Unsupervised Machine Learning for the Classification of Astrophysical X-ray Sources

Autor
Víctor Samuel Pérez Díaz

Trabajo presentado como requisito para optar por el título de
Profesional en Matemáticas Aplicadas y Ciencias de la Computación

Director, Tutor
Juan R. Martinez-Galarza

Codirector
Alexander Caicedo-Dorado

Escuela de Ingeniería, Ciencia y Tecnología
Matemáticas Aplicadas y Ciencias de la Computación
Universidad del Rosario

Bogotá - Colombia
2021

UNIVERSIDAD DEL ROSARIO

UNDERGRADUATE THESIS

Unsupervised Machine Learning for the Classification of Astrophysical X-ray Sources

Author:
Víctor Samuel PÉREZ DÍAZ

Supervisor:
Rafael MARTÍNEZ-GALARZA

Cosupervisor:
Alexander CAICEDO-DORADO

*A thesis submitted in fulfillment of the requirements
for the degree of Professional in Applied Mathematics and Computer Science*

Center for Astrophysics | Harvard & Smithsonian
Applied Mathematics and Computer Science

November 2021

UNIVERSIDAD DEL ROSARIO

Abstract

School of Engineering, Science and Technology
Applied Mathematics and Computer Science

Professional in Applied Mathematics and Computer Science

Unsupervised Machine Learning for the Classification of Astrophysical X-ray Sources

by Víctor Samuel PÉREZ DÍAZ

Context. The Chandra Source Catalog (CSC), which collects the X-ray sources detected by the Chandra X-ray Observatory through its history, is a fertile ground for discovery, because many of the sources it contains have not been studied in detail. In CSC we could find several types of sources, from young stellar objects (YSO) and binary systems, to even very far quasars (QSO) or active galaxies with supermassive black holes in their cores. Among the potentially paradigm changing sources that we could look for in Chandra data are compact object mergers, extrasolar planet transits, tidal disruption events, etc. However, only a small fraction of the CSC sources have been classified. In order to conduct a thorough investigation of the CSC sources, and to be prepared for the coming very large X-ray surveys, we need to classify as many catalog sources as possible.

Aims. This work proposes an unsupervised learning approach to classify as many Chandra Source Catalog sources as possible, first exploring the advantages and limits of using only the X-ray data available. Unsupervised learning is particularly suitable given the vast amount of detections that have not been independently classified yet. Clustering the source observations by their similarities, and then associating these clusters with objects previously classified spectroscopically, we aim to propose a new methodology that could provide us with a probabilistic classification for a numerous amount of sources.

Methods. We employ unsupervised learning methods, first K-means, then focusing on Gaussian Mixtures, applied to a list of X-ray properties, to probabilistically classify high energy sources in the Chandra Source Catalog (CSC). We achieve this by associating specific clusters with those CSC objects that have a classification in the SIMBAD database, and then assigning probabilistic classes by association to unclassified objects in each cluster with an algorithm based on the Mahalanobis distance.

Results. We are able to successfully identify clusters of previously identified objects that likely belong to the same class, and even within groups that were identified as having predominantly a type of source, such as "galaxies", "QSO", "YSO", we find sub-classes related to their unique variability and spectral properties. The result of this exercise is a robust probabilistic classification (i.e. a posterior over classes) for 10090 of CSC sources. The tables for each cluster and respective code is available at `https://github.com/BogoCoder/astrox`.

Conclusions. We developed a methodology to provide probabilistic class assignation to numerous X-ray sources of the Chandra Source Catalog. Through this process we have seen that it is possible to construct a pipeline based on unsupervised machine learning for this task. We have seen that our approach works well for particular general type of sources, such as a *YSO*, or *extra-galactic* sources. In other cases, we have ambiguity in the number of classes presented in a particular cluster, having very different predominant types within them. This ambiguity might be solved by an addition of other wavelength regime data, such as optical from SDSS (Sloan Digital Survey Summary). This analysis is planned for a future work. This thesis present an early approach for the final goal of classifying all possible CSC sources that lacks of a class.

Acknowledgements

There are not enough words to express the gratitude to many people that made my undergraduate career an unbelievable experience. I hope that these short paragraphs honours that grateful feeling.

First of all, I would like to thank my advisors. Thank you Prof. Rafael Martínez-Galarza, for his patience and guidance, particularly in the knowledge transmitted to me in the field of astrophysics, which at the start I found exciting and a dream accomplishment opportunity, but honestly knew nothing about. Thank you Prof. Alexander Caicedo-Dorado, for his encouragement, important life lessons and constant accompaniment. I am thankful about the compassion and empathy that they always showed through this process, even in the hardest times, always believing in the things that I am capable of.

I am grateful about having such excellent professors through my career, which provided me with rigorous knowledge but as well with lessons that I think made me a better person. Thank you to all my university classmates, for the laughs, friendship, and support when I needed an urgent hand. Without them, finishing the career would have been hundreds of times harder. I am grateful as well of having the opportunity to study MACC, particularly with the people that made this possible. MACC have been an excellent career choice, where I learned so many exciting things that made me dream about new possibilities, but as well where I collected so many new experiences that made me the person who I am now. Thank you Nicolás, for being the best career friend and classmate. Thank you Catalina, for listening me, accompanying me and encouraging me to continue; since the start of this process, even in the most challenging times, you were there.

Finally, I want to thank my family, brothers and sister, who always saw potential in me and encouraged me to follow my dreams.

Mamá, gracias por siempre creer en mí. Papá, gracias por enseñarme tantas cosas. Este logro es dedicado a ustedes.

Contents

A mi mamá y papá, Yeny y Álvaro, por motivarme siempre a soñar más allá de las estrellas.

Chapter 1

Introduction

1.1 The Chandra Observatory

The Chandra X-ray Observatory is NASA's flagship mission for X-ray astronomy. The telescope, launched to space in 1999 and currently still under operation, has been observing the X-ray sky with two instruments, the Advanced CCD Imaging Spectrometer (ACIS) and the High Resolution Camera (HRC). The two instruments register X-ray sources with unprecedented sensitivity and resolution in X-ray wavelengths. Among the most common targets of Chandra are X-ray binaries, accreting black holes in the center of galaxies, supernova remnants, and young, rapidly rotating magnetic stars. Over its 22 years of scientific life, Chandra has made some of the most remarkable discoveries in high-energy astrophysics. See Fig. 1.1.

Chandra is one of the NASA's Great Observatories, sharing a name with the Spitzer Space Telescope (already retired), the Hubble Space Telescope, and the Compton Gamma Ray Observatory (already retired). These observatories were focused in infrared, optical-UV and gamma-ray wavelengths, respectively, having Chandra as its X-ray representative. See Fig. 1.2.

The *Chandra X-ray Center* (CXC), which is in charge of all the operations of Chandra X-ray Observatory, is managed for NASA by the Smitsonian Astrophysical Observatory (SAO), part of the Center for Astrophysics | Harvard & Smithsonian, located in Cambridge, Massachusetts.

The Chandra Source Catalog (CSC) lists the X-ray sources detected by the Chandra X-ray Observatory through its history. This catalog is the base and focus data of this work. We will cover more details in Chapter 2.

1.2 Astrophysical X-ray Sources

Chandra have made some of the most remarkable discoveries in X-ray astronomy. These discoveries cover different kind of sources, which we are interested in classifying. For the work at hand, we are basing our classification in the SIMBAD object classification [1]. This source can be consulted in order to obtain more information about the different types available. Some of the source classes that we mostly mention in this work are listed here:

1.2.1 Young Stellar Objects

A Young Stellar Object (YSO) is a star in its early stage of life or evolution. A star is constructed accumulating material from a circumstellar disk. This general class cover several subclasses that are of interest:

[1] http://simbad.u-strasbg.fr/simbad/sim-display?data=otypes

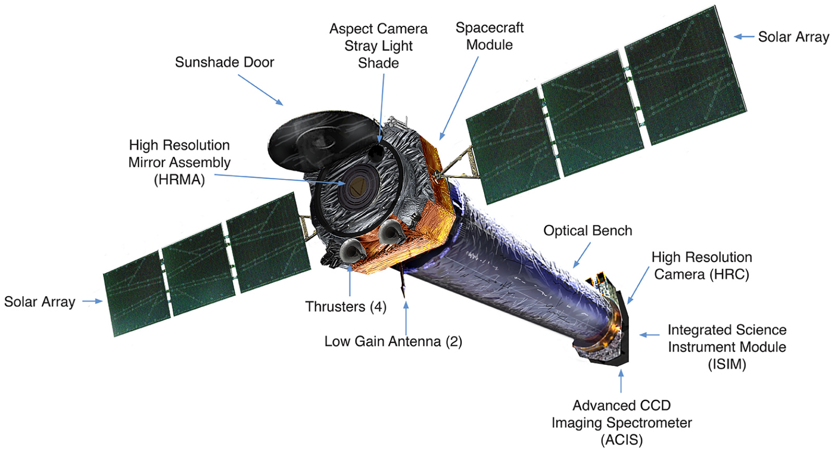


FIGURE 1.1: Illustration of Chandra and its components. *Credit: NASA/CXC*

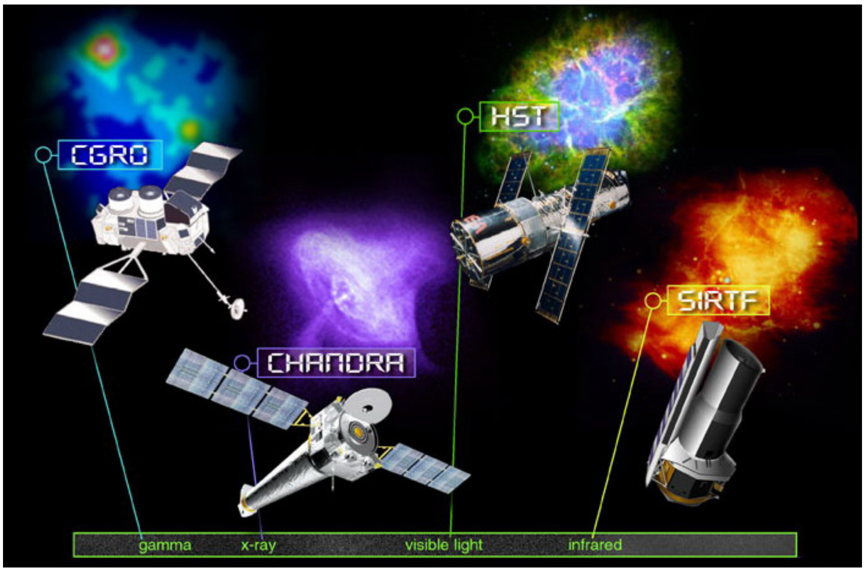


FIGURE 1.2: The Great Observatories, with their representative spectrum. *Credit: NASA*

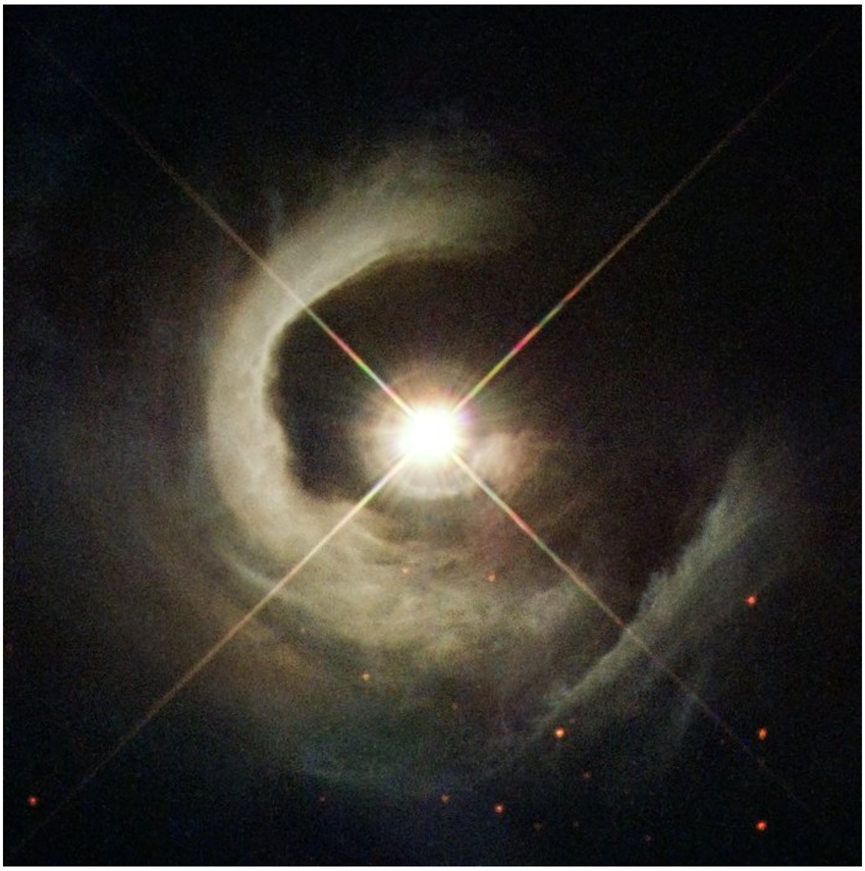


FIGURE 1.3: The star C1331 Cyg is a T Tauri star in the dark cloud LDN 981. The circumstellar disk is clearly visible. *Credit: ESA/Hubble, NASA, Karl Stapelfeldt (GSFC), B. Stecklum and A. Choudhary (Thüringer Landessternwarte Tautenburg, Germany)*

T Tauri star

A T Tauri star (TTauri* in SIMBAD) is a type of variable star [2] that shows both periodic and random changes in their luminosity, and has less than about ten million years in age. As its name says, T Tauri stars are named after the prototype star T Tau. This kind of YSO have started a process to become a main sequence star, having a mass of the same order which is similar to our Sun. See Fig. 1.3 for an example. Chandra observations have been of great importance in order to identify young stars, along with optical and infrared data (Wilkes and Tucker, 2019).

