Astronomy & Astrophysics manuscript no. output ©ESO 2024
September 13, 2024

Mapping Hα-Excess Candidate Point Sources in the Southern Hemisphere Using S-PLUS Data

L. A. Gutiérrez Soto[1, 2, ⋆], R. Lopes de Oliveira[2, 3, 4], S. Akras[5], D. R. Gonçalves[6], L. F. Lomelí-Nuñes[6], C. Mendes de Oliveira[2], E. Telles[4], A. Kanaan[7], T. Ribeiro[8], W. Schoenell[9]

[1] Instituto de Astrofísica de La Plata (CCT La Plata - CONICET - UNLP), B1900FWA, La Plata, Argentina
e-mail: gsotoangel@fcaglp.unlp.edu.ar
[2] Departamento de Astronomia, IAG, Universidade de São Paulo, Rua do Matão, 1226, 05509-900, São Paulo, Brazil
[3] Departamento de Física, Universidade Federal de Sergipe, Av. Marechal Rondon, S/N, 49100-000, São Cristóvão, SE, Brazil
[4] Observatório Nacional, Rua Gal. José Cristino 77, 20921-400, Rio de Janeiro, RJ, Brazil
[5] Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing, National Observatory of Athens, GR 15236 Penteli, Greece
[6] Observatório do Valongo, Universidade Federal do Rio de Janeiro, Ladeira Pedro Antonio 43, 20080-090, Rio de Janeiro, Brazil
[7] Departamento de Física, Universidade Federal de Santa Catarina, Florianópolis, SC, 88040-900, Brazil
[8] NOAO, P.O. Box 26732, Tucson, AZ 85726
[9] GMTO Corporation 465 N. Halstead Street, Suite 250 Pasadena, CA 91107

Received September 15, 1996; accepted March 16, 1997

ABSTRACT

Context. We leverage the Southern Photometric Local Universe Survey (S-PLUS) Fourth data release to identify and classify Hα-excess sources in the Southern Sky. This approach combines extensive photometric data with advanced machine learning techniques to enhance source classification.

Aims. We aim to enhance the classification of Hα-excess sources by integrating multi-wavelength photometric data with advanced machine learning methods. Our goal is to accurately distinguish between extragalactic and Galactic sources and to address the challenge of incomplete photometric data.

Methods. We selected Hα-excess candidates using a $(r - J0660)$ versus $(r - i)$ color-color diagram from both the main survey and Galactic disk of S-PLUS. Dimensionality reduction was performed using UMAP, followed by clustering with HDBSCAN. Initially, clustering was performed using only the S-PLUS filter colors. Subsequently, we incorporated WISE data to evaluate improvements in classification performance. Finally, a Random Forest model was employed, utilizing the combined S-PLUS and WISE filter colors to identify key distinguishing features.

Results. Combining multi-wavelength photometric data with machine learning techniques has substantially enhanced the identification and classification of Hα-excess sources. We identified a total of 6 956 sources with excess in the $J0660$ filter, indicative of Hα-excess. Among these, we classified various object types, including cataclysmic variables, quasars, young stellar objects, and different types of stars and galaxies, in agreement with the SIMBAD database. Notably, our sample also includes objects with Hα in absorption, primarily RR Lyrae stars. Using only the 12 S-PLUS filters, UMAP and HDBSCAN effectively clustered the data, distinguishing between Hα-emission objects and those with Hα in absorption. Incorporating additional WISE filters further refined this clustering, enabling successful separation of extragalactic sources from Galactic ones and improving the differentiation between cataclysmic variables and QSOs. The Random Forest model, based on HDBSCAN results, identified key color features that effectively distinguished between the different classes of Hα-excess sources.

Key words. surveys – techniques: photometric – stars: novae, cataclysmic variables – galaxies: dwarf – quasars: emission lines

1. Introduction

Atomic excitation followed by recombination in Balmer hydrogen emission lines may be ignited in different ways, thermal and nonthermal collisional excitation in shock-heated gas and energetic photons acting over a diffuse gas. Universe ~~is~~ being hydrogen-abundant, the observation of those electronic transitions offers an important window into the study of astrophysical objects. Among all the possible electronic transitions, the Balmer series represents an extremely useful tool in ~~Astronomy~~. In particular, the Hα emission line – rest-frame wavelength of 6564.614 Å at vacuum – that corresponds to the electron transition from the $n = 3$ to the $n = 2$ energy levels, is the strongest one, in both emission or absorption, and the most widely used to identify various types of objects (e.g star-forming regions, H II regions, planetary nebulae (PNe), supernovae, novae, young stellar objects (YSO), Herbig-Haro objects, circumstellar disks, post-asymptotic and asymptotic giant stars (AGB), red giant stars (RGB), active late-type dwarfs). Amongst massive stars, emission lines are observed in Be stars with decretion disks, Wolf-Rayet (WR) stars, and interacting binary systems that are experiencing mass exchange, like symbiotic stars (SySt), cataclysmic variables (CVs), among others.

In high-redshift sources, such as starburst galaxies and quasi-stellar objects (QSOs), Hα emission is present but is redshifted to longer wavelengths, ~~beyond 6563 Å~~. However, when we detect emission near 6563 Å in these sources, it is typically not due

⋆ E-mail: gsotoangel@fcaglp.unlp.edu.ar

Summary of comments: Halpha_SPLUSDR4_edu.pdf

Page:1

Author: ET Subject: Squiggly Underline Date: 2024-09-17 09:46:50

i don´t know if the term ´clustering´or ´cluster the data´ is the technical term. I would rather use ´classification´. But if it is a technical term, then , maybe somewhere just explain the meaning of clustering the data in the methodology section.

Author: Eduardo Telles Subject: Insert Text Date: 2024-09-17 08:16:16

survey of

Author: ET Subject: Squiggly Underline Date: 2024-09-17 08:23:11

Author: ET Subject: Squiggly Underline Date: 2024-09-17 08:23:17

Author: Edu Telles Subject: Note Date: 2024-09-17 08:31:03

Radiative excitation and ionization are much more important for hydrogen in most astrophysical cases. Collisional excitation and de-excitation coefficients are much lower than the Einstein coeffients in common nebulae. Unless in very dense environments.
I think you should rephrase to express the relative importance more precisely, and check so that you can give numbers, if you want to use this sentence.

Author: ET Subject: Strikeout Date: 2024-09-17 08:31:49

Author: ET Subject: Squiggly Underline Date: 2024-09-17 08:35:01

Author: Eduardo Telles Subject: Replace Text Date: 2024-09-17 08:33:57

Astronomy, because it falls in the commonly used optical spectral range.

Author: ET Subject: Strikeout Date: 2024-09-17 08:37:04

Author: ET Subject: Squiggly Underline Date: 2024-09-17 08:39:55

to Hα recombination but rather the result of UV emission lines that have been redshifted into the visible spectrum.

Most of the aforementioned classes of objects are not homogeneous and remain far from complete, even in the local universe. Some classes are highly populated, while others are significantly underrepresented. For example, there are more than 300 known SySts in the Milky Way but only 65 in nearby galaxies (Akras et al. 2019b; Merc et al. 2019) with constantly new discoveries every year ~~since then~~ (e.g. Merc et al. 2020; Akras et al. 2021; Merc et al. 2021, 2022; Munari et al. 2021, 2022; Akras 2023). The number of known PNe in our galaxy is on the order of ~3500 (Parker et al. 2016), which may represent only 15-30% of the total population (Frew 2008; Jacoby et al. 2010).

Hα surveys have been conducted with varying angular resolutions, sky coverage, and sensitivity. Some surveys, despite having modest spatial resolutions, have successfully resolved extended nebular emissions, enabling the study of supernova remnants, galaxy groups, and star-forming regions (e.g. Davies et al. 1976). Others, with higher spatial resolution, disclosed compact emission-line sources in the Milky Way and nearby galaxies. Examples of them are the INT Photometric Hα survey (IPHAS; Drew et al. 2005; Barentsen et al. 2014), the SuperCOSMOS Hα survey with the UK Schmidt Telescope (UKST) of the Anglo-Australian Observatory (Parker et al. 2005), and the VST Photometric Hα Survey (VPHAS+; Drew et al. 2014).

