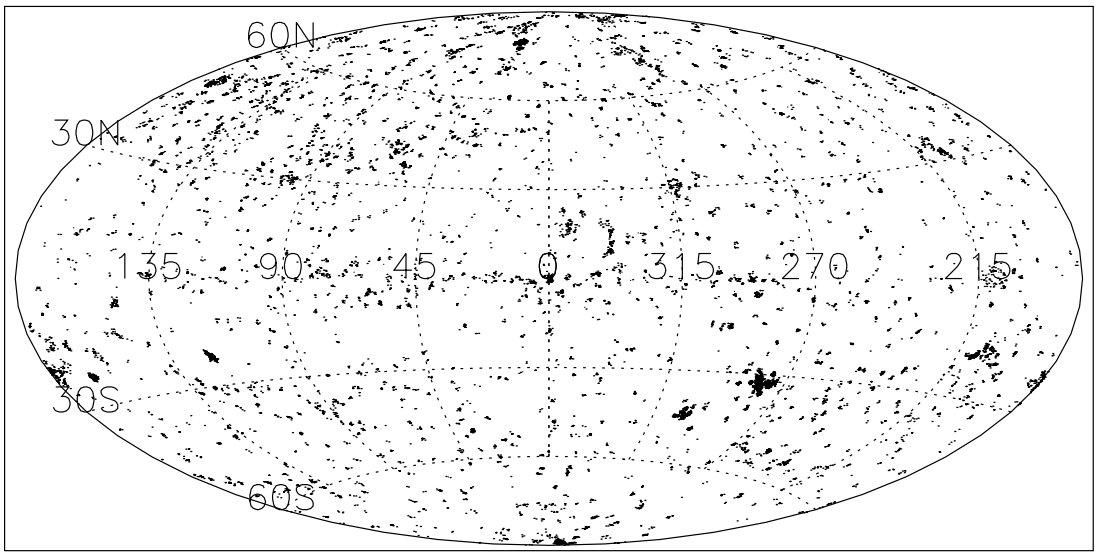

FIG. 3.—Galactic distribution of classified WGACAT sources.

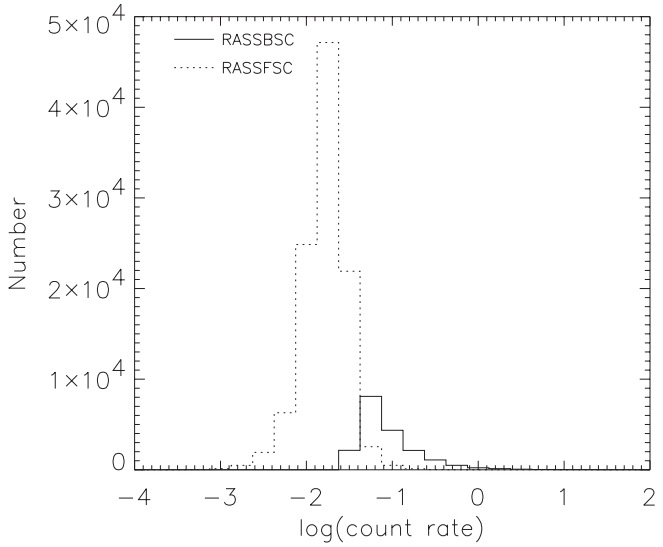

FIG. 4.—Photon count rate distribution for RASS BSC (*solid line*) and RASS FSC (*dashed line*) sources.

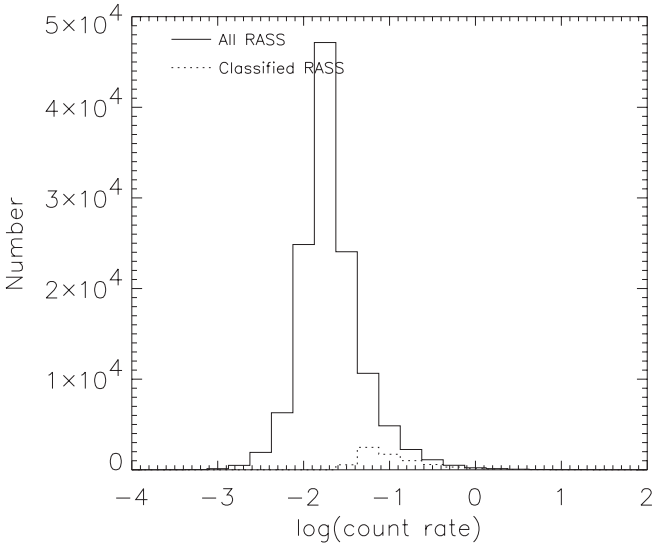

FIG. 5.—Photon count rate distribution for all RASS (*solid line*) and classified RASS (*dashed line*) sources.

boundaries for the SUMSS surveys is nontrivial, data in the 8% of the region not covered were treated as having no counterpart. Even in classes in which radio counterparts are most frequently found, most objects do not have a radio counterpart. So this treatment of 8% of the sky should not cause a significant bias.

Rather than use the radio flux density as a classification parameter, we have chosen to use instead a simple binary flag indicating whether the X-ray source has an associated radio object. There are two reasons why we adopted this approach. First, by using only a detection flag we are able to combine both radio detections and nondetections into a single larger training set, which improves the quality of the classification over what one would get from two separate training sets. The classifier algorithms we have used do not provide any natural way of including upper limits as classification features. In principle, it is possible to provide two separate parameters, a detection flag and the flux or upper limit, and to try to train the classifier to identify the implied connection between those two parameters. We have tried this approach for this and other classification problems, but we find that it often does not improve the classification accuracy because the additional parameter makes the optimization problem more difficult. It would be very useful to have generalized classifiers that did explicitly allow for parameters that have associated uncertainties and upper or lower limits, but such tools have not yet been developed.

The second reason for using the radio detection flag is that the simple knowledge of the presence of a radio source already contains most of the information from the radio band. This statement is empirically supported by projects such as the FIRST Bright Quasar Survey (White et al. 2000), which selects candidates based on radio-optical detections and finds that the radio flux density is not a strong discriminator between source types. For example, Figure 7*b* in White et al. shows that the fraction of radio-optical candidates found to be quasars changes by only a factor of 2 when the radio flux varies by 4 orders of magnitude, from 1 mJy to 10 Jy. This weak dependence on the radio brightness is due at least partly to the enormous distance in frequency between the radio and optical (or X-rays). The radio-optical spectral index α (where the flux as a function of frequency $F_\nu \propto \nu^\alpha$) changes by only 0.1 when the radio flux varies by a factor of 4. Two-thirds of NVSS catalog sources have fluxes within a factor of 4 of the catalog limit, and more than 90% are within a factor of 16 of that limit. That implies that for most sources the spectral index can be estimated within 0.1 without knowing the flux at all.

The other radio characteristic that makes a source detection flag useful is the slowly changing nature of the radio source populations with flux. Such a flag would be significantly less effective in the optical because the star-galaxy and star-quasar ratios change dramatically with optical magnitude. In the radio, the fractions of radio galaxies, AGNs, and star-forming galaxies varies relatively little over the flux range covered by the NVSS and SUMSS catalogs.

Classifiers generally work best when they are not presented with redundant information. If our classifiers were intended for classification of extragalactic sources, the hydrogen column density, $n_{\rm H}$, at the position of the X-ray source might be a useful discriminator, since absorption hardens the X-ray emission in the *ROSAT* band. One can imagine approaches where the observed flux is corrected to accommodate for absorption. However, our classifiers try to classify both Galactic and extragalactic sources so that such approaches are inappropriate. Rather the classifiers use the Galactic latitude (and to a lesser extent the longitude). The position and absorption are strongly coupled, so that our classifiers can distinguish objects appropriately in high and low column regions. The Galactic position also allows the classifier to accommodate the diminishing density of some Galactic sources (e.g., X-ray binaries) at high latitudes. We have tried classifiers that use both position and $n_{\rm H}$ as input parameters but have not seen any substantial differences in behavior compared with position alone. In the future we anticipate exploring classifiers that only use $n_{\rm H}$.

As a last step before the data are used by the classifier, information from all tables was combined. Only objects for which all parameters required by the classifier were available (either from the table or by use of a default value) were included in the final sample.

