Communications of BAO, Vol. 71, Issue 2, 2024, pp. 377-382

Machine Learning Classification of Young Stellar Objects and Evolved Stars in the Magellanic Clouds Using the Probabilistic Random Forest Classifier

Sepideh Ghaziasgar [*1], Mahdi Abdollahi [†1], Atefeh Javadi [‡1], Jacco Th. van Loon [§2], Iain McDonald [¶3], Joana Oliveira [‖2], and Habib G. Khosroshahi [**1,5]

[1]School of Astronomy, Institute for Research in Fundamental Sciences (IPM), P.O. Box 19568-36613, Tehran, Iran
[2]Lennard-Jones Laboratories, Keele University, ST5 5BG, UK
[3]Jodrell Bank Centre for Astrophysics, Alan Turing Building, University of Manchester, M13 9PL, UK
[5]Iranian National Observatory, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

Abstract

The Magellanic Clouds (MCs) are excellent locations to study stellar dust emission and its contribution to galaxy evolution. Through spectral and photometric classification, MCs can serve as a unique environment for studying stellar evolution and galaxies enriched by dusty stellar point sources. We applied machine learning classifiers to spectroscopically labeled data from the Surveying the Agents of Galaxy Evolution (SAGE) project, which involved 12 multiwavelength filters and 618 stellar objects at the MCs. We classified stars into five categories: young stellar objects (YSOs), carbon-rich asymptotic giant branch (CAGB) stars, oxygen-rich AGB (OAGB) stars, red supergiants (RSG), and post-AGB (PAGB) stars. Following this, we augmented the distribution of imbalanced classes using the Synthetic Minority Oversampling Technique (SMOTE). Therefore, the Probabilistic Random Forest (PRF) classifier achieved the highest overall accuracy, reaching 89% based on the recall metric, in categorizing dusty stellar sources before and after data augmentation. In this study, SMOTE did not impact the classification accuracy for the CAGB, PAGB, and RSG categories but led to changes in the performance of the OAGB and YSO classes.

Keywords: *stars: classification - stars: AGB, RSG, and post-AGB - stars: YSOs - galaxies: spectral catalog - galaxies: Local Group - methods: machine learning*

1. Introduction

The Magellanic Clouds (MCs), as nearby dwarf galaxies with the distance of 50 kpc and 60 kpc (Pietrzyński et al., 2013, Subramanian & Subramaniam, 2009, 2011) and metallicities of 0.5 and $0.2Z_{\odot}$ (Russell & Dopita, 1992), offer an ideal environment to study stellar contributions to dust production (Ruffle et al., 2015).

The life cycle of stars is represented by different stellar classes, each with distinct physical characteristics and processes. Dusty stellar objects enriched chemically during evolution can be classified into young stellar objects (YSOs) and evolved stars (Boyer et al., 2011).

Young stellar objects (YSOs) are in the early phases of star formation. They are surrounded by gas and dust and offer a window into the physical processes driving star formation and galaxy evolution (Kokusho et al., 2023, Sewiło et al., 2013, Suh, 2016). Also, the luminosity of YSOs can vary from optical to IR depending on their mass and evolution stage (Miettinen, 2018, Oliveira et al., 2013).

[*]sepide.ghaziasgar@gmail.com, Corresponding author
[†]m.abdollahi@ipm.ir
[‡]atefeh@ipm.ir
[§]j.t.van.loon@keele.ac.uk
[¶]Iain.Mcdonald-2@manchester.ac.uk
[‖]j.oliveira@keele.ac.uk
[**]habib@ipm.ir