Colour-colour diagrams from photometric surveys are also used to identify possible Hα emitters. For example, the (r - Hα) versus ($r - i$) colour-colour and similar diagrams has been used to find CVs (Witham et al. 2006, 2007), YSOs (Vink et al. 2008), SySt (Corradi et al. 2008; Corradi & Giammanco 2010; Corradi et al. 2011; Miszalski & Mikołajewska 2014; Mikołajewska et al. 2014, 2017; Akras et al. 2019c), early-type emission-line stars (Drew et al. 2008), and PNe (Miszalski et al. 2009; Viironen et al. 2009; Sabin et al. 2010; Akras et al. 2019a).

Witham et al. (2008) developed a method to select Hα emission line sources in the IPHAS survey by implementing the aforementioned color-color diagram (r - Hα) versus ($r - i$). Hα excess line objects are identified by iteratively fitting the stellar locus and considering those objects as candidates that fall several sigma above this stellar locus in the r - Hα color. This conservative method leaves a total of 4 853 point sources that exhibit strong photometric evidence for Hα emission. They obtained spectra from around 300 sources, confirming more than 95 percent of them as genuine emission-line stars.

citetMonguio:2020 developed the INT Galactic Plane Survey (IGAPS) by merging the IPHAS and UVEX optical surveys. The IGAPS catalog includes 295.4 million photometric measurements in the i, r, narrow-band Hα, g, and U$_{\rm RGO}$ filters. It identifies 8,292 candidate emission line stars and over 53,000 variable stars with confidence greater than 5σ.

More recently, Fratta et al. (2021) introduced a technique using Gaia data to identify Hα-bright sources in the IPHAS catalog. They partitioned the data based on Gaia color-absolute magnitude and Galactic coordinates to minimize contamination and then applied the strategy from Witham et al. (2008) to these partitions.

Two ongoing multi-band surveys are observing the sky in a systematic, complementary way, with 5 broad and 7 narrow-band filters, including Hα: the Javalambre Photometric Local Universe Survey (J-PLUS[1]; Cenarro et al. 2019), covering the Northern celestial hemisphere, and the Southern-Photometric Local Universe Survey (S-PLUS[2]; Mendes de Oliveira et al. 2019), covering the southern sky with a twin 83 cm telescope and filter system. The first one is paving the way for an even more ambitious survey, the Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS; Benitez et al. 2014 and miniJ-PAS; Bonoli et al. 2021), which will observe the Northern sky with 56 narrow-band filters. As source hunters, the spectral energy distributions provided by these surveys enable an unprecedented source classification using photometry only. However, in the Big Data era, efficient investigation tools are required to deal with their massive imaging and catalogues production, and machine learning techniques have been increasingly used to explore these data sets.

Here we present a census of Hα-excess point-like sources from the S-PLUS DR4, utilizing the (r - J0660) versus (r - i) color-color diagram. We leverage the S-PLUS DR4 dataset and employ advanced machine learning techniques to enhance the identification and classification of these sources. Specifically, we use Uniform Manifold Approximation and Projection (UMAP; Becht et al. 2018; McInnes et al. 2020) for dimensionality reduction followed by Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN; Campello et al. 2013) clustering to group sources based on their multi-wavelength photometric signatures. This approach allows us to handle high-dimensional data effectively and uncover patterns that traditional methods might overlook. Additionally, we incorporate Wide-Field Infrared Survey Explorer (WISE; Wright et al. 2010) data and apply a Random Forest (Breiman 2001) model to refine our classification and identify key features that distinguish different types of Hα-excess sources.

Section 2 describes the observations related to the S-PLUS project, including important information on the fourth data release, photometry, and data. Section 3 presents the technique implemented to select the Hα-feature sources and includes the analysis of the results. In Section 4, we present the machine learning methods used to analyze and make a more accurate classification of the Hα sources. Finally, Section 5 discusses our main results and conclusions.

2. Data and Observations

2.1. S-PLUS Survey Overview

This manuscript uses data from S-PLUS DR4 (Herpich et al. 2024). DR4 encompasses 171 fields at very low galactic latitudes (|b| < 15°), an additional 341 fields carried over from DR3 spanning the Main Survey footprint (with |b| > 30°), and 150 fields within the Magellanic Clouds region. This accumulation results in a total of 1629 fields in DR4, covering an expansive area of 3022.7 square degrees. Notably, this coverage includes 347.4 square degrees within the disk regions and 289.5 square degrees within the Magellanic clouds. S-PLUS is conducted using a dedicated 0.83 m robotic telescope located at Cerro Tololo, Chile (Mendes de Oliveira et al. 2019).

S-PLUS surveys the southern sky using the 12 filters from the Javalambre filter system (Marín-Franch et al. 2012), spanning the wavelength range from 3 000Å to 10 000Å. This system comprises seven narrow-band filters (J0378, J0395, J0410, J0430, J0515, J0660, and five broad-band Sloan-like (Fukugita et al. 1996) filters (see Fig. 1). The narrow-band J0660 filter used in S-PLUS is centered at λ 6614Å and has a width of approximately 147Å (Table 2 of Mendes de Oliveira et al. 2019).

[1] https://www.j-plus.es

[2] http://www.splus.iag.usp.br

Page:2

Author: Edu Telles Subject: Note Date: 2024-09-17 08:45:16

when we detect emission near 6563 from high redshift sources,the recombination of Hα is not the cause, instead, it is the outcome of UV emission lines that have shifted towards the visible spectrum.

Author: ET Subject: Squiggly Underline Date: 2024-09-17 08:40:18

Author: Eduardo Telles Subject: Insert Text Date: 2024-09-17 08:47:00

the lists or catalogues of the aforementioned...

Author: José Eduardo Telles Subject: Strikeout Date: 2024-09-17 08:48:01

Author: Edu Telles Subject: Note Date: 2024-09-17 08:50:36

this reference seems too old for all of this!!

Author: José Eduardo Telles Subject: Oval Date: 2024-09-17 08:52:57

The identification of objects is based on the method successfully applied by Witham et al. (2006, 2008) to the IPHAS catalog, since similar filters are also available in S-PLUS: *r*, *J*0660, and *i*. Similar technique was also used by Scaringi et al. (2013); Wevers et al. (2017); Monguió et al. (2020); Fratta et al. (2021) to reveal Hα excess sources.

We first generated (*r* - *J*0660) versus (*r* - *i*) diagrams for each magnitude bin in each field and then attempted to fit the regions predominantly occupied by main-sequence and giant stars using a linear regression model. Subsequently, we applied an iterative σ-clipping technique to the data. Objects with Hα emission typically exhibit an excess in (*r* - *J*0660), causing them to appear above the main stellar loci in these plots. Therefore, it is expected that objects with Hα signatures will be located above these fitted lines. For fields in the main survey with low stellar density, mostly those outside the Galactic plane, this initial fit often works well (as illustrated in Figure 4). However, many fields of the Galactic disk display (at least) two distinct stellar loci in the color–color plane, resulting from differential reddening and/or contributions from both main-sequence stars and giants, where the fit is likely to align with the reddened locus (also illustrated in Figure 5).

To address this challenge in the Galactic disk, we followed the procedure implemented by Witham et al. (2008), we selected the objects above the initially fitted line and iteratively adjusted the fit, moving it upwards towards the uppermost locus of points in the color–color diagram. This upper locus generally corresponds to the unreddened main sequence, see Figure 5. In cases where the final fit is poorer than the initial one (e.g., in fields containing only a single stellar locus), we reverted to the initial fit. Once the appropriate fit for each magnitude bin was established, we identified objects significantly above the fit as likely Hα line excess candidates. During this process, we examined the color–color diagram for each field and bin to ensure the fit was suitable, and found that, in general, 2 to 3 iterations were sufficient to locate the upper locus.

This method ensures that objects exhibiting excess in Hα emission should adhere to the specified criterion:

$$(r - J0660)_{\rm obs} - (r - J0660)_{\rm fit} \geq C \times \sigma_{\rm est} \qquad (1)$$

$(r - J0660)_{\rm obs}$ denotes the observed color difference between the *r* and *J*0660 bands, $(r - J0660)_{\rm fit}$ represents the color difference predicted by the linear regression fit, *C* is a constant parameter set to 5, and $\sigma_{\rm est}$ is the estimated standard deviation of the residuals around the fit, defined as:

$$\sigma_{\rm est} = \sqrt{\sigma_s^2 + (1-m)^2 \times \sigma^2_{(r-J0660)} + m^2 \times \sigma^2_{(r-i)}} \qquad (2)$$

where σ_s represents the root mean squared value of the residuals around the fit, $\sigma_{(r-i)}$ denotes the error in the color index between the *r* and *i* bands, $\sigma_{(r-J0660)}$ denotes the error in the color index between the *r* and *J*0660 bands, and *m* represents the slope of the linear regression fit. The fits were performed using the `astropy.modeling` library [7].

