Classification of Variable Stars
Machine Learning, Report 1
Dr Raeisi/ Dr Rahvar

Mahdi Abdollahi, Ariana Haghju, Nooshin Torabi

March 2020

1 Introduction

1.1 Variable Stars

A variable star is a star whose brightness as seen from Earth (its apparent magnitude) fluctuates.
This variation may be caused by a change in emitted light or by something partly blocking the light. Many, possibly most, stars have at least some variation in luminosity.
Variable Stars consist of two classes based on the reason of their variability. We are going to introduce the classes which we are dealing with in our data.

1.1.1 Intrinsic Variable Stars

Stars where the variability is being caused by changes in the physical properties of the stars themselves. This category can be divided into three subgroups.
Here is the group 1 that we have to deal with.
Pulsating variable stars: One calls pulsating variables the stars showing periodic expansion and contraction of their surface layers. Pulsations may be radial or non-radial[1].

- Classical Cepheid variables
Classical Cepheids (or δ Cephei variables) are population I (young, massive, and luminous) yellow super giants which undergo pulsations with very regular periods on the order of days to months.

- Type II Cepheids
Type II Cepheids have extremely regular light pulsations and a luminosity relation much like the δ Cephei variables. Type II Cepheids stars belong to older Population II stars, than do the type I Cepheids. The Type II have somewhat lower metallicity, much lower mass, somewhat lower luminosity, and a slightly offset period verses luminosity relationship.

- RR Lyrae variables
RR Lyrae stars are radial pulsators with periods in the approximate range 0.2 to 1.0 day. Their metallicities span a wide range, from about the solar value to a hundred times less. As with the Type II Cepheids, which are probably a different evolutionary stage of the same kind of star, their spectra show evidence of shock waves being propagated outwards through their atmospheres once per cycle[2].

- Anomalous Cepheids
A group of pulsating stars on the instability strip have periods of less than 2 days, similar to RR Lyrae variables but with higher luminosity. Anomalous Cepheid variables have masses higher than type II Cepheids, RR Lyrae variables, and our sun.

It may be difficult to unambiguously classify an individual star as a Type II Cepheid, an anomalous Cepheid, or a classical Cepheid, on the basis of the light curve alone, and that other information (e.g., galactic position, radial velocity, luminosity, chemical composition) is often used together with the light curve to classify Cepheids[2].

1.1.2 Extrinsic variable stars

Variable Stars where the variability is caused by external properties like rotation or eclipses. There are two main subgroups.

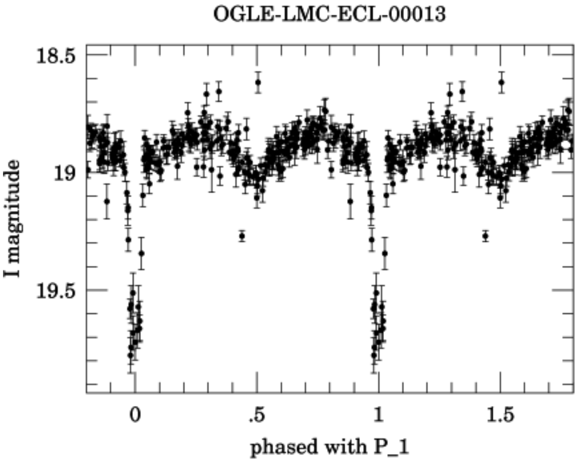

Figure 1: Light curve of a Eclipsing binary, the minimum points happen when one star passes in front of the other. Lighcurve from OGLE site.

Again some more details about the ECL stars that exist in our data.
Eclipsing binaries:
These are binary systems with orbital planes so close to the observer's line of sight (the inclination i of the orbital plane to the plane orthogonal to the line of sight is close to 90 deg) that the components periodically eclipse each other. Consequently, the observer finds changes of the apparent combined brightness of the system with the period coincident with that of the components' orbital motion[1], figure(1).

