Multiwavelength classification of Galactic X-ray sources using machine-learning

Oleg Kargaltsev[1], Jeremy Hare[2], Hui Yang[1], Igor Volkov[1], Blagoy Rangelov[3], George Pavlov[4]

1. - The George Washington University 2. - NASA GSFC 3. - Texas State University 4.- Pennsylvania State University

Abstract: The second CXO source catalog (CSCv2) contains over 300,000 X-ray sources and a vast majority of these sources have been observed serendipitously. A reliable automated classification of these sources is a challenging task that can greatly increase the populations of rare objects of known astrophysical types, and reveal new unusual and interesting sources. Compared to XMM-Newton, the unprecedented angular resolution of CXO allows for a much more reliable optical/IR counterpart identification for X-ray sources in the Galactic plane, especially within crowded environments. We discuss the recent developments to our machine-learning classification pipeline (MUWCLASS) driven by CSC v2 and recent sensitive surveys at optical/IR wavelengths. We will present the updates to the training data set and pipeline, showing how these improvements have impacted the pipeline's performance, and demonstrate the applications of the pipeline to the Galactic fields.

This work is supported by the Chandra X-ray Observatory grant AR9-20005A and NASA APAD grant 80NSSC19K0576.

Methods: Supervised Machine Learning (ML)

- ML algorithms can be used to classify these objects in an automated way;
- We currently use **Random Forest** (RF) algorithm, similar in structure to decision trees (e.g., C4.5, CART) but we have tested others as well;
- RF works by taking a sample of the training dataset with replacement (i.e., bootstrapping), building a decision tree by randomly selecting a subset of parameters to optimize on, and then repeats the process;
- RF algorithm builds an ensemble of decision trees, each one different from the rest. This method is much less prone to overfitting when compared to a single decision tree (Breiman et al. 2001) and is generally more robust due to the larger number of decision trees;
- Features (source properties) currently used for classification:

EP052Flux	X-ray flux in 0.5−2 keV
EP27Flux	X-ray flux in 2−7 keV
HR2	Soft band hardness ratio
HR4	Hard band hardness ratio
BPmag	Gaia BP band magnitude
RPmag	Gaia RP band magnitude
Gmag	Gaia G band magnitude
Jmag	2MASS J-band magnitude
Hmag	2MASS H-band magnitude
Kmag	2MASS K-band magnitude
W1mag	WISE W1-band magnitude
W2mag	WISE W2-band magnitude
W3mag	WISE W3-band magnitude
G−B	G minus BP band color
G−R	G minus RP band color
R−W3	RP minus W3 band color
W1−W2	W1 minus W2 band color
W2−W3	W2 minus W3 band color
H−W2	H minus W2 band color
J−H	J minus H band color
J−K	J minus K band color
B−J	BP minus J band color

Methods: Training Datasets

- There are two independently constructed training datasets, 3XMM-DR8-based (XMM-TD) and CXCv.2-based (CXC-TD), with confidently classified X-ray sources. For 3XMM-DR8 the verification and validation phase has been completed, for CXC-TD it is ongoing. The XMM-TD and CXC-TD currently contain 9892 and 3693 sources, respectively.

3XMM-TD	AGN	NS	PSR_BIN	CV	HMXB	LMXB	STAR	WR	YSO
# of sources	7067	88	10	144	24	58	1486	37	978

CXCv2-TD	AGN	NS	PSR_BIN	CV	HMXB	LMXB	STAR	WR	YSO
# of sources	1691	102	20	32	24	53	497	22	1252

- X-ray properties are insufficient to classify most of the sources, hence additional multi-wavelength (MW) information must be used; In each catalog up to 18 features (see Table above).
- Each TD has its own advantage and disadvantages. A large fraction of X-ray sources are present in both TDs and are the same sources but observed at different times and independently cross-correlated with MW catalogs. XMM-TD is larger but positional accuracies of X-ray sources are larger and chance of confusion (within the r=2″ matching circle) with unrelated IR/NIR/optical sources is larger (although X-ray sources from crowded environments have been omitted in XMM-TD). CXX-TD is smaller but positions are much more precise (matching circle r=1″) and confusion probabilities are lower (allowing to use sources from more crowded environments).
- There are currently 9 distinct object classes: Active Galactic Nuclei (AGN), Neutron Stars (NS), Neutron Star Binaries (NS_BIN), Low Mass X-ray Binaries (LMXB), High Mass X-ray Binaries (HMXB), Cataclysmic Variables (CV), Stars, Wolf-Rayet Stars (WR), and Young Stellar Objects (YSO). Although some classes are rather heterogeneous one has to balance the desire to have more classes with the need to populate them with enough confidently classified objects.
- Both TDs are strongly imbalanced being dominated by Stars, YSOs, and AGN. To combat this severe problem the SMOTE procedure is used (synthetic sources are created using the properties of the real ones) to increase the numbers of sources for underpopulated source classes.

