Astronomical Data Analysis Software and Systems XV
ASP Conference Series, Vol. 351, 2006
C. Gabriel, C. Arviset, D. Ponz and E. Solano, eds.

Bayesian Classifiers for Variable Stars

Mauro López

Centro de Astrobiología, CSIC-INTA, Madrid, Spain

Concha Bielza

Artificial Intelligence Department, Technical University of Madrid, Spain

Luis M. Sarro[1]

Spanish Virtual Observatory, Madrid

Abstract. In this contribution we report on the development of a Bayesian network classifier trained on the Hipparcos catalogue plus complementary information obtained from SIMBAD and Vizier catalogues that aims at separating the different variability classes present at the Hipparcos sensitivity level.

1. Introduction

The Optical Monitoring Camera onboard INTEGRAL was the first instrument in space after Hipparcos to deliver a number of photometric time series far too large for a human based classification of variable objects. In this context, the Spanish Virtual Observatory has developed software for the detection and classification of light curves of eclipsing binary systems (Sarro et al. 2005). There obviously remained the necessity to further refine the classification of those systems identified as variables of a non eclipsing nature, a necessity that is becoming even more apparent with the forthcoming launch of the CoRoT[2] mission that will add new photometric time series to the Virtual Observatory databases.

The classification of periodic variables is a typical pattern recognition problem for which there is a wealth of methodological strategies taken from the field of Machine Learning. Amongst them, neural networks is by far the most popular approach in astronomy (see e.g. the volume edited by Tagliaferri et al. (2003) for a summary of recent applications in the field) although other alternatives such as Support Vector Machines and Bayesian network classifiers are becoming marginally significant.

Previous attempts to solve this classification task (Pojmanski 2002) did not make use of photometric colours and thus, considered only a limited number of

[1]Artificial Intelligence Department, UNED, Madrid, Spain

[2]`http://smsc.cnes.fr/COROT/`

classes. Furthermore, his work was based on an *ad hoc* definition of polygons in attribute space.

2. Preprocessing

The raw photometric time series in the V band are subject to a number of preprocessing steps. In order to obtain reliable, robust estimates of the amplitude and the Fourier coefficients of the folded light curves, the latter had to be regressed, rebinned and the missing phase bins completed. Hipparcos periods were used in the training set for consistency. The complete procedure is described elsewhere in detail (Sarro et al. 2005) but can be summarized as the following sequence of steps:

1. The folded light curve is rebinned (bin width $\Delta\phi = 0.02$)
2. The missing bins in the rebinned light curve are completed by using curvature terms of the closest light curves retrieved from a Kohonen Self Organized Map (SOM) constructed with complete Hipparcos light curves
3. The photometric values inferred in the previous step are also used to complete the gaps in the original unbinned light curve. The resulting data set is regressed in order to minimize the effect of noise and outliers.
4. The result of the regression is Fourier analysed in order to obtain the first four sine and cosine coefficients.

3. Bayesian Network Classifiers

Bayesian networks (BN) are probabilistic graphical models which represent a set of random variables and the relationships between them. The information is presented in a very condensed and human readable way: an acyclic directed graph of nodes (variables) and arcs (conditional (in)dependence probabilistic relationships), see e.g. Korb & Nicholson (2004).

The structure of a BN can be automatically learned from data. Score based learning algorithms need a scoring function and a (local or global) strategy to search for a high scored network. There are basically two types of networks especially suitable for classification purposes: Naïve Bayes (NB) and Tree Augmented Naïve Bayes (TAN). The class attribute is the single parent of each node of a NB network. TAN networks allow for a second parent, modelling more complex relationships among attributes.

Bayesian networks have been used successfully in the last years to solve classification problems. They represent a good alternative to neural networks in this field, mainly because of their semantic power and their lack of black box features.

4. Analysis

The stellar database has some features that make the classification harder. First, data probability distributions are heterogeneous. Highly populated regions of the parameter space can coexist with vast empty regions. In this context, atypical data are very well hidden. Some of them may be errors. We decided not to

remove them in view of any further information, if we want our classifier to prove its flexibility. Second, some *attributes* have missing values, being even almost empty for all samples of a given star class. Furthermore, some *classes* are very populated, but the sample size for others is not enough to make assumptions about them. In this first approach to a classification, imputation techniques to fill these missing values in, did not appear to be appropriate and therefore instances and attributes with missing data were removed from the database. Also, we only considered classes with at least 15 stars in our database. Third, some attributes are continuous data and many of the best algorithms designed for BN are of no use with them. Fourth, non trivial dependencies need to be detected among attributes. Fortunately, BN are expressly designed for this task.

