Are large training samples always necessary for machine learning classification problems?

S. Akras

Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing, National Observatory of Athens.

Abstract

Over the last 2 decades machine learning (ML) algorithms have become increasingly popular in astronomy. Several photometric sky-surveys have been conducted and even more are planned for the near future covering a large spectral range. To explore this large amount of data is necessary to apply automatic techniques. In this talk, I will present the results of the application of ML algorithms to identify new symbiotic stars (SySts) and planetaty nebulae (PNe) candidates. Despite the small training samples, the performance of the models turns out to be sufficient. For specific problems in astronomy, the combination of data from different spectral wavelengths(different surveys) may be more crucial than the size of the training samples. The detection rate of symbiotic stars spectrocopic discoveries has increased by a factor of three compared to previous attempts.

The problem

ML algorithms require large dataset for training good models and derive meaningful insight. For instance, star/galaxy classiication is based on photometric dataset with million of sources (e.g. Costa-Duarte +19).

However, there are many astrophysical problems for which limited observational data are available. In particular, the small numbers of known stellar objects such as symbiotic stars, cataclysmic variables, planetary nebulae, among others lead to very small and highly unbalanced training samples.

The question is how much data are needed in order to get good ML models? Are the results reliable?

Dataset

For the classification problem of SySts and PNe and their separation from other Hα emitter (mimics), the training sample was built combining sources from various lists (Table 1) and photometric data from 2MASS and WISE.

Due to the limited number of known SySts (~323; 263 S-type, 60 D-type; Akras +19a), all individual sources have similar size (~200). This leads to a highly unbalanced training sample with ~200 sources of interest (SySts or PNe) and ~1800 mimics (the rest of them).

In case of PNe (~2500 known; Parker +16), we focused only on the compact ones with angular diameter < 6 arcsec as we were interested in finding new candidates in the IPHAS and VPHAS+ photometric surveys.

Class of Object	Sample	References
PNe	188	Ramos-Larios & Phillips 2005
Post-AGB	180	Vickers et al.2015,Akras et al.2017 Suarez et al. 2006, Yoon et al. 2014
Wolf-Rayet	162	van der Hucht 2001
Be	185	Chojnowski et al. 2015
AeBe	173	Vieira et al. 2003, Herbst & Shevchenko 1999 Rodrigues et al. 2009
CV	191	Hoard et al. 2002
Mira	316	Huemmerich & Bernhard 2012, Whitelock et al. 2008
K giants	240	Carlberg et al. 2011, Gray et al. 2016
M giants	210	Tabur et al. 2009, Gray et al. 2016
Classical T Tauri	183	Galli et al. 2015, France et al. 2014 Grankin et al. 2007, Herbst & Shevchenko 1999
Weak T Tauri	213	Grankin et al. 2008, Galli et al. 2015 Cieza et al. 2007, Herbst & Shevchenko 1999
YSO	260	Rebull et al. 2011, Harvey et al. 2007
SySts	220†	Paper I and references therein

Table 1. List of references for the selected types of stellar sources.

Results

Symbiotic stars

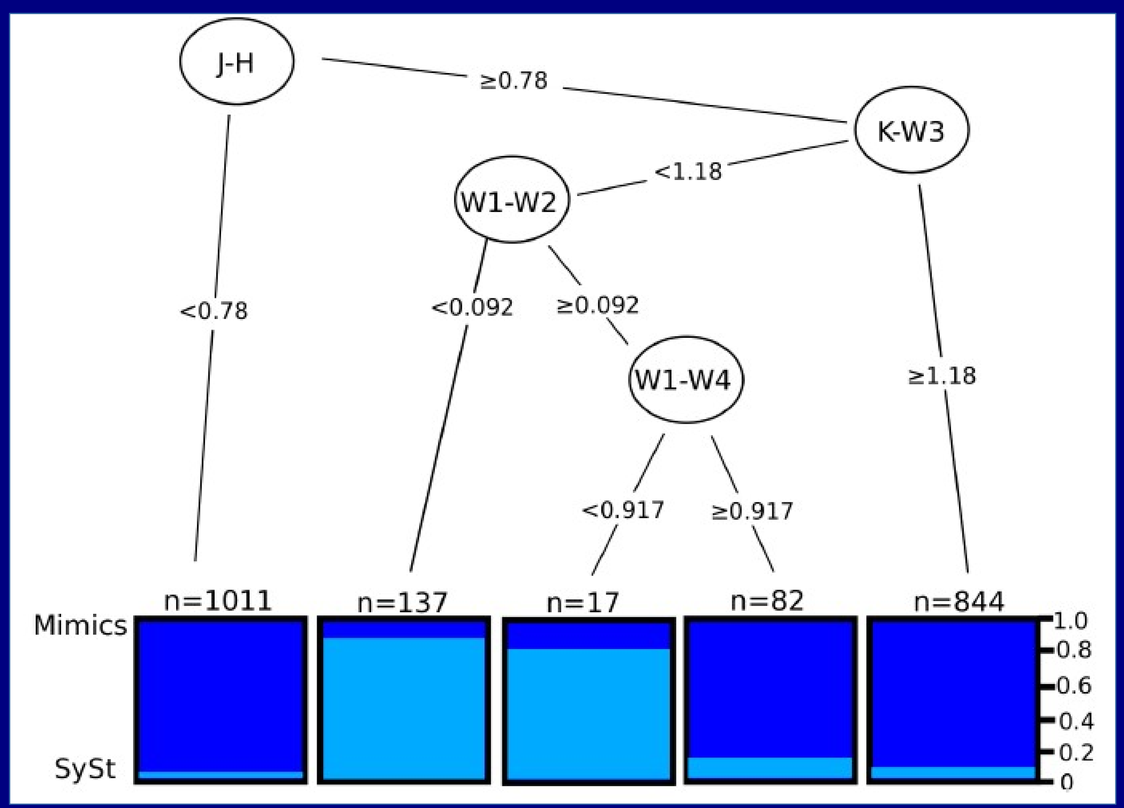

Figure 1. Best decision tree model for S-type Symbioic stars classification.