2.3. *Counterpart Validity*

The errors in the X-ray positions of the objects in the WGACAT and RASS samples are relatively large compared to

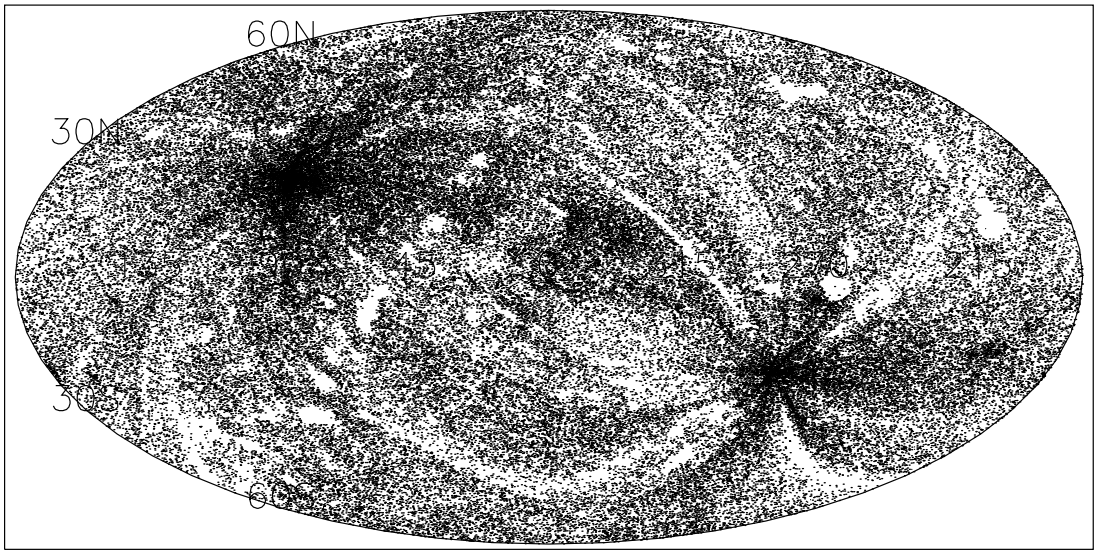

FIG. 6.—Galactic distribution of all RASS sources.

the typical separation of objects detected in the USNO-B survey (Monet et al. 2003). When we search for optical counterparts to the X-ray sources, we have used the brightest object within 30″ of the nominal X-ray position. Most objects have at least one candidate counterpart within 30″, and on average about five objects are seen within the limiting radius. How much confidence can we have in the validity of our cross-identification with optical and radio sources? Since we do not perform follow-up observations, this question can only be addressed statistically.

One powerful check on the validity of the identifications is to look for counterparts at positions near but slightly offset from the nominal positions. In addition to the actual correlation of each object in the WGACAT and RASS samples with the USNO-B survey, we correlated a point 1° away in Galactic longitude. If our cross-correlations were dominated by spurious cross-identifications—i.e., the optical counterparts had no relation to the X-ray sources—then we would expect the statistics of cross-match between the nominal target positions and the offset positions to be similar.

The selection of the offset angle is a compromise. A smaller offset angle ensures a closer matching of the selection of the data. However, if the offset angle is smaller than the radius of the *ROSAT* field of view, then the "background" for a serendipitous source may be offset to the center of the *ROSAT* observation. Since the *ROSAT* WGACAT observation centers are very atypical, with a population of very bright stars and galaxies, the smallest radius that avoided this was selected.

Comparing the counterparts selected near the X-ray positions with the offset sample in Table 3, we find that nominal counterparts are on average 1 mag brighter than the offset counterparts. Not unexpectedly, for the BSC subset of the RASS survey the difference is even greater, about 4 mag.

If the counterparts and offset objects both reflect a uniform distribution of objects in space, then we would expect the surface density of objects, $N(b)$, at given apparent brightness, b,

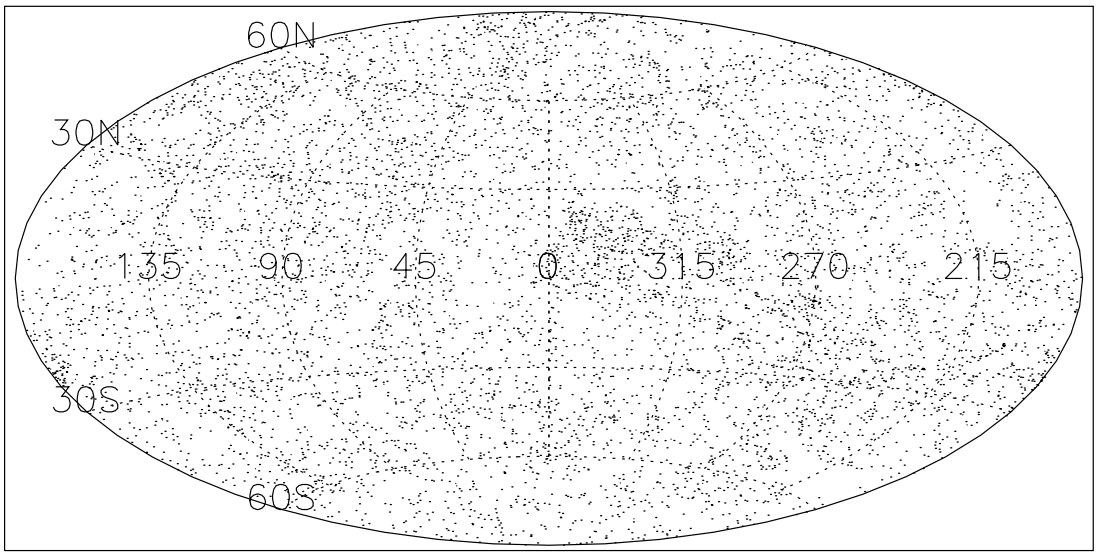

FIG. 7.—Galactic distribution of classified RASS sources.

TABLE 2
Class Attributes (Object Parameters) Used by the Basic ClassX Classifiers

Attribute Name	Attribute Source	Classifier[a]	
		XOR	X
Galactic longitude, l_{II}	Input	y	y
Galactic latitude, b_{II}	Input	y	y
X-ray brightness[b]	X-ray data	y	y
Hardness ratio 1, HR1[c]	X-ray data	y	y
Hardness ratio 2, HR2[c]	X-ray data	y	y
X-ray extent (source size)[d]	X-ray data	y	y
Blue magnitude, B[e]	Optical data	y	n
Red magnitude, R[e]	Optical data	y	n
Radio counterpart flag[f]	Radio data	y	n

[a] X for WGACAT-X and RASS-X classifiers, XOR for WGACAT-XOR and RASS-XOR classifiers. Parameter required by a classifier is indicated by "y," otherwise "n."
[b] Defined as -2.5 log (count rate).
[c] From RASS or computed from WGACAT.
[d] From RASS or ROSPSPC (for WGACAT) if available, else 0.
[e] From the USNO B1 catalog.
[f] If counterpart is found in NVSS or SUMSS, then 1, else 0.

to go as $N(b) \sim b^{-1.5}$. A 1 mag shift then corresponds to a factor of about 4 in the surface density of objects. Such a simple analysis would imply that about 80% of the counterparts are real.

This naive estimate ignores two important effects. First, sources are not necessarily distributed uniformly: Galactic objects will show a preference to the Galactic plane, and there may be effects from the finite size of the Galaxy. Also, the sources we detect are selected on the basis of the X-ray brightness. If a class of sources has relatively bright optical counterparts, then the cutoff in X-ray luminosity may correspond to a relatively high optical brightness. Any faint optical counterparts would have X-ray emission that is too low to be detected.

In addition to looking at the brightness of the selected counterparts, one can also look at the total number of candidate sources near the X-ray positions and compare them to the number of sources at the offset positions (see Table 3). X-ray sources are rare compared to optical sources. The average density of RASS sources is about 3 deg^{-2}. The WGACAT sources have an average density of about 12 deg^{-2} when we account for the coverage of the *ROSAT* pointed observations. If we look at a random 30″ radius circle in the sky, the expectation of finding an X-ray source is less than 0.002. Even where the density of X-ray sources is much higher than average, randomly chosen positions are unlikely to include an X-ray object.

So when we look for candidate counterparts in the offset positions, we expect on average to find one fewer optical candidate counterpart than we do when we look at the nominal X-ray position whenever there is a real X-ray counterpart in the USNO survey at near the nominal position. Table 3 shows that for the RASS FSC and WGACAT there is indeed about one extra candidate per object near the X-ray positions. For the BSC there are, on average, several excess candidates corresponding to the X-ray source. This likely results from observations of bright clusters of galaxies and X-ray objects in star clusters where there can be many optical counterparts associated with the X-ray source. This indicates that most X-ray sources have optical counterparts in the USNO survey. However, even if the real counterpart is among the candidate counterparts, it may not be the brightest candidate and thus would not be selected for our analysis.

To address these uncertainties we have undertaken a more rigorous analysis of the distribution of offset counterparts versus the counterparts at the nominal X-ray positions. Suppose we have some selection function, s, that can be calculated for any potential counterpart to a given X-ray source. The candidate counterpart with the maximum value of s will be chosen as the counterpart for use by the classifier. Each X-ray source is assumed to have a real counterpart with some value of s. There is a probability distribution for the selection function of $p(s)$ for real counterparts.

In addition to the X-ray counterparts, there is a population of background objects. For each background object we can also calculate the selection function s. In the absence of real counterparts we would measure a distribution of background counterparts with a probability of $b(s)$. At the sensitivities we are exploring, the density of X-ray sources is only a few per square degree. For a random element of sky our 30″ radius circle (about 2×10^{-4} deg^2) is very unlikely to have any X-ray sources. By looking for counterparts at the offset positions we can measure $b(s)$ directly.

For a given X-ray source, we find a counterpart with a selection function s, unless there happens to be a background candidate with a larger value of s. The observed distribution of the selection function will be

$$o(s) = [1 - B(s)]\, p(s) + [1 - P(s)]b(s),$$

where $P(s)$ and $B(s)$ are the cumulative probabilities that a counterpart or background source, respectively, has a selection function greater than s.

