Highlights of Spanish Astrophysics VII, Proceedings of the X Scientific Meeting of the Spanish Astronomical Society held on July 9 - 13, 2012, in Valencia, Spain. J. C. Guirado, L.M. Lara, V. Quilis, and J. Gorgas (eds.)

OMC automatic variable star classification

M. López del Fresno[1,2], L.M. Sarro Baro[3], and E. Solano Márquez[1,2]

[1] Centro de Astrobiología (CSIC-INTA), Departamento de Astrofísica, PO Box 78,
E-28691, Villanueva de la Cañada, Spain
[2] Spanish Virtual Observatory, Spain
[3] Departamento de Inteligencia Artificial, UNED, Juan del Rosal, 16, 28040 Madrid, Spain

Abstract

The Optical Monitoring Camera (OMC), on-board the ESA mission INTEGRAL, has stored more than 190.000 light curves for almost 10 years. Among the targets included in its input catalogue there is a relevant amount of variable stars. In many cases, OMC has gathered photometric information of sufficient quality to enable a stellar variability analysis of those stars. In this contribution we show the full pipeline of our classification system, from the period calculation to the final membership classification. We also include relevant points of the system such as the parameters used for filtering good quality light curves, specific optimizations applied to the classification algorithms and an overview of the results obtained.

1 Introduction

OMC instrument has been gathering photometric information for about 10 years. Until very recently there has not been available a catalogue of the variable stars contained in its catalogue [1]. The processing of that information has been a very time consuming task due to the size of the OMC database.

One of the many implications that we can infer of that experience is that we are on the limit of what human, non automatic, processing can achieve. The amount of data that current missions like Gaia are going to provide us in the next few years are forcing us to develop new techniques to confront this fact.

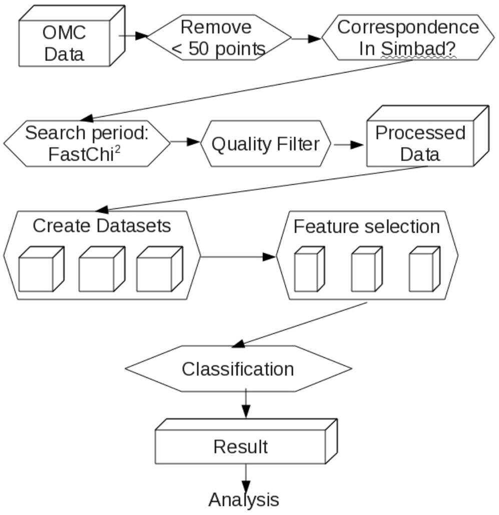

Figure 1: Classification system pipeline

2 OMC automatic classification

The effort to detect and classify automatically variable stars is not knew. A very well known example is the CoRoT mission, where a procedure for light curve analysis was developed by [6]. The results of the automatic classification pipeline can accessed here. Another example is [3], where a new methodology was applied to detect and classify periodic variable stars in the Trans-Atlantic Exoplanet Survey[1]. Finally, Gaia, an ESA mission, which will be launch next year, will monitor about one billion stars during five years and has a team dedicated to develop software and algorithms to provide reliable automatic classification results to the scientific community.

There are lots of automatic classification algorithms that have been proved successfully in several fields of knowledge (as bayesian and neural networks, random forests, or decision trees). In this section we will only outline the different techniques used in this experiment. Fig. 1 shows the pipeline used in this classification.

Supervised classification systems, as the one used here, need a training set that teaches the system the characteristics of the classes. We have modified the CoRoT training set, mostly by removing all the multiperiodic attributes as OMC data quality allows us to search for only one period. We have 959 instances in the training set corresponding to 12 classes divided in the following way:

- ECL (Eclipsing)
- Cepheids: CEPHEID (classical) and DMCEP (double mode)
- Long period: MIRA and SR (semi-regular)

[1] http://exoplanetarchive.ipac.caltech.edu/docs/datasethelp/EETSS_TrESLyr1.html

- RR Lyrae: RRAB, RRC and RRD
- Other: DSCUT (delta-scuti), SPB (slow pulsating blue), GDOR (gamma doradus), BCEP (beta cephei)

Each instance has 18 attributes. All of them, except for two 2MASS colours ($J - H$ and $H - K$) are derived from the period, in particular we have phase differences, coefficient ratios and amplitudes.

