Classification of Evolved Stars with (Unsupervised) Machine Learning

J.Welsh[1], C.Bordiu[2], J.Tan[1], M.Krokos[1], F.Bufano[2], E.Sciacca[2]
[1] *University of Portsmouth,* [2] *Instituto Nazionale di Astrofisica*

Abstract

Next-generation observational astronomy instrumentation is expected to generate unprecedented volumes of data that can not be feasibly processed without machine learning. In this work, we analyse the effectiveness of unsupervised machine learning techniques for the for the classification of evolved stars using multi-wavelength photometric data. Research outcomes include a dataset of 16,000 sources including asymptotic giant branch (AGB), red supergiant (RSG), Wolf Rayet (WR) and Lumninous Blue Variable (LBV) sources. The dataset features 8 independent colours, computed from photometric magnitudes from Gaia, WISE and 2MASS. Our results indicate that density-based clustering algorithms can utilise colours effectively, attaining an accuracy of 98.8% when not considering classification of outliers. We further investigated the use of dimensionality reduction techniques with improve clustering and get rid of possibly redundant data. Our results show that these techniques can significantly improve the effectiveness of the chosen clustering algorithms, and additionally result in the clear separation of oxygen-rich and carbon-rich AGB stars within the feature space, improving upon previous literature results. We envisage that our findings can be replicated across other datasets containing photometric data, towards achieving even higher accuracies - we plan to perform a future systematic experimentation and to make our ML pipeline available within the NEANIAS cloud-based science gateway as an easy-to-use interactive testbed environment for domain scientists.

Data Preparation

For the application of unsupervised machine learning, a custom photometric dataset was prepared. Target sources were parsed from available stellar catalogues and photometric magnitudes were retrieved from catalogues Gaia EDR3, 2MASS and WISE. SIMBAD was used to verify the spatial coordinates of stars.

The dataset contains 15,991 total sources, with five unique classes of evolved stars distributed as follows: 54.46% Oxygen-rich AGB; 38.36% Carbon-rich AGB; 4.86% RSG; 2.12% WR; 0.17% LBV.

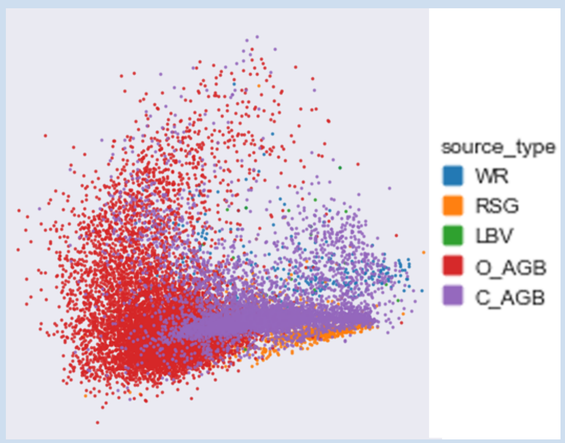


Fig 1. Visualisation of photometric dataset with original labels.

1. Parsing Target Source Coordinates Spatial coordinates from available stellar catalogues of Asymptotic Giant Branch (AGB), Wolf Rayet (WR), Luminous Blue variable (LBV) and Red Galactic Supergiant (RSG) were retrieved.

2. Retrieving photometric magnitudes. a CDS Upload X-Match operation was run in TOPCAT to retrieve wavelength bands of target sources from catalogues; **Gaia EDR3** (G_BP = 532 nanometers, G_RP = 797 nanometers), **2MASS** (J = 1.25 microns, H = 1.65 microns, K = 2.17 microns) and **WISE** (W1 = 3.4 microns, W2 = 4.6 microns, W3 = 12 microns, W4 = 22 microns.

3. Computing Colour Values. By computing and using colours as opposed to the photometric magnitudes we gain more physical information about the stars that may be identified by unsupervised machine learning algorithms. In addition to this, the dimensionality of the features used will be lower, meaning computation speed of said algorithms will be faster. Colours were computed as follows: **Gaia EDR3** (BP_RP = G_BP - G_RP), **2MASS** (JH = J - H, HK = H - K), **WISE** (W12 = W1 - W2, W23 = W2 - W3, W34 = W3 - W4), **MISC** (RJ = G_RP (Gaia) - J (WISE), KW1 = K (2MASS) - W1 (WISE)).

Acknowledgements

The research leading to these results has received funding from the European Commission's Horizon 2020 research and innovation programme under the grant agreement No. 863448 (NEANIAS).

Application of Unsupervised Machine learning

The unsupervised machine learning classification algorithms analysed were: Centroid-based - KMeans clustering; Density-based - DBSCAN and HDBSCAN clustering. DBSCAN showed poor results, but HDBSCAN showed much better results, attaining an accuracy of 64%. However, this result was inferior to the centroid-based clustering algorithm KMeans, which attained an accuracy of 66%.

