Variable Stars Classification with the Help of Machine Learning

Kirill Naydenkin[1], Konstantin Malanchev[2,3], and Maria Pruzhinskaya[2]

[1] Physics Faculty, Lomonosov Moscow State University, Leninskii Gori 1, 119234 Russia, rtut654@gmail.com
[2] Lomonosov Moscow State University, Sternberg Astronomical Institute, Universitetsky pr. 13, Moscow, 119234, Russia
[3] National Research University Higher School of Economics, 21/4 Staraya Basmannaya Ulitsa, Moscow, 105066, Russia

Abstract. With the appearance of modern technologies such as CCD-matrices, large telescopes and computer networks the precision of our observations increased immensely. On the other hand, such accurate and complex data formed TBs large data bases which are very fragile and unattainable for the treatment by classical methods. The scales of this problem can be seen especially in variable star sky surveys. For many terabytes of data one has to classify all the stars in catalog to find stars of particular type of variability. This problem is known as very important since almost every part of modern astrophysics is interested in new objects to study. In some fields like cosmology, this question is very vital due to high demand for new data of model-anchors like Cepheids or supernova stars. To facilitate this task many machine learning based algorithms were proposed (Richards et al., 2011 [2]). In this study we perform a way to classify the Zwicky Transient Facility Public Data Release 1 catalog onto variable stars of different types. As the priority classes we set Cepheids, RR Lyrae and $\delta Scuti$. "One vs all" classification technique revealed highly accurate results on validation data, concretely 0.90–0.95 with ROC-AUC metrics.

1 Introduction

The Zwicky Transient Facility (ZTF) is a 48-inch Schmidt telescope with a 47 sq. deg. field of view at the Palomar Observatory in California. This large field of view ensures that the ZTF survey can scan the entire northern sky every night. The ZTF survey started on 2018 March 17. During the planned three years survey, ZTF is expected to acquire $\sim$ 450 observational epochs for 1.8 billion objects. Its main scientific goals are the physics of transient objects, stellar variability, and solar system science (Graham et al. 2019 [3]; Mahabal et al. 2019 [5]).

In this paper we made an attempt to classify the first data release of ZTF survey (DR1) which contains data acquired between March and December 2018, thus covering a timespan of around 290 days. The first data release includes more than 800 thousand light curves observed in both zr and zg passbands.

The further steps of classification ZTF DR1 imply highly accurate data treatment in the beginning. To have more clear awareness about the stage of data treatment it is important to divide it onto small parts. First of all, in demand of supervised machine learning algorithms it is important to prepare precise labeled data which in our case consist of star-catalog with coordinates and types of 55 thousands variable stars of the General Catalog of Variable Stars (GCVS, Samus et al. 2017 [4]). Even though this number is relatively small compared to the size of ZTF data release, there are some ways to generate the data on the basis of accurate initial frames.

1.1 Data Preparation

Since ZTF DR1 marks the same objects observed in different passbands and/or different sky fields with different identifiers (IDs), cross-match can yield more than one ZTF DR1 ID for given GCVS object. Taking into account that ZTF ranged in $\approx 12^m - 21^m$, it is reasonable to remove those stars which do not belong to that range (with slight offset due to star magnitude fluctuations with time). After this filtering, 43 thousands of total 55 thousands GCVS stars remained.

Fig. 1 represents the distribution of number of cross-matched ZTF DR1 objects. To find the real range of difference in coordinates between catalogs we measured distances for the case when the only ID was found for a given GCVS object and for the case of two IDs separately in both filters. As the result, we received that only relatively small amount of objects has more than 0.1" difference. This coordinates distinction can be explained by catalogs' positional inaccuracy, which for GCVS is 0.1".

