Supervised Star Classification System for the OMC Archive

M. López, L.M. Sarro, E. Solano, R. Gutiérrez, and J. Debosscher

Abstract INTEGRAL, COROT and, in the near future, KEPLER and GAIA are missions that are producing, or will produce, light curves as we have never seen before, because of their quantity or quality. The exploitation of the scientific potential hidden in these datasets is limited by size of the dataflows. For example, visual classification of the $\approx 10^8$ GAIA light curves is infeasible for any research team. Supervised classification of light curves has been one of the main research lines in the Spanish Virtual Observatory, and we are now co-leading the data analysis work group for the CoRoT mission. In this contribution we show the development of an automatic multistage classification system based on bayesian networks for the OMC (Optical Monitoring Camera) data. OMC is an optical camera on board ESA's INTEGRAL, whose data archive is managed in the LAEFF Scientific Data Center (http://sdc.laeff.inta.es/omc)

Supervised Classification

The impossibility to manually extract knowledge from huge datasets has lead to the development of many fields under the common name of Machine Learning. In

Mauro López, e-mail: mauro@laeff.inta.es
Enrique Solano e-mail: esm@laeff.inta.es
Raul Gutiérrez e-mail: raul@laeff.inta.es
SVO / LAEX-CAB (INTA-CSIC), Postal address.- LAEFF, European Space Astronomy Center (ESAC), P.O. Box 78, E-28691 Villanueva de la Cañada, Madrid, Spain

Luis M. Sarro
Dpto. Inteligencia Artificial. ETSI Informática - UNED, C\ Juan del Rosal 16 - 3, E28040, Madrid - Spain, e-mail: lsb@dia.uned.es

Jonas Debosscher
Instituut voor Sterrenkundem Katholieke Universiteit Leuven. Celestijnenlaan 200D BUS 2401 3001. Leuven - Belgium, e-mail: jonas.debosscher@ster.kuleuven.ac.be

particular, automatic classification has proved to be very useful for the astronomical community.

Supervised classification is suitable for problems in which the classes are known in advance. A model is built using a so called training set of well known instances. There are several different methods to build the model, but the underlying goal is always to produce probabilistic class assignments for unclassified new cases.

Bayesian networks (BN) are probabilistic graphical models for supervised classification which represent a set of random variables and the relationships among them. They present the variables (nodes) and the conditional dependence probabilities relationships among them (arcs) in a very condensed and human readable way. BN use prior/expert knowledge for computing class membership probabilities and are able to explain not only what class is more likely, but also why, as opposed to other Machine Learning algorithms that operate like black boxes with no explanation of the inference process. It is also very important to remark that their learning speed and flexibility allows their use in many fields.

Classification in classification schemes with large numbers of classes are problematic and tend to produce poor performance results. Joining classes into several groups can be useful to build more specialized classifiers, which leads to an improvement of the classification efficiency. A hierarchy of classifiers can progressively classify the instances into more specific groups. This idea is specially useful when the classes are not totally independent but make up groups in the problem domain, because we would then represent class relationships in a natural way.

Different hierarchies represent various approaches to the problem, and it is possible to define measures of how well a hierarchy describes a particular problem. Usually, since we are grouping classes that share some common properties, hierarchies suggested by the experts in the field are good starting points for the search in the space of groupings.

Experiment description

OMC detects the optical emission from the prime targets of the gamma-ray instruments on-board INTEGRAL. The OMC Input Catalogue was created gathering potential targets such as gamma- and X-ray sources, AGNs, variable stars, and Hipparcos and Tycho reference stars (for astrometric and photometric calibration). The actual OMC catalogue comprises more than 540000 sources, 154000 of which have at least one photometric measurement.

In order to build the classification model we will use a training set that describes the following variability classes according to the frequency content of their time series:

- Eclipsing: EA, EB, EW
- Cepheids: classical, double mode, RVTAU, PTCEP
- Long period: MIRA and semiregulars

- RR Lyrae: RRAB, RRC, RRD
- Other: mostly irregulars and multiperiodic

We searched the entire catalogue for good quality light curves. Those with more than 50 photometric points, and classified in Simbad as stellar objects were considered to fall into this category. The process for characterizing the variability, i.e. detecting the frequencies and related parameters (amplitudes and phase differences), is described in depth in [1]. Although OMC was not designed to provide light curves suitable for the analysis of multiperiodicity, light curves with more than 50 photometric points are good enough to include attributes describing multiperiodicities in the time series in our general setup.