1.2 A brief introduction to Julian Date

The Julian Day Number (JDN) is the integer assigned to a whole solar day in the Julian day count starting from noon Universal time, with Julian day number 0 assigned to the day starting at noon on Monday, January 1, 4713 BC, proleptic Julian calendar (November 24, 4714 BC, in the proleptic Gregorian calendar), a date at which three multi-year cycles started (which are: Indiction, Solar, and Lunar cycles) and which preceded any dates in recorded history.
The Julian date (JD) of any instant is the Julian day number plus the fraction of a day since the preceding noon in Universal Time. Julian dates are expressed as a Julian day number with a decimal fraction added.

2 Data

2.1 OGLE

The Optical Gravitational Lensing Experiment (OGLE) is a Polish astronomical project based at the University of Warsaw that runs a long-term variability sky survey (1992-present). Main goals are the detection and classification of variable stars (pulsating and eclipsing), discoveries of the microlensing events, dwarf novae, studies of the Galaxy structure and the Magellanic Clouds[1].
We chose to work with OGLE data because it provides the most complete set of data in most (if not all) classes of variable stars.

2.2 Cleaning the data

The data we used was cleaned before by OGLE team. To construct the Objects located in the gaps between neighboring detectors in the OGLE-III mosaic camera, they used extended size of the reference images.
It can be expected that non-photometric or just cloudy nights should result in enormously large scatter of the measurements obtained during these nights. Thus, for every star brighter than $I = 18.0$, the mean magnitude and the standard deviation were derived, and then, for each point of each star the deviation from the mean magnitude was calculated and normalized by the standard deviation[3].

[1] Wikipedia.com

2.3 Details of the data we are working with

We are using OGLE data of I-band photometry for 619,308 stars. We have concluded OGLE data about each star's period, target (SMC, LMC, BLG and disk), Mean I-band magnitude and amplitude of I-band magnitude. They may help us in the feature selection, though we are going to fold the data to obtain more probable features ourselves. We have uploaded the csv file of the data in the following address:
`https://drive.google.com/open?id=1bDJoDuati-qK6JhTKSZLHvwDU0ou7EvD`

3 Plotting the Photometry Data

We chose a random star from each class and plotted the data from their photometry. The results can be seen in figure(2).

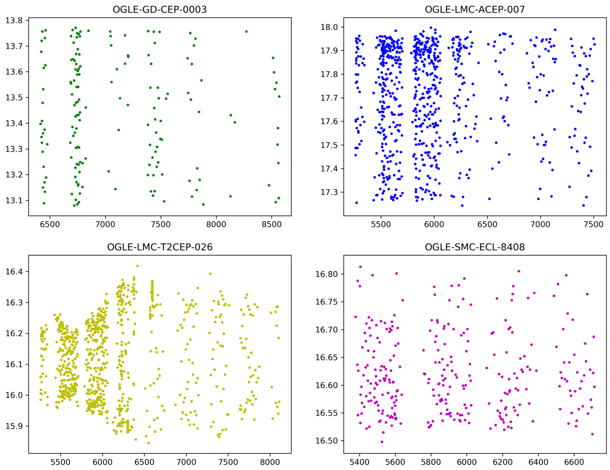

Figure 2: Plot for Photometry data, the vertical axis is the magnitude and the horizontal axis is time in Julian format

In figure(3) the data of five chosen stars is shown on one plot to compare the range of magnitudes, which can be a useful feature to classify variable stars. However, these five chosen stars may be (are) a biased representation of our data and we are going to show histograms that will include all the data, later in the report.