TABLE 3
X-Ray Sources and Optical Candidate Counterparts

Sample	Size	N_X[a]	N_{offset}[b]	$(N_X - N_{offset})$/Size)	$m_{b,X}$[c]	$m_{b,offset}$[d]	$m_{r,X}$[e]	$m_{r,offset}$[f]
BSC	18,806	156,505	95,811	3.2	13.84	18.23	12.71	17.05
FSC	105,924	741,603	587,215	1.5	16.79	18.12	15.63	16.92
WGA	88,579	598,932	517,786	0.92	17.08	18.34	15.90	17.12

Note.—Characteristics of optical candidate counterparts near the actual X-ray position and offset by 1°.
[a] Total number of candidate counterparts within 30″ of the nominal X-ray position.
[b] Total number of candidate counterparts within 30″ of the offset position.
[c] Average B magnitude of the selected candidate counterpart near the nominal X-ray position.
[d] Average B magnitude of the selected candidate counterpart near the offset position.
[e] Average R magnitude of the selected candidate counterpart near the nominal X-ray position.
[f] Average R magnitude of the selected candidate counterpart near the offset position.

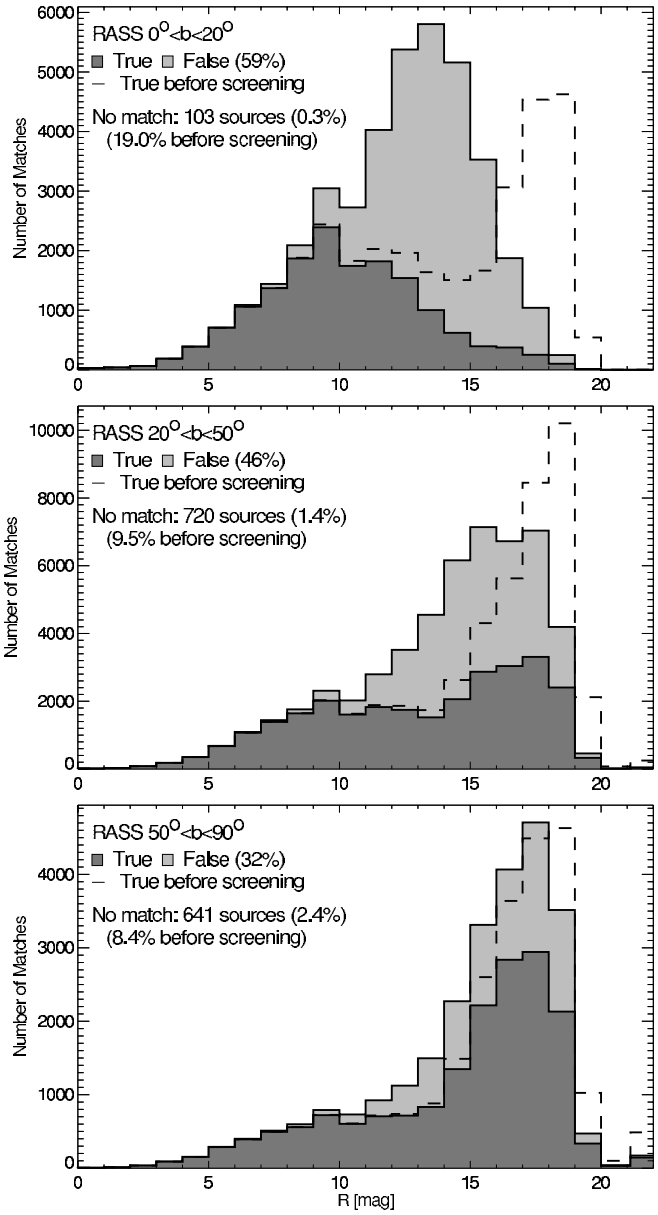

FIG. 8.—Magnitude distribution of real and background counterparts for the RASS sample. The dashed line gives the counterparts that would be detected if there were no background objects.

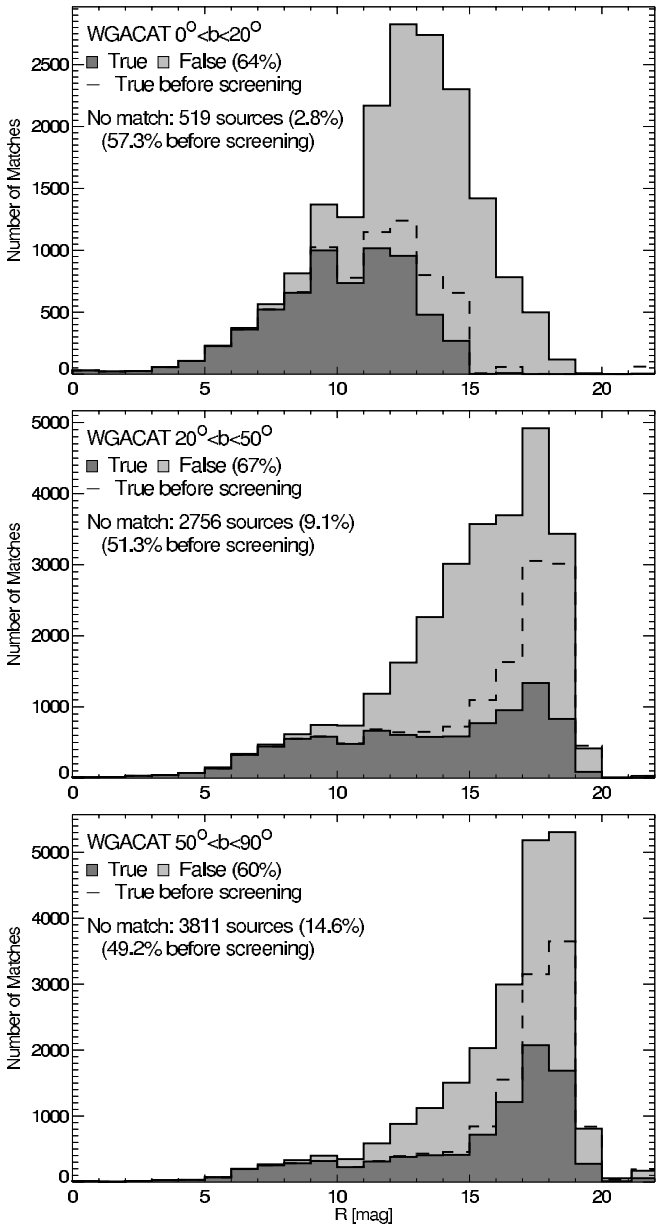

FIG. 9.—Magnitude distribution of real and background counterparts for the WGACAT sample. The dashed line gives the counterparts that would be detected if there were no background objects.

For our classifiers, s is the brightness of objects within $30''$ of the nominal position or 0 for objects beyond $30'$. Other formulations for finding the counterpart can use combinations of the brightness, positional uncertainty, and position offset from the X-ray position. So long as measurements of the selection function for counterparts and background objects do not interact, a comparable analysis can be done.

We have measured $b(s)$ and $o(s)$ and can solve for $p(s)$ by using a Lucy (1974) style deconvolution, which ensures the functions are positive definite. Figures 8 and 9 show the results of this analysis.

For each of several Galactic latitudes the distribution of counterpart luminosities is broken into real and background components. For low Galactic latitudes, where most counterparts are stars, there is a strong peak of real counterparts at bright magnitudes. Farther from the Galactic equator, where X-ray sources are more likely extragalactic sources, a lower brightness population of counterparts emerges. At fainter than 15–16 mag most counterparts are not correct. The true X-ray counterpart is masked by the brighter background objects. Since the RASS sources are brighter, the fraction of correct identifications is generally higher.

A dashed line is included in these figures to show what the number of counterparts would have been had there been no background objects. This corrects for the "screening" of the real counterparts by the background. The figures also indicate the number of objects for which no matching object was found in the USNO catalog, as well as the number of objects that we estimate would not have had a match if there had been background objects, i.e., for which there was no real counterpart in the USNO data. This includes at least three classes of objects: objects where the counterpart is too faint to be included in the USNO catalog, objects where the counterpart is outside the $30''$ radius limit, and cases in which the X-ray object itself is spurious.

For RASS data the total fraction of correctly identified sources is about 50%. However, when we correct for screening, we find that only about 10% of sources do not have counterparts within the USNO. Looking at the sharp cutoff in the dashed line in Figure 8 near the luminosity cutoff of the USNO catalog, it seems clear that many of these objects are simply too faint to have been included in the USNO catalog.

For WGACAT the number of correct counterparts is only about 35%. While there is also some screening of fainter sources, just under 50% of WGACAT objects appear to have counterparts in the USNO sample. Since the WGACAT X-ray sources are fainter than the RASS, we would anticipate that a larger fraction of the counterparts will lie below the magnitude limit of the USNO catalog. Since some sources in the WGACAT sample are far offset from the center where the position determination is less reliable, the WGACAT sample will also find a somewhat larger fraction of objects outside the 30″ limit. While we could increase the search radius at larger off-axis angles to try to get the "correct" counterpart for these sources, it is clear that screening due to the background is already a serious issue at 30″, and so we have used a constant search radius.

Overall, under half of the objects in the RASS and WGACAT samples are identified with the correct counterpart. At fainter magnitudes the optical magnitude works essentially as an upper limit to the actual magnitude of the counterpart. The classifier techniques described in the following sections adapt to the changing role of the counterpart magnitude automatically, so long as we can properly train the classifier.

3. CLASSIFICATION TECHNIQUES

3.1. *Introduction*

"Classification" is the process of mapping the observable characteristics of an object to a set of classes that typically represent different physical types; a "classifier" is the implementation of a classification algorithm to perform this mapping. We consider here methods for "supervised classification," meaning that a human expert has both determined into what classes an object may be categorized and also provided a set of sample objects with known classes. This set of known objects, called the training set, is used by the classification programs to learn how to classify objects. The process of creating such a classifier for a particular data set is usually called "training."

