Comparison of classification methods for XMM sources

F-X. Pineau, S. Derriere, L. Michel and C. Motch

Observatoire Astronomique de Strasbourg, France

Abstract : The statistical identification of all serendipitous X-ray sources detected by the EPIC camera is one of the tasks devoted to the Survey Science Centre (SSC) of XMM-Newton. Using a probabilistic cross-correlation of the 2XMMi catalogue with others like the SDSS DR7 or the 2MASS, we have built several samples of multiwavelength data for which various thresholds on the number of spurious associations can be applied. We create a learning sample of classified XMM sources from the SDSS spectroscopy and from the Archival Catalogue and Database Subsystem (ACDS) which is the part of the SSC pipeline that performs the cross-correlation of EPIC sources against a large collection of archival data including Simbad. This allowed us to apply both supervised or unsupervised classification methods. We tested a range of classification algorithms : k-Nearest Neighbours, Mean Shifts, Kernel Density Classification, Learning Vector Quantisation, oblique decision tree (the OC1 algorithm) and Support Vector Machines. Advantages and disavantages of each method are briefly reviewed, and their respective performances are compared. We also show an example of the application of the kernel density classification with several classes on the results of the correlation of the 2XMMi with the SDSS DR7.

Introduction

The Incremental Second XMM-Newton Serendipitous Source Catalogue (2XMMi) is the largest catalogue of X-ray sources ever published so far. It has been compiled by the XMM-Netwon Survey Science Centre (SSC) on behalf of ESA. One of the responsibilities of the SSC is to provide the community with statistical identifications of all 2XMMi sources using a multi-wavelength analysis.
We presented two years ago (Pineau et al. 2008) an original tool that we used to cross-correlate the 2XMMi source list with various other catalogues. Last year, we showed that the correlation of the 2XMMi catalogue with the SDSS DR7 (DR7) and 2MASS catalogues can be used to distinguish and classify different classes of objects.
We now test and compare various classification algorithms on two samples of multiwavelength data.

Samples

We have built two samples from a Bayesian cross-correlation of the 2XMMi with the SDSS DR7 and the 2MASS catalogue and from a cross-correlation of the 2MASS with the GSC2.2.1 :

- XS : contains all 2XMMi/DR7 associations having an individual probability of identification > 0.9. All optical sources have their magnitudes $<$ 22.2, are extended, not blended, not saturated.
- XGT : contains all 2XMMi/2MASS/GSC2.2.1 associations for which the 2XMMi/2MASS association has a probability of identification > 0.6. The GSC source must lie within a radius of 5" and 1" from the XMM and the 2MASS source respectively and must have both the r and b magnitudes defined. We only keep here sources having a galactic latitude $> 10°$.

The associated learning samples (LS) have been built from the Tycho 2 catalogue, the DR7 spectral identifications, Simbad, the Véron and other ACDS catalogues. They only contains two classes : Star and Extragalactic source (Galaxies, AGNs and QSOs). Table 1 provides the number of sources of each class in the two samples.

Table. 1: Distribution of object classes in our two samples.

Sample\Class	Unknown	Star	ExtraGal	TOTAL
XS	7 686	532	1 587	9 805
XGT	7 117	501	1 311	8 929