Evolved stars, including asymptotic giant branch (AGB) stars with low– and intermediate mass (0.8-8 $M_{\odot}$) and red supergiants (RSGs) with high mass (M$\geq$ 8 $M_{\odot}$), are dust producers that enrich the ISM with heavy elements (Herwig, 2005, Höfner & Olofsson, 2018). The significant brightness of AGBs ($\sim 10^{3-4}$ $L_{\odot}$), along with their radial pulsations, makes these stars detectable in the infrared (Goldman et al., 2017, McDonald & Zijlstra, 2016). Many evolved AGBs are long-period variables (LPVs) (Javadi & van Loon, 2019, 2022, Javadi et al., 2013, Marigo et al., 2017, Navabi et al., 2021, Saremi et al., 2021, Torki et al., 2023, Whitelock et al., 2003). AGB stars are classified into oxygen-rich (OAGB) and carbon-rich (CAGB) subcategories based on their chemical abundance, while RSGs represent massive stars in the final stages of their lives (Javadi et al., 2011, Levesque, 2010, Massey & Olsen, 2003, Yang et al., 2020), often culminating in supernovae or compact remnants. Post-AGB (PAGB) stars mark a transitional phase, shedding their outer layers before evolving into white dwarfs, revealing unique chemical signatures (Kamath, 2020, Kamath et al., 2014, 2015, van Winckel, 2003).

Machine learning algorithms are powerful tools for classifying stellar objects (Baron, 2019, Ghaziasgar et al., 2022, 2024, Kinson et al., 2021, 2022). The availability of more spectroscopically and photometrically labeled data enhances the ability of these algorithms to classify dusty stellar classes with greater accuracy, improving the overall reliability of stellar classification.

2. Data

The dataset used in this study is derived from the Surveying the Agents of Galaxy Evolution (SAGE) project for tracking dust and gas in Magellanic Clouds (Meixner et al., 2006). This dataset comprises multiwavelength spectroscopically labeled near-infrared and mid-infrared filters. From the SAGE spectral catalog, as shown in Table 1, we selected 12 multiwavelength filters for each spectral class (SpClass) and 618 dusty stellar objects in the MCs (Jones et al., 2017, Kemper et al., 2010, Ruffle et al., 2015, Woods et al., 2011). The features selected for training include UMmag, BMmag, VMmag, IMmag, J2mag, H2mag, Ks2mag, IRAC1, IRAC2, IRAC3, IRAC4 and [24].

We augmented our dataset using the SMOTE (Synthetic Minority Oversampling Technique) approach (Chawla et al., 2011) that addresses the class imbalance in datasets as presented in Table 1. Instead of simply duplicating instances from the minority class, SMOTE generates synthetic samples by interpolating between existing data points. This method selects a random data point from the minority class and creates new samples along the line segments connecting the sample to the nearest neighbor (Chawla et al., 2011). The SMOTE algorithm, applied to the training datasets, balances the population of minority classes with the majority class, potentially improving classifier performance. The Simple approach represents the original imbalanced dataset, while the SMOTE approach refers to the augmented dataset, where class imbalance has been addressed using SMOTE.

Table 1: This is a spectral classification of dusty stars in the Magellanic Clouds. Based on SMOTE methodology, the "Augmented Data" column represents the population after data augmentation.

Classes	LMC	SMC	Total	*Augmented Data
CAGB	136	38	174	200
OAGB	88	19	107	193
PAGB	33	4	37	183
RSG	72	22	94	190
YSO	157	49	206	206

3. Classification Models

We employed supervised learning algorithms to classify samples of YSOs and evolved stars (Ghaziasgar et al., 2022, 2024). The algorithms were trained on spectroscopically labeled data and evaluated on a test dataset to assess their accuracy. The models we used included Probabilistic Random Forest (PRF) (Baron, 2019, Kinson et al., 2021, 2022, Reis et al., 2019), Random Forest (RF) (Baron, 2019, Baron & Poznanski,

doi: https://doi.org/10.52526/25792776-24.71.2-377

2017, Breiman, 2001, Carliles et al., 2010), K-Nearest Neighbor (KNN) (Altman, 1992), C-Support Vector Classification (SVC) including SVC-poly and SVC-rbf (Baron, 2019, Vapnik, 1995), and Gaussian Naive Bayes (GNB) (Wilson et al., 2023). Among all the classifiers, the PRF model performed best before and after data augmentation with the SMOTE method, as detailed below.

