Proc. 28th Ann. Conf. GfKl2004, submitted

Astronomical object classification and parameter estimation with the Gaia Galactic survey satellite

C.A.L. Bailer-Jones

Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany

Abstract. Gaia is a cornerstone mission of the European Space Agency (ESA) which will undertake a detailed survey of over 10^9 stars in our Galaxy. This will generate an extensive, multivariate, heterogeneous data set which presents numerous problems in classification, regression and time series analysis. I give a brief overview of the characteristics and requirements of this project and the challenges it provides.

1 The Gaia Galactic survey mission

Gaia is a future satellite mission which will study our Galaxy in unprecedented detail (Perryman et al. (2001)). Its objective is to study its composition, origin and ultimate evolution by determining the properties of over one thousand million stars in different populations across our entire Galaxy. One of the major contributions of Gaia is that it will measure distances to stars with much higher precision than is currently possible. Distance measurement is a very important (and difficult) task in astrophysics, as only with distances can we properly map structure in the Galaxy and determine fundamental stellar properties (e.g. absolute brightness). Gaia will also measure the space motions of stars in exquisite detail, which will be used in sophisticated dynamical models to map out the distribution of matter and is an important component in testing models of Galaxy formation.

2 Astrophysical data

Much of this so-called astrometric data from Gaia would be of little value if we did not know the intrinsic properties of the stars observed, quantities such as the temperature, mass, chemical compositions, radius etc. (collectively referred to as *Astrophysical Parameters*, or APs; see Bailer-Jones (2002b)). For this reason, Gaia is equipped with two photometric instruments which sample the stellar spectral energy distributions (or spectra) at discrete locations, producing *photometric* data. The first of these instruments measures the spectra at five locations, the second at about ten. (The optimization of these filter systems is ongoing; for more details on this and the sampling of stellar spectra, see my other contribution in these proceedings.) Together these data provide

a 15-dimensional data space from which we need to determine at least four APs for a very wide range of types of stars. For many of these stars we have (or can gather or simulate) reasonable quality pre-classified data which may be used as templates in a supervised classification/regression model, such as neural networks or minimum distance methods (Bailer-Jones et al. (1998); Bailer-Jones (2002a)).

The problem is, however, considerably more complex. First, each star is observed about 100 times over the course of the mission, and many of these stars are variable, i.e. their photometric measures vary over time on a range of time scales. This is both a problem and a benefit, because for some objects the way in which they vary is a significant source of information for determining their intrinsic properties (both the primary APs and additional characteristics). Second, not all of the objects which Gaia observes are stars. Gaia observes 'blind', that is, it observes every single object in the sky brighter than some level without any prior selection or information on what the objects are. Many of these objects will be single stars, but many other types of objects will be observed, including galaxies, quasars, asteroids and unresolved binary stars. Therefore, before we can even try to determine APs, we must perform a discrete classification to see whether the object is a type of star we are interested in. In some cases we can use morphological information, i.e. we get an image which is not simply a point source (typically for some – but not all – galaxies). Using this is of course an involved image classification problem in its own right (Naim et al. (1995)). In many cases we have no such morphological information, so we must perform the classification using the photometric data.

The classification problem is complicated further by the presence of a third instrument which will measure the entire stellar spectrum of each star over a narrow wavelength region. The spectrum covers some 500 elements. While we can certainly apply dimension reduction techniques to these data, it nonetheless provides considerably more independent information on the primary APs. Moreover there are several additional astrophysical parameters which we want to determine from these spectra. A particular challenge is combining these data with the two sets of photometric data.

3 Classification challenges

Gaia will produce a complex data set, the proper exploitation of which presents us with a number of significant challenges. The objectives can be summarised as follows:

- Discrete classification of objects: discriminate between single stars, multiple stars, galaxies, quasars, supernovae, asteroids etc.
- For single stars, determine their astrophysical parameters (APs), the exact number of which and the precision with which they should be established depending on their type. There are four primary APs and several

subsidiary ones. We will probably also want to perform a (discrete) classification of these stars into astrophysically relevant groups.
- Provide for the efficient identification of new types of objects for which we have no (or little) prior knowledge, i.e. employ unsupervised methods or outlier detection techniques.

The practical requirements of the classification system may be summarized as follows:

- Cope with missing data (e.g. due to partial instrument failure or 'down time') and deal with censored data (i.e. upper or lower limits on a measure due to the limited sensitivity or dynamic range of an instrument) in an unbiased fashion. In some cases an upper limit on a non-detection is an important indication of the type of star.
- Quantify uncertainties via probabilities of class membership; this must take account of the fact that some stars may be members of more than one class.
- Provide accurate estimates of AP uncertainties. The input data and the resulting APs will sometimes have correlated errors. Typically we have a good noise model for our instruments although the correlations between them are harder to characterize.
- The APs are not independent and they do not have an isolated effect on the data. For this reason, we cannot independently infer each AP in a multivariate regression.
- Cope with degeneracies. A degeneracy means that different objects can appear the same in the data space (within the expected measurement errors), especially at low signal-to-noise ratios. Some degeneracies are intrinsic and known to exist but have not been mapped out in detail. Degeneracies must be recognised and different classifications/sets of APs provided where appropriate (along with associated probabilities).
- Make efficient use of variability (time series) information. The classification systems should be insensitive to variability where it is not relevant (e.g. due to noise or errors) but recognise and exploit it where it is relevant (for certain types of stars).
- In some cases we have prior information on the APs of specific objects; making efficient use of this is a challenge.

Clearly, classification and parameter estimation with Gaia cannot be solved with a simple one-step approach. It will probably have to employ many different techniques operating in a hierarchical or iterative fashion.

4 Outlook

Gaia will be launched in 2010 at a cost of some 450 million Euro. The data analysis – including the classification – will be undertaken by a dedicated

but geographically distributed consortium of astronomers, computer scientists and statisticians. For more information on the mission, see the Gaia web site at `http://www.esa.int/science/gaia`. The classification issues are being addressed by a dedicated working group, ICAP, which stands for *Identification, Classification and Astrophysical Parametrization.* Its web site, which gives more details on the problem, data and techniques currently being used, is `http://www.mpia.de/GAIA`

References

BAILER-JONES, C.A.L. ET AL. (1998): Automated classification of stellar spectra. II: Two-dimensional classification with neural networks and principal components analysis. *Monthly Notices of the Royal Astronomical Society*, 298, 361–377

BAILER-JONES, C.A.L. (2002a): Automated stellar classification for large surveys: a review of methods and results. In: R. Gupta, H.P. Singh, C.A.L. Bailer-Jones (Eds.): *Automated Data Analysis in Astronomy.* Narosa Publishing House, New Delhi, 83–98.

BAILER-JONES, C.A.L. (2002b): Determination of stellar parameters with GAIA. *Astrophysics and Space Science*, 280, 21–29

NAIM, A. ET AL. (1995): Automated morphological classification of APM galaxies by supervised artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, 275, 567–590.

PERRYMAN, M.A.C. ET AL. (2001): Gaia: Composition, formation and evolution of the Galaxy. *Astronomy & Astrophysics*, 369, 339–363.