Figure 4 illustrates the procedure applied to one field in the main survey (STRIPE82-0142). The iterative approach was used for each individual field, with red lines indicating the initial fit. Sources showing an excess in the *J*0660 filter, or outliers from the stellar locus, are identified as those deviating more than 5σ from these fitted lines. The selection of these sources involved applying Eq.1 to the preselected data, with σ estimated using Eq.2. The large orange star in panel c of Figure 4 represents a known Hα emitter (CV, FASTT 1560, Abril et al. 2020) that lies significantly above the stellar locus, with $(r - J0660) > 0.5$. Figure 5 shows the same procedure applied to the Galactic disk. The red lines indicate the initial fit, while the black dashed lines represent the final iterative fits.

3.2. Results and Analysis

This should be a new section

Our objective is the identification of Hα excess sources within the S-PLUS footprint, leveraging the unique filter system of the survey. This effort resulted in 3 637 outliers for the main survey and 3 319 for the Galactic disk. The distribution of the sources with excess Hα emission in the $(r - J0600)$ versus $(r - i)$ color-color plane is depicted in Figure 6. Square-shaded orange symbols represent objects with Hα excess identified in the main survey, while green circle symbols denote those found in the Galactic disk. All the sources placed above the locus of the main and giant stars exhibit an excess in the *J*0600 filter, attributed to the Hα line. The broad distribution of sources on the color-color diagram of $(r - J0660)$ and $(r - i)$ indicates the selection of several types of Hα emitters. These sources are likely associated with PNe, CVs, SySt, YSOs, Be stars, as well as extragalactic compact objects like QSOs and galaxies, among others (see Figure 2 of Gutiérrez-Soto et al. 2020).

The fractional contribution of different classes of sources to the overall sample was evaluated by cross-matching the objects' list with the `SIMBAD` database[8]. Optical spectra available in the Sloan Digital Sky Survey (SDSS; York et al. 2000) and in the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST; Wu et al. 2011) were also explored. In all cases, positive matches between the different catalogs were conisdered those sources that have an angular distance on the skyplane within a given limit ($d_{max,proj}$). Verification of the photometry and assessment of Hα excess in the selected objects within the disk area were conducted by cross-matching the Hα source list identified in S-PLUS with photometric data from VPHAS+ DR2.

3.2.1. Matches with SIMBAD sources

We identified a total of 1 263 positive matches between our catalogs of Hα compact excess sources and the `SIMBAD` database, assuming a search radius of $d_{\rm max,proj} = 2$ arcsec for the main survey and 1 arcsec for the Galactic disk. In the main survey, the identified objects primarily fall into categories such as variable stars, predominantly cataclysmic variables and/or candidates (CataclyV*), eclipsing binaries and/or candidates (EB*), RR Lyrae Variables (RRLyr), as well as various kinds of stars including normal stars, white dwarfs, and/or candidates (WD*). Additionally, extragalactic compact sources that exhibit redshifted lines that coincide with the *J*0660 filter, simulating the Hα emission line, are also present, encompassing AGN, Seyfert galaxies, QSOs, and various types of other objects (see Table 2 for details).

For the disk, the identified categories include emission-line stars (Em*), young stellar objects (YSO) and candidates, which encompass T Tauri (TTau*) and Herbig Ae/Be (Ae*) star candidates. Additionally, variable stars such as cataclysmic variables

[7] `https://docs.astropy.org/en/stable/modeling/index.html`

[8] `http://simbad.u-strasbg.fr/simbad/`

Page:6

Author: Edu Telles Subject: Note Date: 2024-09-30 09:18:07

maybe another sentence describing a bit more clearly what you are fitting.

also not clear what you are clipping the lower 3 sigmas to find the upper envelope?!!

Author: ET Subject: Oval Date: 2024-09-30 09:19:23

Author: ET Subject: Typewriter Date: 2024-09-30 09:20:07

This should be a new section

Author: Edu Telles Subject: Note Date: 2024-09-30 09:24:36

/Figure 6 is too far down the text. reorder so that the figures are near the where they are referred to.

it would help if youhad only 3 bins. In fact, your results now, independ of the bins!!

Author: Edu Telles Subject: Note Date: 2024-09-30 09:10:03

Here, I begin to think that the sample could have been divided in 3 magnitude bins. The four bins are not being instructive and also occuppies much more space.

Author: ET Subject: Underline Date: 2024-09-30 09:25:29

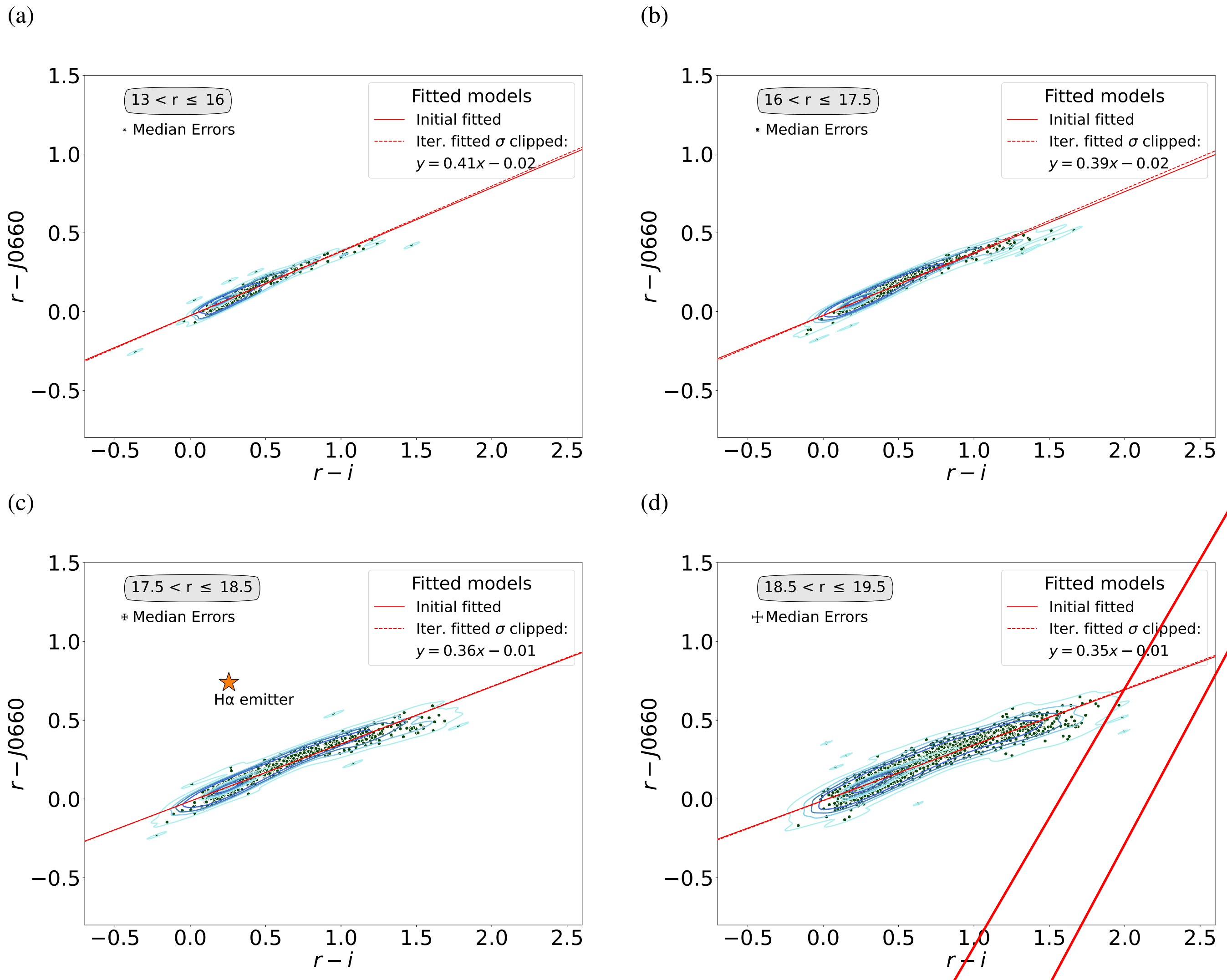

Fig. 4. An illustration of the selection criteria used to identify strong emission-line objects via colour-colour plots. The data shown here are all from the S-PLUS field STRIPE82-0142. The data are split into four magnitude bins, as shown in the four panels. Objects with Hα excess should be located near the top of the color-color diagrams. The thin red continuous lines illustrate the original linear fit to all data (green points). The fit line equation is provided in the legend. Objects selected as Hα emitters must be located above the dashed line. The orange star in the plot (*c*) is the CV namely FASTT 1560 and S-PLUS ID DR4_3_STRIPE82-0142_0021237.

not visible! maybe not value the original but only the final fit after iterations!

