

Gaia spectra classification using self-organizing maps models

Pallas-Quintela, L.¹, Garabato, D.¹, Manteiga, M.², Dafonte, C.¹, Pérez-Couto, X.¹ and Álvarez, A.¹

¹ CIGUS CITIC - Department of Computer Science and Information Technologies, University of A Coruña, Spain

² CIGUS CITIC, Department of Nautical Sciences and Marine Engineering, University of A Coruña, Spain

Abstract

The 3rd Gaia Data Release (GDR3) published 219 million BP/RP spectra, and this number is expected to grow significantly for GDR4. Our research group, within the Outlier Analysis (OA) working package, classified at GDR3 56 million outlier BP/RP spectra using Self-Organizing Maps (SOM).

Going one step further, our goal is to classify all available sources based solely on BP/RP spectra through SOM Maps. However, this approach faces two major challenges. The former one is the algorithm's performance. If we want to efficiently train a model, it is needed a distributed environment that reduces both time and resources consumption, even if only DR3 sources are processed. The latter one is the classification resolution loss, as the satellite gathers sight more stars than galaxies or quasars, making the training dataset to be imbalanced. To tackle such issues, we developed a lightweight classification approach.

We have built a set of classification models based on SOM Maps, considering well-known and representative stars, galaxies, and quasars sources. Following this approach, the time needed at the training phase is exponentially reduced. Moreover, as we can compute the labels of the training dataset, it is possible to perform a finer-grain labelling of the neurons. To determine the class of each neuron, we count the spectral types of the sources within each one and assign the most representative class accordingly.

Once the models are trained, the classification process for each spectrum proceeds as follows: first, it is computed the neuron of each of the maps where it lays the best. Afterwards, it is determined in which map lays the best and if a such spectrum does not fit under a threshold in the best map, it will be marked as an outlier. In the latter case, such source will be part of the sources to be analysed by OA.

My poster in zenodo.org can be found here