To this point we had matched labeled GCVS with objects from ZTF. As we mentioned before, ZTF DR1 contains photometry in two passbands (zg, zr) which we can be used either separately or in combination. We have to note that regardless shrinking the search radius to 1.5" we still can receive multiple results — from a light curve in only one filter up to a few series in both (Fig. 2). Taking

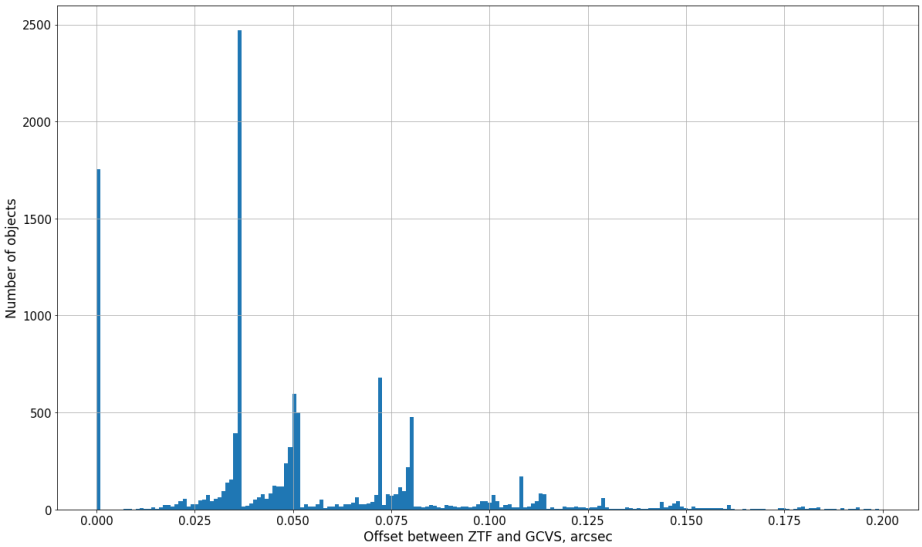

Fig. 1. A number of cross-matched ZTF DR1 IDs distributed by the angular separation from GCVS objects.

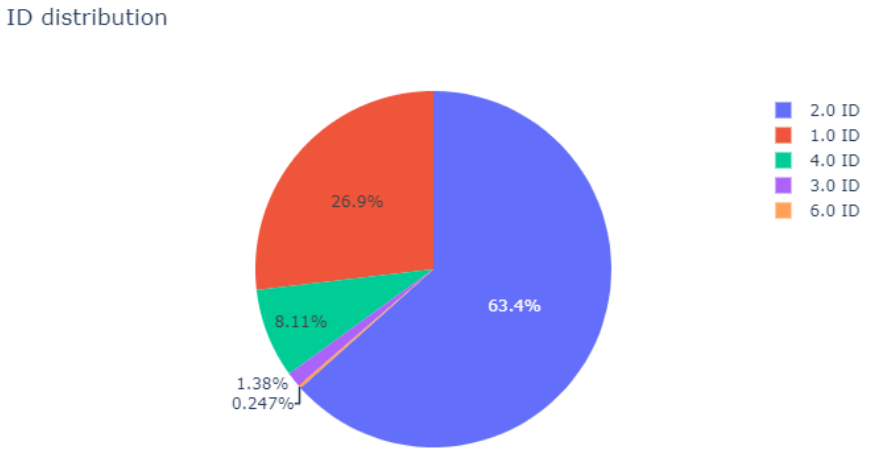

Fig. 2. Distribution of ID-numbers among all responses.

into account such asymmetry we made up three homogeneous sets: 19k objects in g band, 14k in r band, and 13k in combination (setting up threshold for minimal amount of observations in a pair).

2 Objects of Interest

In our study we took some particular classes of interest among the variable stars: Cepheids, RR Lyrae and δ Scuti. It is important to note that we do not use multi-class classification for chosen types, every type goes through one-vs-all approach.

