Identification of ultracool dwarfs in J-PLUS DR2 using Virtual Observatory tools and machine learning techniques

P. Mas-Buitrago[1,2], E. Solano[1,2], A. González-Marcos[3], et al.

[1]Centro de Astrobiología (CAB), CSIC-INTA, Madrid, Spain

[2]Spanish Virtual Observatory

[3]Universidad de La Rioja, La Rioja, Spain

Summary

We present the search for ultracool dwarfs (UCDs, spectral types later than M7 V) performed across the entire J-PLUS second data release. With a methodology driven by the use of multiple Virtual Observatory (VO) tools and services, we identified a total of 7829 new candidate UCDs.

One of the main objectives presented in P.Mas-Buitrago et al. (2022, in review) is to explore the ability to reproduce this search with a purely machine learning (ML)-based methodology.

VO-based search

We start with a pre-screening process in which we use three different approaches to obtain a shortlist of candidate UCDs. In two of these approaches, we only keep J-PLUS objects with reliable parallax and proper motion in Gaia EDR3, respectively. We apply a colour cut of $G - G_{rp} > 1.3$ mag to shortlist the candidate UCDs.

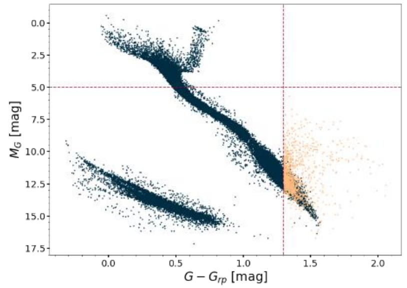

Fig. 1. Location of the objects shortlisted as candidate UCDs (yellow) via astrometric selection in a colour-magnitude diagram.

In the third approach we shortlist the candidate UCDs using only a J-PLUS colour cut of $r - z < 2.2$ mag. Finally, we use the tool VOSA [1] to estimate the effective temperature of the candidates and keep only those with $T_{eff} \leq 2900\ K$.

ML-based methodology

Using the J-PLUS DR2 data from a 20x20 deg^2 region of its sky coverage, we built seven different colours as input features: $i - z, r - i, J0861 - i, z - J0861, (i - z)^2, (r - i)^2$ and $r - z$. Then, we labeled the instances as positive or negative class using the candidate UCDs obtained with the VO-based methodology for the same region.

Because the sample is strongly imbalanced, we proposed a first filtering step using the principal component analysis (PCA) algorithm. When training the PCA model, we obtained that 94% of the sample's variance lied along the two first principal components. Projecting the data onto the hyperplane defined by these two principal components, it is possible to make a first cut in the identification of UCDs.

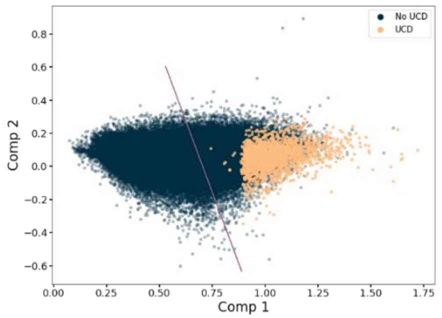

Fig. 2. Projection of the sample used in the ML methodology onto the hyperplane defined by the first two principal components. The purple line represents the decision threshold used to make a first cut at identifying UCDs.

Using the reduced sample obtained in the PCA filtering, we developed a support vector machine (SVM) classifier using the seven J-PLUS colours as input features in the training step. Using k-fold cross validation, we conducted a search for the optimal hyperparameters and achieved a total recall of 98% on the test set.

Results

- Using the fitted PCA and SVM models to predict on unseen data from another 20x20 deg^2 region, we were able to recover 96% of the candidate UCDs found with the VO-based methodology.

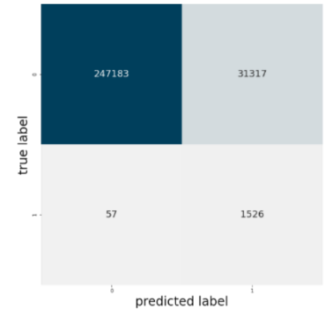

Fig. 3. Confusion matrix for the blind test on unseen data, with a recall of 96%. 1 and 0 labels represent UCD and non-UCD objects, respectively.

- We found crucial the preliminar PCA filtering to deal with the strong imbalance of the data and discard the hottest objects.
- The developed ML-based methodology is able to discard a larger number of true negatives (non-UCD objects) before the analysis with VOSA. This is a significant achievement, since the main bottleneck of the VO-based methodology is the large number of objects to be analysed with VOSA, as it is a very time consuming task.

References

[1] http://svo2.cab.inta-csic.es/theory/vosa

Watch the 2-minutes presentation here

Email: pmas@cab.inta-csic.es

Acknowledgements. This research has made use of the Spanish Virtual Observatory Project funded by MCIN/AEI/10.13039/501100011033/ through grant PID2020-112949GB-I00.