Automated classification of Gaia sources

Kester Smith, Carola Tiede, Coryn Bailer-Jones

Max-Planck Institute for Astronomy
Koenigstuhl 17
69117 Heidelberg, Germany

Email: smith@mpia-hd.mpg.de

Introduction

Gaia is the next generation astrometric mission from ESA, due to launch in 2011. It will survey the entire sky down to a magnitude of approximately V=20, detecting about one billion stars, or about 1% of the Galactic stellar population. Positions, proper motions and parallaxes will be determined with unprecedented accuracy, and radial velocities will also be determined for the brightest objects, making a genuine three dimensional survey of the local environment. The mission will also obtain astrophysical information on the properties of the stellar sample, providing an insight into the formation and subsequent dynamical and chemical evolution of the Milky Way.
Together with the stellar sample, several million galaxies, perhaps half a million quasars and many solar system objects will also be detected. No input catalogue will be used, so automated classification of detected sources is a key part of the data processing process. Here, we describe progress on the discrete source classifier (DSC) which is being developed at MPIA.

The Gaia mission

The astrometric accuracy will be 12-25μas at V=15 and 100-300μas at V=20 (see Figure 1 for historical context). The main astrometric instrument will be complemented by two other onboard instruments. The radial velocity spectrometer (RVS) will measure radial velocities with a precision of 10-15 km s^{-1} or better for the brightest 150-200 million stars. A low dispersion integral field spectrophotometer covers the wavelength range 330–1000nm. This instrument in fact provides two overlapping spectra, one at the blue end and one at the red. These are referred to as BP and RP spectra. Each is expected to comprise about 48 bins.

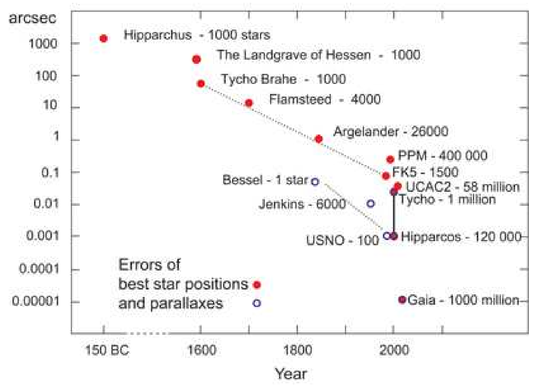


FIGURE 1: Astrometric performance of Gaia compared to previous measurements

Discrete source classifier

The discrete source classifier (DSC) takes as its main inputs the BP and RP spectra, together with the parallax and proper motion measurements. The position is not used for classification.
Algorithm development is currently being carried out with simulated data. The underlying spectra are drawn from libraries (the stellar data is drawn from the Basel libraries) and passed through the instrumental response function to produce the observed spectrum. Noise is then added appropriate to the required magnitude. At present, all classification takes place at an assumed Gaia magnitude of G=15 (corresponding approximately to V=17). Parallaxes and proper motions are assigned for the stars based on a model of a homogeneous spherical distribution, whilst for extragalactic objects the parallax and proper motion is generated purely from the anticipated astrometric error function.
At the current stage of development, four classes of astrophysical object are considered: single stars, physical binaries, point-like galaxies and quasars. Other classes, such as asteroids and non-physical binaries (i.e. chance pairs) have been temporarily excluded because the astrometric quantities are not easily defined for them. To test the classification algorithms, samples of sources are selected from the simulated data grids. Two data samples are used, each containing two thousand sources of each astrophysical class, making eight thousand sources in total. One data set is used as a training or reference set, and the other test set is then classified. Since the astrophysical classes of the test set are in fact known, the statistical performance of the classifier can be assessed.

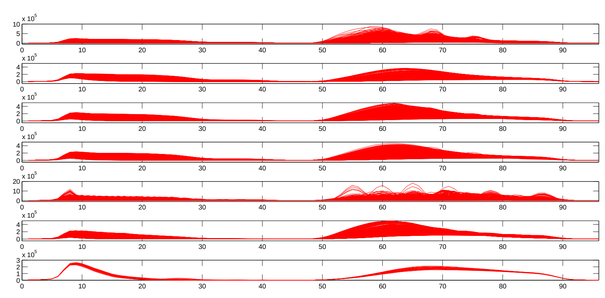


FIGURE 2: BP and RP spectra for several astrophysical classes. From top: single stars, white dwarfs, physical binaries, non-physical binaries (i.e. chance superpositions), quasars, galaxies, and asteroids

Algorithms

Several algorithms have come under consideration so far. Some of these are discussed briefly below.

k Nearest Neighbour

With k=1, the Nearest Neighbour method classifies a source to be the same class as the nearest object in the training set. 'Nearest' is naturally interpreted to mean a Euclidean distance in data space if we consider only the 96 bins of the BPRP spectrum, and this method is the one that generated the results presented here. A more sophisticated (and easy to implement) method would be to weight the input data according to the relative measurement errors. When other measurements, such as parallax, are to be introduced, the choice of error function requires more thought. For $k > 1$, the output consists of a probability estimate for each class based on the k nearest points in the training grid.
The main advantage of nearest neighbour techniques is that they are relatively easy to implement. Their drawbacks are that they tend to be computationally expensive and it is difficult to achieve a representative density of training sources in the 96 dimensional data space.
Table 1 shows the confusion matrix for the nearest neighbour method with $k = 1$ and excluding parallax and proper motion. Rows correspond to the true class of the test objects, and columns show the classification results as a percentage of the total input sources of that class. The leading diagonal indicates sources that are correctly classified, off diagonal elements show misclassification rates. The rows each add to 100% of input sources.