First of all, let's describe the physical nature of chosen types. Cepheids are pulsating stars, which radius and brightness (as well as the temperature) change with time. Cepheids stars are well known for the dependency between luminosity and pulsating period, this property makes them important indicators of cosmic distances. RR Lyrae stars can also be used as standard candles for distance measurements, though they do not follow a strict period-luminosity relationship at visual wavelengths. δ Scuti as well as Cepheids are important standard candles and have been used to establish the distance to many large clusters around the center of our galaxy.

3 Analysis

The next step toward the data classification is to engineer features out of light curves. To start we created three most valuable sources of information (Richards et al., 2011 [2]): magnitude amplitude range, the main peak period and power of Lomb–Scarge periodogram (Lomb 1976 [7]; Scargle 1982 [8]). For this purpose we used astropy [13] library-based LombScargle() function which allows us to define both positional coordinates of Lomb–Scargle periodograms peak.

One possible way to select a set of variable stars out of ZTF DR1 is to use the Lomb–Scarge periodogram. Such attempts were made recently and yielded to the strong results (Chen et al 2020 [6]).

On the one hand, one could possibly observe the importance of different features for different types of variable stars from (Richards et al., 2011 [2]). On the other hand, ZTF light curves can be slightly different from the data studied in the article and this could possible change the picture.

After choosing metric for result evaluation we worked with validation set to find an optimal list of features for the classification task

in ZTF DR1. Richards et al., 2011 [2] presented a table of pairwise random forest feature importance for all basic variable types. As one can see coordinates of first peaks of Lomb–Scarge periodogram and their ration (ratio of periodogram's frequencies) have a significant influence on majority of classes. Basic characteristics of a signal such as amplitude, std, skew, median absolute deviation also have a strong correlation with correct classification of all classes. For our one-vs-all approach we took 17 the most strongest features out of Richards et al., 2011 [2] — coordinates of first four periodogram's peaks, frequencies ratio, mentioned basic signal properties and few additional such as trend angle. To implement these features we used python libraries — ASTROPY and statistical analysis from SCIPY. It is obvious, that application of this approach to one-vs-all classification leads to imbalance in labeled data because even the largest types consist out of less then 15% of the train set. To deal with imbalanced data few possible options are available. First of all, we can use one of the ways to generate rare sample, for example random sampling from a chosen distribution. On the other hand, we can divide the more abundant class into N distinct clusters and train each of N classifiers on one of the distinct clusters and on all of the data from the rare type. After that ensemble the result from all models. As far as we address the imbalanced problem with significantly low number of minor class examples, it is more efficient to use over-sampling techniques instead of under-sampling. For this purpose we took Synthetic Minority Oversampling Technique, or SMOTE. SMOTE first selects a minority class instance A at random and finds its k-nearest minority class neighbors. The synthetic instance is then created by choosing one of the k-nearest neighbors B at random and connecting A and B to form a line segment in the feature space. The synthetic instances are generated as a convex combination of the two chosen instances A and B. Using SMOTE from imblearn python library we updated the minority class by oversampling to have 10 percent the number of examples of the majority class, then used random undersampling to reduce the number of examples in the majority class to have 50 per cent more than the minority class.

For the first trial of binary classification we used Cepheids. The performance is shown in Table 3.

Table 1. Different metrics for Cepheid stars on validation set.

Passband	Accuracy	Precision	ROC AUC	F1
zg	0.869	0.921	0.873	0.857
zr	0.823	0.891	0.811	0.788
Overall	0.814	0.855	0.778	0.734

4 Discussion

4.1 Comparison with the Previous Studies

Comparing our approach with recent study of Chen et al 2020 [6], which performed a large number of new RR Lyrae, Cepheid and δ Scuti we used machine learning classification instead of directly measured distances in parameter space. Machine learning way of data classification can perform better results because it relies on both successes and failures to estimate the reliability of certain candidate. Moreover, applying hierarchical classification we have an ability to track mistakes of our model on different levels — end up with highest possible accuracy on large types which generally have some physical difference like binaries, pulsating, eruptive or rotating stars. On the next step large types break into sub-types.