BY Draconis variable

A BY Draconis variable (BYDra in SIMBAD) is a type of variable star of late spectral types, usually a YSO. Particularly, this class refers to a rotating variable, which is a star which changes its luminosity when it spins around. It also shows quasiperiodic light changes, which could range from hours to several months.

[2] **Variable star**: A variable star is a star whose significance or brightness fluctuates, i.e., is constantly changing.

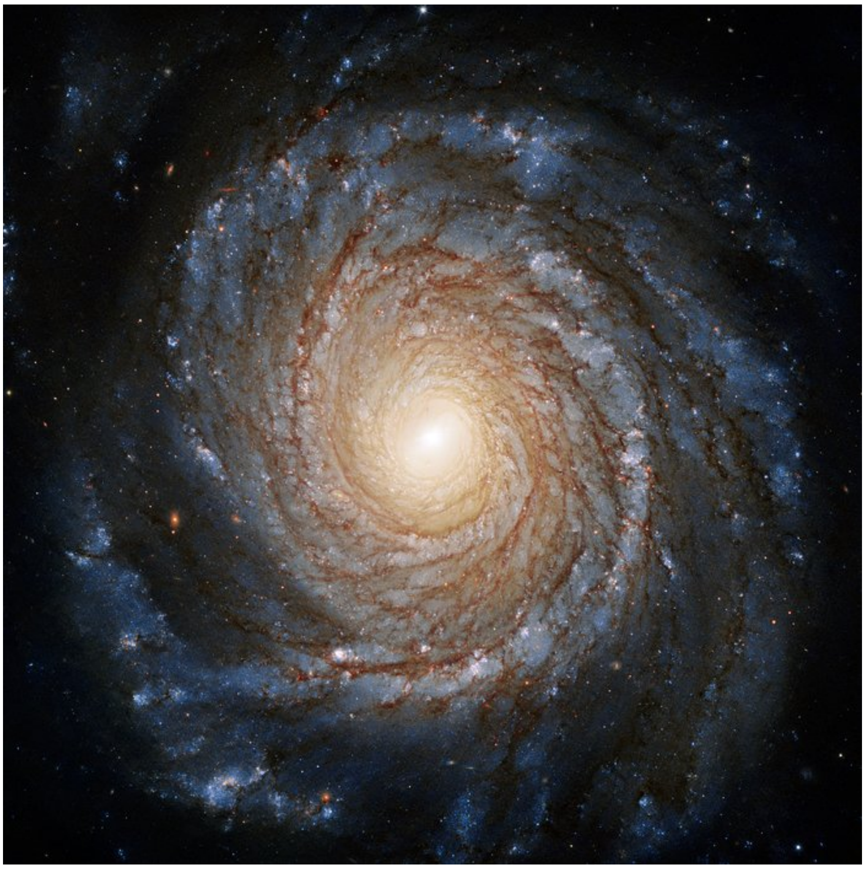


FIGURE 1.4: Galaxy NGC 3147, a type 2 Seyfert galaxy located in the constellation Draco. *Credit: ESA/Hubble & NASA, A. Riess et al.*

1.2.2 Quasar

A quasar, or quasi-stellar object (`QSO` in SIMBAD) is an active galactic nuclei (AGN)[3] that is extremely bright and has a supermassive black hole in its center.

1.2.3 Seyfert galaxy

A Seyfert galaxy, named after Carl Seyfert, who first described the class in 1944, is an active galaxy that hosts an active nuclei very similar to quasars in luminosity and distance, however, the galaxy host is detectable. These galaxies tend to be sources of powerful emissions of X-ray and radio energy, although they seem normal in ordinary light. See Fig. 1.4. There are two generally recognized classes. Type 1 Seyfert galaxies have broad emission lines on their spectra, suggesting a central and very rapidly expanding concentration of hot gas in its center. They are very bright in ultraviolet and X-rays wavelengths. Type 2 Seyfert galaxies (`Seyfert_2` in SIMBAD) have emission lines that suggests slower expansion velocities. They are very bright in the infrared wavelength, and have a characteristic bright core.

[3] **AGN:** This are the most bright sources of radiation in the universe, powered by accretion over a compact region in the center of a galaxy, theorized as a supermassive black hole (SMBH).

1.2.4 Pulsars

A Pulsar (`Pulsar` in SIMBAD), or pulsating radio source, is a source that has a regular periodicity, particularly a compact rotating star, emitting bursts of radio emissions from its magnetic poles. In general, it is accepted that this type of source is highly-magnetised and rapidly rotating. X-ray pulsars emits burst of x-rays in intervals usually regular.

1.2.5 X-ray binaries

X-ray binaries are systems of binary stars that are particularly bright in the X-ray wavelength. In these systems we have an interaction between two objects: a neutron star or a black hole (accretor) is accreting matter that comes from an usual normal star (donor). X-rays can be produced by a number of reasons, including the acceleration of charged particles during accretion, shocks of hot gas, or thermal emission from a compact central object. In a similar scenario, replacing the neutron star or black hole to a white dwarf, we have a Cataclysmic Variable system. X-ray binaries are subdivided into subclasses, that could be provided by mass in low-mass X-ray binaries (`LMXB` in SIMBAD), intermediate-mass X-ray binaries, and high-mass X-ray binarties (`HMXB` in SIMBAD).

1.2.6 Supernova remnants

A supernova remnant (`SNR` in SIMBAD) is the structure composed of gas, dust and particles that results from a supernova explosion. Chandra have been determinant in the identification of supernova remnants, allowing to study and understand new phenomena never seemed before. Chandra's first light pointed towards the Supernova remnant Cassiopeia A, and allowed astronomers to have the first idea of the compact object at the center of the structure. See Fig. 1.5.

1.3 This thesis

This thesis uses data obtained from the Chandra Source Catalog (CSC), provided by the Chandra X-ray Center (CXC) as part of the Chandra Data Archive.

The CSC is a fertile ground for discovery, because many of the sources it contains have not been identified or studied in detail. In the Chandra dataset we can find sources such as compact object mergers, extrasolar planet transits, tidal disruption events, etc. However, only a small fraction of the CSC sources have been classified, based either on their X-ray properties, or on the properties of their matching registers in other wavelengths. In order to make a systematic study of the CSC sources, we need to classify as many as possible.

The goal of this project is to provide probabilistic classification labels to the CSC sources, by employing unsupervised machine learning techniques and information from already labeled sources in matching data archives. Unsupervised learning is more suitable for this task because of the lack of a large training set of labelled X-ray sources. By associating specific groupings (clusters) of Chandra data with specific objects that have been previously classified spectroscopically, we hope to significantly enhance the number of sources for which a label is available. The main

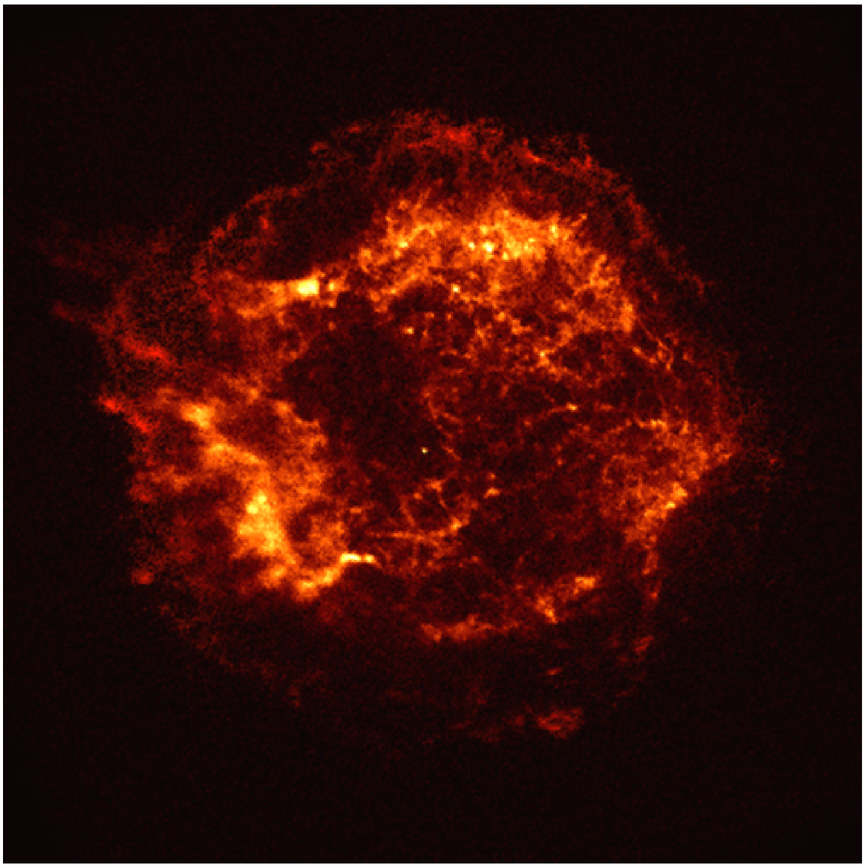


FIGURE 1.5: Cassiopeia A. Chandra's first light. *Credit: NASA/CXC/SAO*

techniques used in this thesis are based in Gaussian Mixture Models and the Mahalanobis distance, we will provide further details in Chapter 3.

The goal of the project is to apply a suitable method to the CSC dataset, in order to provide the most likely classes for thousands of X-ray astrophysical sources. We want to research on the opportunities and limitations of an approach like this, and analyze if it is suitable for a possible probabilistic classification.

1.4 Related work

Given the accelerating release speed of new larger surveys, a trend is happening in astronomy, which has lead researchers to use more sophisticated statistical models, and specifically, machine learning and data science methods, which have become state-of-art procedures to perform classification, which is the topic that we want to cover in the work at hand. We could say that efforts of classification of X-ray sources with the use of machine learning methods is relatively new. It is crucial to join efforts on this task in order to prepare as best as we can for the arrive of new, state of the art surveys, such as the eROSITA All Sky Survey (eRASS), which is the first all sky image in 2-10 keV band (Merloni et al., 2012). Some important progress have been made on this task, both using unsupervised and supervised learning techniques. In (Pineau et al., 2010) a comparison between different supervised and unsupervised learning methods was performed for the statistical identification of XMM-Newton sources. This research used a probabilistic cross-correlation of the 2XMMi (XMM-Newton Serendipitous Source Catalog) catalogue with catalogs such as SDSSDR7 [4] or 2MASS [5]. In the algorithms compared we encounter: : k-Nearest Neighbours, Mean Shifts, Kernel Density Classification, Learning Vector Quantisation and Support Vector Machines. In (Lo et al., 2014), a method of supervised learning is presented in order to automatically classify variable X-ray sources that are present in 2XMMi-DR2, specifically Random Forest. They used 10 fold cross validation and obtained an accuracy of approximately 97% for a 7 class data set. It is emphasized here that machine learning classification and detection of anomalies will help in the scientific discoveries of the future. In (Farrell, Murphy, and Lo, 2015) a catalog of variable sources in 3XMM is presented, which is autoclassified using Random Forest. In order to train the classifier, they used variable stars manually classified from 2XMMi-DR2, at the end obtaining an accuracy of approximately 92%. We can observe that interest in this research field is still in progress, but everyday it is more relevant. In (Rostami Osanloo et al., 2019), an automated machine learning tool for classification of extra-galactic X-ray sources is proposed by using multi-wavelength data, particularly data taken with Hubble Space Telescope. More focus in unsupervised learning techniques have come up in the later years. In (Ansari, Zoe, Agnello, Adriano, and Gall, Christa, 2021) a probabilistic assignment is performed using mixture density networks (MDN). Training data is composed of magnitudes from the SDSSDR15 and WISE[6]. An approach using infinite Gaussian mixture models is used in order to classify the objects in the dataset as stars, galaxies, or quasars, and to adjust the optimal parameters of the MDN. As a result, they had an accurate split into stars, galaxies, and quasar of 94%. Finally, in (Logan, C. H. A. and Fotopoulou, S., 2020) an alternative unsupervised machine learning method is presented in order to separate stars, galaxies and QSO using photometric data. This approach uses

[4] Sloan Digital Sky Survey

[5] Two Micron All-Sky Survey

[6] Wide-field Infrared Survey Explorer

HDBSCAN[7] in order to find the different classes in a multidimensional color space. Using a constructed dataset of approximately 50000 spectroscopically labelled objects, they obtained for star, galaxy and QSO classes, an F1 score of 98.9, 98.9, and 93.13, respectively.