There are also "unsupervised classification" algorithms (e.g., clustering, mixture models) that attempt to both determine the types of objects and how to separate them directly from the parameter-space distribution of the unclassified sample. We have chosen to work primarily with supervised classification methods, however, since we understand much of the underlying physics for the electromagnetic emissions that are measured, and we can thus choose intelligently from among the many measured parameters to build the best training sets and select the best classes.

There are two steps to construct a supervised classifier. In the training phase, the training set is used to decide how the parameters ought to be weighted and combined in order to separate the various classes of objects. In the application phase, the weights determined in the training set are applied to a set of objects that do not have known classes in order to determine what their classes are likely to be.

If a problem has only a few important parameters, then classification is usually an easy problem. For example, with two parameters one can often simply make a scatter plot of the feature values and determine graphically how to divide the plane into homogeneous regions in which the objects are of the same classes. The classification problem becomes very difficult, however, when there are many parameters to consider. Not only is the resulting n-dimensional space difficult to visualize, but there are so many different combinations of parameters that techniques based on exhaustive searches of the parameter space become computationally infeasible. Practical methods for classification then involve a heuristic approach intended to find a good enough solution to the optimization problem.

3.2. *Oblique Decision Trees*

There are several "dimensions" that we can vary in building classifiers. The input observational characteristics and the output physical classes can be varied. We can use different sets of training information, and we can vary the basic algorithm for classification. In this paper we report on results using only a single classifier algorithm, the Oblique Classifier 1 (OC1) system of oblique decision trees (Murthy et al. 1994) for a fixed set of output classes. We have chosen the OC1 algorithm because it is freely available,[14] its accuracy is comparable to the best available algorithms, and it is sufficiently fast (in both training and application). An additional benefit is that the decision tree can be examined after it has been trained to determine the key criteria for classification; this is difficult with, for example, neural networks.

Conceptually, the oblique decision tree classifier is rather straightforward. It considers the n-space defined by the set of n-input observational characteristics, where each characteristic is treated as a continuous variable. A binary tree is constructed in which at each node a plane in the n-space (described by a linear combination of the parameters) divides the objects into two groups. The first node represents a plane that divides the space into two regions. Objects are sifted down the left or right branches of the tree depending on which side of the plane they fall. The next node represents another plane that further divides the two subspaces. Ultimately, one reaches a leaf node of the tree where all the objects in the region are assigned to a single class. Some parts of the parameter space may be well delineated by only a few planes, while other parts might require many planes in order to separate complex distributions.

Oblique decision trees are difficult to construct because there are many possible planes to consider at each tree node. OC1 includes a flexible and efficient algorithm for creating a decision tree given a training set. See the Murthy et al. (1994) paper for full details; we describe here some key features of the algorithm.

OC1 uses a "greedy" algorithm in the initial tree construction. It first attempts to find the plane in the n-space that most cleanly divides the training set sample into two samples having distinct sets of classes. Various impurity measures are available for determining the quality of a particular split. It then repeats the process recursively for the subspace on the two sides of the dividing plane. The algorithm continues until each remaining subregion is perfectly classified, with all included training set objects having the same class.

In most cases this initial tree divides the parameter space too finely. For example, some leaf nodes may contain only a single object, picked out by planes that separate it from a mass of nearby objects having different classes but with similar

[14] See http://www.tigr.org/~salzberg/announce-oc1.html.

parameters. The tree overfits the training set data, tracking details much more closely than is justified. To address this, OC1 prunes its decision tree. A fraction of the training set objects is reserved during the initial tree construction. This pruning sample is used to test the decision tree; decision nodes are eliminated if their removal does not reduce the classification accuracy for the pruning sample. The final tree does not classify the training set perfectly (some subregions contain multiple classes of objects), but it has higher overall accuracy than the original overfitted tree.

Oblique decision tree classifiers are not the only possible choices: other commonly used algorithms include neural networks, nearest-neighbor methods, and axis-parallel decision trees. See Salzberg et al. (1995) and White (1997, 2000) for discussions of some astronomical applications and more detailed comparisons of these algorithms.

3.3. *Voting Decision Trees and Classification Probabilities*

We have improved on the accuracy of the classification by using not just a single tree, but rather a group of 10 trees that vote (White et al. 2000). This multiple-tree approach has been shown to be effective at improving the accuracy of classifiers (Heath et al. 1996). OC1 uses a complex search algorithm that includes some randomization to avoid the classic problem of getting stuck in local minima in the many-dimensional search space. Thus, one can run OC1 many times using different seeds for the random number generator to produce many different trees.

Heath et al. (1996) used a simple majority voting scheme: classify the object with each tree and then count the number of votes for each class. We have improved on this by using a weighted voting scheme, where each tree splits its vote between classes depending on the populations of the classes from the training set at that leaf. (Recall that after pruning a leaf may contain objects of several different classes.) If an object winds up at a leaf node with N training set objects of which L_i are of class i ($i = 1. \ . \ .C$), the tree's fractional vote in favor of classification i is $(L_i + 1)/(N + C)$. (The particular form used for the ratio was derived from the binomial statistics at the leaf.) The votes from all 10 trees are averaged to produce a vector of probabilities that an object belongs to each of the possible classes in the training sample. We associate the largest element of this vector with the ''class'' of the source.

3.4. *The Output Classes*

There are many distinct classes of X-ray sources, and one of the goals of this research is to understand the level of detail to which we can successfully distinguish such sources with the information we have at hand. In practice, in this initial effort we have tended to be conservative, using only six basic classes (Table 1).

A problem that needs to be addressed in the classifier design is that the same astronomical object may legitimately belong to very different object types, especially as viewed from different wavelengths. While the X-ray properties of an X-ray binary are likely to be dominated by the accretion onto the compact companion, the optical appearance of the system may be that of, say, a normal B-star—and it may be categorized as such in some catalogs. Similarly, while the X-ray emission of a cluster of galaxies originates mostly in the intracluster gas, the cluster optical or infrared counterpart would typically be a cluster galaxy.

Such ambiguities can complicate all phases of the classification, including construction of training sets, the training process itself, and interpretation of the results. Depending on the use of the classifications, all classification errors are not equal. If a user is interested in distinguishing Galactic and extragalactic sources, then misclassifying an AGN as a galaxy is not as bad as misclassifying it as a star. Indeed, the distinction between a ''normal'' galaxy and an AGN is fairly arbitrary. The usage here is determined by the classification of objects in the training sets that have been used, rather than specific markers.

3.5. *ClassX Classifiers*

We introduce here four ''basic'' ClassX classifiers derived from the WGACAT and RASS BSC data (Table 2). They are used in the subsequent discussion to illustrate how the amount and the nature of the information fed into a classifier affects classification results. The RASS-X and WGACAT-X classifiers use *ROSAT* data only, including positional information. In addition, the RASS-XOR and WGACAT-XOR classifiers use optical data for the optical counterparts and a flag indicating whether the source has a radio counterpart in the NVSS (Condon et al. 1998) and SUMSS (Mauch et al. 2003) surveys; objects for which no optical counterpart could be found were not used in the training of these classifiers. The radio cross-correlation was done using a $30''$ radius. The radio catalogs have a surface density of about 50 sources per square degree, so that on average we would expect fortuitous counterparts for about 1% of our sources. In fact, about 17% of the RASS BSC sources and 7% of the RASS FSC and WGA sources have radio correlates. All four illustrative classifiers are trained to distinguish the same basic set of classes: stars, WDs, X-ray binaries (XRBs), AGNs, galaxies, and clusters of galaxies.

Many more classifiers are available at the ClassX Web site. These include classifiers using correlations with other optical and infrared catalogs, other input parameters, and different sets of output classes.

Unlike many classification methods that involve multiwavelength data and rely on total reliability of the multiwavelength counterparts, ClassX can be quite effective in situations in which counterpart reliability is low. ClassX classifiers can learn to efficiently use the counterpart information even if a counterpart is physically not the same object as the X-ray source. A simple illustration of this capability is as follows. Given the upper limit for stellar X-ray luminosity, *ROSAT* can detect stellar X-ray emission only from relatively close, hence visually bright stars, typically brighter than 12–13 mag. At the same time, QSOs are fainter than 13 mag. ClassX learns this distinction from the training sample. When applied to a source that has an optical counterpart fainter than 13 mag, the ClassX classifier then knows that such a source is unlikely to be a star. It would classify it in most cases as something else, for example, as a QSO, and only significant pressure from the rest of the available information can force it to change its decision in favor of a star. The important point here is that the optical counterpart does not have to be the physical QSO counterpart to the X-ray source; the optical information used by the classifier is that the source is *not* a star, and then the classifier uses the rest of the available information to decide which of the nonstellar classes to assign to this particular X-ray source. Thus, the issue of counterpart reliability in ClassX is not as crucial as one might think. It becomes crucial, of course, when the multiwavelength information needs to be assigned to the source, but this task is different from classification.