The PRF classifier, the developed version of RF, is designed to handle noisy and uncertain datasets (Reis et al., 2019). The RF is a machine learning algorithm that builds an ensemble of decision trees, each trained on a randomly selected subset of features and data, to avoid overfitting and generalize well to new data. In RF, predictions are made through majority voting for classification studies (Baron, 2019, Breiman, 2001). However, RF assumes that data and labels are fixed and accurate, making it less effective when dealing with noisy or uncertain inputs. The PRF overcomes this limitation by introducing a probabilistic framework that treats both features and labels as probability distributions rather than fixed values. PRF routes data points probabilistically across tree branches, accounting for uncertainties in both features and labels. This probabilistic framework enhances its ability to handle noisy inputs effectively (Kinson et al., 2021, 2022, Reis et al., 2019).

4. Results and Ongoing works

We presented the classification results, as can be seen in Table 2, Fig. 1 and Fig. 2, using two approaches, Simple and SMOTE, named based on the distribution of each dataset. As shown, in comparison to other classifiers, SMOTE outperforms Simple in PRF and RF classification. Based on the recall metric, the PRF classifier demonstrated the highest total accuracy, achieving 89%. Using the SMOTE technique, the performance of the best model for the CAGB, PAGB, and RSG classes did not improve, with accuracy remaining at 100%, 100% and 88%, respectively. However, SMOTE led to some variations in the OAGB and YSO classes.

In the following, we can compare photometrically labeled data with spectroscopically labeled data with similar features. Additionally, incorporating multiwavelength data as model inputs could refine label determination for each object. More multiwavelength and spectroscopic observations are needed to improve dusty stellar classifications, especially for less populated classes like PAGBs and RSGs.

Table 2: Classification report; contains the model's precision, recall, and f1-score values for each class. The f1-score is calculated by averaging. The macro average f1-score represents an average of the f1-score over classes. The weighted average f1-score is calculated as the mean of all per-class f1-scores while considering each class's support.

(a). Classification report, Simple PRF.

Class	Precision	Recall	F1-score
CAGB	0.95	1.00	0.97
OAGB	0.80	0.73	0.76
PAGB	0.50	1.00	0.67
RSG	0.78	0.88	0.82
YSO	0.95	0.88	0.91
accuracy			0.89
macro avg	0.80	0.90	0.83
weighted avg	0.89	0.89	0.89

(b). Classification report, SMOTE PRF.

Class	Precision	Recall	F1-score
CAGB	0.86	1.00	0.92
OAGB	1.00	0.64	0.78
PAGB	0.50	1.00	0.67
RSG	0.70	0.88	0.78
YSO	1.00	0.92	0.96
accuracy			0.89
macro avg	0.81	0.89	0.82
weighted avg	0.91	0.89	0.89

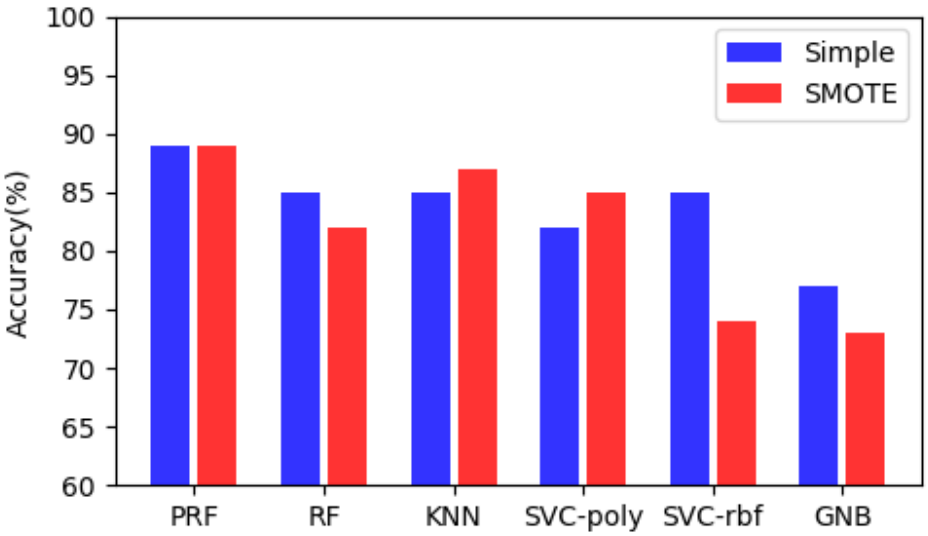

Figure 1: This plot presents the performance of Simple classifiers and SMOTE classifiers based on their accuracy scores.