1.5 Outline

In Chapter 2 we describe the main dataset used in this work, the Chandra Source Catalog, with the query and preprocessing needed in order to make the data suitable for a first approach using unsupervised learning methods.

In Chapter 3 we describe the theoretical fundamental of the unsupervised learning methods used, and we introduce a probabilistic classification algorithm that was used for the final output of this thesis.

In Chapter 4 we present the partial and final results of the thesis, along with a discussion of the implications, limitations and possible improvements of these.

Finally, in Chapter 5 conclusions of the thesis are summarized, along with the outcomes and possible future work.

[7] Hierarchical Density-Based Spatial Clustering of Applications with Noise

Chapter 2

Data

In this study we will use an unsupervised classification pipeline to classify as many catalog sources as possible in the Chandra Source Catalog 2.0. It is clear that any analysis could not be performed without appropriate data. Fortunately, we do have amazing catalogs in X-Ray astronomy, and the Chandra Source Catalog is one of them. This chapter presents the data used for this thesis as well as the preprocessing pipeline employed in order to make the data suitable for the application of unsupervised learning methods.

2.1 The Chandra Source Catalog

The Chandra Source Catalog (CSC)[1] is a catalog that collects and summarizes the X-ray sources detected by the Chandra X-ray Observatory through its history. In its version 2.0 (CSC2), which is the second major release of the catalog, it includes properties for 317, 167 X-ray sources in the sky, with a total size close to 36 TB. In these properties we can find measurements related with the source photometry (brightness), spectroscopy (energy), and variability (changes of the source over time). Properties are available for 928, 280 source observation detections, which were detected in 10, 382 Chandra observations until 2014. Along these properties, we have approximately 1700 columns of tabular data, usually presented across the source 5 energy bands (broad, hard, medium, soft, and ultra-soft) for ACIS, and in 1 band for HRC (wide).

Through its history, Chandra has observed the universe in a X-ray band of 0.5-8 KeV. It is of general interest to detect and release organized properties of all the sources observed by Chandra, which are valuable to all astronomers and non-astronomers, in order to have already carefully processed data that is ready to be used in scientific research (Wilkes and Tucker, 2019).

2.1.1 Chandra Source Catalog Description

The latest release of the Chandra Source Catalog, v2.0, has available measurement properties that cover different features of a source behaviour. Particularly, the catalog takes advantage of data aggregation in order to increase sensitivity and be able to detect more sources clearly. Improvements in sensitivity and sky coverage results in having measured properties for 317167 X-ray sources in CSC2, three times the number of sources from CSC1, allowing statistical and data science analysis in an individual source observation scale or over large samples. These sources were detected in approximately 7200 stack observations, and a total exposure time greater

[1] https://cxc.cfa.harvard.edu/csc/

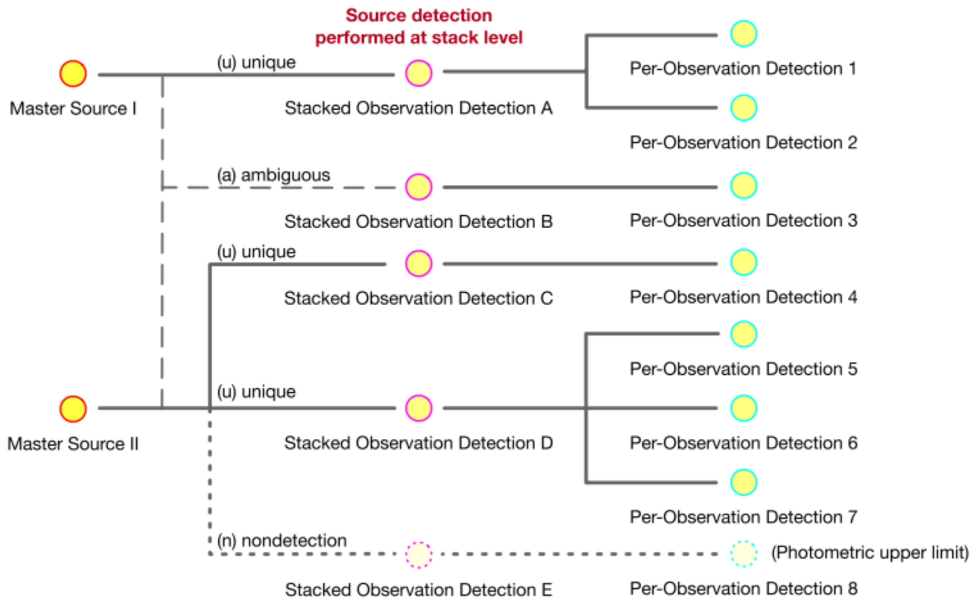

Source detection
performed at stack level
(u) unique
Master Source I
Stacked Observation Detection A
Per-Observation Detection 1
Per-Observation Detection 2
(a) ambiguous
Stacked Observation Detection B
Per-Observation Detection 3
(u) unique
Stacked Observation Detection C
Per-Observation Detection 4
Per-Observation Detection 5
(u) unique
Master Source II
Stacked Observation Detection D
Per-Observation Detection 6
Per-Observation Detection 7
(n) nondetection
(Photometric upper limit)
Stacked Observation Detection E
Per-Observation Detection 8

FIGURE 2.1: CSC2 hierarchy for source/detection. *(Dataset identifier: ADS/Sa.CXO#CSC. Courtesy of NASA/CXC. Evans et al., 2010.) Taken from (Wilkes and Tucker, 2019).*

TABLE 2.1: Summary of source properties for master, stack and individual observations. *Adapted from (Wilkes and Tucker, 2019).*

Property type	Properties
Astrometry	Source position, extent, significance, likelihood
Photometry	Energy fluxes and aperture photon in each energy band, spectral model fluxes
Spectral	Hardness ratios, spectral fit parameters
Variability	Inter-observation and intra-observation variability probability

than 245.8 Ms (Wilkes and Tucker, 2019).

The master source properties are summarized from the stack observations in which a detection of the source is presented. Properties for detected sources are measured for both the stacked detection and the individual observations in the wide band for HRC[2], and in the broad[3], soft[4], medium[5], and high[6] energy bands for ACIS (Wilkes and Tucker, 2019). A hierarchical structure of the catalog is shown in Fig 2.1.

A summary of the properties of the catalog (astrometry, photometry, spectroscopy and variability), particularly the relevant properties for this work, is presented in Table 2.1.

These, and many more source properties, are available through the CSCview

[2] (w, 0.1–10 keV)
[3] (b, 0.5–7.0 keV)
[4] (s, 0.5–1.2 keV)
[5] (m, 1.2–2.0 keV)
[6] (h, 2.0–7.0 keV)

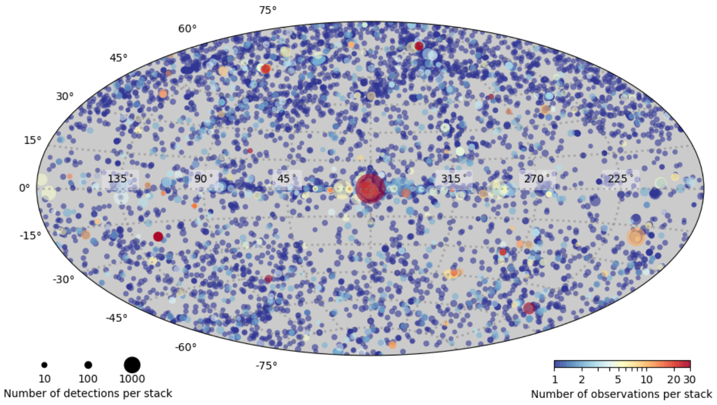


FIGURE 2.2: *The locations of the CSC 2.0 detections in the galactic plane. The size of each circle is proportional to the logarithm of the number of detections per stack. The color is determined by the number of close (in distance) observations. All the little points shown here represent a possible type of source that we could classify. For example, these could be X-Ray binaries, quasars, young stellar objects, galaxies, etc. However, most of the sources have not been classified yet. This represents a fertile ground for discovery, and a clear motivation for the work at hand. Taken from the CSC documentation webpage.*

interface[7]. Detailed descriptions and specific properties are available in CSC documentation webpage[8]. In Figure 2.2 a visualization of the CSC sources over the entire sky is presented.

2.1.2 Science with the Chandra Source Catalog

Usually, the most energetic and violent phenomena in the universe are very bright in X-rays. For example, most supermassive black holes in the universe, such as the famous SMBH at the core of M87 that gained recognition because of the image of its horizon (Event Horizon Telescope Collaboration et al., 2019), are frequently first detected by the X-ray bursts of hot gas and particles accreting around it (Perlman and Wilson, 2005), or the very X-ray energetic jets propulsed from them (Wilkes and Tucker, 2019). See Fig. 2.3. Thus, using the CSC is a great choice in order to construct paths to discovery, given the carefully collected properties of all the sources observed by Chandra. The CSC is also very suitable to perform a crossmatch with other wavelength catalogs, that could help in the identification and classification of astrophysical sources (Wilkes and Tucker, 2019), such as we do in this work. In order to give probabilistic classifications to non-classified sources, we had to perform a crossmatch with SIMBAD Database for obtaining the existing classes. A detailed explanation of this process is explained in Chapter 3.

The Chandra Source Catalog is a rich resource of information and for potential discovery. Most of the objects in CSC2 are unexplored Chandra detections, and most of these detections have not been studied in detail. As we mentioned earlier, this catalog includes several types of sources that range from young stellar objects and binary systems, to even very far active galaxies with supermassive black holes in their

[7] http://cda.cfa.harvard.edu/cscview/

[8] https://cxc.cfa.harvard.edu/csc/

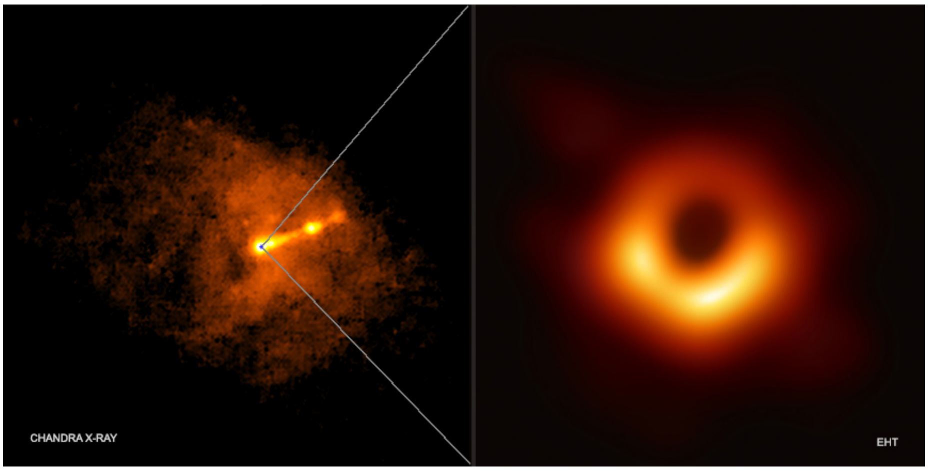


FIGURE 2.3: Left: Chandra X-ray image of the M87 galaxy. A clear visual of the X-ray jet is provided. Right: the zoom toward the central supermassive black hole of the galaxy, obtained by the Event Horizon Telescope. *Credit: X-ray: NASA/CXC/Villanova University/J. Neilsen; Radio: Event Horizon Telescope Collaboration*

cores. These different sources are presented in the Chandra Source Catalog by their X-ray properties, which cover spectral properties, variability measurements, hardness ratios, etc. These properties allow us not just to analyze the data in the search of a particular type of object, but to be open for anomalies or possible unknown phenomena from a X-ray source, as well to finding rare objects such as compact object mergers, extrasolar planet transits, tidal disruption events, etc. The recent trend that boosted new data science and machine learning techniques will help in the expansion of the horizons for discovery and investigation in high-energy astrophysics, and the work at hand will present a new idea that could support this idea based on an unsupervised learning technique.

