Comparation 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 Science Survey 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 multilayered data for which various thresholds on the number of spurious associations can be applied. We create a learning sample of learned 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: Nearest Neighbours, Mean Shift, Kernel Density Classification, Learning Vector Quantization, d-dimensional decision tree (the OCl algorithm) and Support Vector Machines. Advantages and disadvantages of each method are briefly reviewed, and their respective performances are compared. We also show an example of the applied learning vector density classifier 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-Newton Science Survey 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 multilayered analysis. We presented two years ago (Pineau et al. 2005) 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 multilayered 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 2XMMi with the GSC2.1:

- XS: contains all 2XMMi/2MASS associations having an individual probability of identification > 0.9. All optical sources have their magnitudes < 23.2, are extended, not blended, not saturated.
- 2XMM/XS: contains all 2XMMi/2MASS/GSC2.1 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 r and k magnitudes defined. We only keep here sources having a galactic latitude > 10°.

The associated learning sources (LS) have been built from the Tycho 2 catalogue, the DE200, the SIMBAD, the Vron and other AGCS catalogues. They only contain two classes: Star and Extragalactic source. 3D and XGT indices are separately used to cross-correlate the XMM catalogue with various other catalogues.

The 2XMMi/2MASS sources are extracted from the Simbad database, used to cross-correlate the 2XMMi source list with various other catalogues. They are only kept here sources having a galactic latitude > 10°. The associated learning sources (LS) have been built from the Tycho 2 catalogue, the DE200, the SIMBAD, the Vron and other AGCS catalogues. They only contain two classes: Star and Extragalactic source. 3D and XGT indices are separately used to cross-correlate the XMM catalogue with various other catalogues.

We tested the KDC with 5 classes on the results of the cross-correlation of the 2XMMi with the SDSS DR7 with both unsupervised and extended object. We tested the KDC with 5 classes on the results of the cross-correlation of the 2XMMi with the SDSS DR7 with both unsupervised and extended object. 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.

Table 1: Distributions of objects classes in our two samples.

<table>
<thead>
<tr>
<th>Sample/Class</th>
<th>XS</th>
<th>GSC</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star</td>
<td>7 347</td>
<td>1 68 7</td>
<td>9 909</td>
</tr>
<tr>
<td>ExtraGals</td>
<td>7 437</td>
<td>1 68 7</td>
<td>9 909</td>
</tr>
</tbody>
</table>

Table 5: Results on the XGT and XS samples.

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 (L VQ), have time to train and classify (SVM), work for 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 probabilities it provides.
- KDC probability could be used, together with the probability provided by the Bayesian cross-correlation on position, to define a final probability of identification for each XMM-archival source association.

References

Murphy, S., Kast, S., & Sallberg, M. 2009, ArXiv Comp. Sc