Thus, the classifier uses complete attributes: period, amplitude, V-I, the first four Fourier sine and cosine coefficients and their coefficient ratios. Star classes are: Cepheids, α^2 CVn (ACV), β Cephei stars, ellipsoidal rotating stars, δ Scuti, Mira, slowly pulsating B stars (SPBs), RR Lyrae, RV Telescopii, semirregulars (SR, SRA) and SX Phoenici.

The different classifiers we tried were built with the help of the free and open source software Weka[3]. Weka manages continuous data in NB structures, the first and simplest network we tried. However, correctly classified instances were below 70%, both using normal distributions and kernel density estimators. NB assumes that all attributes are conditionally independent given the class, which does not hold in this case. TAN was the best solution, although we had to prediscretize the continuous attributes. The entropy-based (supervised) discretization method, taking the classes into account, yielded the best results. The search for the TAN net was done using the K2 algorithm, under the *BayesNet* option of Weka (Bouckaert 2004).

There were some classes too difficult to classify with the data available. The confusion matrix showed us that the classifier mixed them together within the same class, without any further error for other classes. Therefore, we aggregated them –ACV & SPB and SR & SRA only– as a first step (in Sect. 5. we mention ongoing work on multilevel classification for the specific refinement of these aggregate classes) and the results raised over 90%. The final TAN net, shown in Figure 1, seems to include reasonable relationships among its nodes.

The final confusion matrix reports the goodness of our classifier in detail. Also, it helps to see that it is crucial, in order to get accurate scores, to have as many samples as possible. 100 samples turned out to be enough to reach an almost flawless classification.

```
Correctly Classified Instances         998               92.7509 %
Total Number of Instances             1076
=== Confusion Matrix ===
   a   b   c   d   e   f   g   h   i   <-- classified as
 257   1   6   3   0   0   0   2   0 | a = ACV & SPB
   1 150   0   0   0   0   0   4   0 | b = DCEP
   3   0  67   0   0   0   4   0   3 | c = DSCT
   6   0   0  20   2   0   0   1   0 | d = ELL
   0   0   0   6  79   0   0   0   0 | e = EW
   0   1   0   0   0 203   0   3   0 | f = M
```

[3] http://www.cs.waikato.ac.nz/ml/weka/

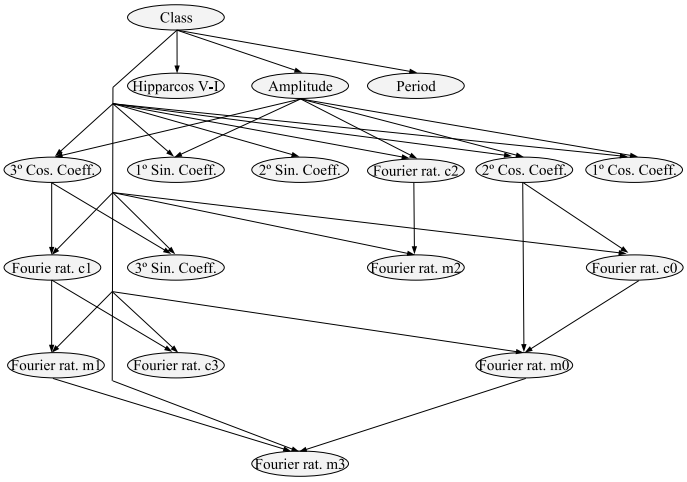

Figure 1. TAN network.

```
0   0   4   1   0   0  71   0   2 | g = RR
3   4   1   2   1   9   0 139   0 | h = SR & SRA
0   0   3   0   0   0   2   0  12 | i = SXAri
```

5. Conclusions and Future Work

Bayesian Networks are a good choice for variable star classification. The correct classification scores obtained are just a lower bound that may be improved taking some considerations. Next steps will be directed towards: filtering atypical data, using a higher number of instances and classes (like γ Doradus, solar-type pulsation and WDs pulsations), incorporating more attributes (B-V, 2MASS colours, Strömgren photometry, and others to characterize multiperiodicity), creating new attributes derived from spectral features, applying attribute selection procedures to remove irrelevances and redundancies, allowing for missing values, using continuous data without discretization, and multilevel classifiers to separate better the aggregated classes, mainly the ACV-SPB group.

References

Bouckaert, R. 2004, Bayesian Network Classifiers in Weka, Working Paper 14/04, Department of Computer Science, University of Waikato, New Zealand.

Korb, K. & Nicholson, A. 2004, Bayesian Artificial Intelligence, Chapman & Hall

Pojmanski, G. 2002, Acta Astronomica, 52, 397

Sarro, L. M., Sánchez-Fernández, C. & Giménez, A. 2005, A&A, accepted

Tagliaferri, R., Longo G., Milano, L., Acernese, F., Barone, F., Ciaramella, A., De Rosa, R., Donalek, C., Eleuteri, A. & Raiconi, G. 2003, Neural Networks, 16, 297