2.2 Acquisition

For each source in the CSC catalog, properties are calculated on stacked images, and also at the individual observation level, as mentioned before. This allows the study of long and short term variability. We centered our first data acquisition in the Per Observations Detection Table of the CSC2, which contains extracted properties in individual observations of each source, many of them observed multiple times. This could give us important information on the lifetime events of different X-ray sources in the universe, and could help us identify several kind of sources, as shown in Fig. 2.4.

2.2.1 Querying over the CSC

In order to have a suitable set which could be well interpreted by an unsupervised learning algorithm, we selected a subset of the CSC2 Per Observations Detection Table. For doing this, we performed a SQL query using CSCView. This query included the following restrictions:

- `flux_significance_b`> 5: flux significance in the broad band to be greater than 5.

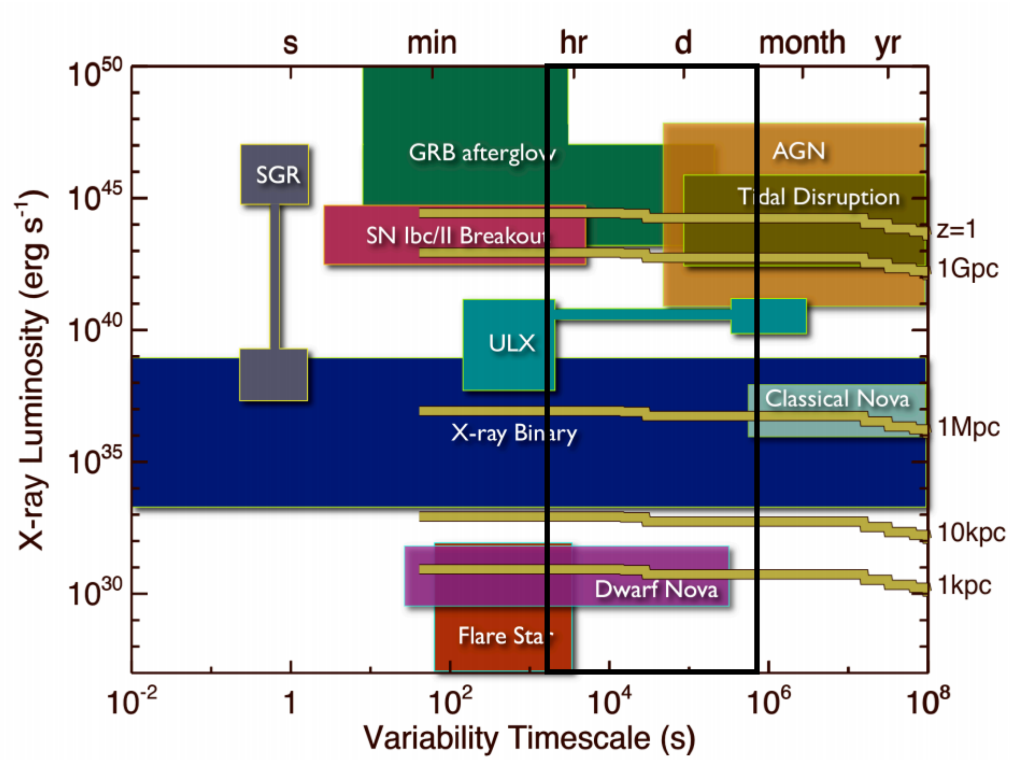


FIGURE 2.4: Peak X-ray luminosity of the most prominent source populations as function of their characteristic variability time scales (adopted from Soderberg et al. 2009). Yellow tracks indicate the an approximate eROSITA 0.5-2 keV sensitivities for various source distances. The black box indicates an approximate of the Chandra sentivity timescale, it is clear that several source classes are present. *Adapted from (Merloni et al., 2012).*

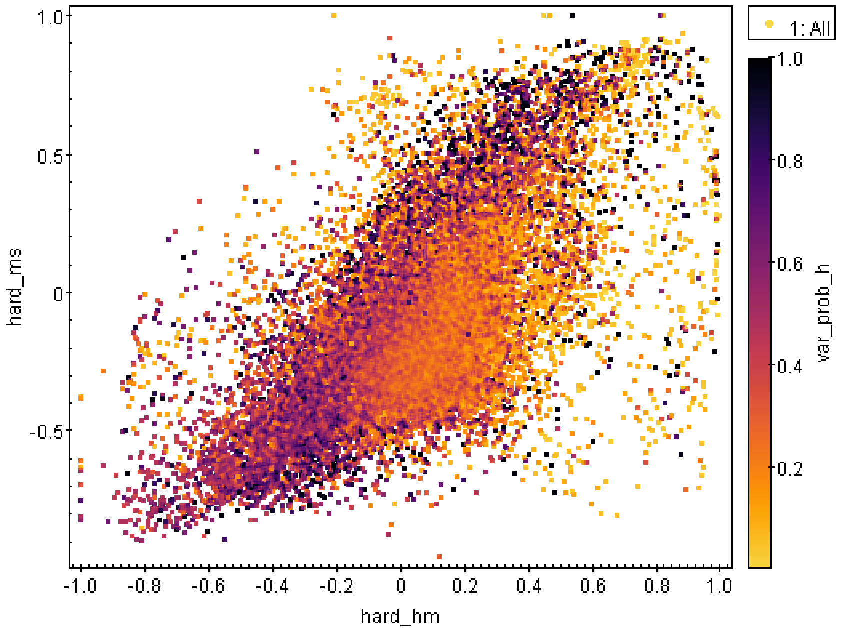


FIGURE 2.5: *hard_ms vs. hard_hm, color coded by var_prob_h. A particular projection of the dimensional space of our data. We can see already a clear functional relation between the hardness ratios at hand and the variability probability in the hard band. For example, at the top and bottom of the distribution we can find highly variable sources. Closer to the mean of the hardness ratios, we can find an elbow dominated by low variable sources. This gives us clues about relations that we can not see yet in our higher dimensional space. We want to explore and understand those relations better and exploit them to find a later classification.*

- Valid values for `o.powlaw_gamma`.
- Valid values for `o.bb_kt`

With these restrictions, we aimed to have enough significant observations for making it possible to adjust their spectra to a model, and thus be able to extract the statistical properties needed for the analysis. As a result, we got a table with 37878 rows of unique observational registers. A visualization of a particular projection of the properties space is shown in figure 2.5.

2.2.2 Preprocessing

Data exploration and properties preprocessing is a key step before every machine learning model usage. For this work, we selected specific astrometry, spectral and variability properties. We summarize the selected properties in the table 2.2.

One of the first steps of machine learning based analysis is to process the data accordingly to the model that we are about to use. Most of the time, this means a normalization over the values of columns in the dataset, in order to have a common scale, which could be presented in different styles, e.g., standardization or min-max normalization. In the work at hand we performed either just a normalization step or a merge between logarithmic transformation and normalization, depending on the distribution and range of the properties selected. The processing selection is shown in the table 2.3.

TABLE 2.2: Properties selected for the analysis at hand. Each energy band is coded as * = *b*, *h*, *m*, *s*. *Adapted from CSC documentation webpage.*

column name	**description**
theta	aperture off-axis angle
src_area_b	area of the deconvolved detection extent ellipse, or area of the detection polygon for extended detections for broad energy band - extent size of the object
hard_hm	ACIS hard (2.0-7.0 keV) - medium (1.2-2.0 keV) energy band hardness ratio - basically the ratio between the hard and medium energy bands
hard_hs	ACIS hard (2.0-7.0 keV) - soft (0.5-1.2 keV) energy band hardness ratio - basically the ratio between the hard and soft energy bands
hard_ms	ACIS medium (1.2-2.0 keV) - soft (0.5-1.2 keV) energy band hardness ratio - basically the ratio between the medium and soft energy bands
bb_kt	temperature (kT) of the best fitting absorbed black body model spectrum to the source region aperture PI spectrum - temperature of the object estimated by a black body model.
powlaw_gamma	photon index of the best fitting absorbed power-law model spectrum to the source region aperture
var_prob_*	intra-observation Gregory-Loredo variability probability (highest value across all stacked observations) for each science energy band - variability probability in a single observation with Gregory-Loredo technique.
var_mean_*	flux variability mean value
var_sigma_*	flux variability standard deviation
var_max_*	flux variability maximum value
var_min_*	flux variability minimum value
ks_prob_*	intra-observation Kolmogorov-Smirnov test variability probability (highest value across all observations) for each science energy band - variability probability in a single observation with Kolmogorov-Smirnov technique.
kp_prob_*	intra-observation Kuiper's test variability probability (highest value across all stacked observations) for each science energy band - variability probability in a single observation with Kuiper's test.

TABLE 2.3: Performed pre-processing on each property. Each energy band is coded as `* = b, h, m, s`.

property name	log	normalization
theta	X	X
src_area_b	X	X
hard_hm		
hard_hs		
hard_ms		
bb_kt	X	X
powlaw_gamma		X
var_prob_*		
var_mean_*	X	X
var_sigma_*	X	X
var_max_*	X	X
var_min_*	X	X
ks_prob_*		
kp_prob_*		

Specifically speaking, we could describe both of the processes as follows:

Normalization

Based on `sklearn.preprocessing.MinMaxScaler` class method from the `scikit-learn` Python library. This transformation is defined as shown in equation 2.1, from the `scikit-learn` library documentation.

$$\begin{aligned} \sigma &= (\mathbf{X} - \mathbf{X}_{\min}) \;/\; (\mathbf{X}_{\max} - \mathbf{X}_{\min}) \\ \mathbf{X}_{scaled} &= \sigma\,(\mathbf{X}_{\max} - \mathbf{X}_{\min}) + \mathbf{X}_{\min} \end{aligned} \tag{2.1}$$

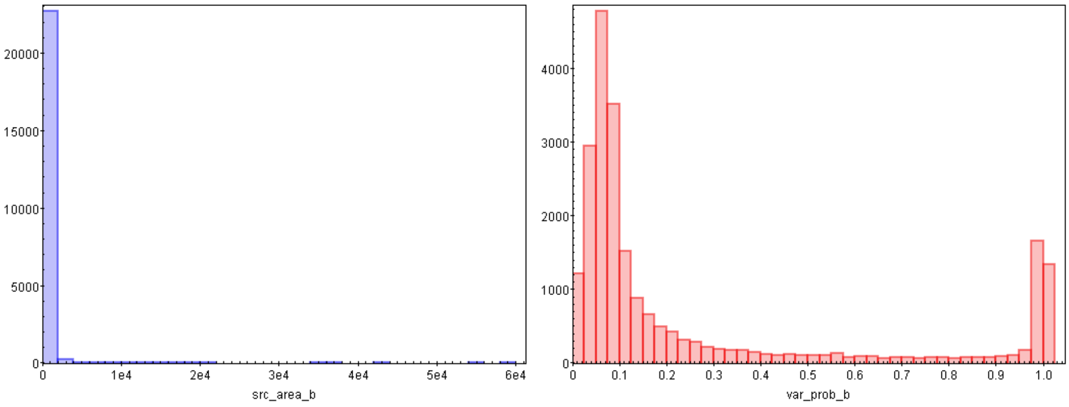
Log transformation

We used mainly `numpy` here, in order to perform a natural logarithm to each of the registers. We overcome the challenge of 0 values (which would make the natural logarithm undetermined) by finding the non-zero minimum value of each descriptor, and then finding the natural logarithm of the data plus the encountered minimum divided by 10. The process could be written formally as shown in equation 2.2.

$$\begin{aligned} \hat{X}_{\min} &= \min_{i=1\ldots n} \{x_i \in \mathbf{X} : x_i \neq 0\} \\ \text{minval} &= \hat{X}_{\min}/10 \\ \mathbf{X}_{log} &= \log(\mathbf{X} + \text{minval}) \end{aligned} \tag{2.2}$$

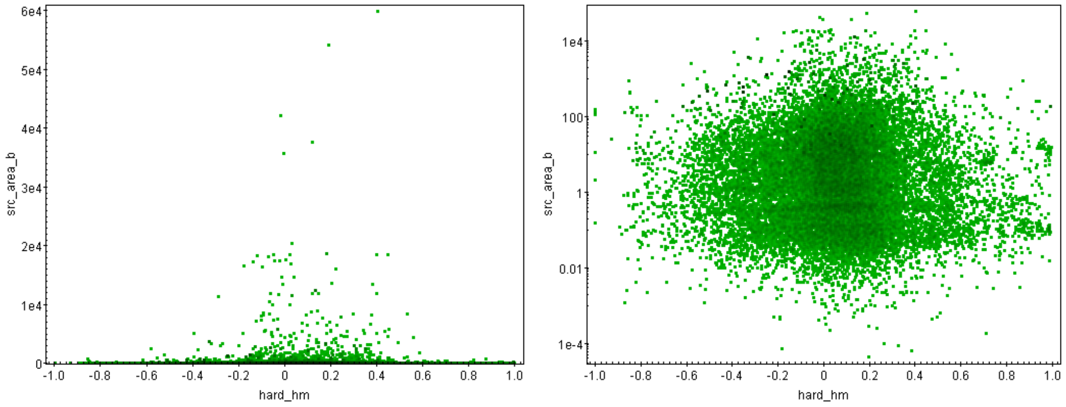
Before doing so, we had to make sure that all observations would have a valid value for their properties, therefore, a removal of invalid values (`NaN`) was performed. After this process, our dataset was reduced to 23025 sources.