(CataclyV*), eclipsing binaries (EB*), and RR Lyrae variables (RRLyr) are found, along with objects exhibiting nebular components, such as planetary nebula (PN) candidates, novae, and reflection nebulae (RfNeb), among others. As shown in Table 2, the highest number of sources in the disk belong to the Em* and young stellar objects category, which is expected, reflecting the active star formation processes present in the Galactic disk.

An important consideration regarding the `SIMBAD` matches is that in the main survey, numerous extragalactic sources with emission lines are selected due to the mapping of high latitudes in the southern sky. Conversely, for the disk, no extragalactic sources have been selected. While the main survey emphasizes extragalactic sources and diverse stellar populations, the disk region primarily showcases young stellar objects and variable stars, indicative of ongoing star formation and stellar evolution processes. In both regions, variable stars such as eclipsing binaries (EB*), among others, are also present. The results are described below and listed in Table 2.

In our analysis of Hα-excess sources, variable stars, such as RR Lyrae stars and eclipsing binaries, are often detected due to their ability to exhibit significant Hα features. It is important to note that RR Lyrae stars, known for their distinctive spectral features, can display Hα absorption lines, occasionally causing them to be identified as outliers in our analysis. This detection is a natural outcome of our selection criterion, which identifies any significant deviation from the expected stellar colors, encompassing both emission and absorption features. Eclipsing binaries often show Hα emission due to complex interactions between their components and surrounding material, as demonstrated in studies of systems like the eclipsing binary VV Cephei, where periodic variations in Hα emission have been observed throughout eclipse phases (Pollmann et al. 2018).

An important observation is that our selection criteria have predominantly excluded extended sources. In the main survey, only 23 AGN and 9 galaxies were identified, making up approximately 3.1% and 1.2% of the total matches, respectively. Additionally, we identified 143 QSOs, representing about 19.6% of the total matches. These percentages highlight the effectiveness of our selection criteria in isolating compact sources with signif-

Page:7

Author: ET Subject: Oval Date: 2024-09-30 09:13:24

Author: ET Subject: Callout Date: 2024-09-30 09:14:31

not visible! maybe not value the original but only the final fit after iterations!

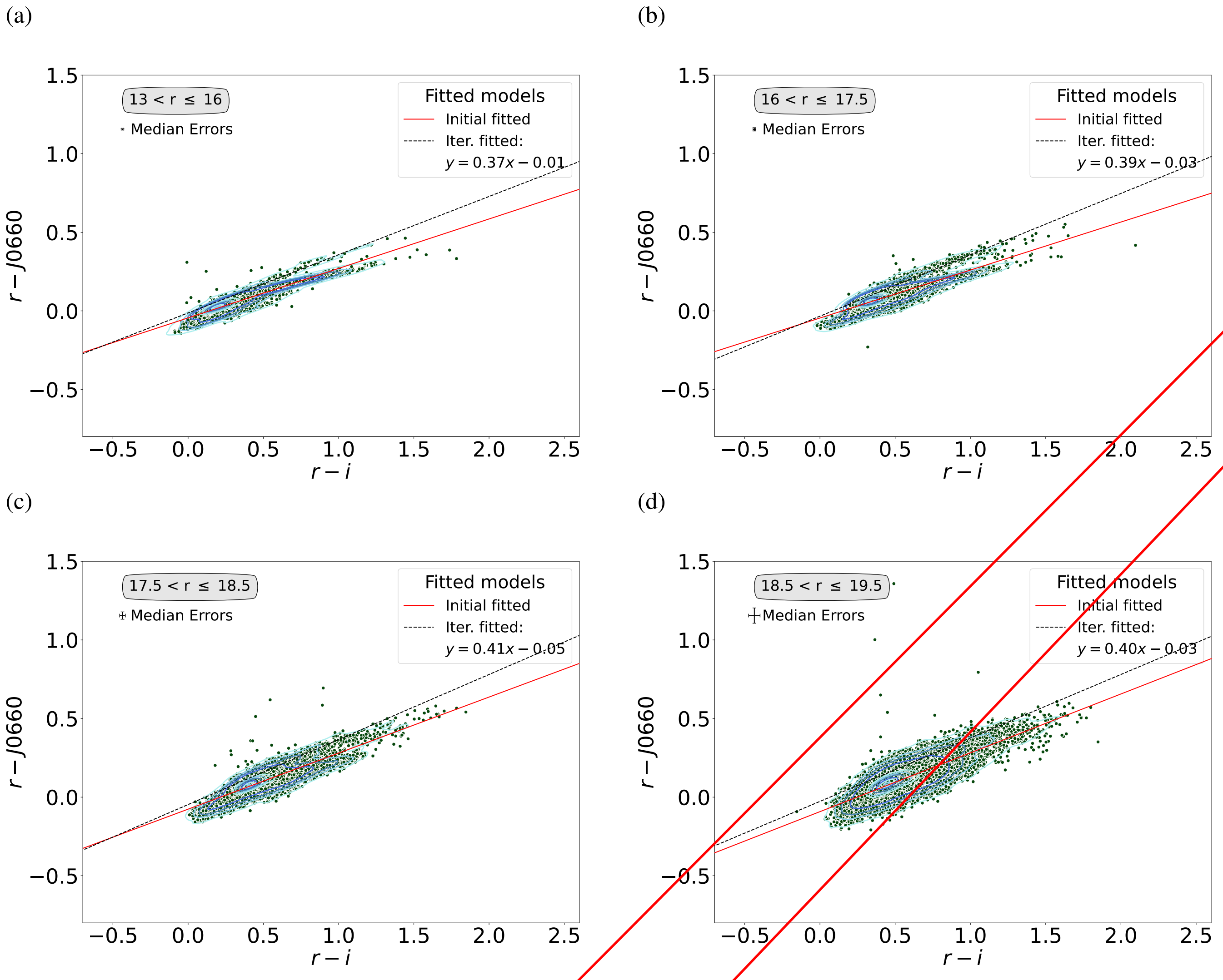

Fig. 5. As in Figure 4, but for the Galactic disk. The red lines represent the original fit to all data, while the black dashed lines represent the final fits to the upper locus of points, obtained by applying an iterative fitting process to the initial fit.

icant Hα excess, while also illustrating the relative proportions of different astrophysical categories identified in our survey.

Redshifted Lines Mimicking the Hα Emission. According to the classification in the literature, a significant portion of the Hα excess sources in our sample are classified as QSOs. It is important to note that the excess observed in the $J0660$ filter for QSOs is due to redshifted emission lines that fall within the wavelength range of this filter, depending on the redshift of the QSOs. For instance, lines such as Hβ, Mg ii 2798 Å, C iii] 1909 Å, and C iv 1550 Å can contribute to this excess (see Gutiérrez-Soto et al. 2020 and Nakazono et al. 2021).

This particular population of apparent Hα emitters includes AGNs, Seyfert 1 galaxies, and other emission-line galaxies. In particular, within the redshift range $0.306 < z < 0.376$, lines such as Hβ and [O iii] 4959, 5007 Å are redshifted into the $J0660$ filter.

matches with SDSS and LAMOST

3.2.2. Spectral Analysis

Our list of Hα-excess sources identified in the main S-PLUS survey was cross-matched with the DR18 SDSS catalog (Ahumada et al. 2020) and the DR9 LAMOST catalog, using a 2 arcsec radius. This cross-matching identified 212 common sources (138 from SDSS and 74 from LAMOST). This procedure was restricted to the main survey due to its overlap with SDSS and LAMOST areas, unlike the S-PLUS Galactic disk survey. It is noteworthy that some Hα-excess sources detected by our algorithm may exhibit transient behavior, meaning that Hα-excess features might be present in spectra from one survey (SDSS or LAMOST) but not in others (S-PLUS), or vice versa. This variability is attributed to differences in observational epochs and conditions across the surveys. Upon spectroscopic examination, approximately 60% of these sources exhibited emission lines, which might include redshifted lines other than Hα, while about 30% showed Hα-related absorption features.