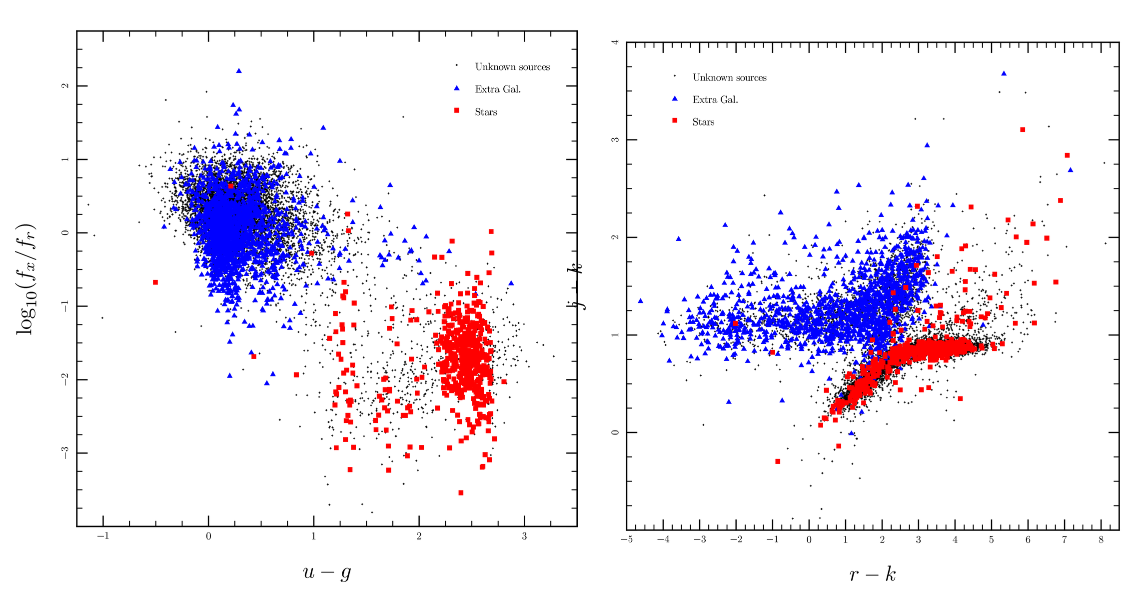

Fig. 1: Samples XS (left) et XGT (right) and their respective LS (coloured dots) in the 2D diagrams where the two classes are best separated.

For each of the samples XS and XGT, we construct two views with a different number of input parameters (see Tab. 2).

Table. 2: Parameters of the 4 samples based on XS and XGT.

Sample	Parameters
XS_3D	Fx/Fr, $u-g$, $g-r$
XS_9D	Fx/Fr, hr1, hr2, hr3, hr4, $u-g$, $g-r$, $r-i$, $i-z$
XGT_3D	$b-r$, $j-r$, $r-k$
XGT_5D	Fx/Fr, Fx/Fh, $b-r$, $j-r$, $r-k$

The 3 parameters of the 3D XGT sample have been selected according to the results of a principal component analysis.

Tested Algorithms

We have tested 6 different algorithms among which 5 are supervised and 1 is unsupervised :

- k-NN : the simplest algorithm, the attributed class is the most frequent one among the k nearest neighbours in the learning sample.
- Meanshift (MS) : the unsupervised algorithm, it is a gradient ascent on local densities estimated by kernel smoothing (KS).
- Kernel Density Classification (KDC) : is a nonparametric Baysian classifier (Richards et al. 2004) using kernel smoothing.
- Learning Vector Quantisation (LVQ) : LVQ is a supervised Kohonen neural network. We test here the OLVQ1 algorithm from the LVQ_PAK package written in C (Kohonen et al. 1996).
- OC1 : is an oblique decision tree (Murthy et al. 1994). We test here the C implementation of the authors.
- Support Vector Machine (SVM) : use the kernel trick to project input data in a feature space. Try then to maximise in the new space the margin between two parallel planes which separate the data in two classes. We test here the SVM_Light implementation in C.

The k-NN, MS and KDC algorithms have been tested using a code we have implemented in JAVA. The three algorithms are based on a kd-tree package also written by us.

Measured Parameters

We use some of the parameters defined in Gao et al. (2008), which are based on the confusion matrix (Tab. 3).

Table. 3: Confusion matrix.

Defined\Predicted	Predicted Star	Predicted ExtG
Defined Star in the LS	TS (true Star)	FS (false Star)
Defined ExtG in the LS	FE (false ExtG)	TE (true ExtG)

$$\begin{aligned}
\mathrm{Accuracy(Acc.)} &= \frac{\mathrm{TS}+\mathrm{TE}}{\mathrm{TS}+\mathrm{FS}+\mathrm{TE}+\mathrm{FE}}\\
\mathrm{TrueStarRate(AccS.)} &= \mathrm{TS}/(\mathrm{TS}+\mathrm{FS}) = \mathrm{Recall}\\
\mathrm{TrueQSORate(AccE.)} &= \mathrm{TE}/(\mathrm{TE}+\mathrm{FE})\\
\mathrm{WeightedAccuracy(WA)} &= \beta \mathrm{xAccS.} + (1-\beta)\mathrm{xAccE.}(\mathrm{Here}\beta = 0.5)
\end{aligned}$$

We applied the train-test method : 1/3 of the LS sources are randomly removed (so we train the classifiers with the remaining 2/3 of LS sources) and classified to compute the confusion matrix.

Results

Sample XS

Table. 4: Results on the XS_3D and XS_9D samples.