We show a visualization of the difference in the range of values adopted by different properties in Fig. 2.6. Additionally, we show an before-after process example in Fig. 2.7.



(A) (B)

Figure 2.6: *(A): src_area_b histogram, (B): var_prob_b histogram. We can see a clear difference in ranges between both properties. For (A) we see a very skeweed distribution, trending to 0, but as well having some extreme values close to* 6e4. *For (B), in the other case, given that it is a probability, it is bounded between 0 and 1.*



(A) (B)

Figure 2.7: *(A): src_area_b vs hard_hm, (B): log(src_area_b) vs hard_hm. There is a huge skewness in the property of src_area_b, as shown in (A), making it to accumulate in the x axis, but having as well some extreme values and outliers. By performing a log transformation processing in (B), we see that the property distribution collapses into an spherical shape relation, making it easier to visualize and for the model interpretation.*

Chapter 3

Unsupervised Approach with CSC properties

3.1 Introduction

In this chapter we will present a theoretical background of the methods used in the analysis performed in this thesis. First of all, we will cover the unsupervised learning techniques; after that, we will explore a crossmatching process in order to extract types, and finally a distance-based similarity algorithm. We will explain some of the concepts behind this algorithm, such as the *Mahalanobis* distance, which is the core metric for this analysis.

Acknowledgement:

The algorithms developed in this thesis were implemented using `Python`. Specifically, the following libraries were used:

- `numpy`: (Harris et al., 2020)
- `matplotlib`: (Hunter, 2007)
- `pandas`: (pandas development team, 2020)
- `seaborn`: (Waskom, 2021)
- `scikit-learn`: (Pedregosa et al., 2011)
- `astropy`: (Astropy Collaboration et al., 2013), (Astropy Collaboration et al., 2018)

3.2 Clustering

In this subsection we will cover the theoretical foundation for the following models: K-means and Gaussian Mixtures. Both of these are methods for clustering data, which corresponds to the task of grouping objects by their similarity. This means that objects in the same cluster tend to be more similar between them than with objects in other clusters. In an astrophysical perspective, we can see this grouping as a tool for finding sources with similar behaviors in their X-ray properties that could either be the same type of source or be in a similar stage of a source stellar evolution. Cluster analysis is of interest in many fields such as computer vision, pattern recognition, bioinformatics, etc.

For the explanation of the methods we want to cover a motivation of the EM algorithm, which is the basis of the Gaussian Mixture Models technique.The expectation-maximization (EM) algorithm is a technique for finding maximum likelihood estimators of parameters in models, specifically those that depends on unobserved latent variables. The latent variable in the case of a mixture model corresponds to assigning a data point to an specific component, i.e., finding the parameters that maximize the log likelihood of the multivariate mixture model. K-means corresponds to a particular case of EM applied to Gaussian Mixtures, with stronger assumptions and limitations. This is a motivation for having K-means as a base case, and later perform a full analysis using Gaussian Mixtures. Thus, K-means was considered in a exploratory way, since the final results and discussion were focused in the Gaussian Mixtures output.

We proved empirically the consistency of the clusters over different iterations of the same model, considering the different variations that could appear given the random factor of the initial conditions of the algorithms. For our cluster number selection (k in K-means and number of components in GMM), we used several techniques such as the elbow method and silhouette analysis, but none of these gave us significant information on the clustering performance. Thus, we based our selection analyzing which model gave the best distinction between astrophysical properties in clusters, and later on, the different types of sources encountered by extracting information from other dataset. Finally, we selected k for K-means and the number of components for GMM to be 6 in both cases. The results shown in Chapter 4 4 explain better the parameter influence and clear separation that made us choose this value.

3.2.1 K-means

K-means is one of the basic and most recognized clustering techniques out there. Nowadays, it is one of the first techniques taught in Machine Learning courses, specifically when we are talking about unsupervised learning techniques. This comes for a variety of reasons, but the most important one is that it is a very simple technique. Nevertheless, do not let this discourage the potential enthusiasm about this algorithm, because, even though it is a simple one, for a potential basic case and a exploration of the potential of unsupervised learning in a dataset it provides an interesting output.

Now, let us go deeper into the technical aspects of this algorithm. First of all, we will assume that we have a dataset that is represented by $\boldsymbol{X} = \{x_1, \ldots, x_n\}$. Remember that each observation of our dataset could have different descriptors, a.k.a, columns, that in a mathematical sense is describing a T-dimensional space, being T the number of descriptors that we have in our observations.

Intuitively, what we want to do in K-means is to divide our data into K groups, or clusters, which are constructed by having that points belonging to the same clusters have the smallest distance between them than with other points in other clusters. As you would infer, and following the basis of machine learning, this could interpreted as a minimization problem. We will define that as follows.

First of all, we want to define a variable ϵ_{nk} that will represent the belonging of the point n to the cluster k. We will do this by having that $\epsilon \in 0, 1$ is a binary variable. Thus, if the point n belongs to the cluster k, we will have that $\epsilon_{nk} = 1$, and $\epsilon_{nj} = 0$

for all $j \neq k$. Let's call the cluster k as C_k.

Having this at hand, we come to the following definition:

$$\epsilon_{nk} = \begin{cases} 1 & x_n \in C_k \\ 0 & otherwise \end{cases}. \tag{3.1}$$

In addition, we will define a representative sample for each cluster as $\boldsymbol{\mu}_k$, which represents the mean value of the samples that belong to the cluster k. This is again a T-dimensional variable that usually is recognized as the centroid of each cluster, given that we will use the euclidean distance and our objective is to find the points in the dataset that minimize the intra-cluster distances. Thus, we want to define an assignation of data points to clusters, as well as $\boldsymbol{\mu}_k$ vectors, such that the sum of squares of the distance of each data point to its closest $\boldsymbol{\mu}_k$ vector is minimized. We can define an objective function that represents the intuition given above as follows:

$$D = \sum_n \sum_k \epsilon_{nk} \left\| \boldsymbol{x}_n - \boldsymbol{\mu}_k \right\|^2 \tag{3.2}$$

This equation represents the sum of the squares of the distances of each data point to the representative sample, or centroid, of its cluster. Our goal is to minimize this function, finding the correspondent parameter ϵ_{nk} and centroid vectors $\boldsymbol{\mu}_k$. In a more descriptive way, we want to find the assignations of the data points to the k clusters and their respective centroids which minimizes the function D. It is good to mention here that, in the context of the EM algorithm, the update of the assignments (ϵ_{nk}) is actually the expectation step, E, and the update of the centroids ($\boldsymbol{\mu}_k$) is the maximization step, M. Let's recall that the K-means model is actually a particular case of the EM algorithm applied to Gaussian Mixtures. We will not go much deeper into the details of this, but further explanation could be reviewed in (Bishop, 2006) or (Alpaydin, 2014).

Now, let's remember that we want to minimize the objective function D. In order to find the optimal assignation for the data points ϵ_{nk}, we can conclude that we need to select the cluster k that provides the minimum squared distance to the kth centroid. In other words, we assign the data points to the cluster represented by the closest cluster centroid. This is formally written as:

$$\epsilon_{nk} = \begin{cases} 1 & \text{if} \quad k = \text{argmin}_j \left\| \boldsymbol{x}_n - \boldsymbol{\mu}_j \right\|^2 \\ 0 & otherwise \end{cases} \tag{3.3}$$

This equation describes the optimal assignments for each data point x_n based on the cluster k which gives the minimum squared distance of the point x_n to its centroid μ_k.

Additionally, for the optimal assignment of the new centroids μ_k for each cluster, we will derive the objective function D with respect to μ_k and then set it to 0. This will give us that the optimal μ_k are defined by:

$$\mu_k = \frac{\sum_n \epsilon_{nk} \boldsymbol{x}_n}{\sum_n \epsilon_{nk}} \tag{3.4}$$

Observe that the denominator of this result is actually the number of points that belongs to a cluster k. Thus, this result tells us that the optimal representatives μ_k are given by the mean of all the data points belonging to the cluster k. A pseudocode for the k-means algorithm is provided in Alg. 1.

```
Initialize $\boldsymbol{\mu}_k$, with $k = \{1, \ldots, K\}$
Repeat until convergence
    for all $\boldsymbol{x}_n \in \boldsymbol{X}$ do
        $\epsilon_{nk} \longleftarrow \begin{cases} 1 & \text{if} \quad k = \text{argmin}_j \left\|\boldsymbol{x}_n - \boldsymbol{\mu}_j\right\|^2 \\ 0 & otherwise \end{cases}$
    end
    for all $\boldsymbol{\mu}_k$, $k = \{1, \ldots, K\}$ do
        $\boldsymbol{\mu}_k = \frac{\sum_n \epsilon_{nk} \boldsymbol{x}_n}{\sum_n \epsilon_{nk}}$
    end
```

Algorithm 1: k-means algorithm.

A more talkative way of describing the k-means algorithm would be as follows:

- Initialize the centroids for the K clusters. This is usually done randomly, but other methods of initialization have been proposed and could be found in the literature (Bishop, 2006).
- Assign all the data points to the cluster represented by its closest centroid. This is performed using euclidean distance.
- Find the new centroids by computing the mean of all the data points that belong to each cluster.
- Repeat this until convergence.

We used the K-means model as a first approach for separating our CSC data in different groups by their similar properties. We applied this model over the preprocessed data. We performed this process several times, changing the values for K and using different evaluation techniques. We eventually found out that a $K = 6$ gave us an interesting separation of sources given their astrophysical properties, since we have found that some clusters tend to have particular trends in the properties of the sources that belong to it, for example, one cluster having particularly high variable sources, or others being dominated by hard or soft observations. Preliminar results of this process can be observed in Fig. 3.1 and Fig. 3.2.

After analyzing the results of K-means clustering, we realized that we could have potential difficulties due to the assumptions of K-means, which come from being a special case of EM applied to Gaussian mixtures with equal variances for all the clusters, and a labeling that is known as *hard*, which basically means that the output of the algorithm is a strict assignation of the data points to a cluster, even though we could have special cases of ambiguous data points. In addition, when we are using K-means, we are assuming that our clusters have an spherical shape.

