Identification and classification of interesting variable stars in the MACHO database

W. Clarke M. Hegland *

(Received 7 August 2000)

Abstract

The MACHO database is an astronomical database of the intensities of about 20 million stars, recorded approximately every night for several years. About one percent of these stars are classified as "variable stars". These variable stars are generally roughly periodic and can have periods ranging from less than one day to hundreds or

* Computer Sciences Laboratory, RSISE, Australian National University, ACT 0200, AUSTRALIA. mailto:Bill.Clarke@anu.edu.au and mailto:Markus.Hegland@anu.edu.au

[0]See http://anziamj.austms.org.au/V42/CTAC99/Clar for this article and ancillary services, Published 27 Nov 2000.

thousands of days. We investigate the application of computationally simple features in order to classify these stars. We present a methodology for extracting potentially interesting stars based on their location in feature-space, and methods for using human interaction to group these interesting stars.

Contents

1 Introduction

The MACHO database [1, 3] is an astronomical database consisting of the intensities of about 20 million stars recorded approximately every night for several years. The main aim of the collection is to search for very rare events called microlensing when a massive compact object passes in between the observer and a star and amplifies the observed signal. The unusually long records of many stars mean it is also useful for observing the behaviour of variable stars.

About one percent of stars are classified as "variable". These variable stars are generally periodic and can have periods ranging from less than one day to hundreds or thousands of days.

The data for each star consists of a matrix of the following values: *time*, *red band spectral intensity*, *red error*, *blue band spectral intensity* and *blue error*. It is worth noting that the time of observation is not regular, and occasionally either red or blue observations are not recorded.

In summary, the dataset consists of a large number of irregularly sampled time series with two weighted observations (*intensities* and *errors*) and missing values.

The large number of stars and irregular sampling present a problem. The size of the dataset motivates us to simplify the problem by working with features which describe each star's behaviour and reduce the dimensional-

ity considerably. Unfortunately, the irregular sampling means that standard methods for Fourier transforms and wavelets cannot be used. We cannot simply interpolate the data on a uniform grid since many stars vary near or less than the Nyquist frequency, which information is lost when interpolating. Scargle [7, 8] has done a lot of work in extending Fourier analysis to the unevenly sampled time domain but the computations are not simple. Reimann's thesis [5] discusses finding the periods of variable stars using a non-linear method, but this is extremely computationally difficult.

There has been a lot of work on variable stars in the MACHO database in astronomy, notably [1, 9]. In the data mining area, Ng, Huang and Hegland [4] attempt to cluster and classify variable stars from the MACHO database using interpolated Fourier techniques with mixed success.

The dataset investigated in this paper is a subset of the MACHO dataset, referred to as the Wood dataset [9] which contains 792 stars that are considered likely to be long-period variable stars.

The red observations of four typical stars in the Wood dataset are shown in Figure 1. The blue observations are similar and are not shown. The errors are shown as error-bars, and some indication is given in each plot of the values of particular features (see §2), notably average, amplitude and timescale. Note the variation in behaviour between stars, and the large gaps where no data was recorded.

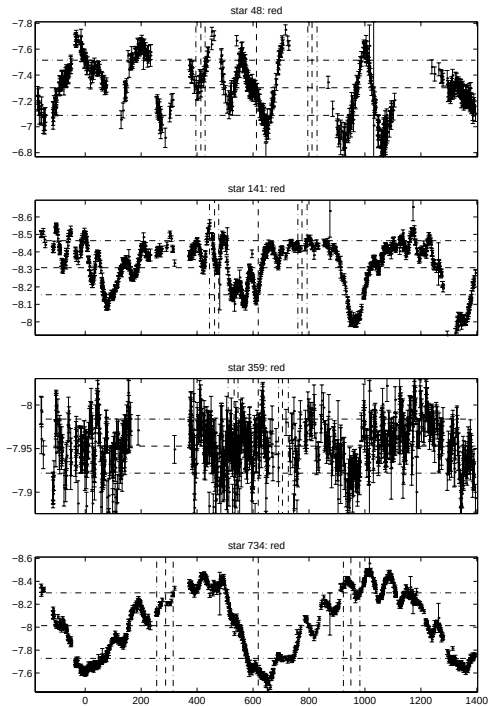

FIGURE 1: The red observations of four typical stars in the Wood dataset. The stars are labelled 77.7306.213, 77.7424.46, 77.7666.748 and 77.8153.64 respectively.