Decision tree model (Fig.1) distinguishes S-type SySt from their mimics combining 2MASS and WISE photometric data (Arkas +19b). The colours J-H, Ks-W3 and W1-W2 turned out to be the most important features. The accuracy of the model is computed ~75% and recoves up to 90% of the known SySts in the IPHAS and VPHAS+ lists. Finally, 36 S-type SySt candidates were found in the VPHAS+ DR2 catalog.

Spectroscopic data of 7 VPHAS+ candidates unvealed the SySt nature for 5 of them (72%, Akras et al 2021), which is three times higher than the previous attempts.

Planetary nebulae

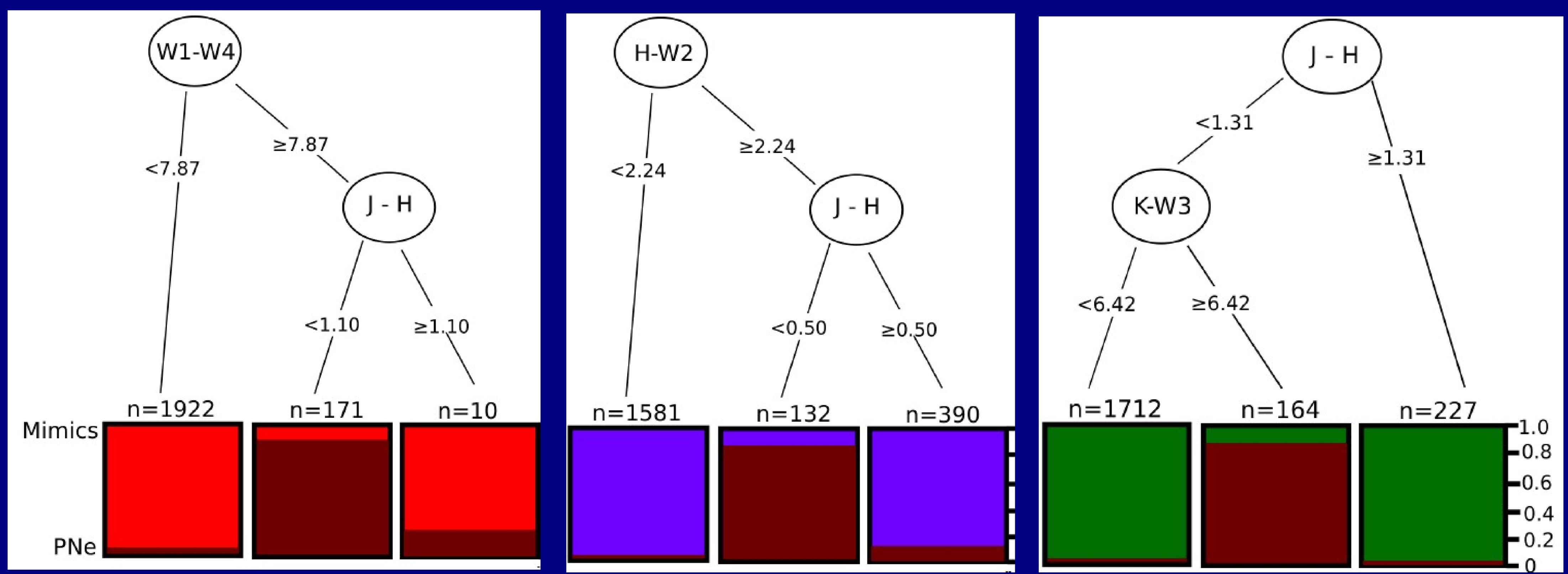

Figure 2. Three best disicion tree models for Planetary nebulae classification.

Decision tree models that separate compact PNe from their mimics combining 2MASS and WISE photometric data are shown in Figure 2 (Arkas +19c). The accuracy of the models range from 80 to 85%. The J-H colour index was found to be important in all three models. 39 new compact PN candidated were found in the IPHAS (24) and VPHAS+ DR2 (15) catalogs. According to these results, approximately 80 new compact PNe are expected to be found in the entire VPHAS+ catalog.

Conclusions

- Despite the small and highly unbalanced traning samples, decision tree algorithm has been very helpful in devising new selection criteria for SySts and compact PNe combining photometric data from 2MASS and WISE.
- The accuracy of the models is up to 75-85% while up to 90% of the known SySts and PNe are recovered.
- More reliable list of SySts and PNe candidates can be obtained with small number of contaminants optimizing the observing time.
- 36 S-type SySts and 39 compact PNe candidates have been found in the IPHAS and VPHAS+ catalogs.
- 5 new S-type SySts were discovered by observing only 7 candidates.

References

Akras et al., 2019a, ApJS, 240, 21A; Akras et al., 2019b, MNRAS, 483, 5077A; Akras et al., 2019c, MNRAS, 488,3238A; Akras et al.,2021, MNRAS, 502, 2513A; Costa-Duarte, M. V., 2019, arXiv190908626C; Parker Q. A. et al., 2016, JphCS728c2008P.