The absence of a *soft* clustering alternative, which could provide a probability of assignation to a cluster instead of a strict labeling, and the lack of flexibility in cluster shapes, orientations and proportions, made us think about a more general approach that could fit well in a multidimensional data space, and that could have potential variety of clusters features. Thus, we selected Gaussian Mixtures to be our

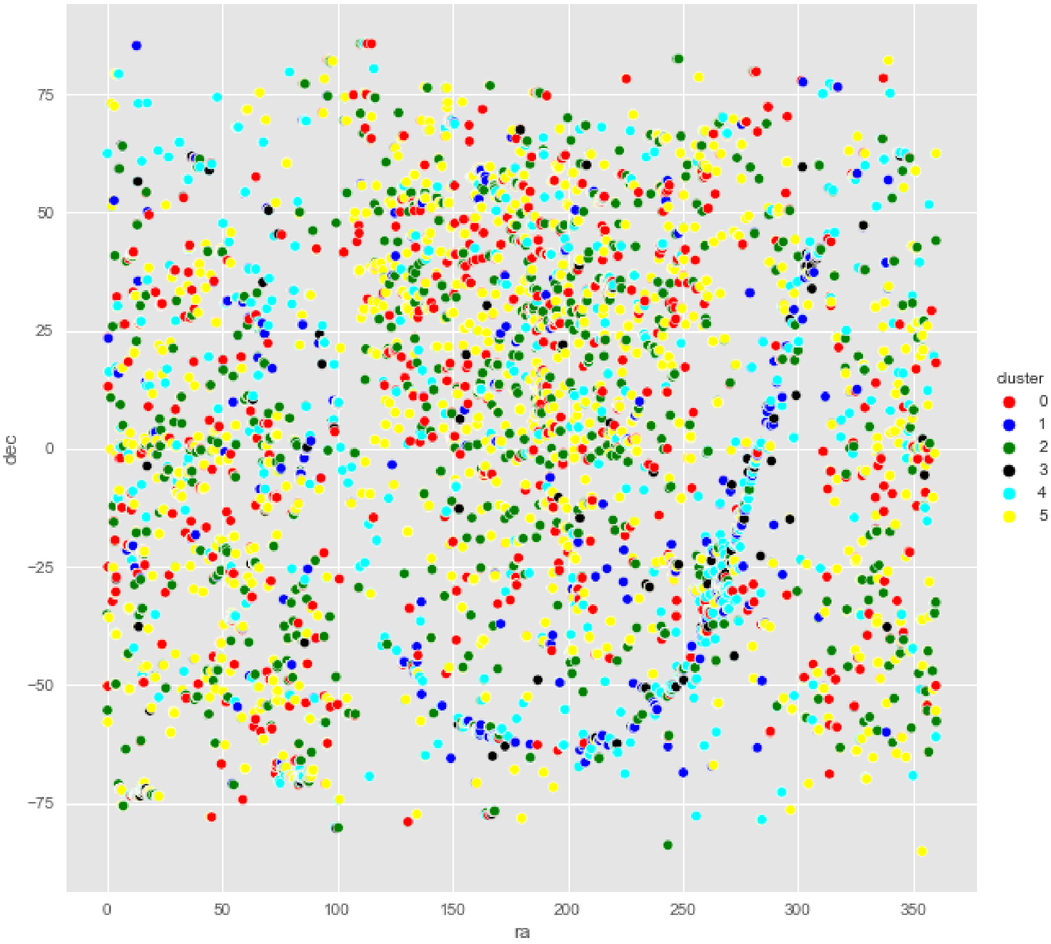


FIGURE 3.1: Results of the clustering method K-means over the celestial plane.Given that we have taken per-observation data, different observations could be related to different clusters. In this case, the color code is given by the cluster assigned to the last observation of a particular source. We can see a clear separation between intra-galactic and extra-galactic sources, which tells us that the model is actually identifying astrophysical differences between the sources.

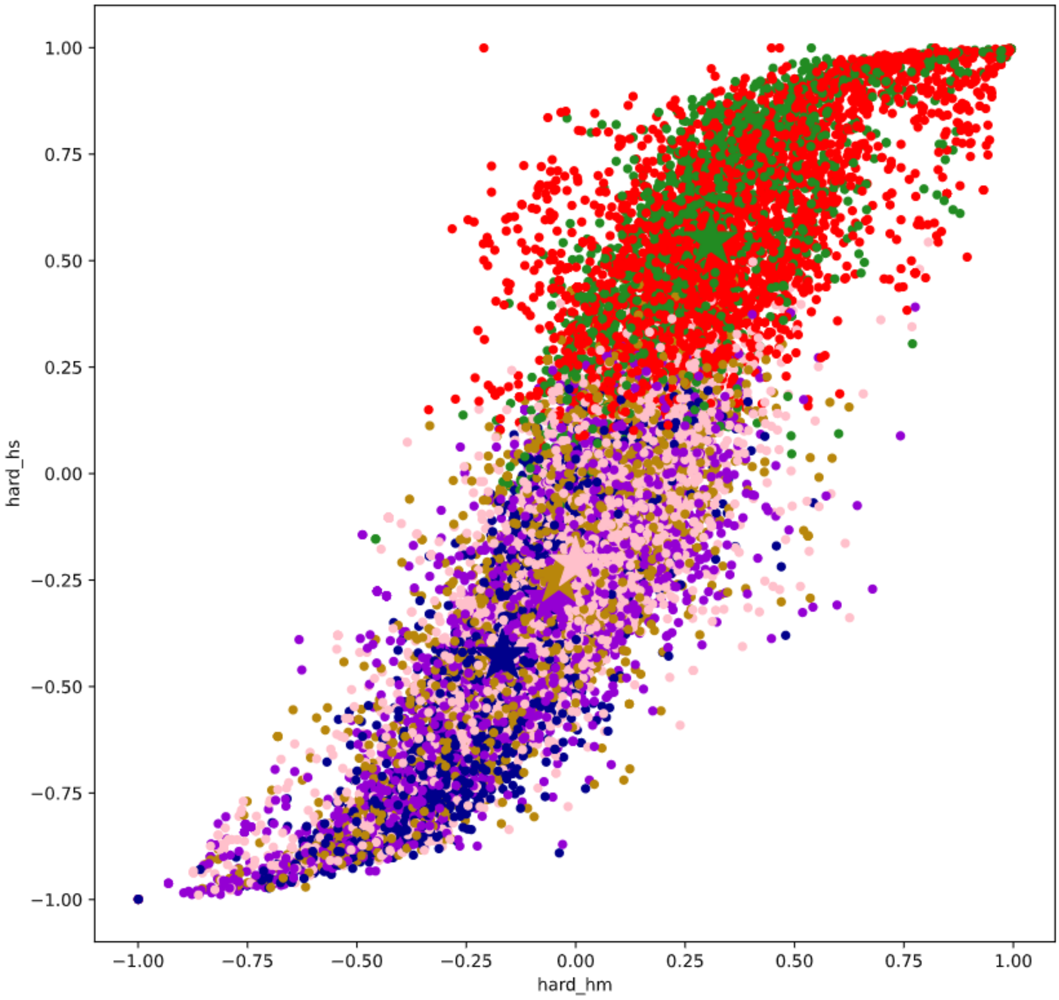


FIGURE 3.2: This visualization corresponds to a different run of the model. Corresponding again to hard_ms vs. hard_hm, this time color coded by the respective cluster of each data point. We can see a clear separation between the different stages of the correlation, having, for example, some clusters dominated by soft or hard sources.

main model.

3.2.2 Gaussian Mixture Models and EM

As we have seen in the previous subchapter, K-means clustering comes with some important difficulties. In order to overcome those particular restrictions, we decided to continue the analysis with a more general approximation to clustering. For now on, the analysis performed in this thesis will be based in the output generated by a Gaussian Mixture Model, given some benefits that we will describe later on.

A Gaussian mixture could be described as a linear superposition, or a convex combination, of K different Gaussian distributions

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{3.5}$$

where π_k is known as *mixture coefficients* or *mixture weigths*, and is restricted to:

$$0 \leq \pi_k \leq 1, \quad \sum_k \pi_k = 1.$$

Each of the Gaussian distributions that construct this mixture is called a *component*. Observe that, as we would expect, this model allows to represents more complex distribution in the subsets present in the data than to fix them all to a singular distribution, which gives us the opportunity to represent different clusters via different components. In order to represent clusters in our data optimally, we want to maximize the likelihood of the parameters of the different Gaussians corresponding to our mixture. An analytical solution for the parameters that maximizes the likelihood in the case of one distribution could be obtained relatively easy. This solution in a one dimensional case actually corresponds, as we would expect intuitively, to the mean and variance of the actual data. However, in the case of Gaussian mixtures, things get a little bit harder, and the maximum likelihood parameters no longer have a closed-form analytical solution. Thus, we could use different techniques in order to approximate those parameters, but the one that is most well adapted to the problems of *unsupervised machine learning* is the *EM algorithm*, which we will cover here briefly.

Let's recall that we want to find the parameters of the Gaussian mixtures that better explain a dataset. Suppose that this dataset has observations in the form $\boldsymbol{X} = \{x_1, \ldots, x_N\}$. Each of the data points $\boldsymbol{x}_n$ are i.i.d (independent and identically distributed random variables). We want to find the means $\boldsymbol{\pi}_k$, the covariances $\boldsymbol{\Sigma}_k$, and the mixture weights π_k that better adjust each Gaussian in the mixture to a cluster in the data.

In order to do that, let's first consider the likelihood function which, given the i.i.d properties, leads to a first representation:

$$p(\boldsymbol{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_n p(x_n \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3.6}$$

having:

$$p(\boldsymbol{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{3.7}$$

Thus, the log-likelihood function is given by:

$$\ln p(\boldsymbol{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{3.8}$$

Let's recall that we cannot produce a closed-form analytical solution for finding the parameters that maximize the log-likelihood defined above. A proposed iterative algorithm is suitable to estimate those parameters, this algorithm is the EM algorithm. We will optimize the log-likelihood with respect of the Gaussian Mixtures parameters. An usual way to do this would be to find their gradients and set it to zero. In this case, we will present an update step formula for each of the parameters. These formulas are dependent of one important variable that we will define here.

Responsibilities

The *responsibility* of the component k for the data point $\boldsymbol{x}_n$ is defined as:

$$r_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \tag{3.9}$$

We could see the responsibility as the posterior probability after observing $\boldsymbol{x}_n$, that the kth component explains the data point $\boldsymbol{x}_n$. This comes from the fact that the responsibilities for a data point is a normalized vector, and as we could see in the numerator of the formula, the responsibilities are proportional to the likelihood of an specific point $\boldsymbol{x}_n$.

Now we will give the formulas for each update of the parameters for fixed probabilities. These steps are executed one at a time, having the other parameters fixed.

Means

The update formula for the mean parameters $\boldsymbol{\mu}_k$ is given by:

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n, \tag{3.10}$$

where we define N_k as:

$$N_k = \sum_{n=1}^{N} r_{nk}. \tag{3.11}$$

Given that this formula depends on the responsibilities r_{nk}, we have that the update of the means actually depends on every parameter of the Gaussian mixture.

We could see intuitively that the quantity N_k is actually the total responsibility of the mixture k for our entire dataset. Thus, we could see that the update for the means is obtained by taking a usual mean of all the points in the dataset, weighted by the respective responsibilities r_{nk} of each datapoint $\boldsymbol{x}_n$ being generated by a mixture k.

Covariances

The update formula for the covariance parameters $\boldsymbol{\Sigma}_k$ is given by:

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T. \tag{3.12}$$

Again, analogous to the mean, we could see that this formula could be interpreted as a weighted covariance estimate for the data points $\boldsymbol{x}_n$, with the weights being the responsibilities for a mixture k to explain a data point $\boldsymbol{x}_n$.

Mixture weights

The update formula for the mixture weights or mixing coefficient parameters π_k is given by:

$$\pi_k = \frac{N_k}{N}, \tag{3.13}$$

where N is the total number of data points.

Given the formula above, we could interpret it as the average responsibility that the component k takes for all the data points, or analogously, as the relative importance of the kth mixture to the dataset.

EM algorithm

The expectation maximization algorithm (*EM algorithm*) is an iterative method for finding estimates of parameters that maximize the likelihood in models dependent on unobserved latent variables, and in particular in this case, mixture models. This algorithm is suitable to our parameter finding problem, given that we don't have a closed-form solution for the parameters, since the responsibilities depend on all the parameters of the Gaussian mixtures, suggesting an iterative procedure in order to find the desire solution. The EM algorithm was proposed initially in (Dempster, Laird, and Rubin, 1977) and a justification of the increase of the log-likelihood function for each step is presented in (Neal and Hinton, 1998).

The EM algorithm, as its name states, consist in two significant update steps: the *expectation* (E) step, and the *maximization* (M) step. In the case of Gaussian mixtures, we will first choose some initial values for the parameters of means, covariances, and mixture weights. These parameters could initially be chosen using different approaches, a basic one would be choosing them randomly, but a more well adapted routine is to perform first K-means clustering, assuming identity covariances, and extract the final parameters. This is suitable for EM since, as we have said before, K-means is a particular case of the EM algorithm applied to GMM.