Algo\Sample		XS_3D				XS_9D			
Algo.	Options	Acc.	Acc.S	Acc.E	WA	Acc.	Acc.S	Acc.E	WA
KNN	k=7	99.57	98.87	99.81	99.34	99.64	99.15	99.81	99.48
MS	fb 0.50[a]	99.43	98.02	99.90	98.96	99.93	99.72	100.00	99.86
KDC	fb 0.50	99.93	99.72	100.00	99.86	100.00	100.00	100.00	100.00
LVQ	nn 500[b]	99.71	99.15	99.90	99.53	99.79	99.15	100.00	99.58
OC1	default	99.50	98.87	99.71	99.29	99.57	99.15	99.71	99.43
SVM	1000.0 1 2[c]	97.29	89.24	100.00	94.62	100.00	100.00	100.00	100.00

Sample XGT

Table. 5: Results on the XGT_3D and XGT_5D samples.

Algo\Sample		XGT_3D				XGT_5D			
Algo.	Options	Acc.	Acc.S	Acc.E	WA	Acc.	Acc.S	Acc.E	WA
KNN	k=7	97.50	95.78	98.15	96.97	98.08	96.39	98.73	97.56
MS	fb 0.50	93.49	96.39	92.39	94.39	98.17	98.49	98.04	98.27
KDC	fb 0.50	98.17	96.69	98.73	97.71	99.42	99.40	99.42	99.41
LVQ	nn 500	98.08	97.59	98.27	97.93	98.50	98.19	98.62	98.40
OC1	default	97.50	95.48	98.27	96.88	97.83	97.59	97.92	97.76
SVM	1000.0 2 5	97.58	91.57	99.88	95.73	100.00	100.00	100.00	100.00

Execution Time

Table. 6: Approximative execution time in second.

Algo.	KNN	MS	KDC	LVQ	OC1	SVM
Exec. time (s)	1	250	2	<1	50	6896

Remarks

All algorithms have a similar accuracy which depends on the input options and on the value of the input parameters. The SVM performs well here only if we set a tradeoff between the errors and the size of the margin that neglect the last one. The quality of the classification depends more on the learning sample than on the classification method. Having for each source the probabilities associated with each class can be useful to put a cutoff to extract the more reliable classified sources. To summarise :

- the fastest : LVQ and KNN (OC1 also, once the training is done) ;
- the one providing probabilities : KDC (others can, but less naturally)
- the unsupervised : MS
- the slowest : OC1 (in training phase), MS, SVM
- the ones best adapted for binary classification only : k-NN, SVM

Classifications made on the 3 first principal component of a PCA give better results than on the 3 physical parameters for the XGT sample. This with all algorithms. It is not the case for the XS sample.

a : fb = fixed bandwidth, followed by the value of the bandwidth.
b : nn = number of neurons, followed by this number.
c : value of the λ parameter, followed by two digits coding the kernel used.

Example of KDC on several classes

We tested the KDC with 5 classes on the results of the cross-correlation of the 2XMMi with the SDSS DR7 with both unresolved and extended objects. The LS was made of 536 stars, 26 CV and X-ray binaries, 673 galaxies, 572 AGNs and 1 425 QSOs. The results (Fig. 2) lead to 1 187 stars, 35 CV, 1 537 galaxies, 3 428 AGNs and 8 291 QSO.