Stars	Binaries	Quasars	Galaxies	
76.10	23.40	0.40	0.10	Stars
18.28	81.62	0.10	0.00	Binaries
2.65	0.60	94.34	2.40	Quasars
0.00	0.00	0.05	99.95	Galaxies
Accuracy: 88.0%				

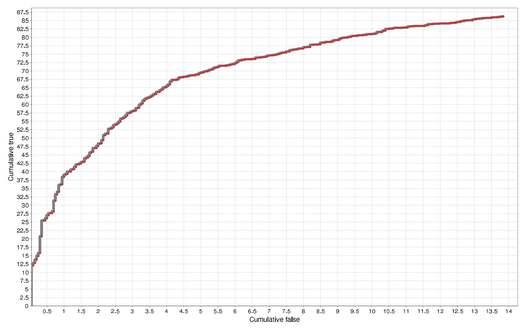
TABLE 1: Confusion matrix for kNN method, k=1 and parallax and proper motion are not included.

Support Vector Machine (SVM)

Support vector machines classify the data by projecting the input space onto a higher dimensional space and then finding optimal linear discriminants between the classes in this higher dimensional space. The solution for the discriminant can be found without prohibitive computational effort by introducing a kernel. We used an implementation called libSVM available online at `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`. A coherent explanation of the workings of SVM's can be found in Bennett & Campbell (2000). The basic reference is Vapnik (1995).
Unlike kNN, the SVM can deal relatively naturally with the introduction of parallax and proper motion. Table 2 shows results for classification with an SVM with the astrometric information included. Figure 3 shows cumulative true positive classification rate against cumulative false negative for the stars and the quasars, with the sources arranged in order of 'most secure' classification (at the left hand side) to most doubtful.

Stars	Binaries	Quasars	Galaxies	
87.55	12.00	0.45	0.00	Stars
8.19	91.76	0.05	0.00	Binaries
0.35	0.00	97.75	1.90	Quasars
0.00	0.00	0.00	100.0	Galaxies
Accuracy: 94.24%				

TABLE 2: Confusion matrix for SVM method including astrometric information.



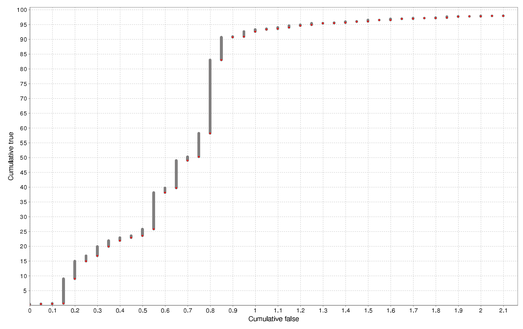


FIGURE 3: (Top) The rate (in percent) of cumulative correct classifications for the input stars (y-axis) versus the rate of false negative classifications, i.e. stars classified as some other class. (Bottom) The same plot for the Quasars.

Neural Network

The results of a neural network method are shown in Table 3. The neural network was implemented within the R statistics package.

Stars	Binaries	Quasars	Galaxies	
90.35	9.65	0.05	0.00	Stars
7.54	92.36	0.10	0.00	Binaries
0.00	0.00	100.0	0.00	Quasars
0.00	0.00	0.15	99.85	Galaxies
Accuracy: 95.65%				

TABLE 3: Confusion matrix for Neural Network method including astrometric information.

Discussion

For all the methods discussed above, the greatest confusion occurs between the single stars and binaries, which is perhaps to be expected.
The quasars class is of particular importance for the astrometric solution, since they form a population of zero parallax objects that could be used as a natural reference frame. It is therefore desirable to obtain clean samples of quasars, uncontaminated by galactic objects.
So far much of the DSC effort has focused on support vector machines, since this method is robust and produces good results. The nearest neighbour method has difficulties either with accuracy or with computational efficiency in the multidimensional data space. The early results from the neural network are promising and may be developed further. In practice, the best approach may be to apply different methods in different regimes.

References

[1] Bennett, K.P. & Campbell, C., "Support Vector Machines: Hype or Hallelujah?" *SIGKDD Explorations,* **Vol. 2**(issue 2), (December 2000): p. 1

[2] Vapnik, V., *The nature of statistical learning theory* (New York, Springer Verlag, 1995)

Conference Milky Way Halo, May 2007, AIfA Bonn, poster number 37