After the initial parameters have been chosen, we can perform the *expectation* (E) step, where we will evaluate the responsibilities given the initialized parameters. Then, in the *maximization* (M) step, we will re-estimate the means, covariances, and mixture weights. We summarize the EM routine in the Alg. 2.

Usually the convergence criterion is given by a threshold either for the log likelihood or the parameters. Even though we know that in each step of the EM algorithm the log likelihood function increments, there are no guarantees for converging to the global maximum, and we may fall in a local maxima. A better chance to fall in a "good" local maximum could be determined by having better initialization parameters (Deisenroth, Faisal, and Ong, 2020).

1. Initialize $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, π_k. Means, covariances, and mixture weights, respectively.

2. E-step: Compute the responsibilities r_{nk} using the current parameter values

$$r_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

3. M-step: Re-estimate the parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, π_k using the already computed responsibilities r_{nk}:

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \boldsymbol{x}_n,$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T,$$

$$\pi_k = \frac{N_k}{N},$$

where

$$N_k = \sum_{n=1}^{N} r_{nk}.$$

4. Check for convergence evaluating the log likelihood or the parameters. If there is no converge, repeat.

Algorithm 2: EM for Gaussian Mixtures

Throughout this chapter, we have described the Gaussian Mixtures and the EM algorithm used to fit this model. However, we have left out details that are not relevant for this thesis, but are provided in (Bishop, 2006) and (Deisenroth, Faisal, and Ong, 2020), where we have taken inspiration. Further information about convergence, proofs for the different update steps, and an important latent variable view of Gaussian Mixtures that justifies the mathematical principles of the results presented here, are present in these references.

We have applied this method to our data, and the resulting clusters gave us important information that we used to design an algorithm for the classification of x-ray sources. Results of these preliminary analysis are presented in Chapter 4.

3.3 Classification Algorithm

The main goal of this thesis is to provide a probabilistic classification for several sources that have not been studied in detail before in the Chandra Source Catalog. In order to do this, it is clear that the output of a clustering model is not enough.

Therefore, in order to successfully provide this classification, we have constructed a procedure based on a similarity analysis using *Mahalanobis* distances. But in order to perform this similarity analysis, we have first to define observations that have been labeled. How do we do that? We perform a crossmatch with a database that could provide a significant amount of labels for the observations in our dataset. We selected the SIMBAD Astronomical Database, operated at CDS, Strasbourg, France, (Wenger et al., 2000) for this task.

For the crossmatch performed in the present work, we used a threshold of 3 arcsec, and we choose the best option from the found sources, i.e., the nearest source match encountered in SIMBAD. Then, for each source, we extracted the `main_type`. The astronomical software TOPCAT (Taylor, 2005) was of great help in this step.

It is important to say that, as we would expect, not every source found a respective label in the SIMBAD repository, or found an ambiguous one, such as *(Unknown, Radio, IR, Red, Blue, UV, X or gamma)*. These source observations are our target for giving a probabilistic classification.

After having the respective classes for all observations in our data, we analyzed their prominence for each of our 6 clusters. We found some classes that tend to appear in specific clusters, or being separated by their unique variability and spectral properties. Further information on the results of this procedure will be provided in Chapter 4.

Having labeled subclasses distributed around clusters allows us to perform a similarity analysis. The sources that did not find any matching label, or the ones that found an ambiguous one, will be our objective sources to classify. In order to explain this procedure, first we will need to cover some basics.

The Mahalanobis Distance

The Mahalanobis Distance is defined as:

$$D_M = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}, \tag{3.14}$$

where $\boldsymbol{x}$ is an observation, $\boldsymbol{\mu}$ are the mean values of the sample set, with a respective covariance matrix $\boldsymbol{\Sigma}$. The Mahalanobis distance is known intuitively as a distance measure of a point and a distribution. It was introduced in 1936 by Prasanta Chandra Mahalanobis (Mahalanobis, 1936).

The great power of the Mahalanobis distance is that it takes into account the correlations of the data set, also not depending on the scale of measurement. This comes as a particular advantage for this distance compared to the classical euclidean distance, since we do not know the actual distribution of each labeled subgroup in their correspondent cluster space. Having the correlation as an influence in the measure, allows us to take into account the distribution of the subgroup and how likely is our objective source to belong to that distribution. It is interesting to note that when the covariance matrix is the identity matrix, the Mahalanobis distance is equivalent to the Euclidean distance. In other terms, if we have spherical shaped subgroups, computing the Mahalanobis distance in one of the belonging points would be equivalent to perform a Euclidean metric. From a distribution perspective, using the euclidean distance corresponds to Gaussian distributions with common spherical covariance matrix, and in the other side, the Mahalanobis distance corresponds to Gaussian distributions with a general common covariance matrix (McLachlan, 2005). From a geometrical point of view, the Mahalanobis distance takes into acccount variances and covariances by transforming the data with an standardization and decorrelation process, and then, having a final transformed spherical data, we could compute the traditional Euclidean distance.

An alternative definition of the Mahalanobis distance using two observations $\boldsymbol{x}$ and $\boldsymbol{y}$ of the same distribution with a respectively covariance matrix Σ is given as follows:

$$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{y})}. \tag{3.15}$$

Softmin function

We will define the softmin function as:

$$\texttt{softmin}(x_i) = \frac{\exp(-x_i)}{\sum_j^K \exp(-x_j)}, \tag{3.16}$$

where x_i are real numbers of the vector $\boldsymbol{x} = \{x_1, \ldots, x_K\}$. The softmin function is defined formally by $\texttt{softmin} : \mathbb{R}^K \rightarrow [0,1]^K$. The softmin function is actually a convention of the widely known softmax function, which has been adopted particularly in neural networks and logistic regression analysis. We could define the softmin function in terms of the softmax function as follows:

$$\texttt{softmin}(x_i) = \texttt{softmax}(-x_i). \tag{3.17}$$

The main goal of the softmin function is to normalize an input real vector into a probability distribution. Thus, after applying the softmin function, every element of the vector will be in the range of $[0,1]$, and $\sum_i^K \texttt{softmin}(x_i) = 1$. In order to do this, in the softmax function, we apply the exponential function to each element x_i of the real vector $\boldsymbol{x}$, and then we normalize it by computing the sum of exponentials applied to all the elements of the vector. In this case we compute the softmin function for the negative value of x_i. We do this in order to obtain the larger probabilities with the smaller input components. We are interested in this particular minimum convention because we will be working with distances, specifically, we want to have larger probabilities of a point to belong to a subgroup, if the point is closer to the distribution that explains that subgroup.

An algorithm for probabilistically classifying sources

Now that we have the basis concepts, we can cover the algorithm that we used in order to provide a probabilistic classification to any undetermined or ambiguous source that we encountered. We partially will call this algorithm the *mahalanobis-softmin* algorithm.

First, we will explain each step in a descriptive way:

1. Takes as an input a cluster $\mathcal{C}_k$.

2. Take the objective label of the objects that we want to classify, i.e., in the particular case of this thesis, we will take as objective labels *NaN*, *Star*, *X* and *Unknown*.

3. We want to bound the number of possible labels in order to have a strong probabilistic classification. For that, we will extract the m most predominant (in size) labeled subgroups in the cluster. We will call each of this subgroups as *predominants*.

4. Iterate over the observations.

   - Compute the *Mahalanobis distance* of the observation to the points of each of the predominant soubgroups. Save those distances in a vector $\boldsymbol{\psi}$. Compute the mean of $\boldsymbol{\psi}$ and save it in a vector $\boldsymbol{\Psi}$, each of the means ψ_p will represent the mean mahalanobis distance of a point to a subgroup.
   - Apply the `softmin` function to $\boldsymbol{\Psi}$, the vector of mean distances to the predominants for a particular point.
   - Save the output of `softmin`($\boldsymbol{\Psi}$). This output will correspond to the classification probabilities for a particular observation. We will call this matrix of probabilities $\mathcal{P}$.

5. Finally and optionally, we could assign a label to each of the observations by looking at the label for which we found the largest probability.

6. Return the matrix of probabilities $\mathcal{P}$.

A formal summary pseudocode of the process above is presented here:

```
Input: A cluster $\mathcal{C}$, an array of labels objective_labels
$\overline{\mathcal{C}} = \mathcal{C}$[objective_labels]
predominants ← predominant_groups($\mathcal{C}$)
probs_matrix ← ∅
for each observation $x_n$ in $\overline{\mathcal{C}}$ do
    mdists ← ∅
    for each subgroup $\mathcal{G}$ in predominants do
        1. Compute the Mahalanobis distance of the point $x_n$ to every point
          observation $y_i \in \mathcal{G}$.
        2. mean_maha ← mean of the vector of distances generated.
        3. mdists.append(mean_maha)
    end
    /* Compute the probabilities using a softmin function.      */
    obs_prob ← softmin(mdists)
    probs_matrix.append(obs_prob)
end
return probs_matrix
```

Algorithm 3: Probabilistic classification algorithm. This algorithm assign a probabilistic classification to specific observations in a cluster, if some labels for some objects are provided.

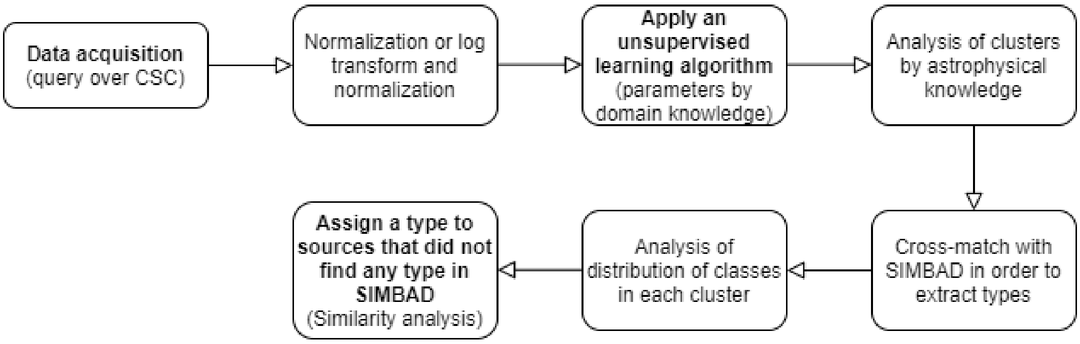

Data acquisition
(query over CSC)
Normalization or log transform and normalization
Apply an unsupervised learning algorithm
(parameters by domain knowledge)
Analysis of clusters by astrophysical knowledge
Cross-match with SIMBAD in order to extract types
Analysis of distribution of classes in each cluster
Assign a type to sources that did not find any type in SIMBAD
(Similarity analysis)

FIGURE 3.3: An intuitive summary of the pipeline performed in the work at hand.

A final summary of the complete procedure presented in Chapter 3 and Chapter 4 is presented in Fig. 3.3.

Chapter 4

Results and discussion

In this chapter we will present and discuss the results of applying the methods presented earlier (Chapter 3) in the data extracted and processed from the Chandra Source Catalog (Chapter 2).

4.0.1 Gaussian Mixtures Model Output

We performed a Gaussian Mixtures Model over the data in order to automatically separate the different observations according to their particular astrophysical properties. This process allowed us to define clusters that are determined by a set of astrophysical features, which could potentially describe as well a particular source class. With this process we already obtained some results that relates to the distribution of astrophysical properties over the different clusters obtained. Particularly, from this analysis we concluded that the model was successful in completing its task of identifying the invariant features of each of the observations, and effectively grouping them by their similarity over the clusters. We present again the plot of the celestial plane in Figure 4.1 in order to highlight, again, the separation of extra-galactic and intra-galactic sources.

Figure 4.2 shows the distribution of the hardness ratios between hard and soft band (`hard_hs`) and the variability probabilities for the broad band (`var_prob_b`) over the resulting clusters. Cluster 4, in cyan, tends to represent softer sources, having both high variable and low variable sources. Cluster 3, in pink, is clearly a harder cluster, representing more energetic sources, but as well tending to have a low variability. Sources in Cluster 0 seems to be less variable than average, but also tend to distribute both in hard and soft X-Ray spectra. More of this invariant features could be observed for the different resulting clusters.

Figure 4.3 shows again a particular projection of our multidimensional space over the hardness ratios, color coded by the clusters. We see here that correlations between variables are determining separations between clusters, having for example, clusters with sources that have a very soft X-Ray spectra, as well others that are in the highest part of the plot, representing harder sources.