2 Features

In order to compare stars, we need to reduce the amount of information from the thousands of data points per star to a short vector of features. These features should be time-invariant (since phase is unimportant in this problem), computationally simple to calculate ($O(N)$ to ensure scalability) and preferably interpretable. The last allows experts to make judgements based on their knowledge of the physics of the problem, without having to learn new paradigms, and also allows non-experts to gain some grasp on the situation.

2.1 Notation

Due to the additional complication of missing values, it is simpler to consider the data for each star to be two time series of intensities and errors: $\{R(t_i^R), i = 1, 2, \ldots, N_R\}$ and $\{B(t_i^B), i = 1, 2, \ldots, N_B\}$. R and B are vectors (red and blue) containing intensities (R_I and B_I) and errors (R_E and B_E). A generic time series of the same form will be referred to as X. The time series X_I is assumed to be the sum of a deterministic signal and random errors:

$$X_I(t_i) = x(t_i) + e(t_i). \tag{1}$$

The error estimate X_E is assumed to have a similar scale to e.

There are cases where we wish to compare the two time series R and B. In these cases we will use a combined time series C, which contains those observations where *both* red and blue intensities were recorded. When it is clear that we are comparing the combined time series C, R and B will refer to the red and blue time series subsets from C.

2.2 Initial features

The average value of a time series is

$$m_X = \frac{1}{N_X} \sum_{i=1}^{N_X} X_I(t_i). \tag{2}$$

Alternatively, one could use weighted average based on the error estimates but in practice the two were found to be very similar, and only the average m_X is used in this paper.

The amplitude of a time series is

$$s_X = \sqrt{\frac{1}{N_X - 1} \sum_{i=1}^{N_X} (X_I(t_i) - m_X)^2}. \tag{3}$$

The correlation of a combined time series is

$$\rho_C = \frac{1}{(N_C - 1) s_R s_B} \sum_{i=1}^{N_C} (R_I(t_i) - m_R)(B_I(t_i) - m_B). \tag{4}$$

Assume we have a derivative of the time series $X'(\equiv \frac{dx(t)}{dt})$, then an estimate of the time scale of the intensity variation is

$$\tau_X = 2\pi\sqrt{\frac{\|X\|}{\|X'\|}}, \tag{5}$$

where the norm $\|X\|$ is defined as

$$\|X\| \triangleq \frac{\int_{t_1}^{t_{N_X}} (X_I(t) - m_X)^2\, dt}{t_{N_X} - t_1}. \tag{6}$$

We can also estimate the "spread" of the time scale in the signal as

$$\Delta\tau_X = 2\pi \Big/ \sqrt{\frac{\|X''\|}{\|X\|} - \left(\frac{\|X'\|}{\|X\|}\right)^2}. \tag{7}$$

Both τ_X and $\Delta\tau_X$ require derivatives to be calculated. This is accomplished using a numerical difference method described in [2], with the integral calculated using the trapezoidal method.

2.3 Transformations

After calculating the above features some transformations were applied to produce the following features. The rationale behind this was to produce less correlated features that were still meaningful.

Colour A standard astronomical conversion, calculated as the difference between the red and blue averages:

$$\mathrm{col}_{RB} = m_R - m_B; \tag{8}$$

Log relative amplitude

$$\mathrm{lra}_X = \log\left(-\frac{s_X}{m_X}\right); \tag{9}$$

Relative delta time scale

$$\mathrm{rdts}_X = \frac{\Delta\tau_X}{\tau_X}. \tag{10}$$

3 Boxing: an Iterative Clustering and Classification Algorithm

While the above features were reasonable at grouping similar stars together, they were not totally successful at separating quite different sorts of stars.