Different cuts in feature space showing the separation between different source classes.

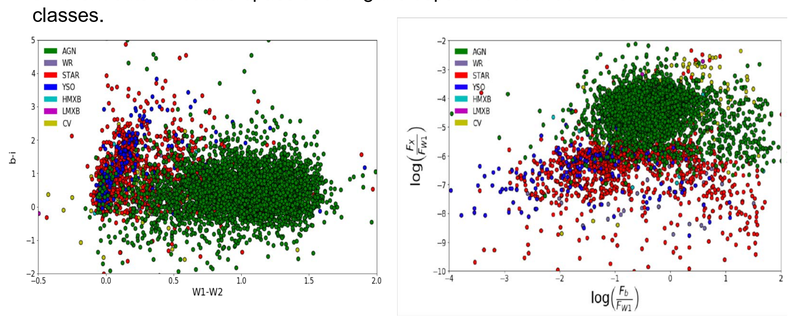
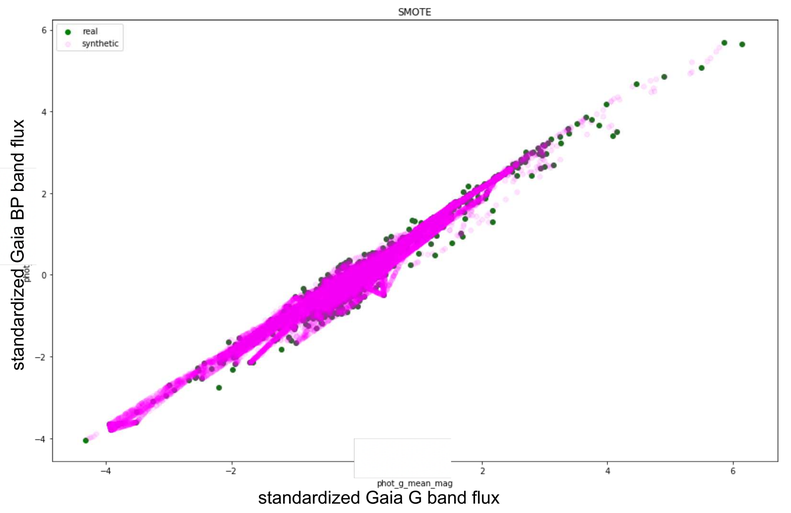


Illustration of the SMOTE (Chawla et al. 2002) procedure with green and purple dots representing real and synthetic data, respectively.

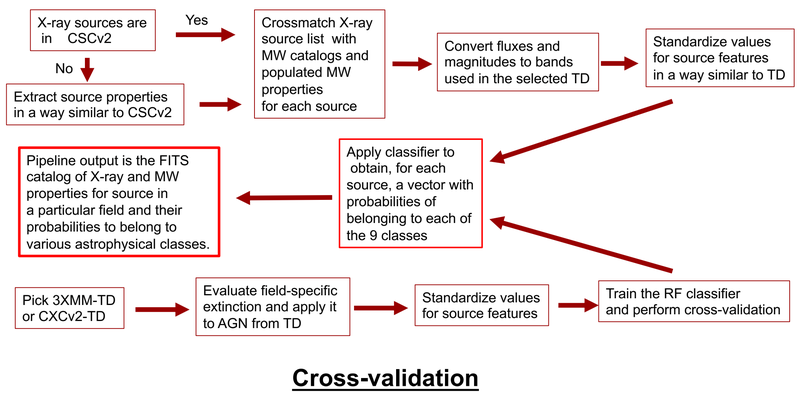


Background

- The CXO and XMM-Newton have been operating for ~20 years;
- For most observations analysis focuses on one primary target, while the other detected sources remain uninvestigated;
- Over time these observations have added up to ~500,000 unique X-ray sources from XMM-Newton and ~350,000 from CXO;
- Likely many interesting sources hiding in these catalogs, but manual classification is tedious;
- To maximize the scientific return and observing power of current (and future) instruments, automated and accurate methods of classification must be developed and tested;
- These ML classification tools will be particularly useful in e-ROSITA era.