We started by cleaning the full OMC dataset. Only those light curves with at least 50 good photometric points (according to the OMC error codes) labelled in Simbad[2] as "star" were selected. We also search on the 2MASS catalogue for the $(J - H)$ and $(H - K)$ colours in order to improve the classification. At this stage we obtained 13189 stars.

The next step was to search for the period of the light curves. We opted to use FastChi^2 method [10] mainly because of its insensitivity to the sample timing. It has also other advantages as statistical efficiency and computational speed. The main drawback we found in our experiment was that sometimes it selects alias frequencies (typically 1/2, 2x or 3x the true period). So the confidence on the folded curve with this method is not absolute.

In order to check if the period is really good we took 2 different complementary approaches: looking at the shape or looking at the data of each light curve. The filter focused on the data was divided in two parts. The first one discarded those curves with a mean noise greater than 10% of the magnitude range of the curve, as they correspond to very noisy curves whose periodicity is very hard to distinguish. For the second one we divided the data in 20 bins according to the phase and accepted only those curves with data points in at least 75% of them. In this way we discard those curves that are not evenly distributed.

The second approach is a bit more sophisticated and consist in the automatic identification of "pretty" light curves. Usually, periodic variable light curves are easy to identify by looking at them, but due to the noise is hard to do it automatically with a simple algorithm. We decided to use a very simple classifier base on Naive Bayes networks [8], using a extremely small training set for achieving this task. We selected 67 samples, divided their graphical representation in a 2x matrix, and assigned them to 4 classes according to their shape: `eclipsing`, `sinusoidal`, `RR-like` and `trash`. We used that training set to feed the Naive Bayes classifier and checked that the results provide very good candidates even in those cases that were not clearly represented in the training set. After applying all the filters and selecting all the light curves which did not fall under the `trash` class we obtained 953 light curves.

For the classification algorithm we decided to take an approach in which every class is confronted with all the other classes in binary classifiers. This decision was taken to simplify the problem of imbalanced datasets [9], multiclass discretization and to maximize the potential of attribute selection methods. The algorithm selected for the classification was the *Support Vector Machines* (SVMs) [5]. SVMs try to find a hyperplane that separates two classes. The main limitation of this algorithm is that basic SVMs cannot handle problems with three or more classes. We also applied the CFS [7] method to the data, which is a feature