The intuitive idea behind the boxing algorithm is to repeatedly choose "interesting" examples (patterns) from the dataset and classify them according to previously selected examples. Clearly, there are three difficulties here:

the ability to choose interesting examples, the classification, and deciding whether the interesting example is deserving of a new class.

Given a set of unboxed patterns $\mathcal{P}$, a set of boxed patterns $\mathcal{B}$ (set of pairs (p, l), where p is a pattern and l is a label), a method of choosing a pattern from the unboxed pattern set $p \leftarrow \text{select}(\mathcal{P})$ and a method of comparing a pattern with the boxed patterns $l \leftarrow \text{compare}(p, \mathcal{B})$, the boxing method is a repeated application of Algorithm 1:

Algorithm 1 $(\mathcal{P}, \mathcal{B}) \leftarrow box(\mathcal{P}, \mathcal{B}, select, compare) =$

1. $p \leftarrow select(\mathcal{P})$,
2. $l \leftarrow compare(p, \mathcal{B})$,
3. $\mathcal{B} \leftarrow \mathcal{B} \cup \{(p, l)\}$,
4. $\mathcal{P} \leftarrow \mathcal{P} \setminus \{p\}$.

The selection algorithm *select* picks a pattern from the unboxed pattern set. Assuming the pattern set contains different classes of patterns, and the comparison algorithm is reasonable at comparing these classes, a good selection algorithm would choose patterns which are "typical" of those different classes (e.g., near the mean, in a k-means sense).

The comparison algorithm *compare* is used to compare a particular pattern with the already boxed patterns. The result is a label identifying with which class that particular pattern belongs. The comparison algorithm can be thought of as a front-end to a similarity measure, which compares two

patterns. The comparison algorithm then chooses the class that contains the most patterns which are most similar. If none of the previously boxed patterns are similar enough to the test pattern, then the pattern is put in a new class.

The comparison algorithm (or similarity measure) must clearly have some "expert" knowledge about the dataset before being able to make good judgements regarding the similarity of two patterns. Automating this is a difficult task. One possible method is to initiate the process using an expert human until the classified base is large enough, such that standard automated methods can be used. The standard automated methods could then use a larger set of explicit information (such as Fourier series or wavelet coefficient features) than the human.

4 Applying Boxing to the Wood Dataset

In applying the above described boxing algorithm to the Wood dataset, we used two selection algorithms: one based on a star's location in feature-space, and one random. In both cases, the comparison was done by a non-expert human who had extensively investigated the dataset.

We reasoned that stars located near the extreme values of feature-space (what we called "extreme cases") would be more interesting, and may exhibit more extreme behaviour [6]. They would hence be better examples of their

respective classes, particularly where there is a smooth variation between two classes (as seems to be the case in the Wood dataset).

Astrophysicists believe that stars change in behaviour over their lifetime, and many of these features reflect this variation. Hence a particular class of star may have a large amplitude (say), and this amplitude may change through its lifetime. It is important to be able to distinguish between these sub-classes (i.e., differentiate young and old stars).

The extreme-case selection in this example was manual, based on each star's locations in various two-dimensional plots of particular features, notably red log relative amplitude, red time-scale, and red relative delta time-scale. In earlier investigations, these features seemed to separate the stars best.

During the extreme-case selection and boxing process 59 stars were extracted and classified, producing a total of 10 boxes (or classes). Some examples of these stars are shown in Figure 2. Box number 9 is especially interesting, since it corresponds to a new class of variable star previously discovered from the same dataset by astronomer Peter Wood [9].

As a justification of the extreme-case selection process, see Figure 3 which shows that although these features are not perfect in extracting clusters, they are good at grouping similar sorts of stars.