Since the instances are represented using continuous values, a discretization stage was needed before the data are used to construct the Bayesian Classification model. We have used the supervised discretization method by [2] in a cross validation framework in order to avoid overfitting in the discretization process.

For every node (classifier) in the hierarchy we have used a Bayesian Augmented Network (a variant of BN), with a structure learnt with the K2 algorithm as applied to the forementioned training set. We have used CFS filtering [3] for selecting an optimal subset of attributes. Classifier models were built using the free and open source software Weka (http://www.cs.waikato.ac.nz/ml/weka/), while the hierarchy system was derived using our own code. As we are using only bayesian classifiers for the nodes we name our system as *Multi-Stage Bayesian Network* (MSBN)

The various subclasses in the category of *Eclipsing* binaries (EA, EB and EW) are almost impossible to separate among themselves, but easy from other classes, due to they are too poorly defined to provide us with a reliable reference ([5]). Thus we decided to use the classification scheme and classifier in [5] (see fig. 1). Furthermore, the *Others* category includes many different classes such as multiperiodic and nonperiodic variables that are our classifiers are unable to separate given the OMC time series characteristics. Thus, in both cases we decided not to further refine the global class assignment.

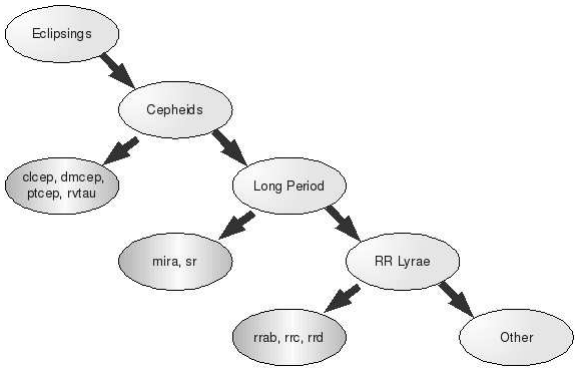

Fig. 1 Hierarchy selected

	MSBN	Single classifier
Eclipsing:	99.7194%	98.9446%
Cepheids:	98.0810%	
CLCEP:	84.9660%	82.5892%
DMCEP:	95.0476%	96.8421%
RVTAU:	46.1557%	22.7273%
PTCEP:	19.6162%	12.9032%
Long Period:	96.2037%	
MIRA:	88.6582%	90.9677%
SR:	71.1070%	38.7096%
RR Lyrae:	92.6176%	
RRAB:	87.6684%	91.6667%
RRC:	73.5493%	57.8947%
RRD:	92.6176%	100%

Fig. 2 Success rate for MSBN and monolithic classifiers

Figure 2 shows the success rate of our classifier. For comparison purposes, we also generated a simple unique bayesian network classifier to assess the improvement obtained from the hierarchical approach.

Conclusions and Future Work

MSBNs are a good choice for variable star classification. By building more specialized classifiers with feature selection taylored to each particular problem, we can improve the quality of our classification system.

Next steps will be directed towards integrating the classification system into OMC's web service (www.sdc.laeff.inta.es/omc) and adding colour information from other catalogues and using the classifiers in [4] (for classifying multiperiodic classes.

This research has made use of the Spanish Virtual Observatory supported from the Spanish MEC through grants AyA2008-02156, AyA2005-04286

References

1. Debosscher, J., Sarro, L.M., Aerts, C. *Automated supervised classification of variable stars I. Methodology*, 2007, A&A.
2. Fayyad U., Irani K.B., *Multi-interval discretization of continuousvalued attributes for classification learning*, 1993, IJCAI– 93
3. Hall M.A., *Correlation-based Feature Subset Selection for Machine Learning*, 1998, Hamilton, New Zealand.
4. Sarro, L.M., Debosscher, J., López, M., Aerts, C. *Automated supervised classification of variable stars II. Application to the OGLE database* , 2008, A&A.
5. Sarro L.M., Sánchez-Fernández C., Giménez A., *Automatic classification of eclipsing binaries light curves using neural networks*, 2005, A&A