4. VERIFICATION

4.1. *Cross-Validation and Classifier Characterization*

The statistical nature of ClassX classifiers means that one has to have some measure of the quality of a classifier to tell if

TABLE 4
Cross-Validation for the Classifiers from RASS BSC and WGACAT

Input Class		ClassX Class					
Name	Number	Star	WD	XRB	Galaxy	AGN	Cluster
RASS-X Classifier							
Star	4694	4505	17	1	138	31	2
WD	78	8	65	0	2	2	1
XRB	192	119	10	52	6	3	2
Galaxy	1015	651	6	4	261	59	34
AGN	726	528	2	0	111	84	1
Cluster	210	56	0	2	40	0	112
Total	6915	5867	100	59	558	179	152
RASS-XOR Classifier							
Star	4675	4629	0	1	45	0	0
WD	75	3	63	2	4	3	0
XRB	173	5	9	61	62	29	7
Galaxy	945	94	3	4	654	170	20
AGN	707	5	1	6	128	561	6
Cluster	188	3	0	4	52	25	104
Total	6763	4739	76	78	945	788	137
WGACAT-X Classifier							
Star	4626	3739	4	10	747	35	91
WD	67	38	11	1	14	0	3
XRB	265	121	1	61	72	2	8
Galaxy	1281	370	0	1	557	166	187
AGN	3012	579	0	8	60	2189	176
Cluster	1496	356	0	3	699	101	337
Total	10747	5203	16	84	364	4278	802
WGACAT-XOR Classifier							
Star	4028	3617	1	7	279	51	73
WD	59	25	4	3	25	0	2
XRB	239	87	0	76	66	2	8
Galaxy	962	267	0	2	307	276	110
AGN	2648	144	0	6	95	2195	208
Cluster	1311	170	0	6	599	67	469
Total	9247	4310	5	100	491	3471	870

the classification results are of any value. To adequately assess a classifier and interpret classification results, one also needs to know the differences between the classes and the relevance of these differences in a particular application of the classification results. In the following, we describe some methods to assess the quality of ClassX classifiers, and we introduce a quantitative characterization of both the classifiers (e.g., reliability and completeness of classification, classifier preference) and classes (e.g., class affinity).

A natural data set to use in order to confirm the quality of a classifier is the training set that was used to develop it. Our classifiers are tested using five fold cross-validation. In this technique the training set is divided into five randomly selected subsets (''folds'') of equal size. Setting aside the first fold, 10 decision trees are constructed by training on the other four folds. Then the trees are tested for accuracy on the first fold, which was not used in the training. This process is repeated five times, each time holding back a different fold. When this is complete, we have classified the entire training sample. This standard technique avoids the overly optimistic results for classification accuracy that one would get if one simply trained the classifier on all the data and then tested it on the same data.[15]

The results of cross-validation can be viewed as a matrix with the input classes as the row headers and the column headers as the output classes (see Table 4). For a perfect classifier, only the diagonal of the matrix would be populated. In practice, the ratio of diagonal to off-diagonal elements gives us an immediate sense of how well the classifier has worked. In most cases the accuracy of the classifier will be higher for the training set sample than for originally unclassified sources, because the population of unclassified sources may differ systematically from known sources (e.g., by being fainter.) On the other hand, some disagreements between the OC1 classifier and the training set classification are the result of classification errors in the (imperfect) training set. There the cross-validation results correspondingly underestimate the classifier accuracy.

[15] Note that classifiers trained with 80% of the data are only used in cross-validation. The classifiers installed at the ClassX Web site and used in this paper for classifications of unknown objects are trained using the entire sample of preclassified sources.

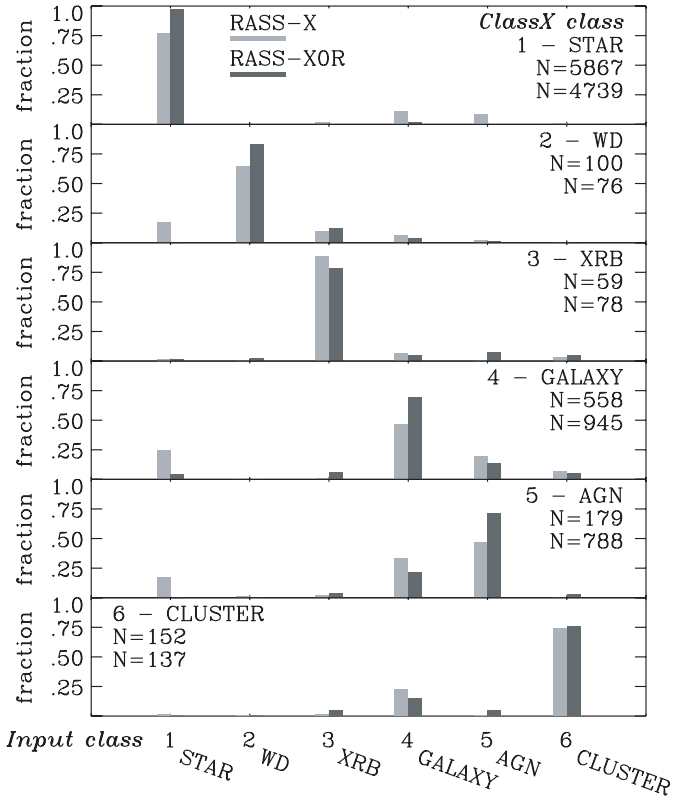

FIG. 10.—Classifier cross-validation. Distribution of input classes within a given ClassX class (class affinity). Light and dark gray scale refer to the results for the classifiers using X-ray data only (from *ROSAT*) and X-ray plus optical and radio data, respectively. Each panel shows the fraction of each of the input classes assigned by ClassX and the class name. Ideally, that fraction is 100% for those times when the input class always yields the correct output class—a 100% *reliable* classification. In practice, ClassX assigns the given class to a fraction of objects whose input class was different. This happens more often for classes whose *affinity* with the given class in the parameter phase space is the largest.

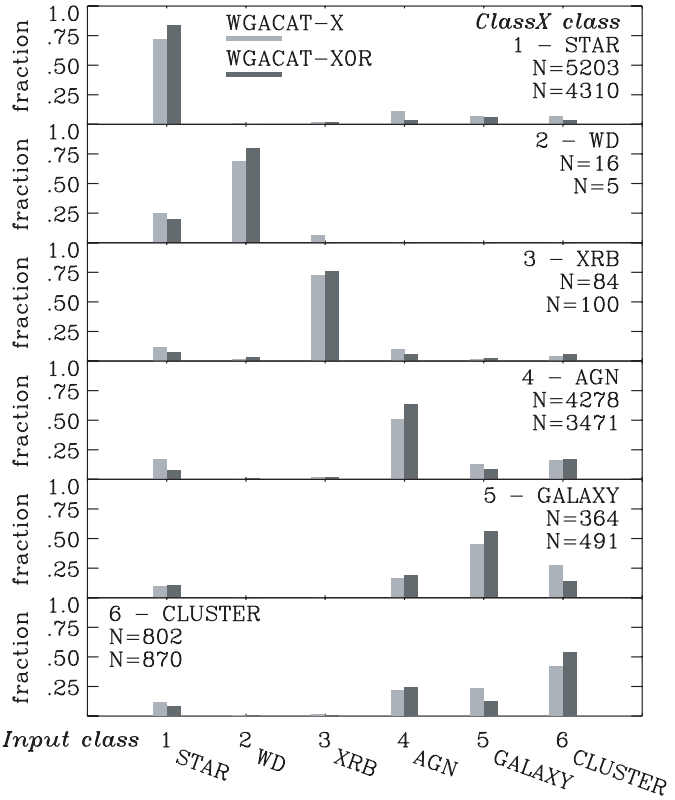

FIG. 11.—Same as Fig. 10, but for the classifiers derived from WGACAT and with the order of the GALAXY and AGN classes reversed.

The cross-validation results are shown in Figures 10–11. Each panel in these figures gives the fraction of objects in input class categories classified by ClassX as objects of a given type. The diagonal across the panels gives, therefore, the fraction of correctly classified sources in each class and thus represents *reliability* of classification. Because of closeness, or *affinity*, of some classes in the parameter phase space (e.g., galaxies and AGNs), the classifier may place some objects of a given input class into a class with similar properties. Figures 10–11 characterize quantitatively such class affinity. One can infer, for instance, from Figure 10 that there is a substantial affinity between the *ROSAT* BSC galaxies and AGNs when only X-ray properties are considered. Addition of optical information decreases that affinity quite noticeably. At the same time, clusters of galaxies are obviously distinctly different from AGNs in the X-ray. The affinity relationships between the classes are somewhat different for objects from WGACAT (Fig. 11).

In Figures 12 and 13, each panel gives the fraction of input objects of a given type classified by ClassX into different class categories. The diagonal across the panels shows the completeness of the placement of sample objects of a given type into the correct class category, giving us a measure of classification *completeness*. In general, Figures 12 and 13 show us the classifier *preferences* as it puts objects of a given type into different class categories.