Most of the objects with available spectroscopic information in SDSS and LAMOST correspond to CVs, QSOs, AGN, and variable stars. A more detailed spectroscopic characterization of these sources is out the scope of this paper. Also, it is worth noticing that there is a number of objects without a conclusive classification.

Figure 7 presents the SDSS (upper) and LAMOST (lower) spectra, along with the corresponding S-PLUS photometry (colored symbols) for two known cataclysmic variables (CVs). The

Page:8

Author: Edu Telles Subject: Note Date: 2024-09-30 09:37:16

I wonder if this is compatible if the luminosity function of quasars at this redshift!
in other words, the observed magnitudes are compatible?!

Author: ET Subject: Callout Date: 2024-09-30 09:38:41

matches with SDSS and LAMOST

Author: Eduardo Telles Subject: Insert Text Date: 2024-09-30 09:45:04

and one eclipsing binary, respectively.

3.2.3. Evaluation of Photometric Color Consistency Between S-PLUS and VPHAS+

We performed a comparative analysis of PSF photometric colors between the S-PLUS data from the Galactic disk and those provided by VPHAS+[9]. For the crossmatching, we considered a radius of 1" and ended up with a number of 793 matches. We computed the differences in two key color indices: $r - i$ and r − Hα. Specifically, we investigated the median difference and the median absolute deviation (MAD) of these colors to assess the consistency and agreement between the two surveys. It is worth noting that VPHAS+, like S-PLUS, employs the r, i, and a narrowband filter (NB-659) designed to detect the Hα line, facilitating a meaningful comparison of Hα emission.

The comparison of colors reveals important insights into the consistency and reliability of S-PLUS photometry (see Figure 11). The median difference in the $r-i$ color between S-PLUS and VPHAS+ was −0.21, with a MAD of 0.07. For the r − Hα color, the median difference was 0.02 with a MAD of 0.27. These results indicate a systematic offset between the photometric colors of the two surveys, which is within the expected range considering differences in instrumentation and filter systems.

A key factor contributing to the differences in the r − Hα color index is the distinct characteristics of the Hα filters used in S-PLUS and VPHAS+. The S-PLUS Hα filter (J0660) has an effective wavelength of 6614 Å and a width of 14.7 nm, whereas the VPHAS+ NB-659 filter has an effective wavelength of 6588 Å and a width of 10.7 nm. These differences can significantly affect the measurement of Hα excess, as the narrower VPHAS+ filter captures a more specific range of wavelengths, potentially leading to higher precision. The broader S-PLUS filter, on the other hand, may include additional continuum emission, affecting the photometric measurement. Additionally, the exposure times in the two surveys differ, with VPHAS+ using a 120-second exposure and S-PLUS using a 290-second exposure. The longer exposure time in S-PLUS allows for greater sensitivity to faint sources and potentially higher signal-to-noise ratios (SNR), contributing to the observed differences in photometric colors.

Despite the observed systematic differences, the MAD values suggest that the photometric measurements from both surveys exhibit good agreement. This consistency is crucial for cross-referencing and integrating datasets from different surveys for comprehensive astrophysical studies. The observed differences in photometric colors may result from various factors, including differences in filter characteristics, photometric calibration, and data processing techniques. Further investigations are warranted to better understand these factors' contributions to the observed discrepancies.

3.2.4. Hα Excess Source Distributions

The upper panel of Figure 12 presents a histogram of the r-band magnitude distribution for all objects in our study from the main survey. The normalized density facilitates comparison between different subsets. The blue curve represents Hα excess objects, while the red curve represents all main stars. The magnitude distribution for Hα excess sources shows a higher concentration at intermediate magnitudes. The lower panel of Figure 12 focuses on the r-band magnitude distribution for the subset of Hα excess objects in the disk. A noticeable large number of source with Hα excess have magnitudes in the r-band between 13 and 13.5, something that we do not see in the stars ~~of Galactic disk. Thi~~s implies that Hα excess objects could be intrinsically more luminous or closer to us than the general population of all stars. However, these stars are closer to the saturation limit. Therefore, we recommend exercising caution with all sources in our sample that have an r-band magnitude less than 13.5.

Figure 13 shows the distribution of all Hα excess sources in Galactic latitude and longitude, along with a zoomed-in view of the Galactic disk in the bottom panel. The distribution of objects in Galactic longitude for the main survey (left panel of Figure 14) indicates that the blue bars, representing Hα excess sources, are relatively evenly spread across the Galactic longitude, similar to the general population of main stars (pink bars). Peaks are observed around Galactic longitudes of 15°, 50°, and 270°, which are also present in the general star population of the main survey.

The bottom panel of Figure 13 and the right panel of Figure 14 show the distribution of objects in Galactic longitude specifically within the Galactic disk. There is a noticeable concentration of Hα excess sources at specific longitudes, particularly around 243°. Additionally, there are small peaks around 225° and 268° in Galactic longitude. While Hα excess sources follow a distribution similar to that of all stars, the peaks are more pronounced for Hα excess sources.

4. Machine Learning Approaches

In this section, inspired by the goal of separating Galactic sources from extragalactic ones in our Hα excess list, we applied machine learning approaches. Our list of Hα excess sources selected in the main survey of S-PLUS naturally includes extragalactic compact objects with redshifted lines detected in the J0660 filter. To classify the sources in our Hα excess list, we utilized the multi-band coverage provided by S-PLUS optical photometry. To achieve this, we employed two unsupervised machine learning algorithms: UMAP and HDBSCAN. UMAP is used to reduce the dimensions of our data and perform a feature extraction, while HDBSCAN classifies the data based on the results from UMAP. We conducted two experiments: one using the 66 colors generated from the 12 S-PLUS filters, and a second one by adding filters from the Wide-Field Infrared Survey Explorer (Wright et al. 2010, WISE). Additionally, we used a Random Forest algorithm to identify important features and construct color-color diagrams to separate the classes of objects identified by HDBSCAN. This methodology is applied to the list of Ha excess sources obtained from the main survey of S-PLUS.

4.1. Dimensionality Reduction and Clustering

4.1.1. UMAP

Uniform Manifold Approximation and Projection (UMAP; Becht et al. 2018; McInnes et al. 2020) is a dimensionality reduction algorithm designed to handle high-dimensional data while preserving its underlying structure. Unlike some other techniques, UMAP is based on a mathematical framework that combines aspects of Riemannian geometry and algebraic topology. This enables UMAP to capture both local and global relationships within the data. UMAP aims to create a low-dimensional representation that retains the intricate nonlinear relationships present in the original high-dimensional features. This process involves constructing a high-dimensional graph representation of the data and then optimizing a low-dimensional graph to

[9] More detailed information about the VPHAS+ survey can be found at: https://www.vphasplus.org/

Page:11

Author: Eduardo Telles Subject: Replace Text Date: 2024-09-30 10:00:41

in the main survey.

Author: Edu Telles Subject: Note Date: 2024-09-30 10:04:01

Figure 14 is not essential

Author: Edu Telles Subject: Note Date: 2024-09-30 09:58:16

done with the 2.6m VST telescope.

so it is a much larget telescope/!

Author: Edu Telles Subject: Note Date: 2024-09-30 11:29:06

did you include W1, W2, W3 and W4 or only W1 and W2?