Despite this, HDBSCAN showed significantly less variance in sources per cluster and even managed to classify a group purely composed of 41 WR stars. It was found that the clustering accuracy of HDBSCAN was suffering due to approximately half of the sources being classified as outliers. When not considering outliers as a cluster, an accuracy of 98.8% was computed.

HDBSCAN	Total Sources	O_AGB	C_AGB	LBV	RSG	WR
Cluster 0	41	0.00%	0.00%	0.00%	0.00%	**100.00%**
Cluster 1	3547	**98.31%**	1.52%	0.00%	0.03%	0.14%
Cluster 2	4384	0.32%	**88.85%**	0.00%	**10.81%**	0.02%
Outliers	8019	64.96%	27.26%	0.35%	3.78%	3.65%

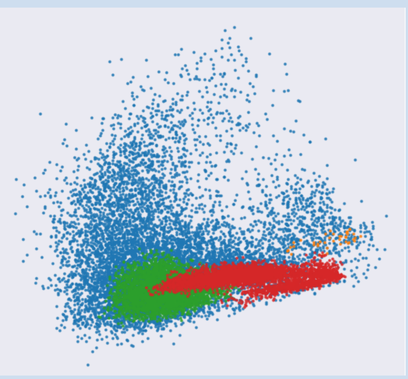


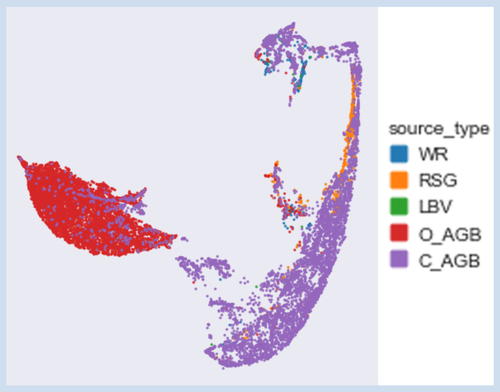
Figure 2. Table: Distribution of sources per cluster with HDBSCAN clustering.
Plot: Visualisation of HDBSCAN clustering, blue data points represent outliers classified by HDBSCAN.

Improving Clustering with Dimensionality Reduction

To get rid of possibly redundant features, improve clustering, and reduce outlier classification, we explored various dimensionality reduction techniques including: Autoencoders (AE); UMAP; T-SNE; and the recent (Not Too) Deep Clustering (N2D) approach suggested by McConville, et al. (2020), which involves applying manifold learning to an autoencoded feature representation.

It was found that by applying HDBSCAN to UMAP's feature representation, accuracy increased from 65% to 86%, and the number of outliers classified was reduced from 8019 to a mere 103. The N2D (UMAP) approach, while only attaining 84% accuracy, resulted in zero outliers being classified. Results were particularly impressive visually with UMAP and N2D (UMAP) as shown in Fig 3. plots, where we can observe the clear separation of O-rich AGB (red cluster) and C-rich AGB (purple) sources, improving upon existing literature results that reported a separation in the near-infrared colour space due to different properties of dust (e.g. Van der Veen & Habing, 1988; Walker & Cohen, 1988)

HDBSCAN	Outliers	Accuracy
Raw Data	8019	0.65
Autoencoder	1938	0.69
UMAP	**103**	**0.86**
T-SNE	1296	0.82
N2D (UMAP)	**0**	**0.84**
N2D (T-SNE)	1609	0.80



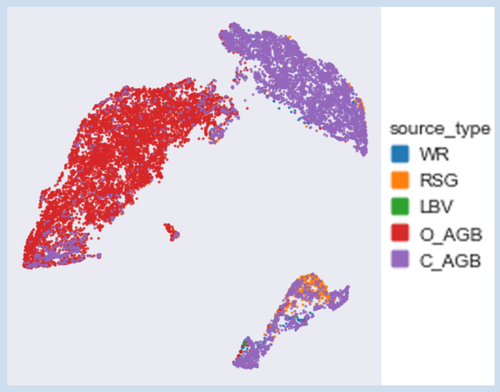


Figure 3. Table: outliers classified and clustering accuracy of HDBSCAN after feature extraction. Centre Plots: UMAP (center) & N2D (UMAP) (right) feature representation, showing clear separation of O-rich and C-rich AGB sources in the featurespace.

Conclusions

Our results indicate that density-based clustering techniques such as HDBSCAN are superior to centroid-based clustering techniques for the classification of evolved stars when using multi-wavelength photometric values as data input. Our results also show that feature extraction can be used to improve clustering accuracy and reduce outliers classified by density-based clustering techniques such as HDBSCAN, particularly with methods such as UMAP that focus on preserving global structure over local structure. Finally, our results show that photometric values can be used for the effective separation of Oxygen and Carbon-rich AGB stars without the need of considering spectroscopic features, improving results found in previous literature.