Affinity and preference plots in Figures 10–13 are useful when one wants to know what outcome to expect from a

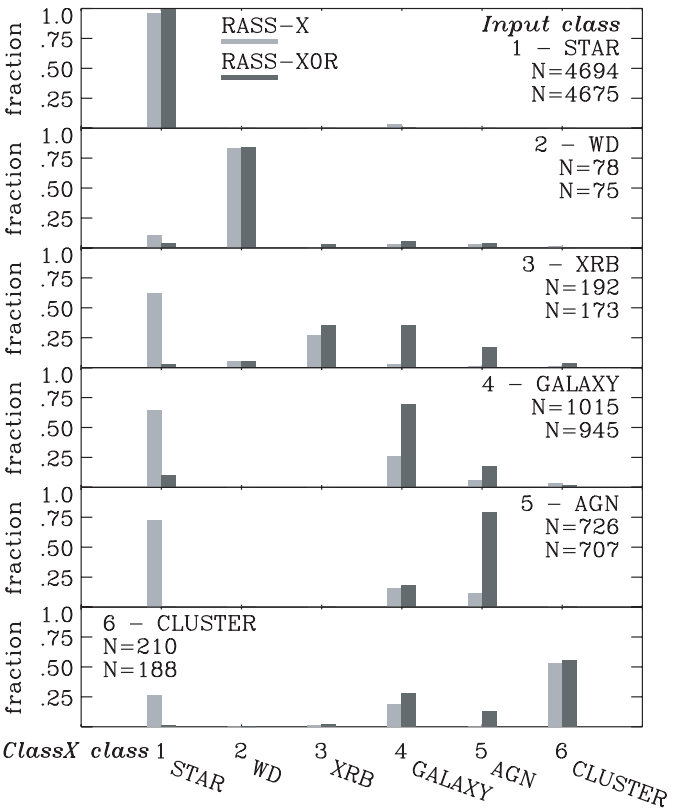

FIG. 12.—Classifier cross-validation. Distribution of ClassX classes within a given input class (classifier preference). Each panel exhibits the fraction of objects of the given input class in each ClassX class. An ideal classifier would assign all objects in a given input class to the same class. The *preference* is the likelihood that a given input class will be assigned to an output class. Gray scale as in Fig. 10.

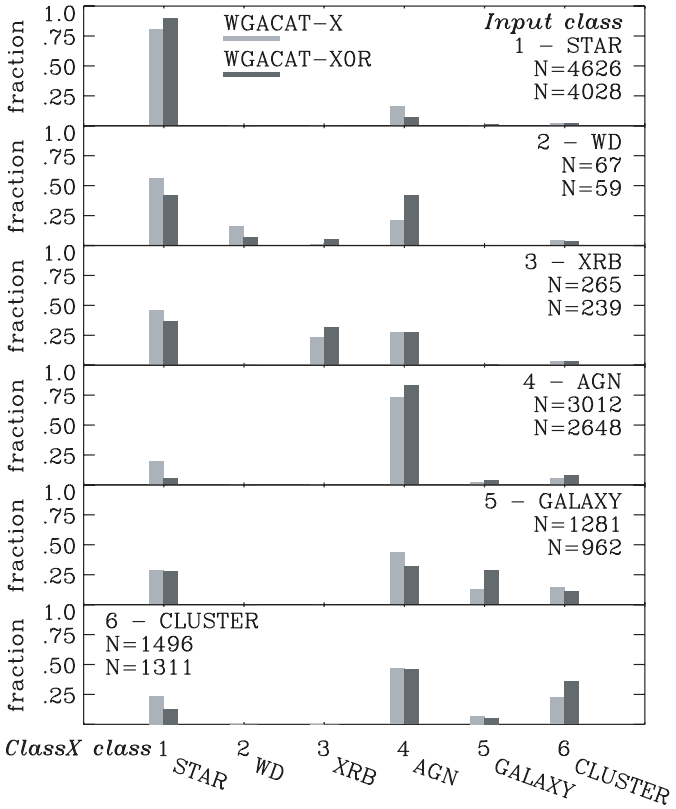

FIG. 13.—Same as Fig. 12, but for the classifiers obtained from the WGACAT training sets.

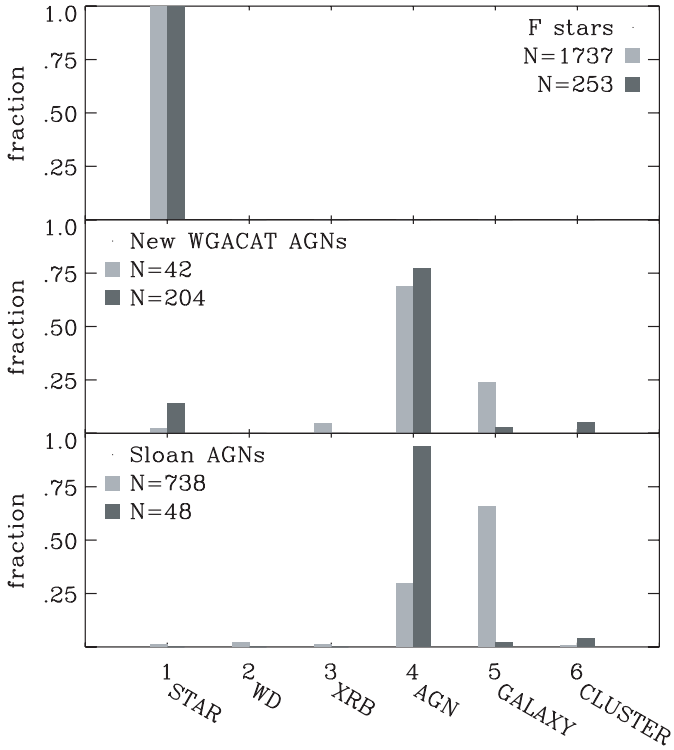

FIG. 14.—Class distribution for a sample of X-ray F stars (*top panel*) and two samples of AGNs (*middle and bottom panels*). Each sample was correlated with the RASS catalog and WGACAT, and the matching sources were classified by the RASS-XOR classifier (*light gray scale*) or the WGACAT-XOR classifier (*dark gray scale*). For each plot, the first value of *N* indicates the number of objects found in the RASS catalog, while the second value gives the number found in WGACAT. Classification shown in light gray in the lower panel is discussed in text. The sample of F stars is from Suchkov et al. (2003). The sample in the middle panel is from P. Padovani (2003, private communication). It comprises mostly the sources from Landt et al. (2001) and Perlman et al. (1998), which were drawn from the previously unclassified WGACAT sources and identified as AGNs. The sample in the bottom panel comprises AGNs from SDSS that were found to have X-ray counterparts in the *ROSAT* All-Sky Survey catalogs (Anderson et al. 2003).

particular classification. For instance, from the XRB panel in Figure 12, one would know that less than half of XRBs are expected to be revealed in a sample of X-ray sources. The same panel in the affinity plot (Fig. 10) would show that 75% or more of sources classified as XRBs are expected to be real XRBs.

The actual counts of objects in both the input and ClassX classes used in cross-validation are given in Table 4. The completeness of the classifier for a particular class is given by the ratio of the diagonal element to the sum of the column. This indicates the fraction of a given class where we recover the correct class. The reliability of the classifier is given by the ratio of the diagonal element to the sum over the row. The normalized row is a measure of affinity of a given class with other classes: for a given input class, what class does the automated classification yield? Of course, in both cases we must assume that the original classification is correct.

The cross-validation matrices immediately show many interesting features. When data are misclassified, they are usually misclassified into related categories. For instance, clusters or AGNs are misclassified as galaxies and vice versa.

The effect of large samples of one class versus smaller samples of another is also evident. Since there are so many stars, they can significantly contaminate samples of galaxies. Even though a classifier may furnish relatively high completeness for a given class, classification reliability for that class would be relatively low when occasional misclassification of a very common type overwhelms the correct classification of a rare type. The smaller the relative frequency of the object, the more distinctively its observational signature needs to stand out against the other classes. For example, WDs are characterized by very soft X-ray spectra. Thus, even though they make up only a small subset, they are still easily distinguished by our classifiers.

4.2. *Verification Using External Samples*

While the cross-validation results are useful, they cannot address any issues involving the selection of data for the training set itself. We can get some insight into that concern by looking at how well the classifiers handle *ROSAT* sources of known class that were not in the training set. Using our standard pipelines, we have classified samples of such sources from a number of catalogs containing identified *ROSAT* objects. Three such samples are discussed below in more detail. In addition to testing the ClassX classifiers, these examples show some of the research areas where the broad classifications the current classifiers provide can be useful.

4.2.1. *Hipparcos F stars*

Suchkov et al. (2003) identified 2011 F stars from the *Hipparcos* catalog as X-ray emitters that have X-ray counterparts in the RASS FSC and, to a lesser extent, RASS BSC. Submission of the list of these stars to the classifier RASS-XOR resulted in an output list of 1737 sources, all of which classified as stars (Fig. 14). Also, a smaller subset of these stars found in the WGACAT by the WGACAT-XOR classifier were all

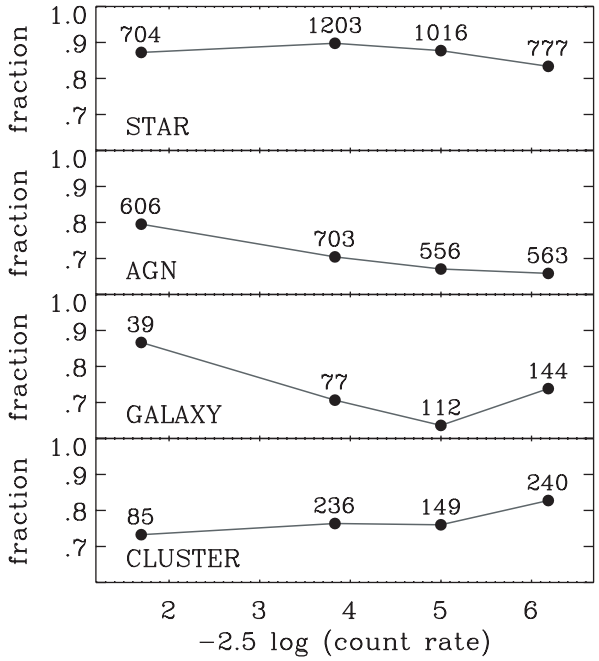

FIG. 15.—Fraction of the WGACAT training set sources correctly classified by the classifier WGACAT-XOR, displayed as a function of X-ray brightness [defined as $-2.5\log(\text{count rate})$]. The actual number of sources of a given class in each brightness bin is also shown.