[2] http://simbad.u-strasbg.fr/simbad/

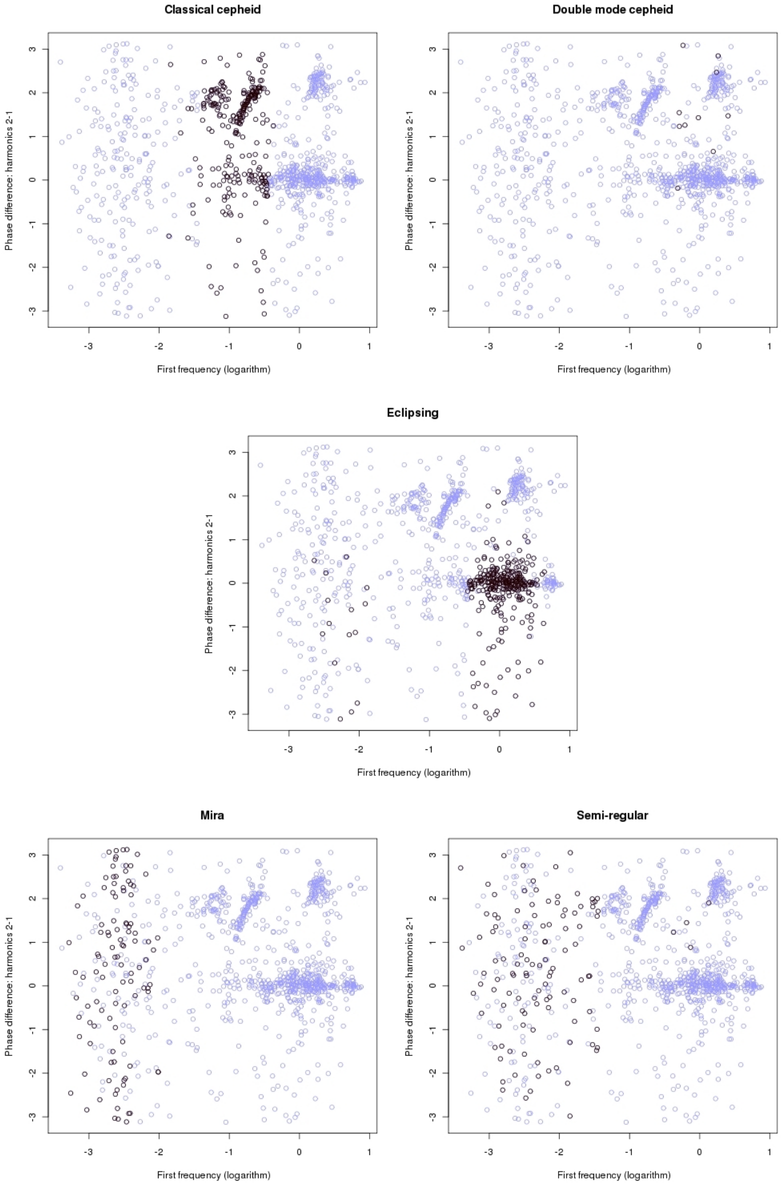

Figure 2: Classification results 1/2. Light colour correspond to the full OMC test set and Dark colour to the class specified in the upper label.