During the random selection and boxing process 30 stars were extracted and classified, producing a total of 9 boxes. Some of the cleanest examples

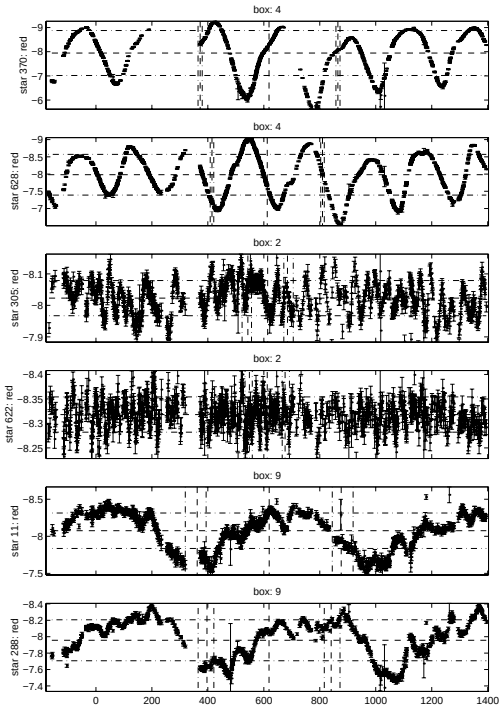

FIGURE 2: Six stars belonging to three boxes, where selection was by extreme cases.

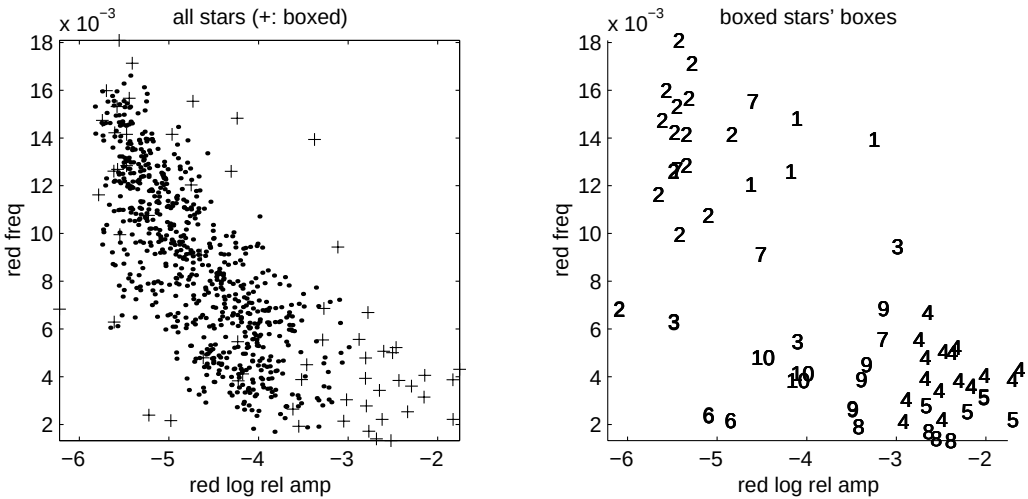

FIGURE 3: Two depictions of the extreme-case boxing process with respect to two particular features: red frequency (inverse of time-scale) and red log relative amplitude. Each point in the left plot is a star in the Wood dataset; the +'s are those that were selected for boxing using the extreme-case method. The right plot shows just those stars that were boxed identified by their label.

of these stars are shown in Figure 4. During the process, it was noted that it became increasingly difficult to make a decision regarding which class the example pattern belonged to since there were usually two potential matches in different classes. It was presumably easier to decide when using the extreme-case selection because it chose stars which were more extreme in behaviour.

5 Conclusion

The computationally simple and interpretable features provide a good stepping stone to classifying the variable stars. In addition, the boxing process was successful in identifying several groups of interesting stars.

The methods used by the astronomers are similar to the boxing, but the former has less intelligence in the selection process and more expert knowledge with regard to the physical behaviour of stars whereas our method makes the work faster since it targets interesting candidates. Astronomers using this process may find it beneficial in locating interesting stars, particularly since their expert knowledge will assist in classification.

There are many directions that can be investigated in the future.