4.2 Future Plans

As the future structure of the research we can mark the following steps. First of all, to obtain some initial results we have to complete the process of validation of our model on GCVS data for our classes of interest. After that we can start to classify ZTF data either after filtrating with Lomb-Scarge periodograms or making features directly for each light curve. As the next step we will use more public available labeled catalogs such as ASAS-SN [9], ATLAS [10], Catalina catalog of periodic variable stars [11], the Gaia catalog of RR Lyrae and Cepheids [12]. At this point, we will also introduce data generation which can significantly improve the accuracy of the results.

At the final step we will take different classification tools of ML: classical Random Forest and XGBoost models with the main hyper parameters to choose and imbalanced learning to make hierarchical classification.

5 Conclusions

In this work we consider the machine learning technique as a perspective approach for the classification task. We use ZTF data release 1 that contains zg and zr photometry of the variable objects. Our preprocessing procedure includes several steps. First, we have to prepare the datasets with labels to use them later for training the model and testing the accuracy of the method. For this purpose the General Catalogue of Variable Stars is used. After cross-matching the catalog with ZTF DR1 we found 19k common objects in zg-band, 14k in zr-band, and 13k in combination of passbands. Then, we have to choose the appropriate features to describe the light curves. As a starting point, we tried the magnitude amplitude range, the main peak period and power of Lomb—Scarge periodogram.

There are many types of variable stars differ by the underlying physical processes or their observational appearance. As the objects of interest we chose RR Lyrae, Cepheid and δ Scuti. We applied binary classification technique to Cepheid stars and found quite good results on the validation dataset. The work done is a preparatory step towards the further thorough machine learning classification of the variable stars in ZTF data.

Acknowledgments. K. Malanchev and M. Pruzhinskaya are supported by RBFR grant 20-02-00779. The authors acknowledge the support by the Interdisciplinary Scientific and Educational School of Moscow University "Fundamental and Applied Space Research".

References

1. Pruzhinskaya, M. V., Malanchev, K. L., et. al.: Anomaly detection in the Open Supernova Catalog. Monthly Notices of the Royal Astronomical Society 489(3):3591–3608 (2019). https://doi.org/10.1093/mnras/stz2362.
2. Richards et al.: On Machine-Learned Classification of Variable Stars with Sparse and Noisy Time-Series Data. The Astrophysical Journal. 733(10) (2011). 10.1088/0004-637X/733/1/10
3. Graham, M. J., et. al.: The Zwicky Transient Facility: Science Objectives. Publ. Astron. Soc. Pac. 131(1001):078001 (2019)
4. Samus, N.N., Kazarovets, E.V., Durlevich, O.V., Kireeva, N.N., Pastukhova E.N.: General Catalogue of Variable Stars: Version GCVS 5.1. Astronomy Reports 61(1):80-88 (2017)
5. Mahabal, A., et. al.: Machine Learning for the Zwicky Transient Facility. The Astronomical Society of the Pacific 131(997) (2019)
6. Chen, et. al. The Zwicky Transient Facility Catalog of Periodic Variable Stars. arXiv:2005.08662 [astro-ph.SR] (2020). 10.3847/1538-4365/ab9cae
7. Lomb, N.R.: Least-squares frequency analysis of unequally spaced data. Astrophys. Space Sci. 39:447–462 (1976). https://doi.org/10.1007/BF00648343
8. Scargle, J. D.: Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. Astrophysical Journal 263:835-853 (1982)
9. All-Sky Automated Survey for Supernovae. https://asas-sn.osu.edu/
10. The Herschel ATLAS. https://www.h-atlas.org/
11. The Catalina Surveys Data Release 2. http://nesssi.cacr.caltech.edu/DataRelease/
12. Gaia Archive at ESA. https://gea.esac.esa.int/archive/
13. Astropy Project. http://www.astropy.org