Using a Learning Tree Decision Algorithm To Classify Celestial Objects

Derek Brehm (GWU), Oleg Kargaltsev (GWU), Blagoy Rangelov (GWU), Igor Volkov (GWU), George Pavlov (PSU)

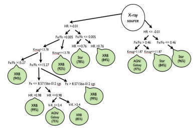
Abstract

With the advent of the latest generation X-ray telescopes there has been a major influx of data associated with the detection of hundreds of thousands X-ray sources. As one can rarely tell a source type from its X-ray properties alone, the full potential of the X-ray catalogs can only be unlocked by correlating multiwavelength (MW) properties via cross-identification with other surveys. However, one would spend an enormous amount of time classifying these objects ``manually'' on a source-by-source basis . Therefore, we are using a supervised learning algorithm to classify sources detected by the Chandra X-ray Observatory. The classifications are based on a training dataset which currently includes about 7,000 X-ray sources of known nature (main sequence stars, Wolf-Rayet stars, young stars, active galactic nuclei, low mass X-ray binaries, high mass x-ray binaries, and neutron stars). For each source, the training dataset includes up to 24 multiwavelength properties. The efficiency and accuracy of the classification is verified by dividing the training dataset in parts and performing cross-validation. The results are also inspected by plotting source properties in 2D slices of the parameter space. As an application of our automated procedure we classified unidentified sources in the fields of HESS J1809-193 and (see AAS poster # 153.21), and in part of the Chandra Source Catalog 1.1. We present the results of the verification tests and the classification results for Chandra Source Catalog 1.1 below.

Learning Tree Decision Algorithm

Following McGlynn et al. (2004) approach, we used a learning decision tree algorithm (LDTA) C5.0 to create a decision tree and classify X-ray sources. Initially the program analyzes the training dataset and then creates a decision tree that allows it to make predictions on unclassified data. For each source, we also calculated probabilities of it to belong to a certain class.

A schematic illustration of a decision tree. The actual tree used in the classifications described below is not shown because of its large size.

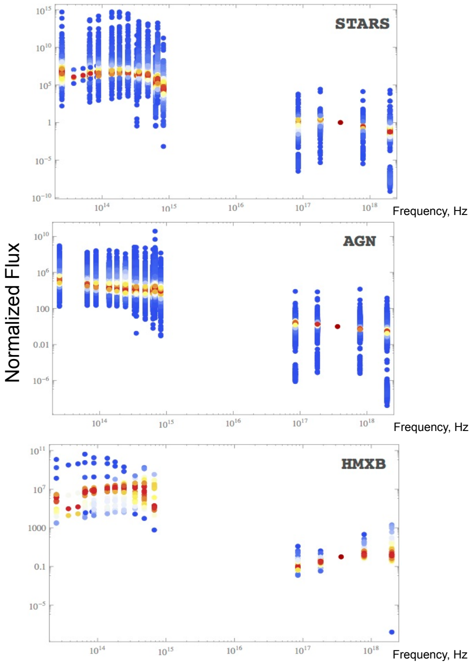


Catalogs used for compiling Training Dataset with identified sources:

- ●Lin et al. 2012 (literature verified sample)
- ●General Catalog of Variable Stars
- ●Pan-Carina YSO catalog
- ●Chandra Orion Ultradeep Project
- ●Veron Catalog of Quasars & AGN13th Edition
- ●Low-Mass X-ray Binary Catalog 4th edition
- ●High-Mass X-ray Binary Catalog 4th edition
- ●VIIth Catalog of Galactic Wolf-Rayet Stars
- ●ATNF Pulsar Catalog
- ●Cataclysmic Variables Catalog.

- ●AGN: Active Galactic Nuclei
- ●Stars: Main Sequence Stars
- ●YSOs: Young Stars
- ●WR: Wolf-Rayet Star
- ●PSR: Pulsars
- ●PSR_BIN: Pulsars in a Binary
- ●LMXB: Low Mass X-ray Binary
- ●HMXB: High Mass X-ray Binary
- ●CV: Cataclysmic variable Binary

Aggregated spectra (normalized to flux at 1 keV) of sources from different classes (flaring stars, AGNs, HMXBs) based on the training dataset:



Astronomical Data used in the training dataset

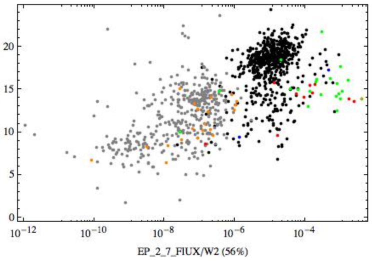
Catalogs used for comparison:

- ● 3XMMDR3 Catalog (2.0” cone search)
- ●2MASS Point Source Catalog (1.5” cone search)
- ●Sloan Digitized Sky Survey DR8 (1.5” cone search)
- ●Spitzer GLIMPSE I + II + 3D catalog (2.0” cone search)
- ●Wise All Sky Data Release (2.0” cone search)
- ●Vista VVV DR1 (1.5” cone search)
- ●USNO-B 1.0 (1.5” cone search)

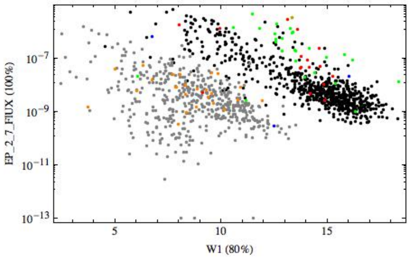
Source parameters used for classification

0.5-2 keV x-ray flux and 2-7 keV x-ray flux from 3XMM

- Hardness Ratio 2 = (f(1-2keV)-f(0.5-1keV))/(f(1-2keV)+f(0.5-1keV))
- Hardness Ratio 4 = (f(4.5-12keV)-f(2-4.5keV))/(f(4.5-12keV)+f(2-4.5keV))
- EPIC Source variability (probability of the source NOT being variable)
- J mag, H mag , and K mag from 2MASS
- R mag from Vista
- B mag and R mag from USNO-B
- U mag, G mag, R mag, I mag, Z mag from SDSS
- Wise 3.4, 4.6, and 12 micron magnitudes
- Spitzer 3.6, 4.5, 5.8, and 8.0 micron mags



0.5-2 keV flux versus 2-7 keV flux for sources from the training dataset



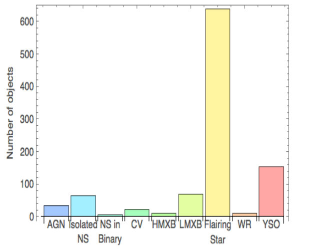
0.5-2 keV flux versus NIR k-magnitude for sources from the training dataset

Source types:
AGN
STARS
CV
ATNF
HMXB
YSO
LMXB

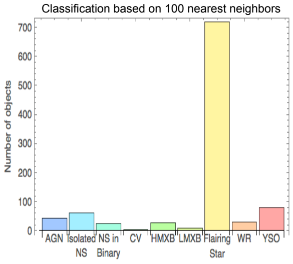
Classifications for Chandra Source Catalog 1.1

When constructing the CSC 1.1 candidate dataset we applied the initial parameters of $|b| < 3$ degrees latitude , $20 < l < 65$, and $295 < l < 340$ degrees longitude. The resulting dataset has 3129 sources.

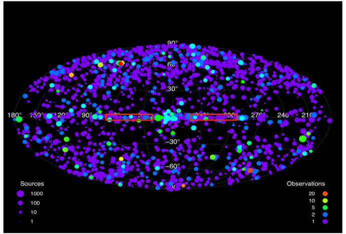
Classification based on C5 LDTA



Classification based on 100 nearest neighbors



32.4% of the total number of sources we classified have > 70% confidence:
AGN = 0.9% NS= 3.6% BIN_PSR = 0.2% CV = 0.7%
HMXB = 0.1% LMXB = 1.2% STARS = 22% WR = 0.2% YSO = 3.5%



A skymap representation of the Chandra Source Catalog (Evans et al. 2010). Red boxes show regions used for the classification.

Future Plans:

We continue to verify our methods and results by performing various LDTA runs and simple nearest neighbors classifications in order to test if the initial LDTA classifications hold up. Also, we are assembling additional training datasets to see if these methods are robust. We plan to classify the entire CSC1.1 catalog and expand classifications to 3XMM catalog We will also produce a user-friendly pipeline in order to make it available to the community.

Questions? Find this guy!

References:

- ●Evans, I. N., et al. 2010, ApJS, 189, 37 J.R.
- ●Chawla et al. SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research 16 (2002) 321–357
- ●McGlynn, T., et al. 2004, ApJ, 616, 1284
- ●Quinlan (1986). Induction of Decision Trees, Machine Learning, (1), 81-106

This research was partially supported by the NASA/SAO award AR3-14017X.