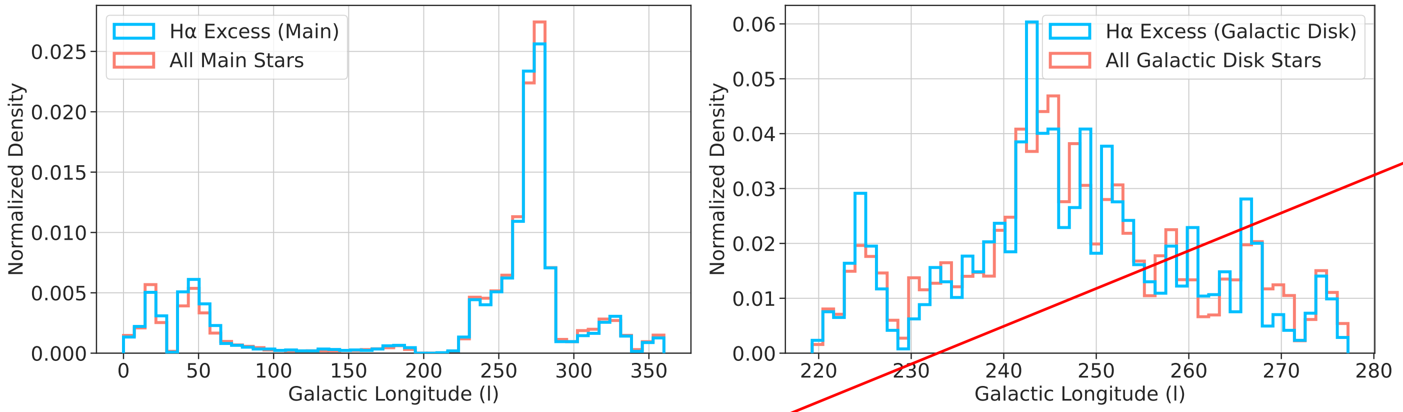

Fig. 14. Distribution of the objects in galactic longitude for Hα excess sources (blue bars) and all stars (pink bars) for the main survey (*left panel*) and the Galactic disk (*right panel*).

For the implementation of the algorithm, we used the Python package `umap`[10]. UMAP has three key hyperparameters: `n_neighbors`, `n_components`, and `min_dist`.

The `n_neighbors` parameter balances local versus global structures in the data by setting the number of neighboring points UMAP considers for each data point when learning the manifold structure. Low values of `n_neighbors` cause UMAP to focus on very local structures, while higher values make UMAP look at larger neighborhoods, potentially losing fine details in favor of capturing broader patterns.

The `n_components` parameter, similar to the parameter used in standard dimension reduction algorithms in the `scikit-learn` package, allows us to set the number of dimensions in the reduced space into which we will embed the data.

The `min_dist` parameter controls how closely UMAP can pack points together in the low-dimensional representation. Lower values result in clumpier embeddings, which are useful for clustering and capturing fine topological structures, while higher values focus on preserving broader topological structures.

4.1.2. HDBSCAN

After obtaining a new system of reduced variables that condenses all the information from the original variables, we utilized HDBSCAN to identify clusters within the data. This clustering approach complements the reduction achieved by UMAP, allowing for a comprehensive understanding of the underlying structure of the dataset.

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN; Campello et al. 2013) is an unsupervised machine learning algorithm for clustering. It builds on the density-based spatial clustering of applications with noise (DBSCAN; Ester et al. 1996) by introducing a hierarchy to the clustering process, which allows for the extraction of "persistent" clusters from the hierarchical tree. HDBSCAN's main advantage over DBSCAN is its ability to find clusters of varying densities and shapes.

For this task, we adopted the Python implementation of `HDBSCAN`[11] (McInnes et al. 2017). The two most critical parameters are the "minimum cluster size" (`min_cluster_size`) and "minimum number of samples" (`min_samples`). The "minimum cluster size" refers to the smallest group size that is considered a cluster. The "minimum number of samples" determines how conservative the clustering will be; larger values result in more points being classified as noise, restricting clusters to denser areas.

HDBSCAN can also classify sources as noise if they do not fit well into any cluster based on these parameters. Additionally, the algorithm relies on a distance metric, such as Euclidean distance, to measure the distance between points and determine their density. The choice of metric can significantly affect the clustering results, as it influences how distances are computed and, consequently, how clusters are formed.

4.2. Classification Results

Our unsupervised UMAP model projects the data, and HDBSCAN subsequently identifies the clusters. To ensure high-quality photometry, we set the criteria for the error to be less than 0.2 in all filters. This results in a list of 2181 objects for the main survey. The reduction of the sample size is crucial to mitigate the influence of noisy or unreliable photometric measurements, which can significantly affect the performance and outcomes of both the UMAP and HDBSCAN algorithms. By focusing on high-quality photometric data, we aimed to enhance the accuracy and robustness of the clustering results, thereby providing more reliable classifications of the Hα excess sources.

To perform cross-validation for selecting the optimal `n_neighbors` and `n_components` parameters in UMAP, we systematically explored a range of values for these parameters. The selection of parameters n_neighbors and n_components in UMAP is critical as it directly influences the quality of the reduced-dimensional representation. Initially, we conducted exploratory data analysis to visualize the dataset in reduced dimensions using various combinations of n_neighbors and n_components. This allowed us to qualitatively assess how well UMAP preserved the underlying structure of the data. We employed quantitative metrics, including the Silhouette Score and Davies-Bouldin Index, to objectively evaluate the performance of different parameter combinations. Silhouette Score measures how well-defined the clusters are in the reduced space, with higher values indicating better separation between clusters, while Davies-Bouldin Index evaluates the average similarity between each cluster and its most similar cluster, with lower values indicating better-defined clusters. A grid of tests was constructed over a range of n_neighbors (5, 10, 15, 20, 30, 50, 70, 100) and n_components (2, 3, 4, 5, 10, 20, 50) values. For each

[10] For more details, see `https://umap-learn.readthedocs.io/en/latest/index.html`

[11] `https://hdbscan.readthedocs.io/en/latest/`

Page:13

Author: Edu Telles Subject: Note Date: 2024-09-30 11:24:57

I just think that the input has an enormous amount of redundancy since you are using all filter colors. what if you tested with only Halpha as input? you would then have 21 colors. Mu question is if the "signal" of Halpha emitters would outstand, in the statistical process. I have no idea.

Group 1 ontains 166 objects, 149 of which have matches in SIMBAD. This group is predominantly composed of RR Lyrae stars (107), followed by eclipsing binaries (19), various types of pulsating variables (9), and a few other stellar objects, including 2 QSOs. This group appears to represent objects with Hα in absorption, as it is well known that RR Lyrae stars exhibit Hα absorption lines.

Group 2 includes 1539 objects, 323 of which have matches in SIMBAD. The majority are eclipsing binaries (275), followed by a few stars (10), QSOs (9), and a small number of cataclysmic variables and RR Lyrae stars. This group is characterized by the significant presence of binary star systems and various types of variable stars.

Group 3 consists of 93 objects, 42 of which have matches in SIMBAD. The group is predominantly composed of QSOs (17) and Seyfert 1 galaxies (10). Other identifications include AGN candidates, radio sources, and a few galaxies. The redshift distribution for extragalactic objects in this group varies in a very narrow range of values from 0.31 to 0.37.

Group 4 includes 325 objects, 143 of which have matches in SIMBAD. This group has a high concentration of QSOs (78) and cataclysmic variables (25). Additionally, it features a mix of blue stars, AGNs, radio sources, and white dwarf candidates. The extragalactic objects in this group show a peak in the redshift distribution around 1.35. It is expected that CVs are located closer to the QSOs than to Galactic sources in the UMAP variable space due to their photometric characteristics, which can resemble those of QSOs in certain features, despite the spectral differences (Scaringi et al. 2013).

In summary, our application of UMAP and HDBSCAN to the Hα excess sources has effectively identified distinct groups with varying astrophysical characteristics using S-PLUS photometry. The classification successfully differentiates extragalactic sources, such as QSOs and AGNs, from galactic sources, including variable stars and binary systems. However, distinguishing Galactic cataclysmic variables from QSOs with redshifts around 1.35 remains challenging. Importantly, our results suggest that objects with ($J0660$ - r) color excess due to emission lines can be distinguished from those with excess caused by Hα absorption lines, mainly RR Lyrae stars. This separation enhances our understanding of the objects in our dataset and provides a solid foundation for further detailed analysis.

4.2.2. Integration of S-PLUS and WISE Photometry

The second experiment included the W1 and W2 WISE filters, adding new colors to the original variable set used for the machine learning models. To do this, we first crossmatched the Hα sources of the main survey with the ALLWISE catalog using a search radius of 2 arcsec, obtaining 3,173 matches. This number was then reduced to 1,910 after applying error criteria to the S-PLUS filters and filtering for objects with errors less than 0.5 in the W1 and W2 filters. The additional colors combined the WISE bands (W1 and W2) with the S-PLUS broadband filters. For instance, we calculated colors such as W1 - W2, W1 - u, W2 - u, W1 - g, W2 - g, W1 - r, W2 - r, and so on. This resulted in 11 new colors being added to the original 66 S-PLUS colors, generating a dataset with 77 variables in total.

