EPJ Web of Conferences **152**, 03011 (2017) DOI: 10.1051/epjconf/201715203011
Wide-Field Variability Surveys: A 21st Century Perspective

Machine learning techniques to select variable stars

Alejandro García-Varela[1,⋆], *Muriel* Pérez[1,2,⋆⋆], *Beatriz* Sabogal[1,⋆⋆⋆], and *Adolfo* Quiroz[2,⋆⋆⋆⋆]

[1] *Universidad de los Andes, Departamento de Física, Cra.1 No.18A-10, Edificio Ip, Bogotá, Colombia*
[2] *Universidad de los Andes, Departamento de Matemáticas, Cra.1 No.18A-10, Edificio H, Bogotá, Colombia*

Abstract. In order to perform a supervised classification of variable stars, we propose and evaluate a set of six features extracted from the magnitude density of the light curves. They are used to train automatic classification systems using state-of-the-art classifiers implemented in the R statistical computing environment. We find that random forests is the most successful method to select variables.

1 Introduction

Machine learning techniques have proved to be quite useful in classification of variable stars. In the existing literature, quantities related to the magnitude density of the light curves and their Fourier coefficients are chosen as features. However, the calculation of Fourier coefficients is computationally expensive for large data sets. In order to perform a supervised classification, we propose and evaluate a set of six robust descriptive statistics that can be calculated efficiently and do not need to be checked externally. We calculate this set of features for OGLE-III variables belonging to the Milky Way and the LMC and SMC galaxies, classified as Cepheids (Ceph), δ Scuti (δ-Sct), Eclipsing Binaries (EBS), Long Period Variables (LPV), RR Lyræ (RRLyr), Type 2 Cepheids (T2Ceph) and a set of Be Star Candidates (BeSC) reported in the literature. We evaluate the performance of the features over the following classifiers: K-nearest neighbors, classification trees, random forests, support vector machines, and gradient boosted trees. We use 10-fold cross-validation to estimate the recall and precision of these classifiers ([1]).

2 Selection of variable stars

In order to avoid quantities sensitive to outliers, we choose six robust estimators. To visualise how data look in the six-dimensional feature space, we use the t-distributed Stochastic Neighbour Embedding (t-SNE) visualisation technique ([2]). Figure 1 shows a t-SNE plot, where the different variability classes are separated in the six-dimensional feature space. Some overlapping of the classes is also evident. After the training process, we validate the classifiers on a data set from the OGLE-IV database. Our classifiers yield correct classifications with high probability, which shows that our proposed set of features can be used to characterise different variability types. We find that the random forest classifier produces the best results on the validation samples (see Table 1).

⋆ josegarc@uniandes.edu.co
⋆⋆ mf.perez648@uniandes.edu.co
⋆⋆⋆ bsabogal@uniandes.edu.co
⋆⋆⋆⋆ aj.quiroz1079@uniandes.edu.co

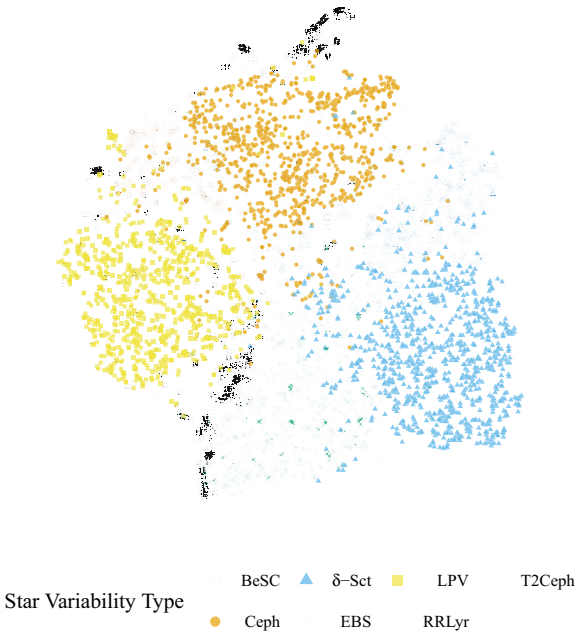

Figure 1. Visualisation of the six-dimensional feature space by t-SNE. The axes are omitted because the scale of this embedding carries no meaning.

Table 1. Confusion matrix. Classification results of the random forest method on the OGLE-IV data set.

Prediction	Ceph	δ-Sct	EBS	LPV	RRLyr	T2Ceph	Other
BeSc	1		21	3			108
Ceph	**126**	1	42		10	1	52
δ-Sct		**146**	209		9		316
EBS		2	**1110**	2	13		676
LPV	1		105	**2790**		1	226
RRLyr	3	10	19		**652**		86
T2Ceph	4		26	4	2	**3**	9
Total	135	159	1532	2799	686	5	1473

Acknowledgments: Authors acknowledge support from Departamento de Física and Facultad de Ciencias, Universidad de los Andes, through CENIF program.

References

[1] Pérez-Ortiz, M. F., García-Varela, A., Sabogal, B., & Quiroz, A., A&A, submitted (2017)
[2] Van der Maaten, L., & Hinton, G., The Journal of Machine Learning Research, **9**, 85 (2008)