Finding rare objects and building pure samples: Probabilistic quasar classification from low resolution Gaia spectra

Coryn A.L. Bailer-Jones, Kester Smith

Gaia, the upcoming mission of the European Space Agency, will perform an all-sky astrometric and spectrophotometric survey complete to $G = 20$, expecting to observe some 10^9 stars, a few million galaxies and half a million quasars (e.g. [3]). Its primary mission is to study Galactic structure by measuring the 3D spatial distribution and 2D kinematic distribution of stars throughout the Galaxy and correlating these with stellar properties (abundances, ages etc.) derived from the spectra. With astrometric accuracies as good as $10\,\mu$as, Gaia cannot be externally calibrated with an existing catalogue. Instead, it must observe a large number of quasars over the whole sky with which to define its own reference frame. This quasar sample must be very clean (low contamination) and is itself of intrinsic interest.

Object classification and estimation of astrophysical parameters is an integral part of the overall Gaia data processing [1] and comprises one of the Coordination Units in the Gaia Data Processing and Analysis Consortium (DPAC) [4] [5].

The Discrete Source Classifier (DSC) is the data processing module responsible for classifying all the objects which Gaia detects. As the name suggests, it assigns objects to discrete classes, e.g. star, galaxy, quasar, binary star, in each case it assigning a class probability. Classification is based primarily on the low resolution BP/RP spectra, because (initially at least) there is no morphological information from Gaia. The subsequent stages in the CU8 data processing are concerned with extracting physical parameters for these classes (e.g. stellar temperatures) and classifying the RVS spectra. DSC is based on machine learning methods for pattern recognition, currently a so-called ”Support Vector Machine”.

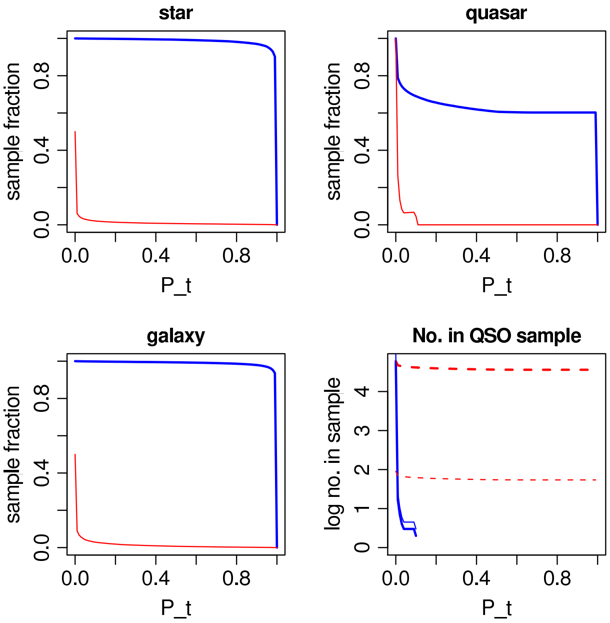

Figure 1: Completeness (blue line) and confusion (red line) of samples of star, galaxies and quasars as function of the classifier threshold

One of the challenges of Gaia is to reliably classify of rare objects, e.g. the expected half million quasars among one thousand million stars. Standard methods for machine learning will often fail to identify them. To address this, the DSC team has developed a method for modifying the output probabilities to accommodate rarity, and applied this in classification experiments on simulated data [2]. Figure 1 shows, for three classes of objects, the completeness (blue

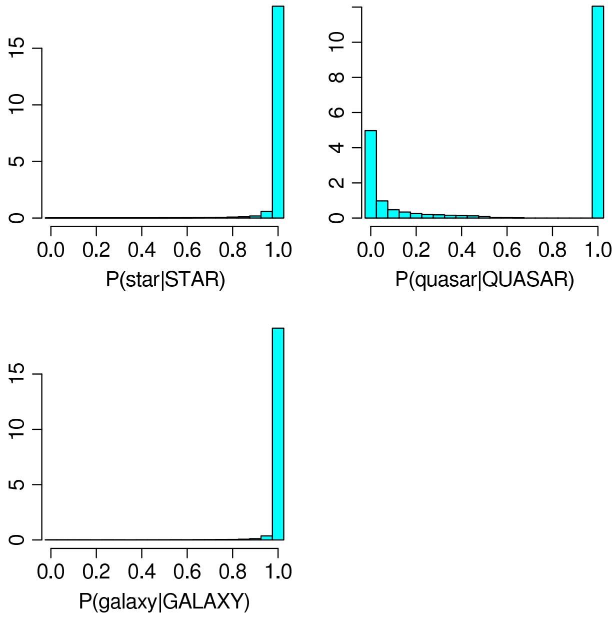

Figure 2: Histogram of probabilites for correctly classified quasars

line) and contamination (red line) of a sample of objects as a function of adjustable probabilty thresholds used to build the sample. We see that we can achieve a zero contamination sample of quasars which still has a completeness of 65%, more than sufficient for Gaia. The corresponding probability outputs from the DSC are shown in Figure 2. With this method we can control the class priors, which allows a single classification model to be applied to any target population without having to tune the training data and retrain the model.

Based on simulations of the universe which Gaia will observe, we find that we are able to achieve a pure sample of quasars (upper limit on contamination of 1 in 40 000) with a completeness of 65% at magnitudes of G=18.5, and 50% at G=20.0, even when quasars have a frequency of only 1 in every 2000 objects. The star sample completeness is simultaneously 99% with a contamination of 0.7%.

For more information see http://www.mpia.de/Gaia

References

[1] Bailer-Jones C.A.L., in *The Three-dimensional universe with Gaia*, C. Turon, K.S. O'Flaherty, M.A.C. Perryman (eds), ESA, SP-576, 393

[2] Bailer-Jones C.A.L., Smith K.S., Tiede C., Sordo R., Vallenari A., 2008, MNRAS 391, 1838

[3] Lindegren L., Babusiaux C., Bailer-Jones C.A.L., Bastian U., Brown A.G.A., Cropper M., Hog E., Jordi C., et al., 2008, Proc. IAU Symposia, vol. 248, 217

[4] Mignard F., Drimmel R. (eds), 2007, *DPAC Proposal for the Gaia Data Processing*, Gaia DPAC Technical note, GAIA-CD-SP-DPAC-FM-030. (Multiple authors including C.A.L. Bailer-Jones)

[5] O'Mullane W., Lammers U., Bailer-Jones C.A.L., et al., in *Astronomical Data Analysis Software and Systems 16*, R.A. Shaw, F. Hill and D.J. Bell (eds), ASP Conf. Ser. 376, 99