Figure 16 shows the results of the reduction in dimensionality and the groups identified by applying UMAP followed by HDBSCAN, using the input parameters described in the previous paragraph. On this occasion, HDBSCAN found five groups and five objects that were classified as noise. Table 3 summarizes these results:

Group 0 contains 1,437 objects, 424 of which match with SIMBAD. Among these, 262 are eclipsing binaries (EB*), followed by 98 RR Lyrae stars (RRLyr). Other objects include EB* candidates, stars, pulsating variables, and a few QSOs. This group predominantly consists of variable stars and a small number of extragalactic sources.

Group 1 includes 59 objects, 23 of which have matches in SIMBAD. The majority are QSOs (20), with a few other objects like a galaxy, a radio source, and a QSO candidate. This group mainly represents extragalactic sources, particularly active galactic nuclei. The redshift distribution has a peak around 2.45.

Group 2 consists of 93 objects, 43 of which have SIMBAD matches. The group is primarily composed of QSOs (18), Seyfert 1 galaxies (10), and AGN candidates, with some galaxies and radio sources. This indicates a strong presence of active galactic nuclei and other extragalactic objects. The redshift distribution for extragalactic objects in this group ranges approximately from 0.31 to 0.37.

Group 3 includes 51 objects, with 36 matches in SIMBAD. The majority are cataclysmic variables (24), with a few CV candidates, hot subdwarf candidates, and white dwarf candidates. This group is largely composed of cataclysmic variables and related stellar objects.

Group 4 contains 269 objects, 100 of which have SIMBAD matches. The majority are QSOs (83), with a mix of blue stars, AGNs, radio sources, stars, and galaxies. This group shows a variety of astrophysical phenomena, both stellar and extragalactic. The redshift distribution has a peak around 1.35.

In summary, the inclusion of WISE filters in our analysis has significantly enhanced the clustering of Hα excess sources. The integration of WISE data has allowed for a more precise differentiation between galactic and extragalactic sources, enriching our understanding of the objects in our dataset. Notably, it has facilitated the separation of cataclysmic variables from QSOs with redshifts around 1.35. For detailed insights, refer to Section 3.2 where the redshifted emission lines of extragalactic objects are highlighted in the $J0660$ filter. However, it is important to note that the addition of WISE data has introduced challenges in identifying the group of RR Lyrae stars using HDBSCAN.

4.3. Extracting Main Features: Color Analysis

Based on the important features extracted from the Random Forest model, we focused on the colors derived from the S-PLUS and WISE filters. These colors are effective in distinguishing the different groups of Hα-excess objects identified by the combined UMAP and HDBSCAN analysis of the S-PLUS data.

In the main survey of the Hα-excess list, we identified extragalactic emitters with red-shifted bluer emission lines resembling the Hα emission line, with sources having a redshift of less than 0.02. By incorporating the WISE filters to create additional colors for the unsupervised machine learning models, we achieved better separation of extragalactic sources from Galactic sources (see Sect. 4.2 for more details).

We used the classifications made by combining UMAP and HDBSCAN to create Random Forest (Breiman 2001) models and identified the most important features, specifically the colors that contribute to the separation or classification of the classes of objects. We implemented the `scikit-learn` package for the Random Forest algorithm, using 66 S-PLUS colors plus 11 additional colors generated with the W1 and W2 filters as input parameters, and labels generated by HDBSCAN. The dataset used in this study exhibited a class imbalance: cluster 0 (1437 points),

Page:15

Author: ET Subject: Oval Date: 2024-09-30 11:26:18

Author: ET Subject: Oval Date: 2024-09-30 11:26:22

Author: Edu Telles Subject: Note Date: 2024-09-30 11:29:32

did you include W1, W2, W3 and W4 or only W1 and W2?

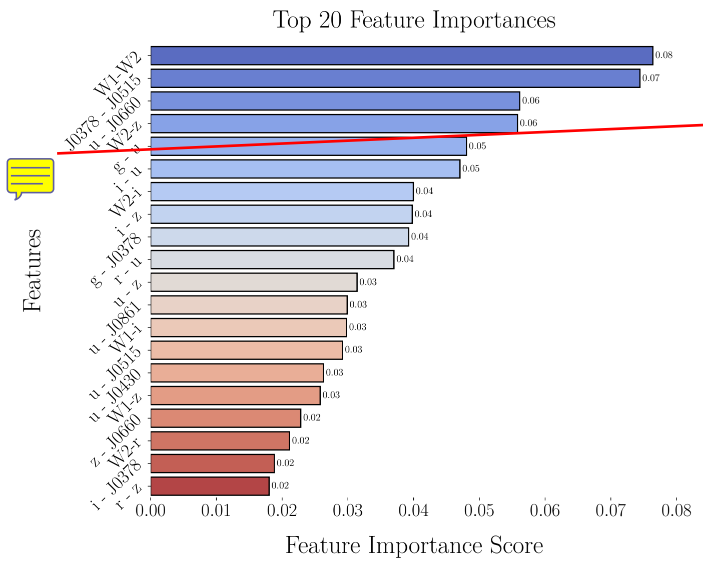

Fig. 17. Top 20 feature importances identified by the Random Forest model, showing the colors that contributed most significantly to the clustering of Hα-excess objects using UMAP + HDBSCAN. The importance scores indicate the relative impact of each color on the classification of different object classes.

This exercise demonstrates that using specific color-color diagrams with selected filters can effectively classify objects clustered using machine learning techniques. By relying on a few key colors instead of all 12 S-PLUS filters and 2 WISE filters required for the machine learning in Section 4.2, we can reduce the number of necessary observations. This approach is advantageous because not all objects have complete photometry in all filters, and some magnitudes may not meet the clean criteria, reducing the number of objects available for classification. Consequently, using a few specific color criteria enables the classification of more objects, as it circumvents the need for complete data across all filters.

5. Conclusions

In this study, we have leveraged the S-PLUS project to analyze and classify Hα-excess sources in the Southern Sky, resulting in the following key conclusions:

1. We identified 6 956 Hα-excess candidates by using the narrow $J0660$ filter in combination with the broad r and i filters from S-PLUS. This included 3 637 candidates from the high-latitude main survey and 3 319 from the Galactic disk.
2. Cross-referencing with the SIMBAD database allowed us to classify these candidates into various emission line objects, such as EM stars, YSOs, Be stars, CVs, PNe, among others. We also identified QSOs, non-local galaxies, and objects with Hα in absorption, including RR Lyrae stars, primarily within the main survey. The higher detection of RR Lyrae stars (111) in the main survey compared to the disk (8), based on SIMBAD, aligns with their expected distribution in older stellar populations.
3. Validation with spectroscopic data from LAMOST and SDSS showed that approximately 60% of the spectra exhibit Hα emission lines, while around 30% show Hα in absorption in the main survey. This comparison indicates the general accuracy of our classifications and supports the reliability of our Hα-excess source identifications. Furthermore, the VPHAS+ data for the Galactic disk are consistent with our findings.
4. Employing machine learning techniques, specifically UMAP for dimensionality reduction and HDBSCAN for clustering, enhanced our analysis of Hα-excess sources. The 12 S-PLUS filters enabled effective differentiation between Hα-emission Galactic objects and extragalactic sources, as well as those with Hα in absorption, such as RR Lyrae stars. However, distinguishing cataclysmic variables from QSOs or AGN with redshifts around 1.35 remained challenging.
5. The integration of WISE filter data refined our clustering process, improving the separation of extragalactic sources from Galactic ones and aiding in the differentiation between cataclysmic variables and QSOs. Despite this improvement, the inclusion of WISE data introduced difficulties in classifying RR Lyrae stars using HDBSCAN, although they were clearly grouped in the UMAP space.
6. Incorporating WISE data into our Random Forest model was crucial for identifying the most significant features for classification. This enhancement led to more effective color-color diagrams and improved our understanding of various Hα-related phenomena.

Our study used observational and analytical techniques to gain valuable insights into Hα-excess sources. Although challenges were encountered, particularly with RR Lyrae stars and certain extragalactic objects, our methods provided a robust framework for understanding Hα-excess phenomena. Future research should focus on expanding sample sizes and incorporating additional spectroscopic data to further refine classifications. Applying these methods to other sky regions or wavelengths could enhance our understanding of Hα-excess sources and their astrophysical contexts.