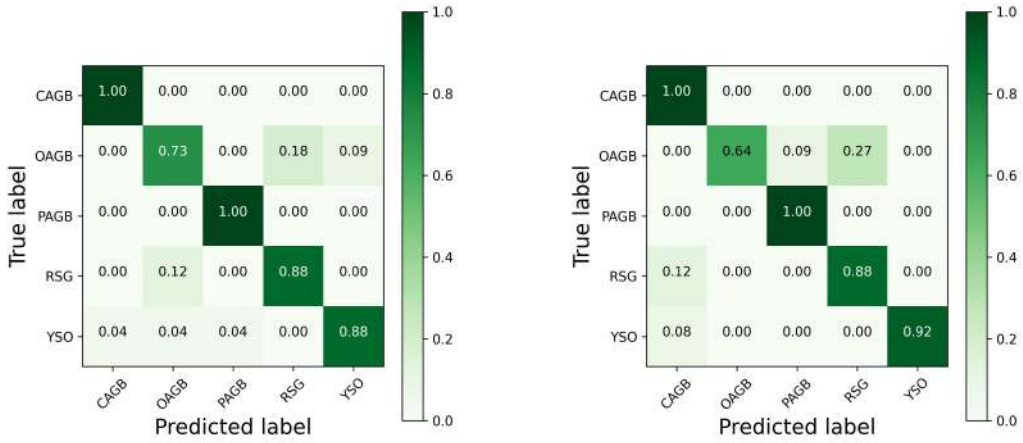

Figure 2: These are the confusion matrices for the Probabilistic Random Forest. The left panel presents the results before data augmentation, and the right panel displays the results after data augmentation with SMOTE. The matrix displays the number of objects predicted by the model in each class. The diagonal elements represent the predicted and actual labels for each class.

Acknowledgements

The authors thank the School of Astronomy at the Institute for Research in Fundamental Sciences (IPM) and the Iranian National Observatory (INO) for supporting this research. S. Ghaziasgar is grateful for the support of the Byurakan Astrophysical Observatory (BAO).

References

Altman N. S., 1992, The American Statistician, 46, 175

Baron D., 2019, arXiv e-prints, p. arXiv:1904.07248

Baron D., Poznanski D., 2017, Mon. Not. R. Astron. Soc. , 465, 4530

Boyer M. L., et al., 2011, Astron. J. , 142, 103

Breiman L., 2001, Machine Learning, 45, 5

Carliles S., Budavári T., Heinis S., Priebe C., Szalay A. S., 2010, Astrophys. J. , 712, 511

Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P., 2011, arXiv e-prints, p. arXiv:1106.1813

Ghaziasgar S., Masoudnezhad A., Javadi A., van Loon J. T., Khosroshahi H. G., Khosravaninezhad N., 2022, arXiv e-prints, p. arXiv:2211.03403

Ghaziasgar S., et al., 2024, in EAS2024, European Astronomical Society Annual Meeting. p. 371

Goldman S. R., et al., 2017, Mon. Not. R. Astron. Soc. , 465, 403

Herwig F., 2005, Ann. Rev. Astron. Astrophys. , 43, 435

Höfner S., Olofsson H., 2018, Astron. Astrophys. Rev. , 26, 1

Javadi A., van Loon J. T., 2019, in Kerschbaum F., Groenewegen M., Olofsson H., eds, IAU Symposium Vol. 343, Why Galaxies Care About AGB Stars: A Continuing Challenge through Cosmic Time. pp 283–290 (arXiv:1811.12025), doi:10.1017/S1743921318006671

Javadi A., van Loon J. T., 2022, in Decin L., Zijlstra A., Gielen C., eds, IAU Symposium Vol. 366, The Origin of Outflows in Evolved Stars. pp 210–215 (arXiv:2204.08944), doi:10.1017/S1743921322001326

