Learning from optically variable stars: the OMC scientific case

Luis Manuel Sarro[1], Albert Domingo[2], J. Miguel Mas-Hesse[2], Enrique Solano[2], Raul Gutiérrez[2]
[1]Dpto. de Inteligencia Artificial (UNED), Madrid, Spain
[2]LAEFF–INTA, Madrid, Spain

Abstract

In this work we present ongoing lines of research carried out by the Group of Operations and Archives at LAEFF (Laboratory for Space Astrophysics and Fundamental Physics) in order to provide "intelligent" tools for the scientific exploitation of the astronomical archive data provided by its Scientific Data Center. We focus here in the problems of detecting non periodic variability and classifying periodic light curves in the ever increasing archive of the Optical Moniting Camera (hereafter OMC) on board INTEGRAL.

Keywords: Artificial Neural Networks, Bayesian Methods, Variable stars

1. INTRODUCTION

The OMC archive is part of LAEFF's (Laboratory for Space Astrophysics and Theoretical Physics) Scientific Data Center (SDC) which at present comprises the IUE, GAUDI-COROT and OMC archives. The Group of Operations and Archives at LAEFF is developing a series of tools based on statistical theory and artificial intelligence to perform preliminary analysis on the data continuosly arriving at the SDC. The rate at which new data are incorporated into the archive rules out the possibility of carrying out tasks such as relevance detection and classification in the traditional human-based fashion. On the contrary, the automated methods developed at the SDC provide a consistent a reproducible way to cope with these overwhelming amounts of data.

The OMC on board INTEGRAL is a refractive telescope plus a large format CCD, with a field of view of $5^\circ \times 5^\circ$ optimized to provide simultaneous V band photometry of the fields covered by the high energy instruments JEM-X, SPI and IBIS (see [3] for details). The main scientific objectives of the OMC are (i) to monitor during extended periods of time the optical emission of all high energy targets within its field of view and (ii) to monitor serendipitously a large number of optically variable sources within its field of view. It is this latter objective that will eventually produce the vast catalogue of thousands of variable sources with well calibrated optical light curves covering periods from minutes to months or years, whose processing in reasonable times is beyond human achievement.

One of the main problems encountered in the automatic processing of time series of photometric measurements is the necessity to disentagle the underlying smooth light curve from the superimposed noise. In the first of the tasks covered in this paper (variability detection) the independent estimation of the noiseless curve provides us with a reference level, the relative variations of which can be assesed by comparison with the variance of the data with respect to this smooth curve. In the second task (light curve classification), the result of the application of artificial neural networks (hereafter ANNs) is greatly improved if the input data are free of noise (although the distributed nature of ANNs provides these systems with remarkable robustness and noise tolerance). Therefore, as part of the data analysis techniques we will need a method to regress the original observations.

In the next section we summarize the bayesian approach to regression upon which the automated procedure for relevance detection and classification is based. In section 3 we describe the

methods used to accomplish the tasks and in section 4 we summarize the results obtained and propose further extensions in the framework of virtual observatories.

2. BAYESIAN LEARNING OF NEURAL NETWORKS FOR LIGHT CURVE REGRESSION.

Let us consider the problem of defining a smooth curve given a set of N observations $\mathcal{D} = \{t_i, V_i\}$ where t_i can be a time or phase variable. We assume as underlying model that the data were generated from a smooth curve $\mu(t)$ with added gaussian noise of zero mean and variance σ

$$P(V|t) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\{-\frac{(\mu(t) - V)^2}{2\sigma^2}\} \tag{1}$$

and seek to find the best predictive distribution for the target variables given a test case and the training set $\mathcal{D}$. If we allow for outliers in our data the gaussian distribution in (3) turns into a t-distribution. It is known that a multilayer perceptron ANN can approximate any function defined on a compact domain arbitrarily closely if enough hidden layers, units and synapses are included in the architecture. We therefore select multilayer perceptrons to approximate the regressed light curve but, in order to avoid the well known problem of overfitting while at the same time specifying a unique architecture complex enough to account for all possible input data, we adopt the bayesian approach instead of using a single set of parameters (weights and biases) obtained by, for example, the backpropagation learning algorithm.