Acknowledgements

LAG-S acknowledges funding for this work from CONICET and FAPESP grants 2019/26412-0. RLO acknowledges financial support from the Brazilian institutions CNPq (PQ-312705/2020-4) and FAPESP (#2020/00457-4). DGR acknowledges the CNPq (428330/2018-5; 313016/2020-8) and FAPERJ (269312) grants. F. A. -F. acknowledges funding for this work from FAPESP grants 2018/20977-2 and 2021/09468-1. FRH ackowledges funding from FAPESP through the project 2018/21661-9. C. C. is supported by the National Natural Science Foundation of China, No. 11803044, 11933003, 12173045. This work is sponsored (in part) by the Chinese Academy of Sciences (CAS), through a grant to the CAS South America Center for Astronomy (CASSACA). We acknowledge the science research grants from the China Manned Space Project with NO. CMS-CSST-2021-A05. AAC acknowledges support from the State Agency for Research of the Spanish MCIU through the "Center of Excellence Severo Ochoa" award to the Instituto de Astrofísica de Andalucía (SEV-2017-0709). The authors would like to thank Amanda Reis Lopes for her useful suggestions and comments.

The S-PLUS project, including the T80-South robotic telescope and the S-PLUS scientific survey, was founded as a partnership between the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), the Observatório Nacional (ON), the Federal University of Sergipe (UFS), and the Federal University of Santa Catarina (UFSC), with important financial and practical contributions from other collaborating institutes in Brazil, Chile (Universidad de La Serena), and Spain (Centro de Estudios de

Page:17

Author: Edu Telles Subject: Note Date: 2024-09-30 11:37:02

Here , there is sth strange.
The result for g -u is the same as u-g?
I would prefer to have only the bluer - redder filter for the colors!

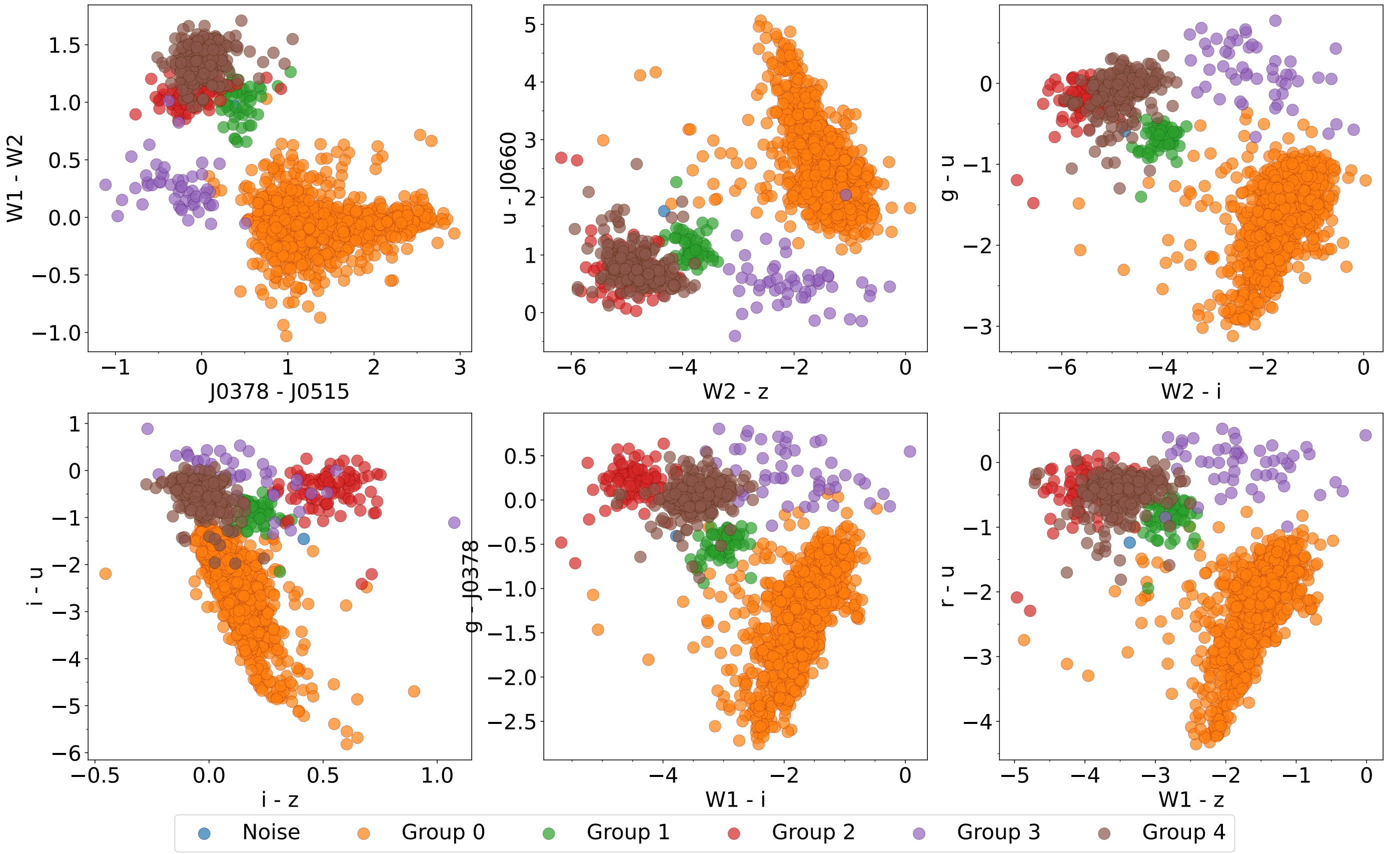

Fig. 18. Examples of color-color diagrams using the top 20 features identified by the Random Forest model. These diagrams illustrate the separation of different classes of objects found in the Hα-excess sources list. The selected diagrams demonstrate effective clustering achieved through UMAP + HDBSCAN, highlighting the key colors that contribute to the classification.

Física del Cosmos de Aragón, CEFCA). We further acknowledge financial support from the São Paulo Research Foundation (FAPESP), the Brazilian National Research Council (CNPq), the Coordination for the Improvement of Higher Education Personnel (CAPES), the Carlos Chagas Filho Rio de Janeiro State Research Foundation (FAPERJ), and the Brazilian Innovation Agency (FINEP).

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is `http://www.sdss.org/`.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

Guoshoujing Telescope (the Large Sky Area Multi-Object Fiber Spectroscopic Telescope LAMOST) is a National Major Scientific Project built by the Chinese Academy of Sciences. Funding for the project has been provided by the National Development and Reform Commission. LAMOST is operated and managed by the National Astronomical Observatories, Chinese Academy of Sciences.

Scientific software and databases used in this work include TOPCAT[13] (Taylor 2005), simbad and vizier from Strasbourg Astronomical Data Center (CDS)[14] and the following python packages: `numpy`, `astropy`, `matplotlib`, `seaborn`, `pandas`, `scikit-learn`, `hdbscan`, `umap`.

References

Abril, J., Schmidtobreick, L., Ederoclite, A., & López-Sanjuan, C. 2020, MNRAS, 492, L40
Ahumada, R., Prieto, C. A., Almeida, A., et al. 2020, ApJS, 249, 3
Akras, S. 2023, MNRAS, 519, 6044
Akras, S., Gonçalves, D. R., Alvarez-Candal, A., & Pereira, C. B. 2021, MNRAS, 502, 2513
Akras, S., Guzman-Ramirez, L., & Gonçalves, D. R. 2019a, MNRAS, 488, 3238
Akras, S., Guzman-Ramirez, L., Leal-Ferreira, M. L., & Ramos-Larios, G. 2019b, ApJS, 240, 21
Akras, S., Leal-Ferreira, M. L., Guzman-Ramirez, L., & Ramos-Larios, G. 2019c, MNRAS, 483, 5077

[13] `http://www.star.bristol.ac.uk/~mbt/topcat/`
[14] `https://cds.u-strasbg.fr/`

Page:18

Author: Edu Telles Subject: Note Date: 2024-09-30 11:44:11

here too.
I prefer the bluer - redder filter for the colors

I think this figure should have the most important color-color plots to separate the classes, no examples.

and maybe eliminate appendix B.