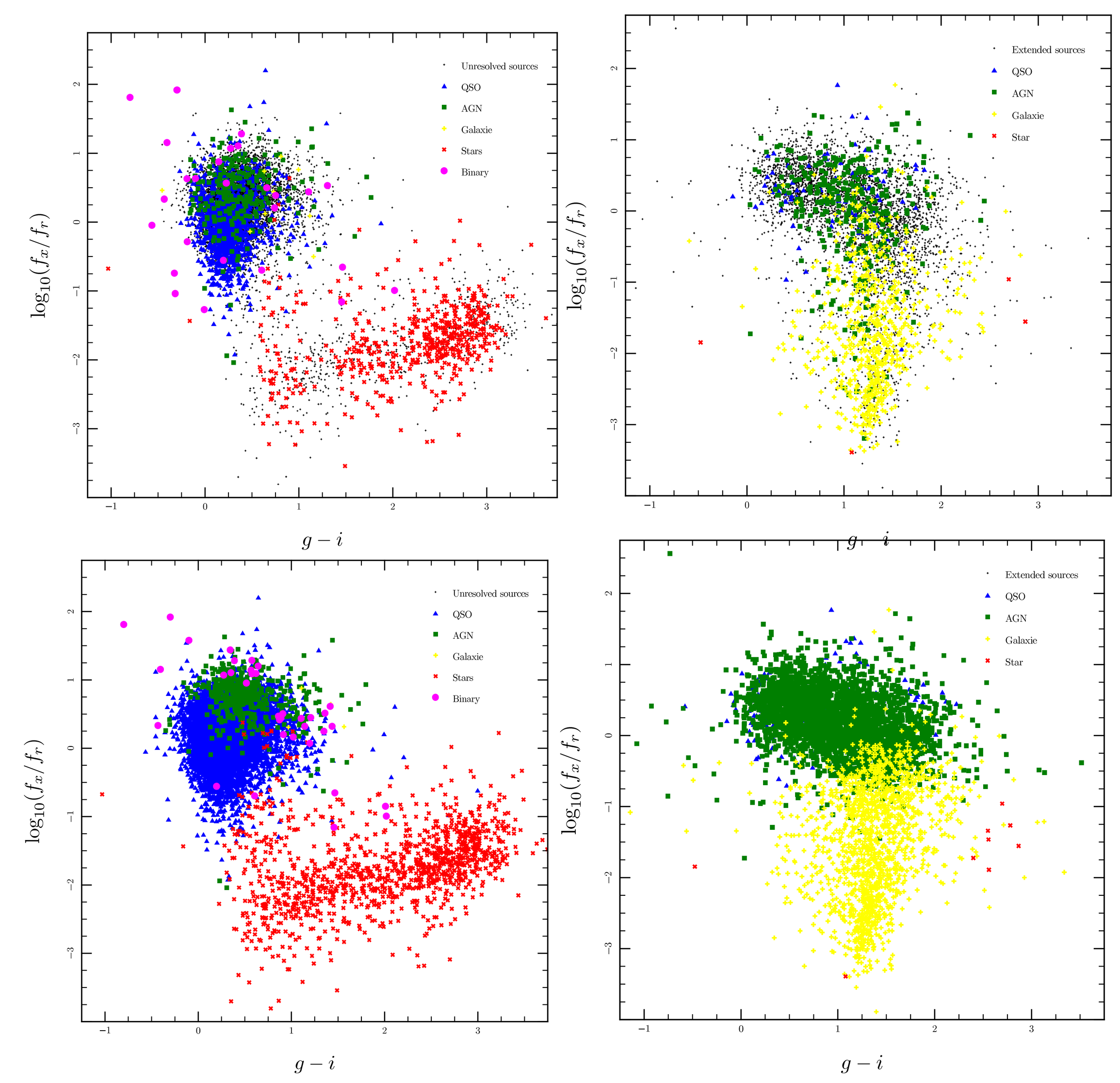

Fig. 2: Results of the KDC on a 2XMMi/DR7 sample with both unresolved (left) on extended (right) objects shown in the Fx/Fr vs $g-i$ diagram.

We provide Tab. 8 the confusion matrix.

Table. 7: Results on the XGT_5D sample.

Define\Assign	Star	XRB/CV	Galaxie	AGN	QSO
Star	535	0	0	0	1
XRB/CV	1	13	0	0	12
Galaxie	0	0	620	45	8
AGN	0	0	75	312	185
QSO	0	0	7	74	1344

99.81% of the stars are well classified and 100% of extragalactic sources. The confusion between groups of extragalactic sources is expected since there is no clear separation between galaxies, AGNs and QSOs, the difference depending mainly on the power of the central engine.

Conclusion

- Classification accuracy depends more on the quality of the LS – and (of course) on the separability of the data – than on the algorithm ;
- The best algorithm depends on the addressed issue : need to be quick to train and to classify (LVQ), have time to train but not to classify (OC1, SVM), LS not well defined (MS), huge input parameter space but not a large number of input sources (SVM), well defined LS and need for probabilities to select most secured cases (KDC), ...
- We plan to use KDC to classify XMM sources :
  - to naturally handle more than 2 classes ;
  - to estimate the reliability of the classification of each individual source, thanks to the probablities it provides.
- KDC probability could be used, together with the probability provided by the Baysian cross-correlation on position, to define a final probability of identification for each XMM-archival source association.

References

Gao, D., Zhang, Y.-X., & Zhao, Y.-H. 2008, MNRAS, 386, 1417

Kohonen, T., Hynninen, J., Kangas, J., Laaksonen, J., & Torkkola, K. 1996, LVQ PAK: The Learning Vector Quantization Program Package, Tech. rep., Helsinki University of Technology, Laboratory of Computer

Murthy, S. K., Kasif, S., & Salzberg, S. 1994, ArXiv Comp. Sc.

Pineau, F.-X., Derriere, S., Michel, L., & Motch, C. 2008, in ASPCS, Vol. 394, ADASS XVII, 369

Richards, G. T., Nichol, R. C., Gray, A. G., et al. 2004, ApJS, 155, 257