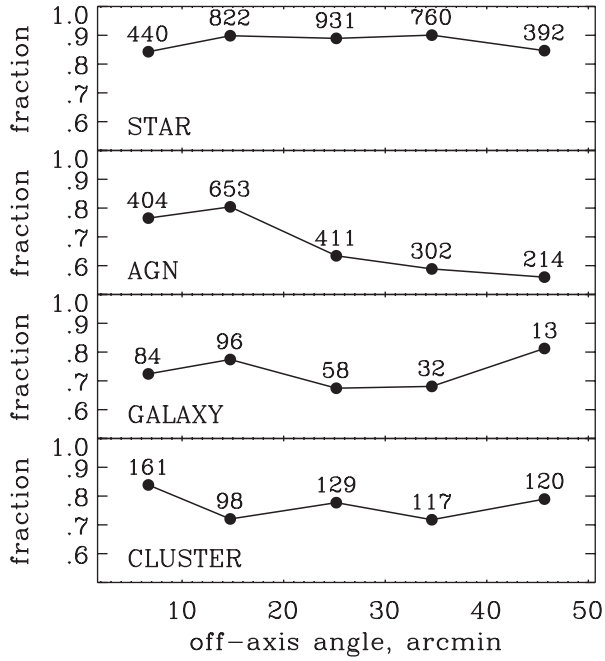

FIG. 16.—Fraction of WGACAT training set X-ray sources correctly classified by the WGACAT-XOR classifier as a function of the source off-axis angle, θ. Sources within $\theta = 3'$, which may include the targets of observations, are excluded. The number of objects within each bin is shown at each point.

identified as stars. This result is consistent with a very high reliability of star classification for these classifiers as inferred from Figures 10 and 12, thus strongly supporting the credibility of the cross-validation results.

4.2.2. *New AGNs from the WGACAT*

P. Padovani (2003, private communication) supplied us with a sample of 251 WGACAT sources that were identified by him and his collaborators as various types of quasars and AGNs (Landt et al. 2001; Perlman et al. 1998; P. Padovani et al. 2004, in preparation). The results of classification of this sample with the WGACAT-XOR and RASS-XOR classifiers are shown in the middle panel of Figure 14. The classifier does a good job distinguishing the AGNs from all other classes.

4.2.3. *AGNs from the SDSS*

SDSS is a deep photometric and spectroscopic optical survey, in which a large number of sources were spectroscopically identified as AGNs. For more than 1200 SDSS AGNs, Anderson et al. (2003) found X-ray counterparts in the *ROSAT* All-Sky Survey. We used a sample of 964 of these AGNs to test the performance of the ClassX classifiers. The results of the classification of that sample are shown in the lower panel of Figure 14. The classifier performance is very good in terms of differentiating the SDSS AGNs from galactic X-ray sources (stars, WDs, and XRBs) and clusters of galaxies. The WGACAT-XOR classifier easily differentiates these AGNs from galaxies; the RASS-XOR classifier is less successful in such a differentiation, likely because it is trained with substantially brighter objects.

4.3. *Classification Accuracy as a Function of X-Ray Brightness*

One clear distinction between the classified and unclassified sources is that the classified sources are generally brighter. One may expect that classification accuracy for fainter sources would be different. As one can see in Figure 15, classification accuracy does indeed vary with X-ray brightness. Interestingly enough, the degree and even the sense of that variation is not the same for different classes. In the case of AGNs, the accuracy drops from 80% at the bright end to below 70% at the faint end of the distribution. In contrast, the classification accuracy of clusters of galaxies tends to increase rather than decrease toward faint sources. For stars, accuracy variation is rather small,

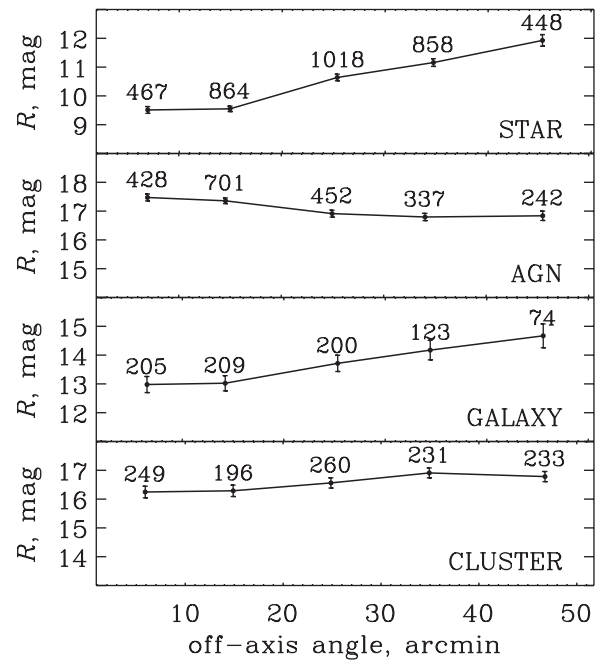

FIG. 17.—Optical counterpart brightness (in the USNO-B1 red band, R) as a function of the X-ray source off-axis angle for training set sources used by the WGACAT classifier (sources within $\theta = 3'$ are excluded). The number of objects within each bin is shown at each point.

TABLE 5
CLASSIFICATION OF RASS BSC BY THE RASS-X CLASSIFIER

RASS BSC Source Name	Class Index	*P*(Star)	*P*(WD)	*P*(XRB)	*P*(Galaxy)	*P*(AGN)	*P*(Cluster)	Class Name
1RXS J000007.0+081653	3	0.290	0.007	0.038	0.445	0.176	0.044	Galaxy
1RXS J000011.9+052318	0	0.666	0.003	0.023	0.142	0.126	0.040	Star
1RXS J000012.6+014621	0	0.702	0.002	0.022	0.125	0.117	0.031	Star
1RXS J000013.5+575628	0	0.871	0.001	0.017	0.060	0.047	0.003	Star
1RXS J000038.4+794037	0	0.888	0.001	0.015	0.049	0.044	0.003	Star
1RXS J000042.5+621034	0	0.888	0.001	0.015	0.049	0.044	0.003	Star
1RXS J000055.5+172346	0	0.819	0.001	0.018	0.076	0.072	0.014	Star
1RXS J000115.6+705535	0	0.871	0.001	0.017	0.060	0.047	0.003	Star
1RXS J000119.8+501659	0	0.696	0.005	0.067	0.111	0.098	0.023	Star
1RXS J000123.3+272241	0	0.888	0.001	0.015	0.049	0.044	0.003	Star

NOTE.—Table 5 is given in its entirety at the ClassX Web site. A portion is shown here for guidance regarding its format and content.

with a slight tendency to accuracy degradation at the faint end. In all cases, however, the accuracy remains relatively high even at the faint end of the training sample.

4.4. *Classification Accuracy as a Function of Optical Counterpart Brightness and Off-Axis Angle*

For the WGACAT sample, classifier performance can also be expected to vary with the source off-axis angle, θ, because of varying properties of X-ray sources and varying reliability of counterpart selection. Because of the PSPC vignetting, objects at large off-axis angles are expected to be somewhat brighter. For all objects this effect is about an order of magnitude. Figure 16 shows that classification accuracy does not seem to vary strongly as a function of the distance of the source from the center of the field of view, except perhaps for AGNs.

Figures 16 and 17 indicate that counterparts selected by the classifier do not seem to be strongly affected by the off-axis angle. Generally, the brightness of the counterparts is substantially greater than what would be found for a random nearby position. Table 3 indicates that we anticipate an average red magnitude of about 17.1. AGNs are very close to the average background luminosity, but they show no dependence on the offset angle.

4.5. *Limits and Issues*

While our validation of the classifiers is not complete and we plan to continue to look at the effects of cross-correlation errors and of selection in the training set, several distinct lines of evidence suggest that these classifiers give reasonable classifications for their sources. The classifiers have been built from heterogeneous data sources, which are likely to have some fraction of incorrect identifications and correlations. Pruned decision tree classifiers seem to be robust in the face of such contamination. One can use the classifiers to attempt to purify the input data set, and we plan to do so in future work.

5. SUMMARY OF RESULTS

Classification results from ClassX for the entire WGACAT and RASS data sets are available at the ClassX Web site.[16] They are illustrated in Tables 5 and 6, which show the first few rows of two selected tables.

In addition to these static classifications, more than two dozen ClassX classifiers, readily accessible for the community for immediate use, are currently deployed at the ClassX Web site. The Web site contains a description of the input data format, which is the list of source coordinates, and the input/output options. All the classifiers are supplied with the information indicating the classifier class categories, parameters (attributes) to be used in classification and returned in the output, databases (catalogs) to be searched for the source information, and other relevant information. In the output, each classified source is supplied with classification probabilities for all classes and is assigned a class name, which corresponds to the class with the highest classification probability. The output also contains the parameter values retrieved for the source and used in classification.

6. CONCLUSIONS

Classification of X-ray (or optical, infrared, etc.) sources into various categories of astronomical object types can rarely,

[16] See footnote 8.

