A BAYESIAN CLASSIFICATION ALGORITHM FOR GAIA

S. Picaud[1,2,3], **A.C. Robin**[2], **U. Bastian**[1]

[1]Astronomishes Rechen Institut, Mönchhofstraße 12-14, D-69120 Heidelberg, Germany
[2]Observatoire de Besançon, BP 1615, F-25010 Besançon cedex, France
[3]Instituto de Astronomia, Geofísica e Ciências Atmosféricas,
Rua do Matão 1226, Cidade Universitaria, 05508-900 São Paulo SP, Brazil

ABSTRACT

We present a classification algorithm based on bayesian probabilities computed from the Besançon model of the Galaxy. The scheme is as follows: A simulation in the direction of the stars to be classified is performed from the Galaxy model. Then, assuming values and errors on the observables (e.g., G magnitude, colours, kinematics,...), probabilities of a star to belong to a given stellar population, and to have such luminosity class, spectral type, and other intrinsic parameters, are deduced, based on the probabilities of having such combination of observables in the simulation. This method can be used further to classify Gaia stars during the mission. It can also be used to identify stars having unexpected observables measured by Gaia instruments and to trigger a Science Alert.

Key words: Gaia; Classification; Science alerts.

1. INTRODUCTION

The Besançon model of the Galaxy (Robin et al. 2003) computes simulations of the expected stars in a given direction with observables properties (apparent magnitudes and colours, kinematics, parallaxes) directly comparable with observations, as well as intrinsic parameters of stars such as absolute magnitudes, spectral type, luminosity class, metallicity, etc. Using the simulations from the model, we can thus build a method allowing to deduce from the observable data of a star the most probable value for its intrinsic parameters. This method can be a useful tool for the Gaia mission, in particular for classification in case of low signal to noise in the Spectro instruments or for detecting science alerts.

The scheme is as follows: A simulation in the direction of the observations is performed from the Galaxy model. Then, for each observed star, assuming values and errors on its observables, one counts the densities of model stars belonging to its confidence interval. In the case where no simulated star is selected, a 'science alert' is triggered, and a study of observations and simulations is performed to settle whether the source has bad measurements or is an object with unexpected properties. The alert treatment is detailed in Section 2. When the simulated stellar density in the confidence interval is not zero, the statistical distribution of these selected simulated stars with respect to their intrinsic parameters are used to classify the observed stars, as explained in Section 3. But before applying the classification method on the observed stars, a calibration must be done to correct possible biases between the observations and the simulations. The calibration algorithm is given in Section 4. Finally, we discuss the question of the rapidity of the program in Section 5.

The observables we choose for building and testing the program are: apparent V magnitude, B–V and V–I colours and parallax. The chosen intrinsic parameters are: absolute magnitude in V, spectral type, luminosity class, log $T_{\rm eff}$, log g and stellar population (thin disc (in 7 sub-components of different ages in the Besançon model of the Galaxy), thick disc, spheroid, bulge). Fake data (i.e., simulations built from the Galaxy model and used in the program as observations to test it) were computed using error laws and limiting magnitude V $\leq$ 8.9 mag similar to Hipparcos. The chosen direction for all the tests was the northern Galactic pole $60^\circ \leq b \leq 90^\circ$, all longitudes), with about 6000 observed stars.

2. SCIENCE ALERTS

A number of sources with unexpected properties might be detected by Gaia: supernovæ, microlensings, rare events, unusual stars or some other objects. Our algorithm allows to extract such sources which are not accounted for in the model predictions. As all simulated stars belonging to the confidence interval of the observed star are counted, we emit an alert for every observation for which no simulated star is selected. In this case, for each observable, the 10 closest stars in the simulations and the 10 closest stars in the observations are extracted, to try to settle whether this difference between model and data is due to a bad measurement in one observable or whether the star has been well measured but has unexpected properties.

2.1. Tests

Tests are going on to check the correct level of alert, in order to get a reasonable number of such events in the case of the Gaia mission.

- First tests were made using fake data and simulations taken from the same model and no alert was emitted. As real alerts (bad measurements, unexpected properties) are avoided in this case, this means that the program does not emit false alerts.
- Seconds tests were run using fake data having some stars with artificial bad measurements (value at 10 sigmas of the good one). In this case, for a lot of these stars no alert was triggered, because some model stars had similar observables. Such badly measured stars but with still realistic observables cannot be detected by our program and will be badly classified.

3. CLASSIFICATION

3.1. Bayesian Method

When simulated stars are present in the confidence interval around the source in the observable space, their statistics are is used to perform the classification: distributions of the simulated stars with respect to their intrinsic parameters are computed taking into account the density provided by the simulations and the probability of belonging to the confidence interval. These density distributions are then considered as probability distributions of the values of intrinsic parameters for the observed star, and used to classify the observed star, i.e., to allocate to it a value for each of its intrinsic parameters.

3.2. Choice of Estimators

Various criteria have been chosen to deduce the best value of intrinsic parameter from the probability distributions, depending on the kind of parameter (discrete or continuous). The estimators, as well as one or two flags per parameter useful to estimate the quality of the classification, are the following:

- Continuous parameters (absolute magnitude in V, log $T_{\rm eff}$ and log g): the estimator of the parameter and the quality value are respectively the mean and the dispersion of the values weighted by their density of probability.
- Luminosity class: the estimator is the most probable class and the quality flag is its probability.
- Spectral type: the estimator is the most probable sub-type (KI,KII,...) of the most probable type (O,B,A,...), and the two quality flags are respectively the probability of the best type and the probability of its best sub-type.
- Stellar population (and age for thin disc stars): the estimator is the most probable population in the case where it is the spheroid, the thick disc or the bulge, and the most probable age component in the case where the most probable population is the thin disc. The first quality flag (in any case) is the probability of the best population, and the second quality flag (only for thin disc) is the probability of the best age component with respect to other age components.