With respect to the selection process, fully automating extreme-case selection would aid the boxing process. It would be interesting to try this selection process on other time-series data.

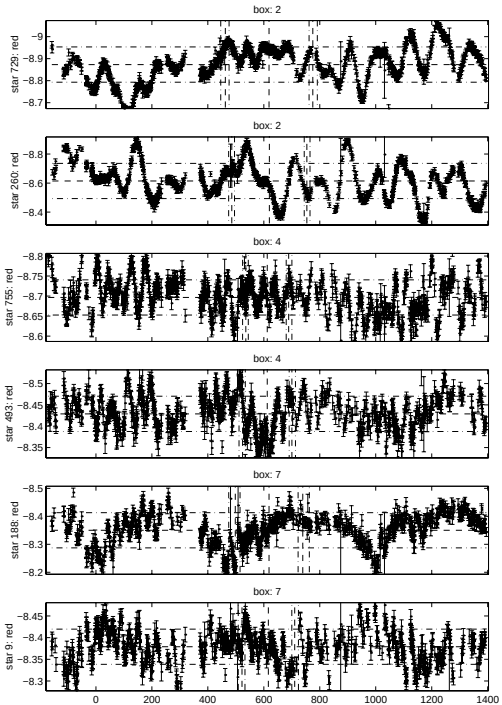

FIGURE 4: Six stars belonging to three boxes, where selection was random.

With respect to the comparison process, the use of "shape" features could be used to precisely describe the particular behaviour of the classes of variable stars. As an aid to automating comparison one could use an expert human as bootstrap mechanism to obtain an initial training set which a standard supervised learning tool could use.

Acknowledgements: Thanks to Peter Wood for the dataset, Tim Axelrod for the expert astronomical advice and the MACHO group for the MACHO database. Also thanks to John Rice for the suggestion to look at extreme cases. The work of the first author was supported by a scholarship from the Cooperative Research Centre for Advanced Computational Systems.

References

[1] C. Alcock et al. The MACHO project LMC variable star inventory. 1: Beat Cepheids-conclusive evidence for the excitation of the second overtone in classical Cepheids. *Astronomical J.*, 109:1653+, 1995. C417, C418

[2] R. S. Anderssen, F. de Hoog, and M. Hegland. A stable finite difference ansatz for higher order differentiation of non-exact data. *Bull. Austral. Math. Soc.*, 58:223–232, 1998. C422

[3] K. H. Cook et al. Variable stars in the MACHO collaboration database. In R. Stobie, editor, *Astrophysical Applications of Stellar Pulsation*, volume 83 of *ASP Conference Series*, pages 221+, 1995. IAU Colloquium 155. C417

[4] M. K. Ng, Z. Huang, and M. Hegland. Data-mining massive time series astronomical data sets: A case study. In X. Wu, R. Kotagiri, and K. B. Korb, editors, *Proceedings of the 2nd Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining (PAKDD-98)*, volume 1394 of *LNAI*, pages 401–402, Berlin, April 15–17, 1998. Springer. C418

[5] J. Reimann. *Frequency Estimation Using Unequally-Spaced Astronomical Data*. PhD thesis, Department of Statistics, University of California at Berkeley, 1994. C418

[6] J. Rice. Private communication, 1998. C425

[7] J. D. Scargle. Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data. *Astrophysical J.*, 263:835–853, 1982. C418

[8] J. D. Scargle. Studies in astronomical time series analysis. III - Fourier transforms, autocorrelation functions, and cross-correlation functions of unevenly spaced data. *Astrophysical J.*, 343:874–887, 1989. C418

[9] P. R. Wood et al. MACHO observations of LMC red giants: Mira and semi-regular pulsators, and contact and semi-detached binaries. In

T. le Bertre, A. Lebre, and C. Waelkens, editors, *IAU Symposium 191, Asymptotic Giant Branch Stars*, pages 151+, Montpelier, France, 1998. Astronomical Society of the Pacific, San Francisco. C418, C418, C426