TABLE 6
CLASSIFICATION OF WGACAT BY THE WGACAT-XOR CLASSIFIER

WGACAT Source Name	*P*(Star)	*P*(WD)	*P*(XRB)	*P*(Galaxy)	*P*(AGN)	*P*(Cluster)	Class Name	ROSPSPC Counterpart?
1WGA J1055.2+5638	0.483	0.039	0.061	0.057	0.254	0.106	Star	n
1WGA J1049.6+5641	0.483	0.039	0.061	0.057	0.254	0.106	Star	n
1WGA J1053.8+5709	0.270	0.039	0.039	0.219	0.348	0.085	AGN	y
1WGA J1053.2+5718	0.265	0.035	0.035	0.209	0.405	0.050	AGN	y
1WGA J1052.9+5725	0.415	0.022	0.026	0.160	0.332	0.045	Star	y
1WGA J1051.3+5725	0.483	0.039	0.061	0.057	0.254	0.106	Star	y
1WGA J1751.8−3450	0.233	0.026	0.028	0.375	0.291	0.047	Galaxy	y
1WGA J1415.2+1119	0.233	0.026	0.028	0.375	0.291	0.047	Galaxy	y
1WGA J1415.2+1119	0.352	0.037	0.141	0.046	0.311	0.115	Star	n
1WGA J1415.0+1119	0.238	0.031	0.032	0.290	0.347	0.062	AGN	y

NOTE.—Table 6 is given in its entirety at the ClassX Web site. A portion is shown here for guidance regarding its format and content.

if ever, be 100% accurate. The presence of uncertainty inherent to classifications based on statistical methods immediately splits the very goal of classification into a set of different goals, which are often incompatible. As a result, any statement about classifier effectiveness would generally make sense only if the classification goal or task, with respect to which the effectiveness is considered, is indicated. For example, one may want to either isolate as completely as possible those objects of a given class in a given sample, even at the expense of a larger fraction of misclassified sources, or deal only with objects of a class identified with the highest possible degree of reliability, even at the expense of rejecting many class objects that the classifier is unable to identify as such at the desired level of reliability. These different goals can be addressed in ClassX with different classifiers. One classifier can be effective in identifying to a high degree of completeness the members of a class, but classification reliability for identified class members may not be high enough. Still another classifier can be effective in delivering highly reliable class members but may miss many actual members of the same class.

Supervised classification techniques are a very powerful way of extending information about well-understood objects in a sample to the entire sample. For X-ray sources, it is possible to do classifications using just a few X-ray parameters as object attributes. Multiwavelength data can substantially improve the quality of the classifications, although adding data without regard to its quality or uniqueness does not necessarily help.

The ClassX classifications are useful for studying classes of objects, but the classification of any individual object should be taken as advisory rather than definitive. Human understanding and judgment is crucial in assessment and interpretations of the results. This is especially true given the statistical nature of ClassX.

In ClassX, a substantial number of input (training) sources are required for each class to effectively classify a sample. This number depends on the degree to which attributes of the class differ from those of other classes. In the case of WDs, the RASS-XOR classifier trained with less than a hundred of these objects proved nevertheless to be quite effective in both detecting the majority of actual WDs in the (training) sample of many thousand objects and ensuring high reliability of WD candidates.

We anticipate that modifications to the classifier algorithm that note when objects do not map well into existing classes will be needed to improve its detection capabilities for previously unknown object types (Laidler & White 2003). Currently, the latter functionality can be emulated through appropriate analyses of classification probabilities provided by ClassX.

Optical information is critical to distinguishing Galactic from extragalactic sources. It is less crucial for classifying clusters of galaxies and WDs. The effect of infrared information in ClassX is generally similar to that from the optical in distinguishing broad classes. This information becomes increasingly useful in finer grained classification. A network of ClassX classifiers, each using a different set of object parameters (attributes) and even a different set of classes, can provide a highly complete and reliable overall classification. Even when the specific counterpart to a given source cannot be accurately identified, the upper limit to the optical brightness given by the ensemble of counterparts is extremely helpful in defining the class of the source. Cross-correlation with optical catalogs is helpful in classification even when it does not lead to a secure identification of the counterpart.

In general, the more detailed and accurate the information available to a classifier, the more precise the classification results. However, information that is not used in distinguishing classes can confuse the classifiers so that careful pruning of the information provided to the classifiers is essential. The phase space of possible classifiers is very large. A substantial fraction of this effort was to learn a reasonable minimum of information to use.

We have begun applying ClassX to more detailed studies of particular object classes. For example, Suchkov & Hanisch (2003, 2004a) find strong evidence for new identifications of low-luminosity low-mass X-ray binaries having hard X-ray spectra, most likely associated with regions of star formation in the Galactic plane. Hanisch et al. (2003) noted up to a five fold increase in the number of known late-type X-ray–emitting stars, suggesting a large pre–main-sequence population of T Tauri stars in active star formation regions. Suchkov & Hanisch (2004b) have continued this study and have been able to correlate the spectral hardness of pre–main-sequence stars with the expected X-ray absorption along the line of sight to different star formation regions (ρ Oph, Orion, the Pleiades, etc.). We are now investigating the use of ClassX on other large databases, such as SDSS, and are planning extensions of ClassX to other X-ray missions (*Chandra*, *XMM-Newton*) with different bandpass coverage and spatial resolution from *ROSAT*.

We wish to thank L. Angelini, P. Fernique, F. Genova, W. D. Pence, M. Postman, and M. Wenger for numerous discussions of the project. The comments and suggestions of the anonymous referee were very helpful in revising and clarifying the results in this paper. This work was funded through NASA's Applied Information Systems Research Program under grant NAG5-11019.

REFERENCES

Anderson, S. F., et al. 2003, AJ, 126, 2209
Condon, J. J., et al. 1998, AJ, 115, 1693
Giommi, P., Menna, M. T., & Padovani, P. 1999, MNRAS, 310, 465
Hanisch, R. J., Suchkov, A. A., White, R. L., McGlynn, T. A., Winter, E. L., Corcoran, M. F., & Voges, W. 2003, in IAU Joint Discussion 8, Large Telescopes and Virtual Observatory: Visions for the Future (San Francisco: ASP), 59
Heath, D., Kasif, S., & Salzberg, S. 1996, in Cognitive Technology: In Search of a Humane Interface, ed. B. Gorayska & J. Mey (Amsterdam: Elsevier), 305
Laidler, V. G., & White, R. L. 2003, in Statistical Challenges in Astronomy III, ed. E. D. Feigelson & G. J. Babu (New York: Springer), 453
Lamar, G., et al. 2003, Astron. Nachr., 324, 156
Landt, H., Padovani, P., Perlman, E., Giommi, P., Bignall, H., & Tsiomis, A. 2001, MNRAS, 323, 757
Lucy, L. B. 1974, AJ, 79, 745
Maccacaro, T., Gioia, I. M., Wolter, A., Zamorani, G., & Stocke, J. T. 1988, ApJ, 326, 680
Mauch, T., Murphy, T., Buttery, H. J., Curran, J., Hunstead, R. W., Piestryzynski, B., Robertson, J. G., & Sadler, E. M. 2003, MNRAS, 342, 1117
Monet, D. G., et al. 2003, AJ, 125, 984
Murthy, S. K., Kasif, S., & Salzberg, S. 1994, J. Artif. Intell. Res., 2, 1
Ochsenbein, F., Albrecht, M., Brighton, A., Fernique, P., Guillaume, D., Hanisch, R., & Wicenec, A. 2000a, in ASP Conf. Ser. 216, Astronomical Data Analysis Software and Systems IX, ed. N. Manset, C. Veillet, & D. Crabtree (San Francisco: ASP), 83
Ochsenbein, F., Bauer, P., & Marcout, J. 2000b, A&AS, 143, 23
Odewahn, S. C. 1995, PASP, 107, 770
Perlman, E. S., Padovani, P., Giommi, P., Sambruna, R., Jones, L., Tsiomis, A., & Reynolds, J. 1998, AJ, 115, 1253
Rutledge, R., Brunner, R. J., & Prince, T. A. 2000, ApJS, 131, 335

Salzberg, S., Chandar, R., Ford, H., Murthy, S. K., & White, R. L. 1995, PASP, 107, 279
Stoughton, C., et al. 2002, AJ, 123, 485
Suchkov, A. A., & Hanisch, R. J. 2003, AAS, 203, 84.03
———. 2004a, ApJ, 612, 437
———. 2004b, AAS 204, 62.15
Suchkov, A. A., Makarov, V. V., & Voges, W. 2003, ApJ, 595, 1206
Voges, W., et al. 1999, A&A, 349, 389
———. 2000, IAU Circ., 7432, 1
Watson, M., et al. 2003, Astron. Nachr., 324, 89
Weir, N., Fayyad, U. M., Djorgovski, S. G., & Roden, J. 1995, PASP, 107, 1243
White, R. L. 1997, in Statistical Challenges in Modern Astronomy II, ed. G. J. Babu & E. D. Feigelson (Berlin: Springer), 135
———. 2000, in ASP Conf. Ser. 216, Astronomical Data Analysis Software and Systems IX, ed. N. Manset, C. Veillet, & D. Crabtree (San Francisco: ASP), 577
White, R. L., et al. 2000, ApJS, 126, 133
Yuan, W., et al. 2003, Astron. Nachr., 324, 178
Zhang, Y., & Zhao, Y. 2003, PASP, 115, 1006