

Data Mining in the Hipparcos Dataset – Summary Document

Dr Nik Whitehead

The Origins of the Data

The Hipparcos (High Precision PARallax Collecting System) was an astrometry mission undertaken by the European Space Agency [1]. It launched in 1989 and collected data until 1993, which was then published as the Hipparcos Catalogue in 1997. The aim of the mission was to identify the positions of nearly 120,000 stars in order to provide an accurate guide catalogue to be used by the then-planned Hubble Space Telescope.

The Dataset

The dataset defines five primary astrometric parameters of 118,218 stars [2]:

Limiting magnitude	$V \sim 12.4$ mag
Position precision ($H_p < 9$ mag)	9.77/9.64 mas RA/Dec
Parallax precision ($H_p < 9$ mag)	0.7 mas
Proper motion precision ($H_p < 9$ mag)	0.88/0.74 mas/yr RA/Dec
Photometric precision (H_p ; $H_p < 9$ mag)	0.0015 mag

It also identified 11597 variable stars (8237 new) and 13882 confirmed or suspected double or multiple star systems.

The complete dataset [3] includes up to 78 attributes for each star, although many of these are such things as flags denoting variability, multiplicity or coverage in other catalogues, or are statistical parameters and errors relating to the reduction of the raw data into the final published form. Only 22 attributes will be used for the analysis; these cover all of the main photometric and resolved astrometric parameters, together with system type and spectral classification data.

Suitability for Analysis

The size of the dataset means that it is sufficiently large for a wide variety of data mining techniques to be applied to it with a high probability of finding any hidden patterns within it. Some data cleansing will need to be performed in order to ensure that the records used for modelling are complete and are not outliers, but with a dataset this large it is expected that it will still be possible to produce large datasets for both modelling and testing processes. The patterns sought are based upon sound astrophysical principles and can therefore be checked against known existing results.

Potential Research Questions

The project aims to address the following three questions:

- Can data mining techniques be used to automatically classify stars within the Morgan-Keenan [4] classification scheme?
- Is it possible to reliably distinguish between main sequence and evolved stars using data mining techniques?
- Are there any relationships between the astrometric parameters and the photometric parameters that can be identified using data mining techniques?

Researcher Background

The author believes that she has sufficient contextual knowledge of the field to undertake this research as she has a B.Sc.(Hons.) in Astronomy & Astrophysics.

References

- [1] European Space Agency, "Hipparcos", Online source: <http://sci.esa.int/hipparcos/> (accessed 18/1/2017)
- [2] European Space Agency, "Summary of the Hipparcos and Tycho Catalogues", Online source: <http://www.cosmos.esa.int/web/hipparcos/catalogue-summary> (accessed 18/1/2017)
- [3] Strasbourg Astronomical Data Centre, "Access to Astronomical Catalogues", Online source: <http://cdsarc.u-strasbg.fr/viz-bin/Cat?l/239> (accessed 18/1/2017)
- [4] Morgan, William Wilson; Keenan, Philip Childs. "Spectral Classification". *Annual Review of Astronomy and Astrophysics*. vol 11, 29–50, 1973.