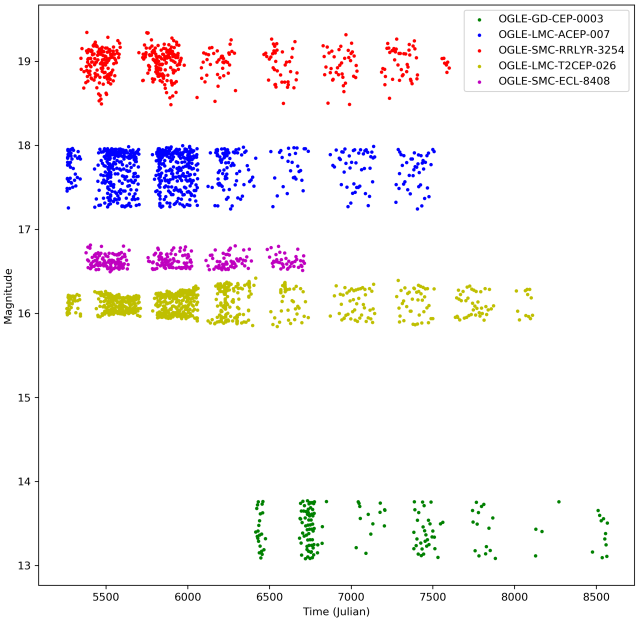

Figure 3: Plot for photometry data of 5 stars, as you see the magnitudes are all in different ranges, but this is not the case for all of the stars(and if it was, classification wouldn't be so hard).

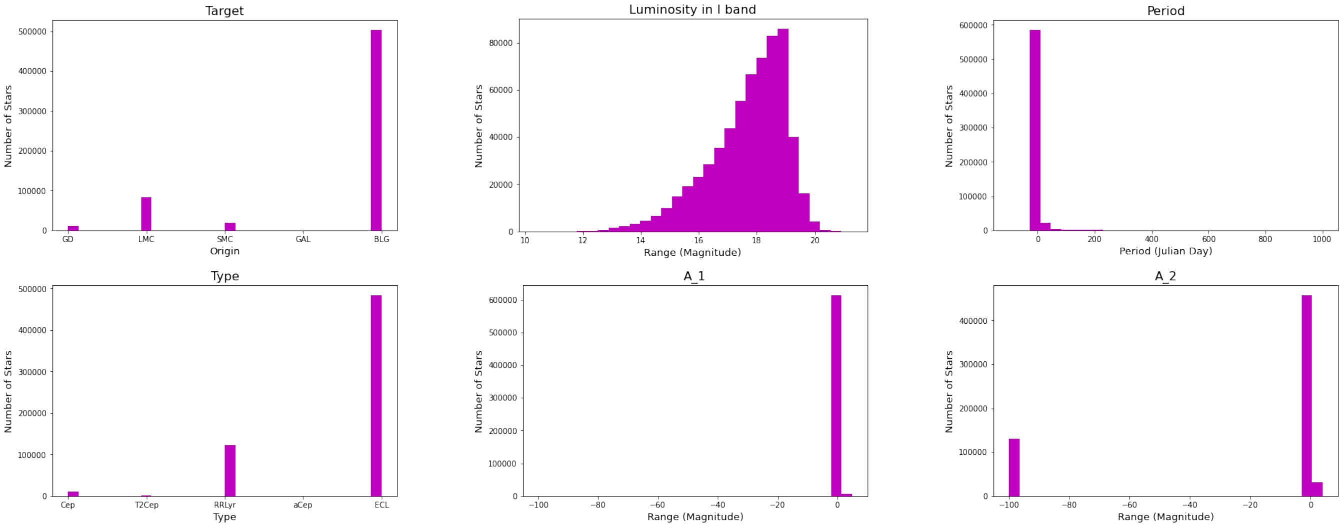

Figure 4: Histograms

4 Statistical analysis

4.1 Histograms of different features

For every column of the data frame that was useful to be plotted, a histogram has been added, figure(4).

4.2 Testing the Correlation of Binary Systems

By using the "corrcoef" function from numpy, we calculated the correlation of two time series (which in this project, it is the photometry time series). This can be used to check whether two stars are form a binary system or not.

4.3 Separating the Classes and Plotting the Magnitude of the stars of Each Class

In this section, we used a conditional structure to separate the classes and counted the number of each class.
The'Cep' class: 11078 stars
The 'aCep' class: 321 stars
The 'T2Cep' class: 1578 stars
The 'RRLyr' class: 122325 stars
The 'ECL' class: 484005 stars
In addition, a function for plotting the histogram of each class is defined. , figure(5).