Let $\{\theta\}$ be the set of parameters (weights and biases) defining one ANN of a predefined architecture that computes the approximation $\mu(t, \{\theta\})$ to the real function $\mu(t)$. In the bayesian approach, the predictive distribution $P(V|t, \mathcal{D})$ is aproximated by the mean of all possible neural networks with the predefined architecture weighted by their posterior probability $P(\{\theta\}|\mathcal{D})$. Thus,

$$P(V|t, \mathcal{D}) = \int P(V|t, \{\theta\}) P(\{\theta\}|\mathcal{D}) \mathrm{d}\theta \tag{2}$$

Once this distribution is computed, the mean, median or mode can be used to define the smooth underlying light curve. The first term in the integral can be easily evaluated for any set of parameters $\{\theta\}$. The ANN architecture together with $\{\theta\}$ defines $\mu(t, \theta)$ and equation

$$P(V|t, \{\theta\}) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\{-\frac{(\mu(t, \{\theta\}) - V)^2}{2\sigma^2}\} \tag{3}$$

for the noise model (or the corresponding t-distribution if outliers are allowed for) provides the means to compute this first term. The second term is the well known posterior probability of the parameters given the training data that assigns higher weights to networks significantly better at explaining these data. This term is computed as the product of the likelihood and the prior for the parameters (Bayes theorem)

$$P(\{\theta\}|\mathcal{D}) \propto P(\mathcal{D}|\{\theta\}) P(\{\theta\}) \tag{4}$$

If instead of the predictive distribution we are interested in the prediction for a given new value of t^{new} using e.g. the mean of the distribution, we would then use the same formalism to compute

$$V^{new} = \int \mu(t^{new}, \theta) P(\{\theta\}|\mathcal{D}) \cdot \mathrm{d}\theta \tag{5}$$

Unfortunately, any reasonable architecture for ANN regression makes it infeasible to sample all the multidimensional parameter space implicit in the integral in equation (2). Therefore, special techniques are needed to approximate the integral by a finite sum of terms drawn from the posterior distribution:

$$V^{new} = \int \mu(t^{new}, \theta) P(\{\theta\}|\mathcal{D}) \mathrm{d}\theta \approx \frac{1}{M} \sum_{l=1}^{M} \mu(t^{new}, \theta^l) \tag{6}$$

where each θ^l is generated by a process that results in the M parameter sets following the posterior distribution. We have used a hybrid Monte Carlo method to approximate the integral.

A full description of the formalism is beyond the scope of this contribution (see e.g. [4] and references therein for a detailed account of the method). We only mention here that the priors introduced in equation (4) are hierarchical, i.e., we use as low level prior for each parameter (a weight for example) a gaussian distribution of zero mean and variance picked from a second level gamma prior that establishes a link between all weights belonging to the same unit. The shape parameter of this gamma prior is specified but the mean is again picked from a higher (third) level gamma prior that binds the distributions of weights in units belonging to the same layer. Finally, the same scheme is repeated with a fourth prior that binds the parameters of all units in the network.

3. VARIABILITY DETECTION AND LIGHT CURVE CLASSIFICATION.

The first task is motivated by the necessity for a rapid detection of variability in light curves of a non periodic nature. Period curves are detected by Fourier or equivalent frecuency space methods. Our method aims to detect known astronomical phenomena such as stellar flares, outbursts or, in general, irregular variability, and consists in three separate steps adjusted to the observing requirements of the OMC. In the first step, a light curve is separated into chunks of contiguous data belonging to the same observation. In the second step, a smooth curve is regressed on the data as described in section 2. Finally, the degree of relevance of the variability found in the smooth light curve is assesed by comparison with the variance of the data around the smooth curve and the variance of the curves $\mu(t^{new}, \theta^l)$ around the mean. The thresholding step is performed via a flexible rule-based system dependent on the target phenomenon, the usual 3σ rule being the simplest example of such rules.

All remaining periodic light curves undergo a systematic classification procedure to extract all possible information from its morphological properties. The preprocessing stage consists in the following steps:

1. Initially, the time series of observations are converted to phase
2. The (often noisy and incomplete) light curve is then regressed using the methods described in section 2.
3. Both the original light curve and the regressed curve are binned in phase bins of width 0.02 (50 bins). The original (often incomplete) light curve is then normalized using the euclidean norm of the regressed curve and the result used to consult a Self Organized Kohonen Map (SOM) constructed with example light curves taken from the Hipparcos catalogue ([1]) selected according to the requirements of completeness and high signal to noise ratio.
4. Hipparcos light curves retrieved from the SOM (a morphological similarity based sample) are then used for pattern completion. Only curvature information is retained while zero and first order terms are adjusted to match the limiting measured phase bins of the light curve.
5. The original (noisy and incomplete) light curve is suplemented with the data obtained in the pattern completion module and regressed again. The second regressed curve is then re-normalized to match the interval $[0, 1]$ and shifted in phase so that the maximum magnitude corresponds to phase 0.75.

