Automated source classification of new transient sources

M. Oertel[1], A. Kreikenbohm[1,2], J. Wilms[1], A. DeLuca[3]
on behalf of the EXTraS collaboration
[1] Remeis Sternwarte & ECAP, Univ. Erlangen-Nürnberg, Germany, [2] ITAP, Univ. Würzburg, Germany, [3] INAF/IASF, Milano, Italy

Introduction

The EXTraS project harvests the hitherto unexplored temporal domain information buried in the serendipitous data collected by the European Photon Imaging Camera (EPIC) onboard the ESA *XMM-Newton* mission since its launch. This includes a search for fast transients, missed by standard image analysis, and a search and characterization of variability in hundreds of thousands of sources. We present an automated classification scheme for new transient sources in the EXTraS project. The method is as follows: source classification features of a training sample are used to train machine learning algorithms (performed in *R*; randomForest (Breiman, 2001) in supervised mode) which are then tested on a sample of known source classes.

1 Sample Definition

The classification is based on the machine learning method. For the training we make use of the 3XMM-DR6 catalog. Spectral, variability and broad band information from the 3XMM sources is used as input for the automated classification algorithm.
The basis for the training sample is derived from Farrell et al. ApJ 813, 28, (2015) with the following modifications:

- Seyfert 1, Seyfert 2, BL Lac (Véron-Cetty & Véron, 2010, A&A, 518, 10)
- HMXB, LMXB (Liu+ 2006, A&A, 455, 1165 and 2007, A&A469, 807, Haberl & Sturm, 2016 A&A 586, 81)
- additional Cataclysmic Variables (CV, Ritter & Kolb catalog)

The remaining sources in the 3XMM-DR6 are used as test sample for the classification algorithm. Fig. 1 shows a sky map of the training sample.

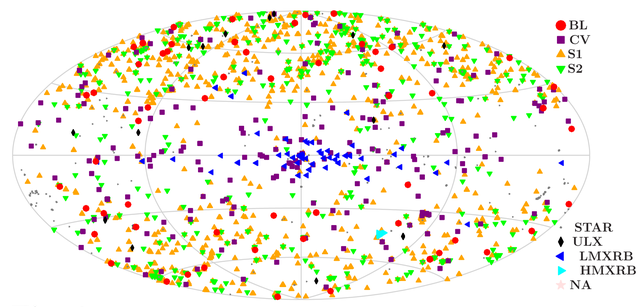

Fig 1: Sky map of the 3XMM training sample based on Farrell+ (2015).

2 Training the classification algorithm

For the definition of the X-ray spectral classification features, the following tasks are performed:

- calculate hardness ratios from different energy bands,
- perform spectral analysis using various models (absorbed powerlaw, thermal emission, plasma emission, etc.)
- counter-part search by cross-matching both samples against more than 40 catalogs from different wavebands.
- determine broadband spectral indices ($\alpha_{\mathrm{r,x}}$, $\alpha_{\mathrm{ir,x}}$, $\alpha_{\mathrm{o,x}}$, $\alpha_{\gamma,\mathrm{x}}$).
- randomize positions in train set to eliminate bias due to multiple detections
- make use of variability information
- balance size of each source class through resampling

The relative importance of each parameter in the machine learning algorithm is shown in Fig. 2. The overall accuracy is 97% (see Table 1).

Table 1: Accuracy and composition of train sample.

Class	Sources	Resampled	Accuracy
BL	60	3536	99%
CV	201	3564	97%
HMXRB	33	3750	99%
LMXRB	66	3502	99%
S1	1486	3773	93%
S2	485	4104	92%
STAR	563	3226	98%
ULX	17	3910	100%

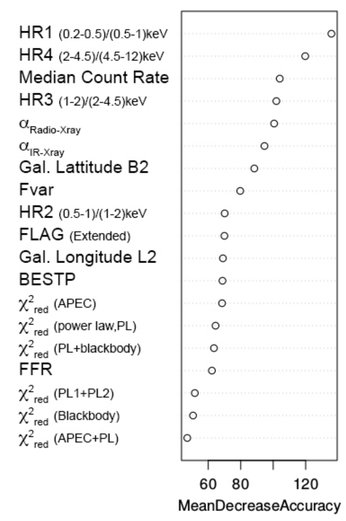

Fig 2: Relative importance of the features for the classification by source class.

