CLASSIFICATION OF VARIABLE STARS

D. Tarczay-Nehéz[1, 2], R. Szabó[1,2] and Z. Szeleczky[3]

Abstract. The most recent space telescopes (e.g. Kepler, K2, Gaia, TESS) and sky surveys (e.g. SSDS, and the forthcoming LSST) provide huge amounts of data, and creating challenges of data processing. These huge amounts of data need to be analyzed with fast and effective robotic computer programming techniques. Machine learning algorithms are becoming popular in astronomy, as they can play a key role in the automatic classification of variable stars. In this work, we present our machine learning algorithm for searching variable stars; it is based on statistical data of light-curves that represent the brightness variability of the Cepheid variables observed by Gaia (see Gaia Data Release 2 in Gaia Collaboration et al. 2018)*.

Keywords: Stars: variables: Cepheids, Methods: data analysis

1 Introduction

The most recent rapid increase in the amount (and quality) of data reveals the importance of applying new automated classification and data processing techniques to the observations. Machine learning techinques have been used for classifications of time-series data since the early 2000s (see e.g. Belokurov et al. 2003; Mahabal et al. 2008; Richards et al. 2011, and the references therein).

To classify variable stars, we have used supervised machine learning techniques, and have compared the accuracy against the Cepheid variables listed in the Gaia DR2. The Gaia archive provides an easy access to all classified variable stars. The Gaia DR2 archive consists of 9575 Cepheid variables (8890 classical, 585 Type II, 100 anomalous Cepheids; see the distribution in Fig. 1)

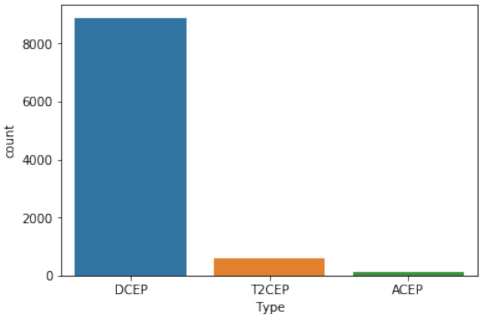

Fig. 1. The distribution of three types of Cepheid variables in the Gaia DR2 archive.

To carry out classifications, we used the Fourier parameters of the folded light-curves listed in the data base:

- period values (fundamental (PF), first overtone (P1O), second overtone (P2O), third overtone (P3O));
- peak-to-peak amplitude (in G band);
- phase differences (ϕ_{21} and ϕ_{31});
- amplitude ratios (R_{21}, R_{31}).

[1] Konkoly Observatory, Research Centre for Astronomy and Earth Science, Konkoly-Thege Miklós 15-17, H-1121
[2] MTA CSFK Lendület Near-Field Cosmology Research Group
[3] University of Liverpool, Liverpool, UK
*https://www.cosmos.esa.int/web/gaia/dr2

2 Classification

We compared the precision of 5 different supervised machine learning algorithms to classify Cepheid variables. For the classification process, we used the scikit-learn Python machine learning library†. In each case, the first step is to split our data into training sets and test sets. We then investigated 3 cases (see Table 1):

- 90% training + 10% test set (T10);
- 85% training + 15% test set (T15);
- 80% training + 20% test set (T20).

Table 1. Accuracy of the machine learning algorithms in the test sample of the Gaia DR2 Cepheids.

Algorithm	**Accuracy**		
	T10	**T15**	**T20**
Logistic regression	0.93	0.93	0.94
Decision tree (depth = 3)	0.93	0.94	0.94
K-Nearest Neighbor	0.93	0.94	0.95
Support Vector Machine	0.93	0.93	0.94
Gaussian Naive Bayes	0.33	0.26	0.26

Fig. 2 shows the scatterplot-matrix of the Fourier parameters of the Cepheid variables used for the classification, which shows the correlations in each parameter, i.e., there is a weak correlation between the amplitude ratios.

3 Example: decision tree

Figure 3 shows the decision tree for our test sample of Gaia DR2 Cepheids. In this case we have set a depth of only 3, meaning that the longest path from the ”root” to the ”leaf” is 3.

The first step in the decision tree algorithm is to analyze one of the given parameters (e.g. X_0 in Fig. 3): is this value smaller than 0.029? The decision can be either true or false. The algorithm moves forward, checking different parameters at each depth. The gini score describes the purity of each leaf/node. If gini = 0, it means that only 1 single class exists in that leaf/node and the decision is pure.

The example describes the number of test sample elements in each node. In the beginning, for the T10 case, our test sample contains 8617 Cepheids. That value represents the number of test sample elements in each category. In fact it contained 91 anomalous Cepheids, 8004 classical Cepheids and 522 Type II Cepheids. The decision tree algorithm may suffer from overfitting if the depth is too large. To minimize errors, one should use a Random Forest method, which is a collection of decision trees.

Technical help from Attila Bódi, László Molnár is gratefully acknowledged. We also acknowledge financial support from the Momentum grant No. LP2018-7/2019 of the Hungarian Academy of Sciences, the Hungarian OTKA Grant No. 119993 and from the Hungarian NKFIH project K-115709.