3.3. Tests

Various tests have been made, or are planned, to estimate the quality of the classification method. Firstly, tests using fake data are made to estimate the quality of the classification by determining the difference between the values of intrinsic parameters found by the program and the real values of the simulated stars. Secondly, the algorithm will be applied on Hipparcos data, by comparing for instance the spectral types given in the Hipparcos catalogue with the ones estimated by the bayesian classification.

Concerning the classification, the results of the first tests with fake data simulating the polar cap at V $\leq$ 8.9 in the condition of observations of Hipparcos are the following:

- Absolute magnitude in V: 80% of stars are included in the confidence interval at ± 1 sigma around the correct value, and 96% are in the confidence interval at ± 2 sigmas. But the dispersion can be high, varying from 0.05 to about 1 mag.
- Luminosity class: 83% of stars are classified in the expected class, and 16% in a neighbour class.
- Spectral type: 88% of stars obtain the correct type, but only 32% of them obtain the correct sub-type.
- log $T_{\rm eff}$: 80% of stars are included in the confidence interval at ± 1 sigma around the expected value, with small dispersions, and 96% are in the confidence interval at ± 2 sigmas.
- log g: 79% of stars are included in the confidence interval at ± 1 sigma around the correct value, with small dispersions, and 95% are in the confidence interval at ± 2 sigmas.
- Stellar population and age: 93% of stars are classified in the expected population, but only 42% of thin disc stars obtain the correct age index.

4. CALIBRATION

The Besançon model being a smooth and global model of the Milky Way, biases may appear between the model and the observations because of local inhomogeneities, observational causes, extinction or model defects. Therefore, a calibration of the simulations is needed before the classification, and this calibration must be automatic.

4.1. Method

The chosen method is as follows: in the observable space, medians, eigenvectors and semi-dispersions on right and left along the eigenvectors are determined for observed stars as well as for simulated stars. Then, translations, rotations and semi-homotecies are applied on simulations to make their medians, eigenvectors and semi-dispersions coinciding with observations. The calibration is iterated until it stabilizes.

4.2. Tests

Tests using fake data showed that the calibration method does not work when the parallax (for which the dispersion is very high for closest stars and small for others) is taken into account with other observables, and works rather well when only the photometric observables (apparent V magnitude, B–V and V–I colours) are used. Therefore, in the tests below, parallaxes have been excluded for the rotations, translations and semi-homotecies applied on simulations. However, as a selection between lower and upper limits for all observables is made before each iteration of calibration and before the classification, the calibration still had effects on the goodness of fit of parallaxes.

The goodness of fit between fake data and simulations were estimated using a χ^2 computed for each observable.

- First tests were made using fake data and simulations taken from the same model to estimate the influence of the calibration on unbiased simulations. Because of the Poisson noise, fake data and simulations are not exactly the same, and the program changed a little the simulations to try to reduce the difference. This difference slightly decreased for some observables, but increased for others. But these changes were small and did not have a great effect on the simulations.
- Second tests are going on using the same fake data but artificially biased simulations. The bias introduced in the simulations was a shift in the apparent V magnitude with a linear dependance on B-V. Only the magnitude has been shifted, but this bias had also an effect on other observables due to the selection between lower and upper limits. Preliminary tests showed that the calibration process enhances deeply the goodness of fit of all observables, especially the parallax, but does not retrieve yet the level of agreement obtained in the case of unbiased simulations.

5. RAPIDITY

Since millions of Gaia sources will be treated every day, the classification and alert detection method have to be fast, as well as the calibration algorithm.

In our tests using a PC of 1.7 GHz, the calibration spent 30s CPU for about 6000 stars. Concerning the selection of simulated stars liable to belong in the confidence interval, we reduced very deeply the computing time by using simulations already sorted in one of the observables (the apparent V magnitude). Taking a distance of 7 (normalized by errors) in the observable space as the frontier of the confidence interval around the observation, the resulting computing times are 1s for an ‘alert treatment’ (only for a very few objects with respect to the whole data) and 0.008 s for a classification.

6. CONCLUSION

We have built an algorithm allowing to detect science alerts and classify the stars observed by Gaia with respect to their intrinsic parameters. A calibration program was also constructed to correct automatically possible biases in the simulations before the classification.

Preliminary tests using fake data and simulations both taken from the Besançon model of the Galaxy showed that no false alert is emitted and that the vast majority of observed stars are well classified. The calibration method enhances deeply the agreement between observations and simulations in case of artificially biased simulations, but does not erase completely the shift yet. Other tests, using real data taken from the Hipparcos catalogue are planned.

The chosen observables were the apparent V magnitude, the B–V and V–I colours and the parallax, but the method can be extented to other observables like kinematics for instance. In the same way, chosen intrinsic parameters were the absolute magnitude, the luminosity class, the spectral type, log $T_{\rm eff}$, log g and the population (and age component for thin disc stars), but other parameters like metallicity could also be taken into account.

ACKNOWLEDGMENTS

This work was supported by a grant from the Ministry of Science, Research and the Arts of Baden-Württemberg (24-727.17/145) to Dr. Sébastien Picaud.

REFERENCES

Robin, A.C., Reylé, C., Derrière, S., Picaud, S. 2003, A&A, 509, 523
erratum: 2004, A&A, 416, 157