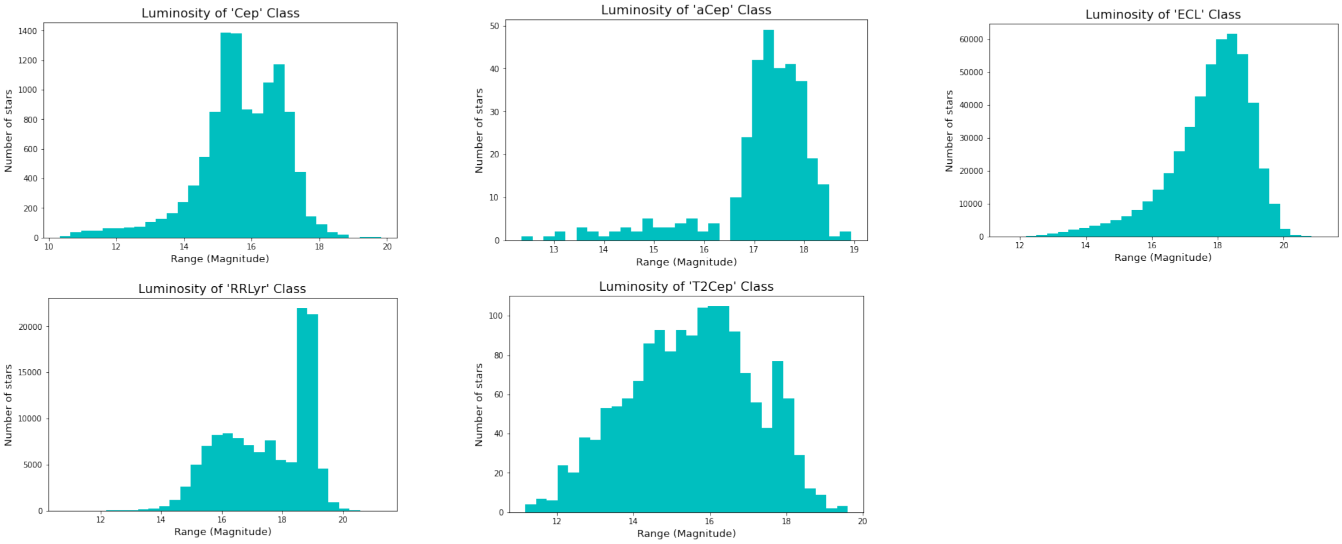

Figure 5: Histograms for Magnitude of each Class

5 Fourier Analysis:

As we saw in “Plotting the Data” section, the data is not evenly distributed. Therefore, we cannot perform the regular Fast Fourier Transform (FFT) on it.
The best algorithm to find the period of a periodic, not uniform data is by using the Lomb Scargle Periodograms. We used the Lomb Scargle function from astropy.timeseries library because we are working with astronomical data and astropy provides many further options to work with.
The results we got for our chosen stars' periods are shown in table(1).

Star ID	Type	Period (by OGLE)	Period by LS
OGLE-GD-CEP-0003	Classical Cepheid	4.6324144	4.632834
OGLE-LMC-ACEP-007	Anomalous Cepheid	0.8963997	0.896390
OGLE-LMC-T2CEP	T2 Cepheid	13.5936111	13.586635
OGLE-SMC-RRLYR-3254	RR LYR	0.6071743	0.607183
OGLE-SMC-ECL-8408	ECL	0.7533662	0.376685

Table 1: Notice the accuracy of LS in finding the period of four types

The fourth column shows the period related to the highest peak we observed on the frequency plot and the third column is the period found by OGLE team.
The related graphs are shown in figure(6). As the table(1) shows, the periods of four types were obtained with a great accuracy. The result for ECL class was significantly different. We are going to explain some of the reasons this method did not work properly for this class in the 'Period finding Algorithms' section.