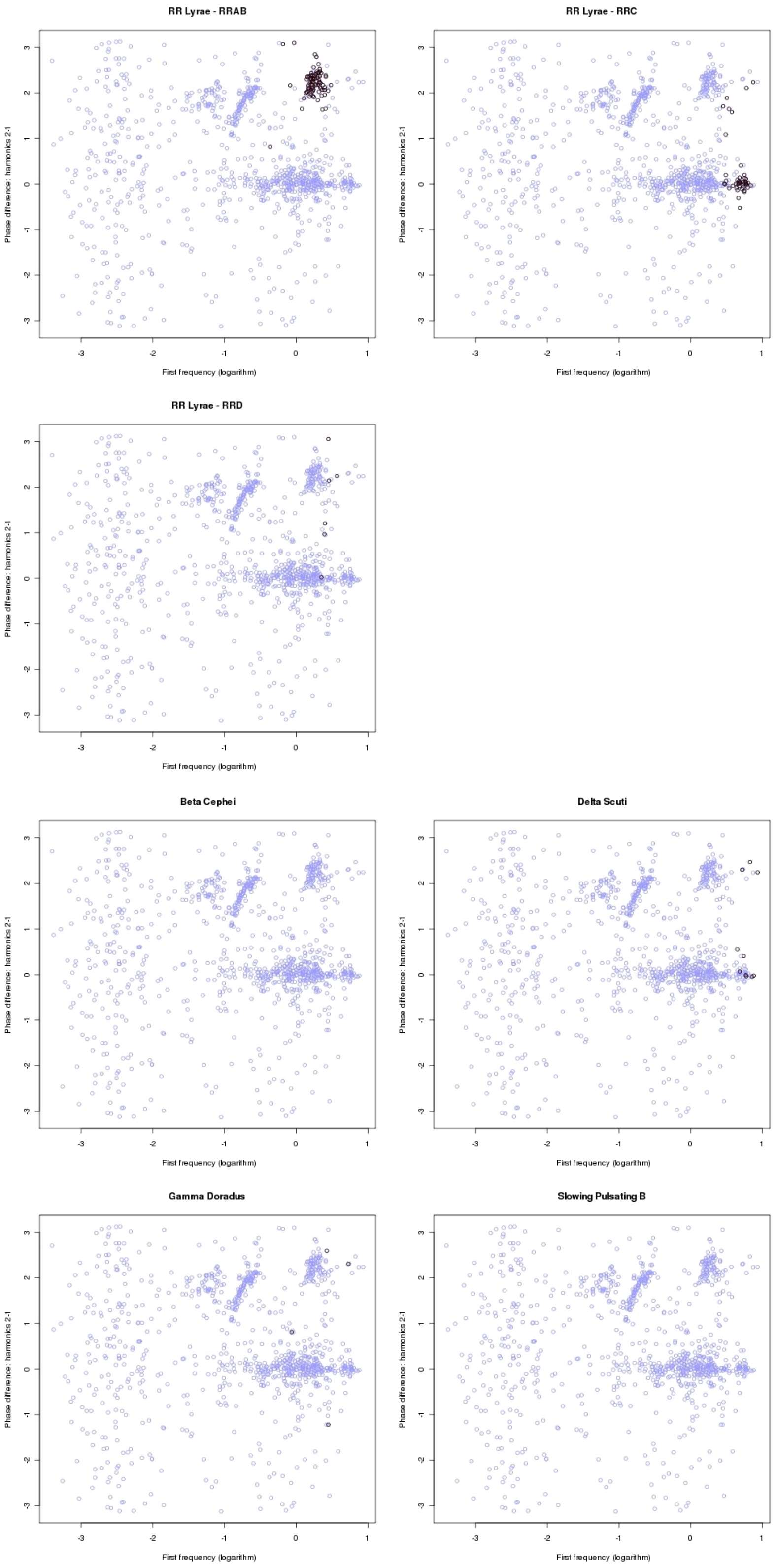

Figure 3: Classification results 2/2. Light colour correspond to the full OMC test set and Dark colour to the class specified in the upper label.

selection method based on selecting attributes/features highly correlated with the class but low correlated with the other attributes.

The combination of the different classifiers has been done using a vote system. For each star we have 66 classifiers, as each one of the 12 class can combine with the other 11. It is expected that the real class will be selected anytime it appears, so ideally it will have more votes than any other (the maximum number of votes for a single class is 12).

3 Results and future work

Fig. 2 and Fig. 3 show the distribution of the classes. In the diagrams we can see that the different classes are grouped quite consistently.

It seems clear from the graphics that most of the confusions are due to the difficulty to distinguish between RRC and contact or semi-detached binaries. Our classifier only use period-related attributes and two colours, and it is not enough for separating those classes.

We have shown that a pure automatic classification system can retrieve lot of information for a dataset. The main bottlenecks in this experiment have been the alias detection in the search for the light curve periods (we could recover more good light curves) and the inability to distinguish between RRC and contact binaries. We are working to solve both issues.

Acknowledgments

This research has made use of the Spanish Virtual Observatory supported through grants AYA2008-02156, AYA2011-24052 from the Spanish MICINN/MINECO. Based on data from the OMC Archive at CAB (INTA-CSIC), pre-processed by ISDC.

References

[1] Alfonso-Garzón, J., Domingo, A., Mas-Hesse, J.M., & Giménez, A. 2012, A&A, 548, A79

[2] Baraffe, I., Chabrier, G., & Gallardo, J. 2009, ApJ, 702, 27

[3] Blomme, J., Sarro L.M., O'Donovan F.T., et al. 2011, MNRAS, 418, 96

[4] Chabrier, G. & Baraffe, I. 2007, ApJ, 661, L81

[5] Cortes, C., Vapnik, V., & Saitta, L. 1995, Machine Learning 20, 273

[6] Debosscher, J., Sarro, L. M., Aerts, C., et al. 2007, A&A, 475, 1159

[7] Hall, M. 1999, University of Waikato, PhD Thesis

[8] John, G., Langley, P., Besnard, P., & Hanks, S. 1995, in *Proceedings of the 11th conference on uncertainty in artificial intelligence*, 1, 338

[9] Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. 2006, Science 30, 25

[10] Palmer, D.M. 2009, ApJ, 695, 496