MUWCLASS pipeline

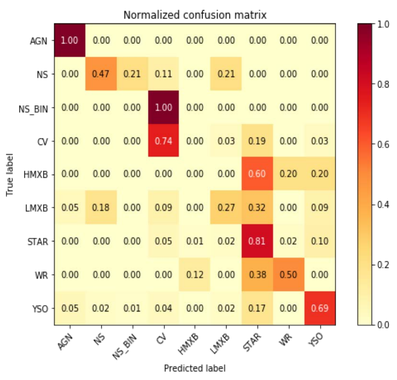
The pipeline flow chart



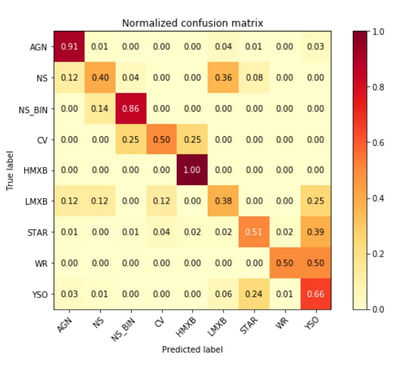
Cross-validation

• We test the classification accuracy by erasing labels (actual source class) for a fraction of sources in the TD and classifying these sources after training the classifier using the remaining sources. The outcome of this procedure is summarized by the Confusion Matrix which for ideal classification outcome is correctly in many cases. However, AGN, Stars. YSOs, and CVs are accurately identified in most cases.

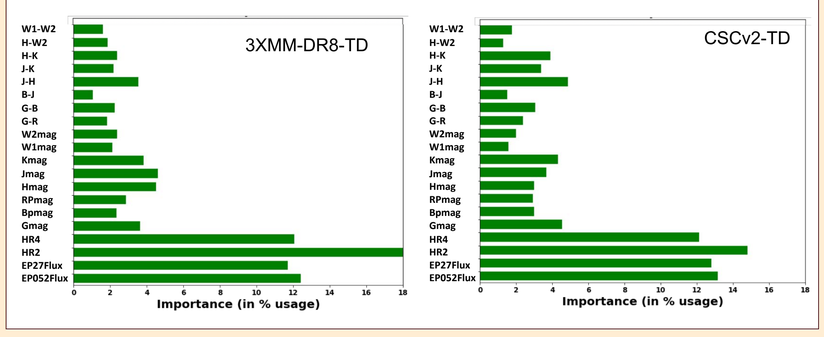
3XMM-DR8-TD



CSCv2-TD



- Another outcome of the cross-validation process is the assessment of the relative importance (usage by the algorithm) of the features



Recent Upgrades and Developments

- The MUWCLASS pipeline code is now fully written in Python with different tasks performed by individual functions
- Cross-matching with external catalogs is automated through remote queries
- Python analog of CSCv2 pipeline to process new observations (after 2014)
- Training dataset built from CSCv2; Conversions between 3XMM and CSCv2 properties of X-ray sources
- Field specific absorption correction is applied to AGNs from TD (which are out of the Galactic plane) while classifying sources in the Galactic plane.
- USNO-B2 replaced with Gaia DR2
- Spitzer GLIMPSE replaces WISE for the Galactic plane sources if coverage exists

Future Plans: Training Dataset

• Training dataset built from CSCv2; Conversions between 3XMM and CSCv2 properties of X-ray sources

• Field specific absorption correction is applied to AGNs from TD (which are out of the Galactic plane) while classifying sources in the Galactic plane.

•Increase the #s of NS, NS binaries (non-interacting), LMXBs, HMXBs.

• Include confusion probability into classification confidence calculation

• Include n_H as an additional feature in both TDs

• Include variability features in both TDs

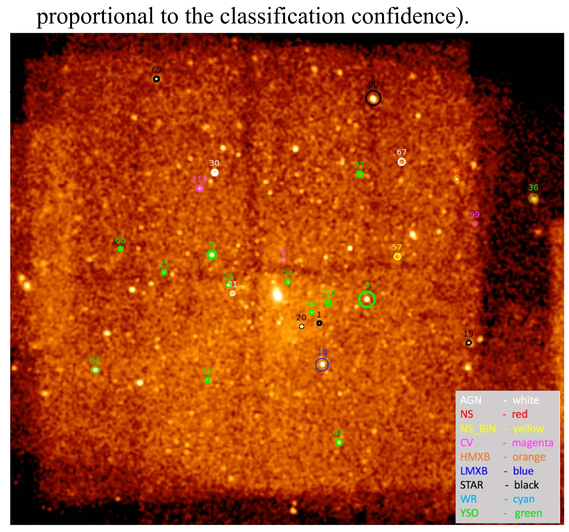
• Replace 2MASS with Pan-STARRS DR2, UKIDSS, and Vista VVV surveys

• Add Y and Z bands from Dark Energy Survey DR1 and UV fluxes from Galex.