Javadi A., van Loon J. T., Mirtorabi M. T., 2011, Mon. Not. R. Astron. Soc. , 411, 263

Javadi A., van Loon J. T., Khosroshahi H., Mirtorabi M. T., 2013, Mon. Not. R. Astron. Soc. , 432, 2824

Jones O. C., et al., 2017, Mon. Not. R. Astron. Soc. , 470, 3250

Kamath D., 2020, Journal of Astrophysics and Astronomy, 41, 42

Kamath D., Wood P. R., Van Winckel H., 2014, Mon. Not. R. Astron. Soc. , 439, 2211

Kamath D., Wood P. R., Van Winckel H., 2015, Mon. Not. R. Astron. Soc. , 454, 1468

Kemper F., et al., 2010, Publ. Astron. Soc. Pac. , 122, 683

Kinson D. A., Oliveira J. M., van Loon J. T., 2021, Mon. Not. R. Astron. Soc. , 507, 5106

Kinson D. A., Oliveira J. M., van Loon J. T., 2022, Mon. Not. R. Astron. Soc. , 517, 140

Kokusho T., Torii H., Kaneda H., Fukui Y., Tachihara K., 2023, Astrophys. J. , 953, 104

Levesque E. M., 2010, in Leitherer C., Bennett P. D., Morris P. W., Van Loon J. T., eds, Astronomical Society of the Pacific Conference Series Vol. 425, Hot and Cool: Bridging Gaps in Massive Star Evolution. p. 103 (arXiv:0911.4720), doi:10.48550/arXiv.0911.4720

Marigo P., et al., 2017, Astrophys. J. , 835, 77

Massey P., Olsen K. A. G., 2003, Astron. J. , 126, 2867

McDonald I., Zijlstra A. A., 2016, Astrophys. J. Lett. , 823, L38

Meixner M., et al., 2006, Astron. J. , 132, 2268

Miettinen O., 2018, Astrophys. Space. Sci. , 363, 197

Navabi M., et al., 2021, Astrophys. J. , 910, 127

Oliveira J. M., et al., 2013, Mon. Not. R. Astron. Soc. , 428, 3001

Pietrzyński G., et al., 2013, Nature. , 495, 76

Reis I., Baron D., Shahaf S., 2019, Astron. J. , 157, 16

Ruffle P. M. E., et al., 2015, Mon. Not. R. Astron. Soc. , 451, 3504

Russell S. C., Dopita M. A., 1992, Astrophys. J. , 384, 508

Saremi E., Javadi A., Navabi M., van Loon J. T., Khosroshahi H. G., Bojnordi Arbab B., McDonald I., 2021, Astrophys. J. , 923, 164

Sewiło M., et al., 2013, Astrophys. J. , 778, 15

Subramanian S., Subramaniam A., 2009, Astron. Astrophys. , 496, 399

Subramanian S., Subramaniam A., 2011, in Astronomical Society of India Conference Series. p. 144

Suh K.-W., 2016, Journal of Astronomy and Space Sciences, 33, 119

Torki M., Navabi M., Javadi A., Saremi E., van Loon J. T., Ghaziasgar S., 2023, in Bisikalo D., Wiebe D., Boily C., eds, IAU Symposium Vol. 362, The Predictive Power of Computational Astrophysics as a Discover Tool. pp 353–355 (arXiv:2204.11530), doi:10.1017/S1743921322001405

Vapnik V. N., 1995, The nature of statistical learning theory. Springer-Verlag New York, Inc.

Whitelock P. A., Feast M. W., van Loon J. T., Zijlstra A. A., 2003, Mon. Not. R. Astron. Soc. , 342, 86

Wilson A. J., Lakeland B. S., Wilson T. J., Naylor T., 2023, Mon. Not. R. Astron. Soc. , 521, 354

Woods P. M., et al., 2011, Mon. Not. R. Astron. Soc. , 411, 1597

Yang M., et al., 2020, Astron. Astrophys. , 639, A116

van Winckel H., 2003, Ann. Rev. Astron. Astrophys. , 41, 391