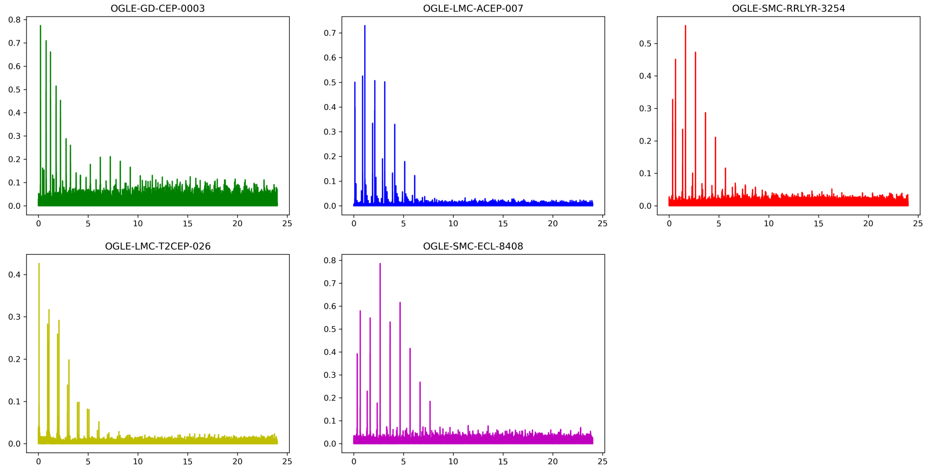

Figure 6: Frequency Plots

5.1 Period finding algorithms for variable stars:

5.1.1 Lomb Scargle (LS):

In Lomb Scargle algorithm, the time series is decomposed into a linear combination of sinusoidal functions. Scargle has derived a formula for transform coefficients that is similar to the Discrete Fourier Transformation (DFT) in the limit of evenly spaced observations.
Therefore, LS fails to give the best model for stars with significantly non-sinusoidal curves and Eclipsing Stars (ECL) usually do not have a perfect sinusoidal light curve.
One particular issue for automated period finders (particularly LS) is that they misidentify a multiple of the period as the 'true' period, i.e. the identified period, $p_i = mp_0$, where m is an integer n or $\frac{1}{n}$, and p_0 is the correct period. This is a common problem for binary systems where the half period is frequently the most significant peak in a periodogram[4].
Input parameters of Lomb Scargle contain Nyquist frequency. Nyquist frequency is half of the sampling rate of a discrete signal processing system. It is sometimes known as the folding frequency of a sampling system. The Lomb Scargle function by astropy, sets the value of Nyquist frequency to 5, by default. To probe higher frequencies we can determine the maximum and minimum frequency and the steps. Therefore, there is the possibility of losing the most repeated frequency.

5.1.2 The Box-fitting Least Squares (BLS):

The Box-fitting Least Squares (BLS) algorithm (Kovacs et al., 2002) fits the time series to periodic box-shaped functions. A box-shaped function consists of the superposition of two step functions with equal amplitude but opposite sign, and offset in time. A periodic box-shaped function alternates between a “low” and a “high” state, with a fixed fraction and phase of each periodic cycle in a given state [2]. This method gives a better period prediction for ECL class. Besides, it provides some of the light curve’s features that may help us in the classifying process.

5.1.3 Other Period finding algorithms:

Another approach is to minimize some measure of the dispersion of time series data in phase space, such as binned means(Stellingwerf 1978), variance (Schwarzenberg-Czerny 1989), total distance between points (Dworetsky 1983) or entropy (Cincotta, Mendez & Nunez 1995), which can often be regarded as an expansion in terms of periodic orthogonal step functions. Bayesian methods (Gregory & Loredo 1992, Wang, Khardon & Protopapas 2012) are also becoming common and there have even been attempts to search for periodicity using neural networks (Baluev 2012)[4].

6 Related Articles:

In this section, we are going to mention the papers related to what we want to do.