References

Belokurov, V., Evans, N. W., & Du, Y. L. 2003, MNRAS, 341, 1373

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, A1

Mahabal, A., Djorgovski, S. G., Turmon, M., et al. 2008, AN, 329, 288

Richards, J. W., Starr, D. L., Butler, N. R., et al. 2011, ApJ, 733, 10

† `https://scikit-learn.org/`

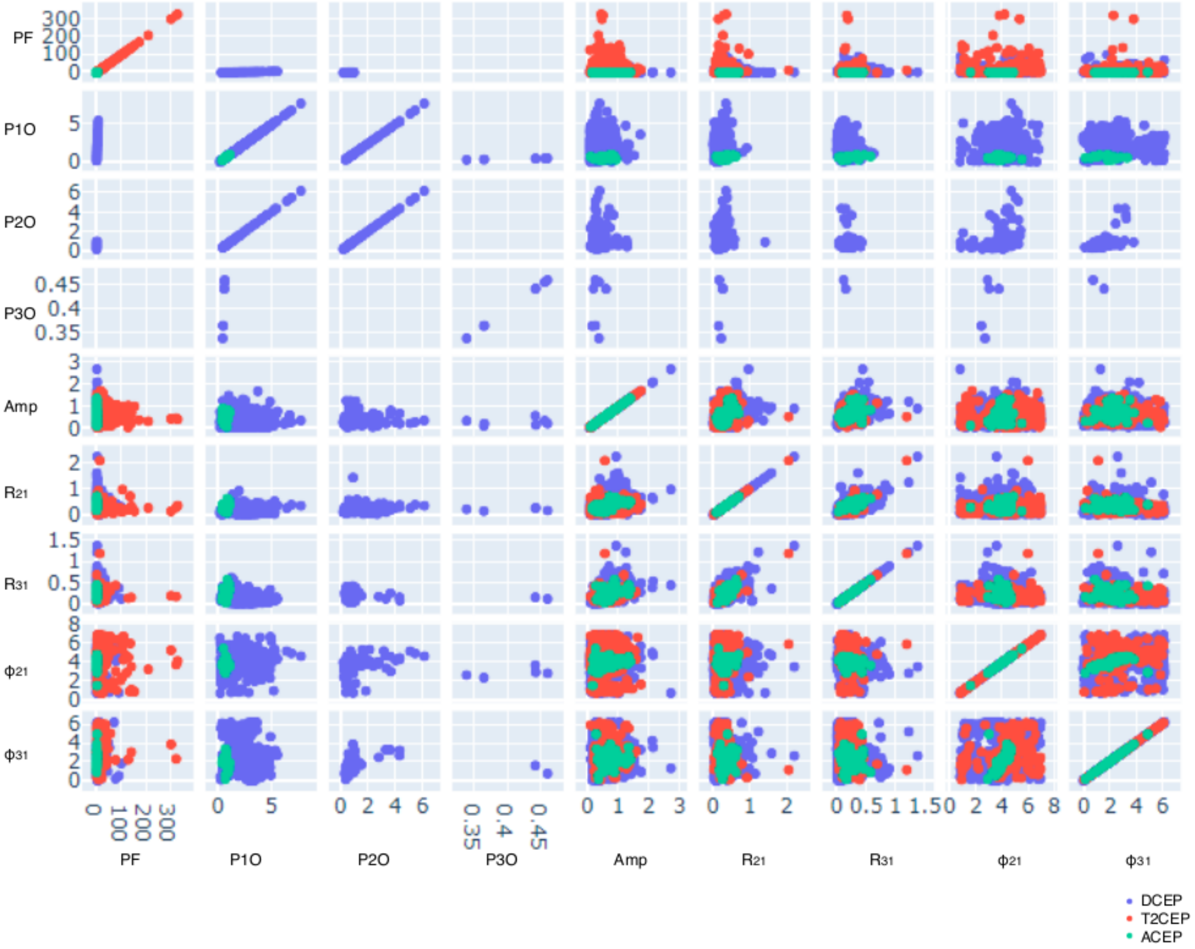

Fig. 2. Scatterplot-matrix of the Fourier parameters on GDR2 Cepheids.

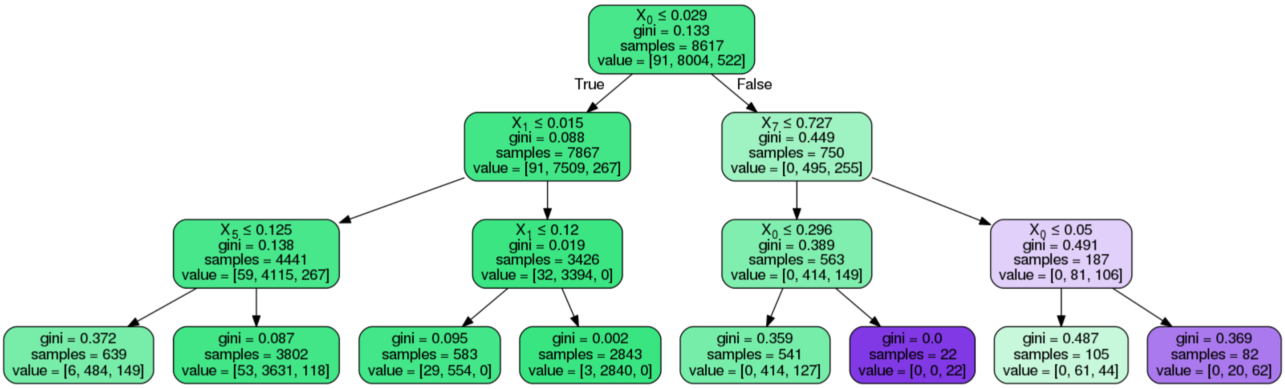

Fig. 3. Decision tree for Gaia DR2 Cepheid variables (with depth = 3) in the case of T10.