3 Results & outlook

Application to the 3XMM-DR6 test sample:
The algorithm is applied to the test sample of 131791 sources. The resulting composition is

BL - 96; CV - 1534; HMXRB - 1055; LMXRB - 480; S1 - 31434; S2 - 11963; STAR - 24869; ULX - 336; Unclassified - 70791;

The distribution (see also Fig. 3) is close to what one would expect from the known distributions of the input sources.

Application to a set of 136 new EXTraS transient sources:
EXTraS found 136 transient sources with unknown classification. Applying the algorithm yields

BL - 0; CV - 21; HMXRB - 2; LMXRB - 6; S1 - 33; S2 - 47; STAR - 27; ULX - 0; Unclassified - 0;

The significance of the classifcation per source is only 30–40% indicating that it is incorrect. A reason for this failure is probably that the transient spectra are consistent with simple powerlaws (see Fig. 4). **A careful analysis of the timing properties of these transients is currently ongoing and shall improve the classification.**

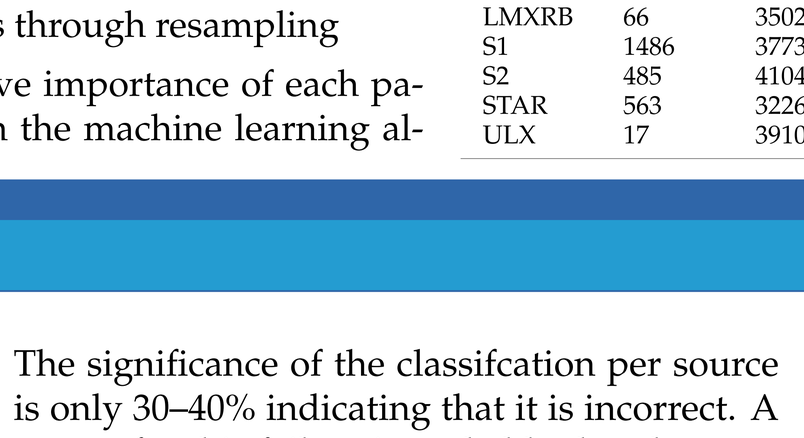

Fig 3: Sky map of classified sources form the 3XMM test sample.

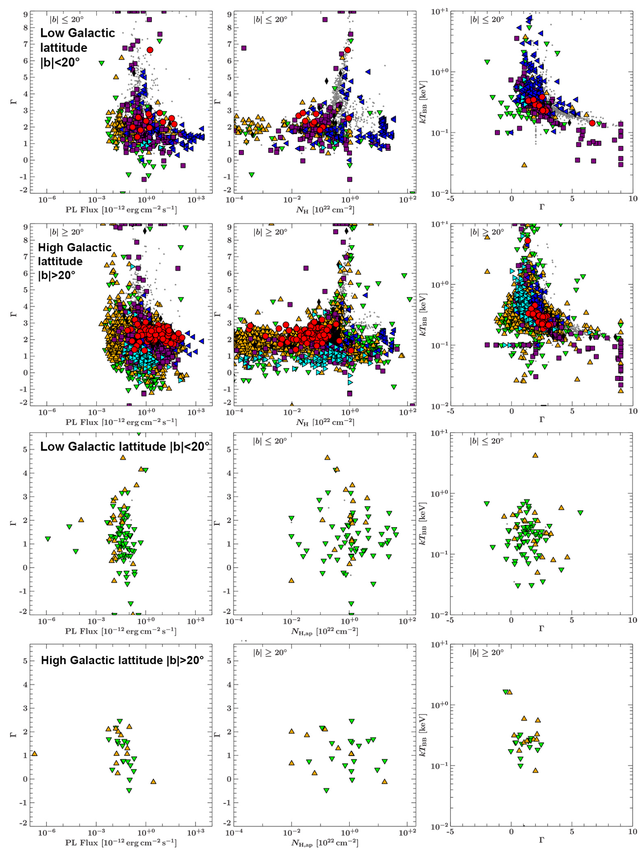

Fig 4: Correlation of various spectral parameters for the sources of the training sample (*upper panels*) and transient sources (*lower panels*).

Contact

For more information, visit `www.extras-fp7.eu`
This research makes use of data obtained by *XMM-Newton*, an ESA science mission funded by ESA Member States and the USA (NASA). EXTraS is funded within the EU/FP7-Cooperation Space framework and is carried out by a collaboration including INAF (Italy), IUSS (Italy), CNR/IMATI (Italy), University of Leicester (UK), MPE (Germany) and ECAP (Germany).