Hierarchical Machine Learning Classification of Eclipsing Binaries in the OGLE Data

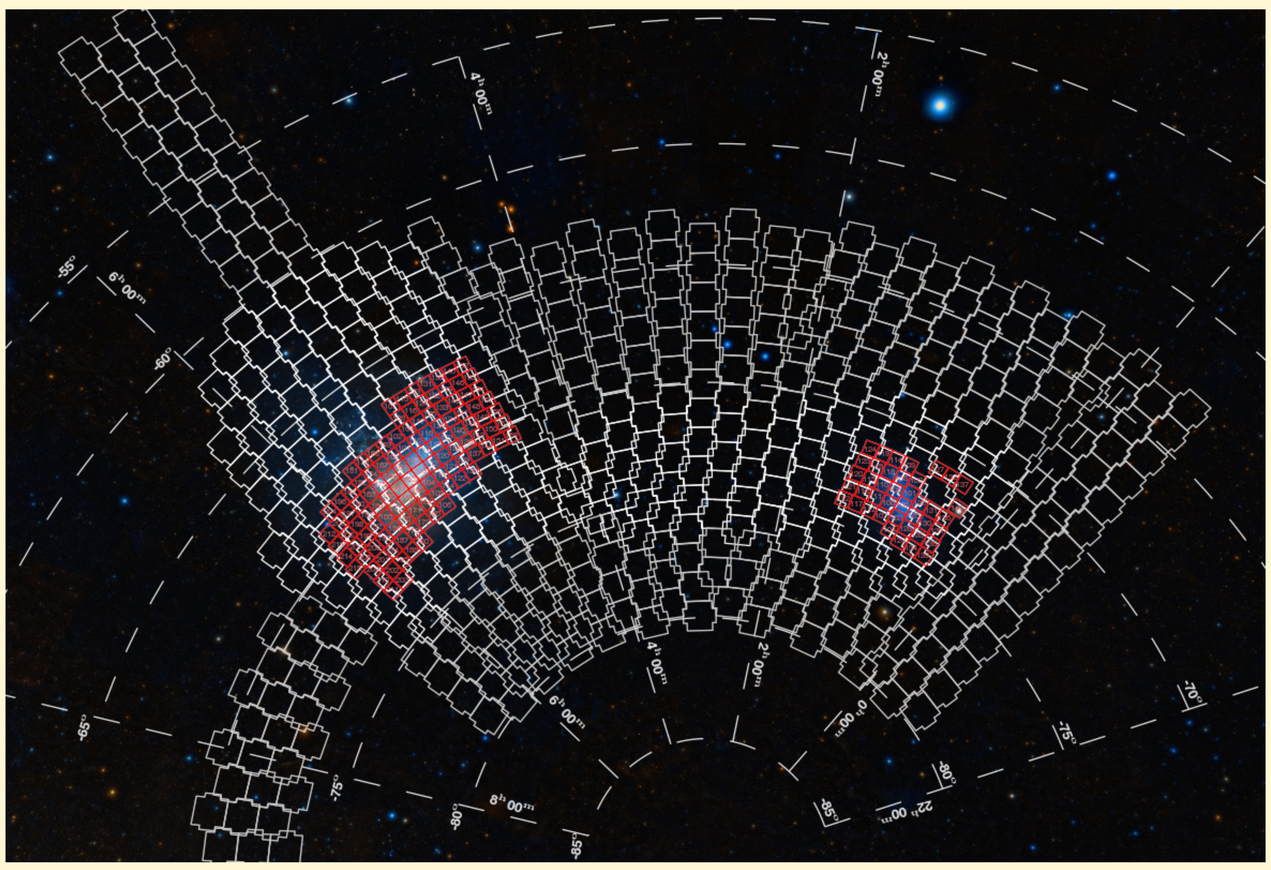
Michał Pawlak

Warsaw University Observatory

mpawlak@astrouw.edu.pl

The OGLE Project

The OGLE project has been operating since 1992. Three phases of the project were successfully completed: OGLE-I (1992-1995), OGLE-II (1997-2000) and OGLE-III (2001-2009). Since 2010 OGLE-IV is in operation. During the first phase of the project, 1.0-m Swope Telescope was used. From the OGLE-II, 1.3-m Warsaw Telescope at the Las Campanas Observatory in Chile is dedicated for the project. The number of monitored object increased significantly from about one million in OGLE-I to about 40 million in OGLE-II and more than 400 million in OGLE-III. In OGLE-IV the number of observed sources reached 1 billion. Most of the observations (about 90%) are taken in *I* filter, while the remaining in *V* filter.
The OGLE project original aim was the detection and analysis of microlensing events, however its scientific results cover much wider field of study. One of the most significant of them is discovery and analysis of hundreds of thousands of variables stars.



OGLE-III and OGLE-IV

The figure above presents the comparison of the OGLE-III (red contours) and OGLE-IV (white contoures) sky coverage in the Magellanic Clouds. The OGLE-III catalogue of variable stars contains more than 26 000 eclipsing binaries in the LMC and 6000 in the SMC, making a perfect training set for machine learning selection and classification of stars observed in OGLE-IV which overs much larger area.