Finally, the result of the preprocessing stage is fed into a classification neural network trained with the same bayesian methods outlined in section 2 but with a softmax model for the output. The classification scheme for eclipsing binary system (the first main class of periodic variables, pulsating variables being described below) is taken from the work by Sánchez-Fernández [6] with an additional class for eccentric binaries. In the work by Sánchez-Fernández it is stressed that a new clasification scheme is needed where each class groups homogeneous dynamical/evolutionary scenarios and their characteristic light curve morphologies. The complete description of the clasification scheme and the ANN based problem solving method will appear in a forthcoming paper. Here we will only summarize its fundamental criteria for classification and the architecture. Class 1 corresponds to detached systems, with well defined beginning and end of both eclipses and flat curves between them; class 2 is defined by semidetached system with the secondary eclipse being less than twenty percent the primary and, in any case, less than 0.1 mag; class 3 groups binary systems close to contact with alternating eclipses without intermediate flat intervals; finally, class 4 corresponds to contact binary systems with maxima

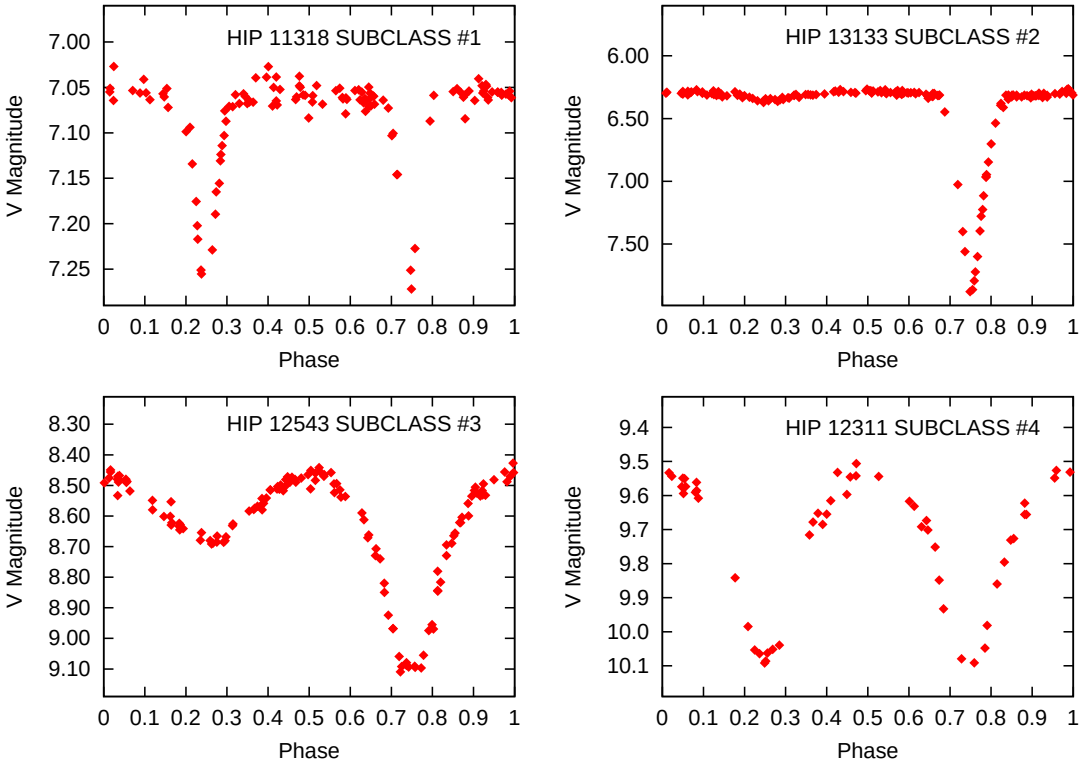

FIGURE 1: Four examples, taken from Hipparcos catalogue, of the eclipsing binary subclasses defined by Sánchez-Fernández in [6].

of the same magnitude or within a 15 percent of the primary maximum. Class 0 was added to separate systems with non circular orbits which, in the original scheme by Sánchez-Fernández were included in class 1. These show eclipses with phase separation different from 0.5. Figure 1 shows example light curves for each of the classes, except class 0.

In the formalism of bayesian training of neural networks, the architecture of the network needs not be critical if computational resources are not a problem. This is so because the hierarchical nature of the priors allows Automatic Relevance Detection (described in [2]) both in the input and hidden spaces. Therefore, we have opted for an unusually large hidden layer of 30 units and let bayesian learning assign higher posterior probabilities to nets with units effectively removed. We have used the complete set of Hipparcos light curves to train the neural network. It must be bear in mind that, since no unique set of parameters is used (such as would be the case with, e.g., backpropagation learning [5]) there is no danger of overtraining (again, see [4] for a thorough discussion on the subject). Nevertheless, full comparison of the performance of the bayesian neural network with several backpropagation networks yielded significantly better results which will be quantified in a separate paper. Only a few light curves that couldn't be classified by a human expert were removed from the set and the remaining curves were labelled according to the criteria by Sánchez Fernández summarized above.