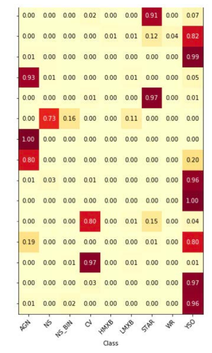
Application Examples:

Field of HESS J1809-193 (18 CXOACIS observations, 334 ks)

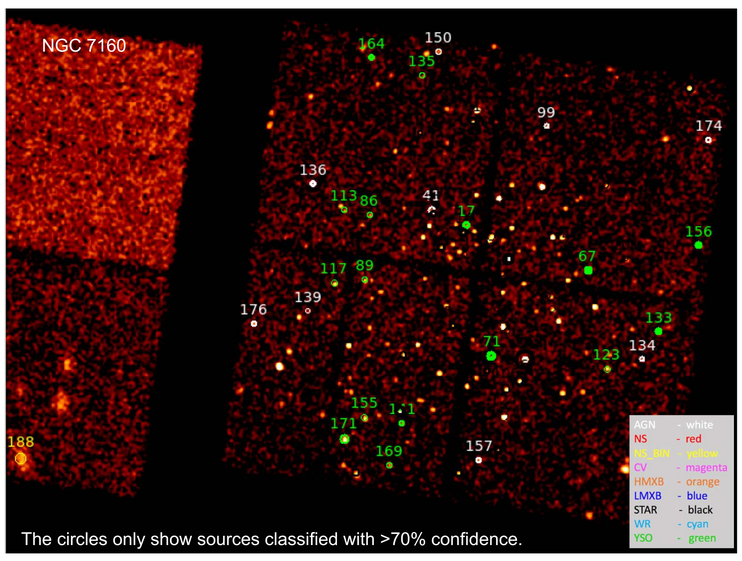
- In many cases GeV/TeV sources are associated with known offset pulsars, especially if those are young and have X-ray PWNe;
- Although these associations are plausible, in many cases the offsets are large and chance coincidences are possible;
- Analyzing other X-ray sources in the field can provide additional information about the associations;
- Below is an example of the pipeline output; the circles show sources that have >70% probability to belong to a particular class (the size of the circle is proportional to the source detection significance while the line thickness is proportional to the classification confidence).



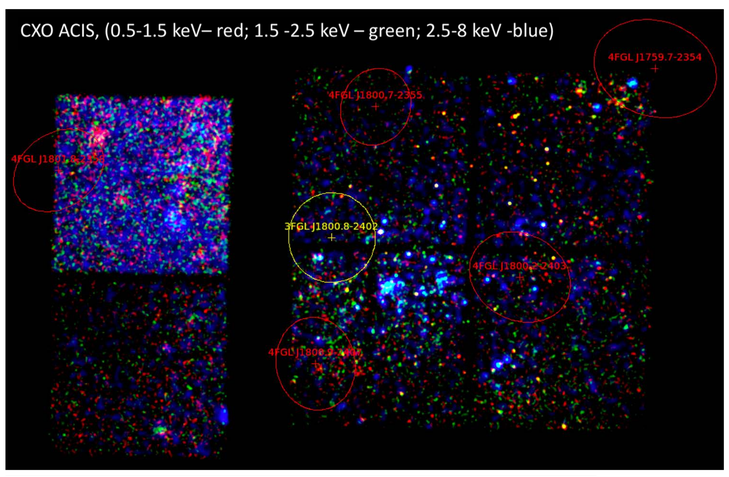
A sample of classification outcome matrix showing the distributions of probabilities between different classes.

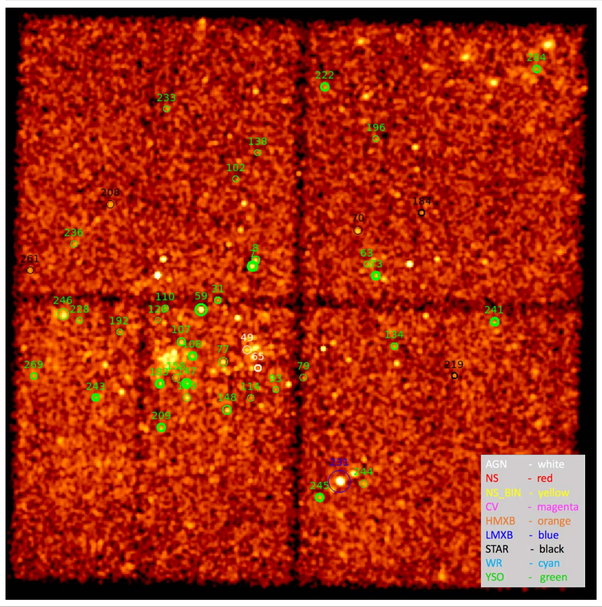


Open cluster NGC 7160 (4 CXO ACIS observations, 69 ks)



3FGL J1800.8-2402 field which now contains four 4 FGL sources and SNR W28-A2 (1 CXO ACIS observation, 78 ks)





Total intensity ACIS Image of the same field with sources classified above 70% confidence shown by circles.

Future plans: MUWCLASS pipeline

- Experimenting with other ML algorithms (neural networks, SVM).
- Incorporating distance information in the classification process.
- Explore having larger number of classes with less sources per class
- Application to nearby open clusters with known distances
- Application to all unidentified 4FGL source fields
- Placing MUWCLASS code and training datasets on GitHub