Hierarchical classification

First Step

The OGLE-III catalg of eclipsing binaries in the LMC (Graczyk et al. 2011) was used to create the training set. The entire sample of known eclipsing binaries from one of the OGLE-IV fields containing about 1000 object was complemented with a matching number of other, randomly selected stars from the same filed. BLS algorithm (Kovacs et al. 2002) is used to determine the period for each of the objects, as well as series of statistics including signal to noise ratio of the periodogram, duration and depth of eclipse, Chi^2 of the fit, white and red noise. The BLS statistics are used as features for Random Forest (Breiman 2001) classification algorithm. To reduce the number of false positives, the prior on probability is set to 80%. Recall of such classification is 93%.

Second Step

For the next step of the classification, the training set is composed of the same eclipsing binaries as previously and other, non-eclipsing stars that passed the first step. The ratio of false to true detections after the firs step is about 6:1. A part from BLS parameters a series of statistical features (standard deviation, skewness, kurtosis, third and first quartile difference) is used. Classification is again performed with Random Forest. Recall of this step, evaluated on the training set is 88% and precision 89%.

Overal Result

The overall performance have been tested independently tested on one another OGLE-IV field not used in the training set. Recall of the entire method in respect to the OGLE-III catalog is estimated to be about **81.5%**. This is consistant with the recall derived from the cross-validation methode which is **83%**.

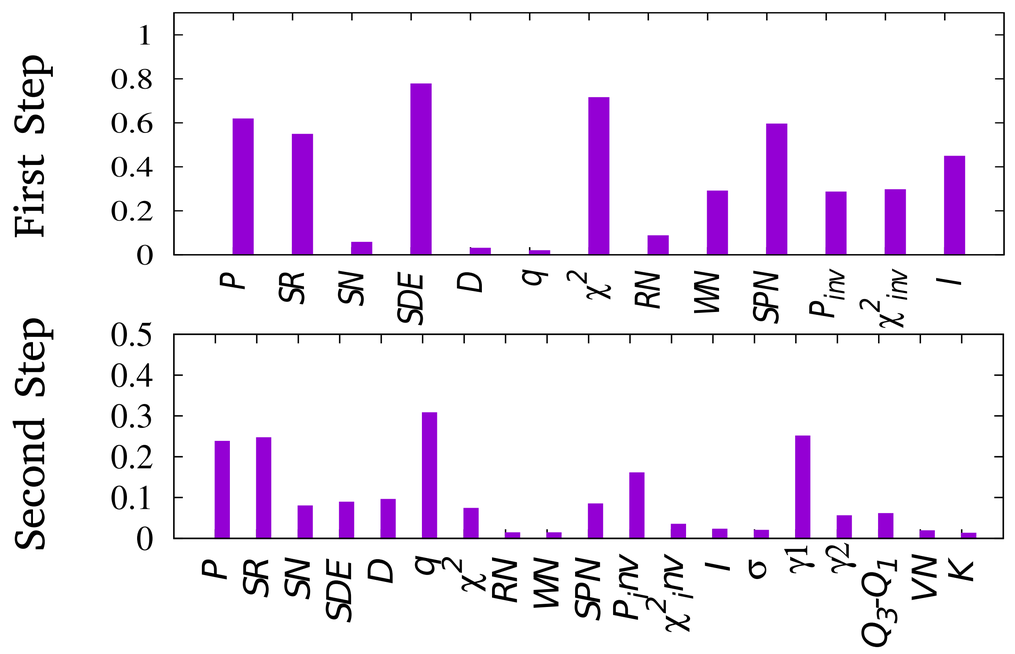
Parameter	Description
P	Period
SR	Singnal Residue
SN	Signal to Noise Ratio
SDE	Signal Detection Efficiency
D	Depth of transit
q	Fraction of the phase in ransit
χ^2	χ^2 of the transit model fit
RN	Red Noise
WN	White Noise
SPN	Signal to Pink Noise Ratio
$P_{\rm inv}$	Period of inverse transit
$\chi^2_{\rm inv}$	χ^2 of the inverted transit model fit
I	mean I-band magnitude

Parameter	Description
σ	standard deviation
γ_1	skewness
γ_2	kurtosis
$Q_3 - Q_1$	difference between third and first quartile
VN	von Neumann index
K	Stetson K index

Parametrization

Set of parameters used in the First Step (left) and Second step (right). Evaluation of the information gain given by the parameters shown in the figure below.

Information gain



References:

Breiman 2001, Machine Learning, 45, 5

Graczyk et al. 2011, Acta Astronomica, 61,103

Kovacs et al. 2002, A&A, 391, 369

Pawlak et al. 2013, Acta Astronomica, 63, 323

Acknowledgement:

This work has been supported by Polish National Science Center grant no. 2014/13/N/ST9/00075