Pulsating variables can only be separated into two broad classes with little physical information. This results from the fact that morphological features of the light curves are not enough to separate different physical scenarios. As an example, figure 2 shows two indistinguishable light curves from totally different pulsating variables, namely HIP48503 (rescaled for clarity, below) and HIP22256 (above). The former is an A9 II RR Lyrae variable with a period of 0.324706 days and a $V-I$ colour of 0.422. The latter, a Mira type variable with a period of 232.6 days and a $V-I$ colour of 1.527.

The fact that morphologically equivalent light curves can be the outcome of essentially different physical scenarios makes it necessary the use of additional information for a more detaild

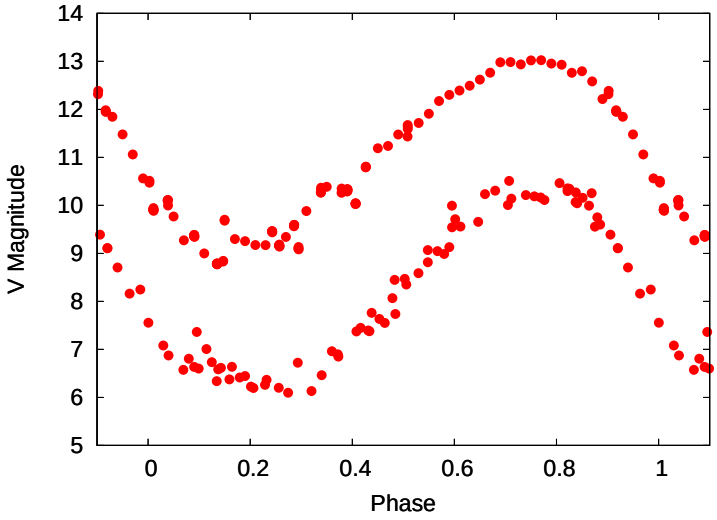

FIGURE 2: Light curves of HIP48503 (below) and HIP22256 (above) taken from the Hipparcos catalogue. The light curve of HIP48503 has been rescaled to match the amplitude of that of HIP22256 to show how two different systems (RR Lyrae and Mira type respectively) exhibit morphologically indistinguishable normalized curves.

classification than a simple sine-like/asymmetric distinction. At present, the ANN is implemented with these two separate classes for pulsating variables (one for sinusoidal light curves and one for asymmetric slopes in the ascending/descending phases).

4. SUMMARY AND FUTURE EXTENSIONS FOR VOS.

In this work we have described a method for (i) the automatic detection of non periodic variability and (ii) the classification of periodicity in visual light curves produced by the Optical Monitoring Camera on board INTEGRAL. The classification, based on connectionist inference, reproduces a scheme of physical motivation with two main categories (eclipsing and pulsation) with 5 subclasses for eclipsing systems (again, with physical significance) and 2 subclasses for pulsating variables that, inevitably, do not correspond to homogeneous systems. It is ongoing work an extension of the expert system to make use of the period plus information distributed accros different virtual observatories in order to produce a complete, physically meaningful classification in the case of pulsation. The complete system including the Self Organized Map for similarity based sample queries will be made available as part of the Spanish Virtual Observatory OMC Data Server.

REFERENCES

[1] *The Hipparcos and Tycho Catalogues*. Number ESA SP-1200. ESA Pubs. Div., 1997.

[2] D. J. C. MacKay. *Bayesian Methods for Backpropagation Networks in Models of Neural Networks III*. Springer-Verlag, 1994.

[3] J. M. Mas-Hesse, A. Giménez, J. L. Culhane, C. Jamar, B. McBreen, J. Torra, R. Hudec, J. Fabregat, E. Meurs, J. P. Swings, M. A. Alcacera, A. Balado, R. Beiztegui, T. Belenguer, L. Bradley, M. D. Caballero, P. Cabo, J. M. Defise, E. Díaz, A. Domingo, F. Figueras, I. Figueroa, L. Hanlon, F. Hroch, V. Hudcova, T. García, B. Jordan, C. Jordi, P. Kretschmar, C. Laviada, M. March, E. Martín, E. Mazy, M. Menéndez, J. M. Mi, E. de Miguel, T. Muñoz, K. Nolan, R. Olmedo, J. Y. Plesseria, J. Polcar, M. Reina, E. Renotte, P. Rochus, A. Sánchez, J. C. San Martín, A. Smith, J. Soldan, P. Thomas, V. Timón, and D. Walton. OMC: An Optical Monitoring Camera for INTEGRAL. Instrument description and performance. A&A, 411:L261–L268, November 2003.

[4] R. M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, 1996.

[5] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning representations by back-propagating errors. Nature, 323:533, 1986.
[6] C. Sánchez-Fernández. Clasificación de curvas de luz de sistemas binarios eclipsantes. Master's thesis, Universidad Autónoma de Madrid, 1998.