1. Udalski, Szymański and Szymański, 2015, Acta Astron.,”OGLE-IV photometry”
2. Udalski, Szymański, Soszyński and Poleski, 2008, Acta Astron., ”OGLE-III photometry”
3. Soszyński et al., 2014, Acta Astron., ”RR Lyr stars in the Galactic bulge”
4. Soszyński et al., 2015a, Acta Astron., ”classical Cepheids in the LMC and SMC”
5. Soszyński et al., 2015b, Acta Astron., ”Anomalous Cepheids in the LMC and SMC”
6. Soszyński et al., 2016, Acta Astron., ”RR Lyr stars in the LMC and SMC”
7. Soszyński et al., 2016, Acta Astron., ”Eclipsing stars in the Galactic bulge”
8. Pawlak et al., 2016, Acta Astron., ”Eclipsing stars in the LMC and SMC”
9. Soszyński et al., 2017, Acta Astron., ”Classical and anomalous Cepheids in the LMC and SMC”
10. Soszyński et al., 2017, Acta Astron., ”classical, type II, and anomalous Cepheids toward the Galactic Center”
11. Udalski et al., 2018, Acta Astron., ”Classical and type II cepheids in Galactic disk and bulge”
12. Soszyński et al., 2018, Acta Astron., ”Type II Cepheids in the LMC and SMC”
13. Soszyński et al., 2019, Acta Astron., ”RR Lyr stars in the Galactic Disk and Bulge”

These 13 papers were all written by OGLE team. The first two papers explain how OGLE gathered its data and cleaned it up. The next 11 papers are about classifying algorithms OGLE team used to classify each class of variable stars. They contain some statistical data that may come useful in feature selection later.

14. Matthew J. Graham et al., ”A comparison of period finding algorithms”, Monthly Notices of the Royal Astronomical Society, 2013

This paper compares the algorithms that were used to find the period and light curves of Catalina Real-Time Transient Survey, MACHO and ASAS data sets. It also discusses the common features used for classifying variable stars, like the shape of light curve, period and the magnitude range. 3 metrics have been introduced in this article to test the accuracy of the model.

15. Dae-Won Kim, Coryn A. L. Bailer-Jones, ”A package for the automated classification of periodic variable stars”, 2015

This paper presents a machine learning package to classify variable stars. They used OGLE and EROS-2 data to train the model. To make it survey independent, they used 16 different features to classify the stars.

16. Becker et al., ”Scalable end-to-end recurrent neural network for variable star Classification”,Monthly Notices of the Royal Astronomical Society, 2020

In this paper, first they discuss why supervised methods are not going to be successful to classify the data LSST is going to observe in 2020. Then they propose a classification model based on RNNs to perform automatic classifica-

[2] https://exoplanetarchive.ipac.caltech.edu/docs/pgram/pgram_algo.html

tion of variable stars. It learns its representation automatically and is designed to work without any pre-computed features.

7 What every member of the group did:

First Mahdi downloaded the photomerty data and put it all in a csv file to reduce the size of the data (22GB → 1.3 GB). Then he wrote the code for reading the data from google drive and data frame and the function that gives the photometry data of a star. He also wrote the ”Data” section of the report.
Ariana worked on the histograms and time correlation of ECL stars. She wrote the ”Introduction” and ”Statistical Analysis” section of the report.
Meanwhile I (which is Nooshin) worked on the period finding algorithms and wrote the code for plotting photometry data and Lomb Scargle. I wrote the ”Period finding algorithms”, ”Plotting Photometry data”, ”Fourier analysis” and ”Related Articles” section of the report. Finally I edited and wrote our report in latex.
and Here is the result.

References

[1] *General Catalouge of Variable Stars.* URL: http://www.sai.msu.su/gcvs/gcvs/.

[2] c. JASCHEK. *LIGHT CURVES OF VARIABLE STARS, A Pictorial Atlas.* Cambridge University Press, 1996.

[3] Udalski A. et al. “The Optical Gravitational Lensing Experiment. Final Reductions of the OGLE-III Data”. In: *Acta Astronomica* 58 (2008).

[4] Graham Matthew J. et al. “A comparison of period finding algorithms”. In: *Monthly Notices of the Royal Astronomical Society* 434.4 (2013), pp. 3423–3444. DOI: https://doi.org/10.1093/mnras/stt1264.