Additionally, we can observe that the model is effectively grouping the clusters by their unique astrophysical properties and the differences presented in each source observation. This came as a first result that encouraged us to continue in the interpretation of the particular properties represented for each cluster, and how that could be related with a particular class distribution over the set. Having partial results that separated sources, first by an extra and intra galactic dissimilarity, and then by particular tendencies in their properties, shows us that a clustering technique

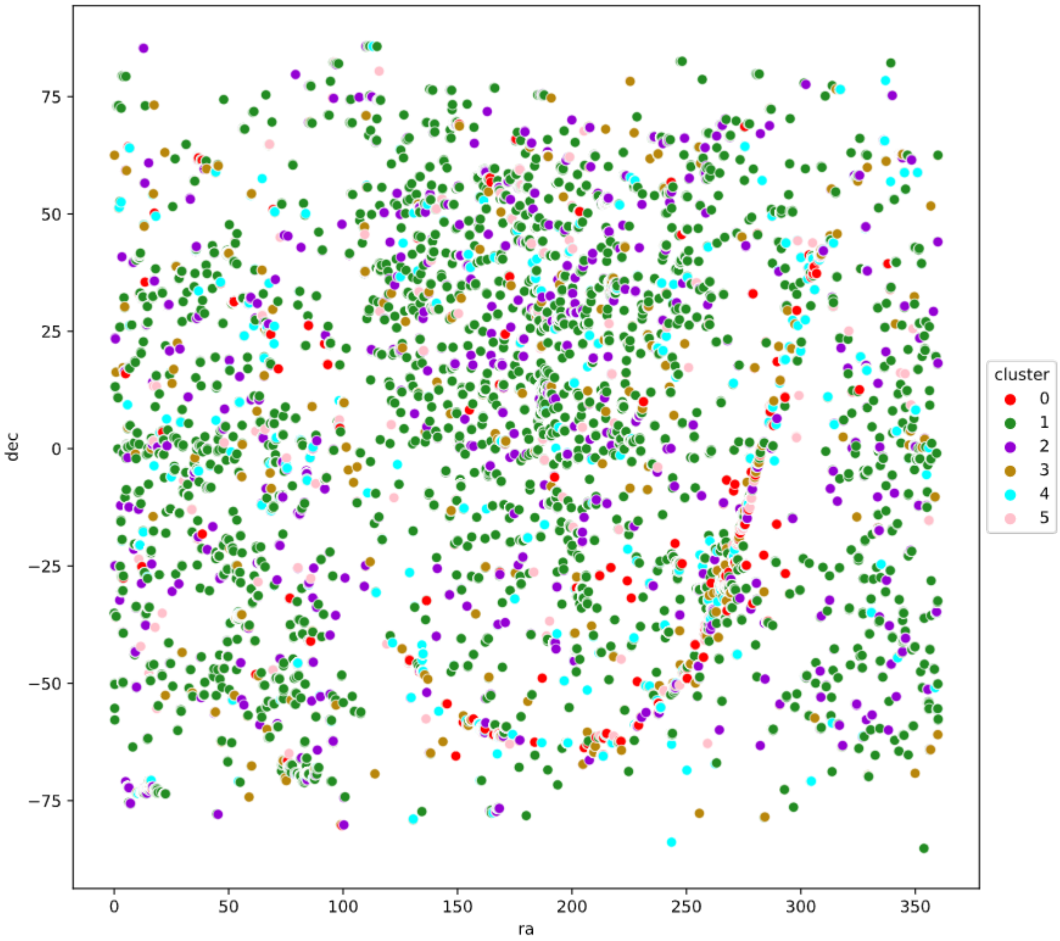


FIGURE 4.1: Clustering result from the Gaussian Mixtures Model over the celestial plane. Given that we have taken per-observation data, different observations could be related to different clusters. In this case, the color code is given by the cluster assigned to the last observation of a particular source. We can see a clear separation between intra-galactic and extra-galactic sources, which tells us that the model is identifying astrophysical differences between the sources. The resulting clusters are much more dispersed over the celestial plane compared to 3.1, but again a predominance of clusters could be analyzed visually over specific regions of the plane.

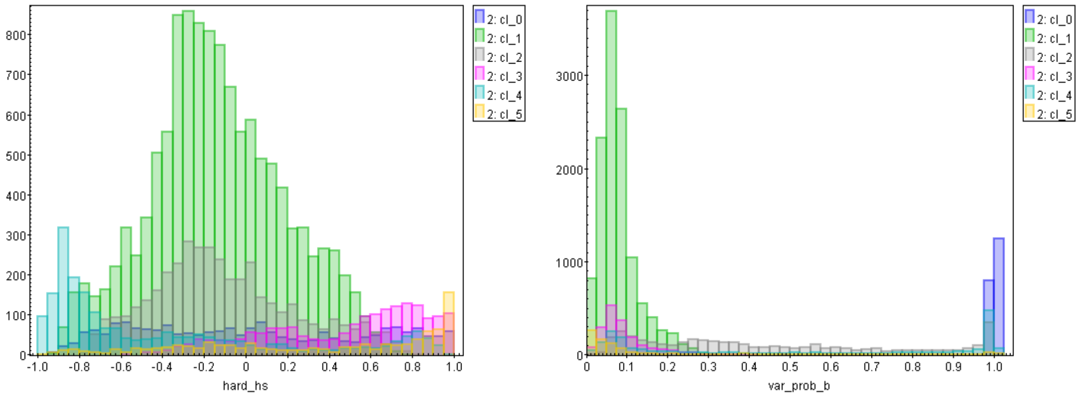


FIGURE 4.2: *Visualization of the distribution of* ***hard_hs*** *(left) and* ***var_prob_b*** *(right) over the clusters encountered by our Gaussian Mixtures Model. Each histogram is color coded by its cluster id. Let's recall that we selected 6 as the number of components.*

,

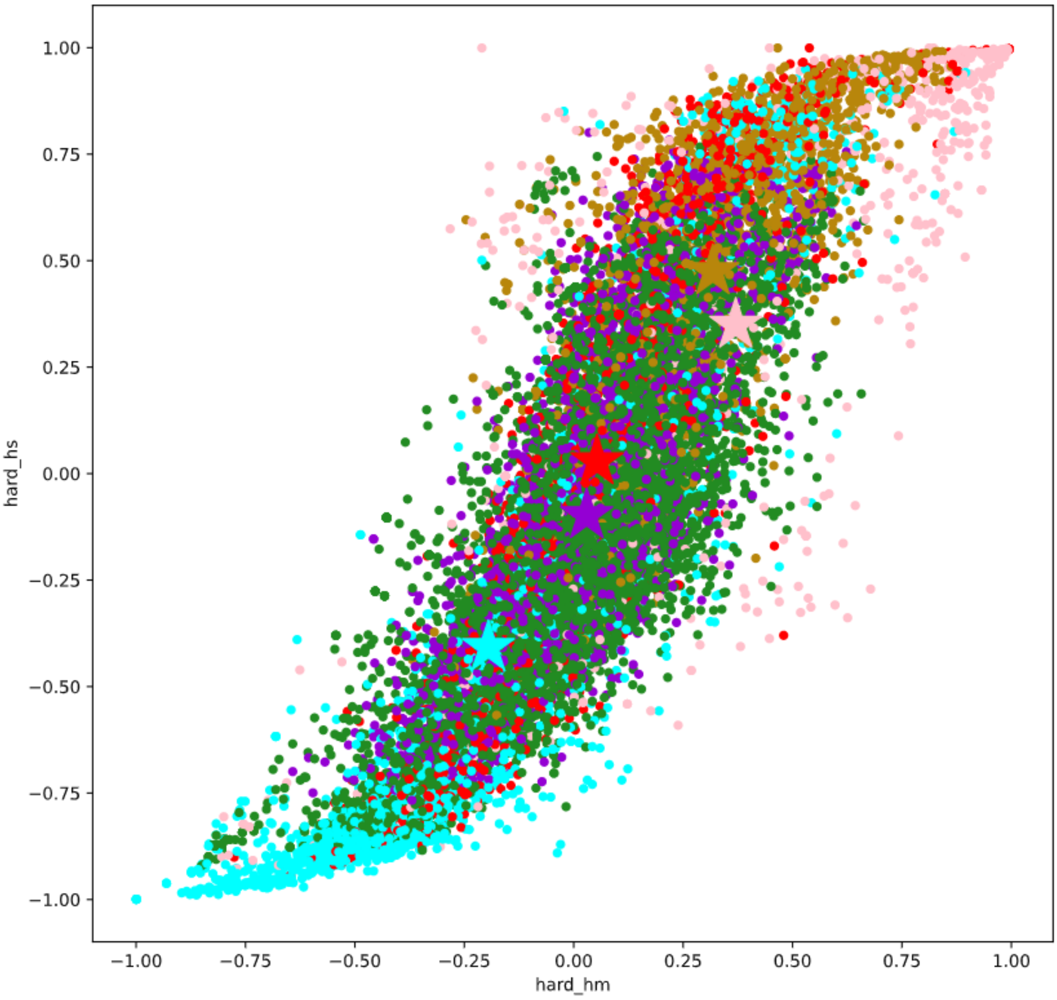


FIGURE 4.3: `hard_ms` vs. `hard_hm`, color coded by the respective cluster for each data point. The stars represents the means of each cluster.

could potentially lead to a contrast between classes and then, to new techniques for classifying sources that lack of counterparts in other wavelengths or have not been studied in detail.

4.0.2 Classification

In order to be able to classify the sources, we needed classes. We extracted these classes with the process described in Chapter 3. After the crossmatch was performed, we decided to analyze if the particular separation based on astrophysical properties was going to be reflected as well in the distribution of classes available. For each cluster, we observed the most predominant classes by number. We present this information in Table 4.1. For example, we could see that in the Cluster 0 we have a clear predominance of YSOs, and even different kinds of these sources, such as Orion_V* or TTau*. In other clusters, such as Cluster 2, we mostly find ambiguous sources or unlabeled sources, but again, after those sources, the predominance comes to YSOs. In Cluster 5 we can find other kind of sources such as Seyfert 2 galaxies, Part of Globulal Clusters or even Pulsars, being clearly dominated by extragalactic sources.

We could argue that the most pure clusters in terms of predominance are Clusters 0 and Cluster 5, covering mostly YSOs and extra-galactic sources, respectively. However, we also find that other clusters are not that pure, such as Cluster 2, where we find *QSO* and *YSO* having a similar predominance. We can find this ambiguity as well in Cluster 1, and in Cluster 3, particularly mixing *HMXB* and *YSO*. These suggests that additional optical information may be needed in order to separate those particular kind of sources that are not well defined by just the X-ray data. Observe in most of the clusters we have a predominance of *NaN* labeled source observations, telling us that most of the sources lack of a classification.

It is of importance to note that we can find the same classes in different clusters, but as we observed previously, this could give us hints of particular sources that belong to a class but have a differentiated behaviour or feature, such as, for example, their variability or their X-ray hardness. Those cases are of interest since we could pinpoint exotic phenomena, or find particularities in the life evolution of sources.

We will remark here too that we have found that in many cases observations from the same source that has been observed several times fell in different classes. This is a particular result that could give us information about the different evolutionary stages of a source, which can look different in different observations. This difference can be due to short term changes in the physical properties of the object in question.

Having in mind that we want to probabilistic classify X-ray sources lacking of a class label, we will focus in those sources labeled with an ambiguous class or unlabeled at all. Approximately 44% of the sources in our data corresponds to these subset.

Now we have everything that we need in order to perform a classification. Some clusters that are well defined by particular astrophysical properties, and classes that are distributed in those clusters. We use the algorithm proposed in Chapter 3 in order to obtain a probabilistic classification for the sources that are labeled as *NaN*, *Star*, *X* and *Unknown*, that corresponds to unlabeled sources (those that did not find any matching type in SIMBAD), and those with an ambigous matching label (for

main_type	size
Orion_V*	495
YSO	418
Star	298
NaN	274
TTau*	103
X	82
BYDra	45
Candidate_YSO	44
HMXB	33
PM*	18

(A) Cluster 0.

main_type	size
NaN	3545
QSO	1494
X	1476
Star	709
YSO	393
Orion_V*	363
AGN	353
Candidate_XB*	346
Seyfert_1	309
GlCl	270

(B) Cluster 1.

main_type	size
NaN	1129
X	401
QSO	379
YSO	358
Orion_V*	347
Star	243
AGN	147
HMXB	132
Seyfert_1	102
TTau*	78

(C) Cluster 2.

main_type	size
NaN	443
X	288
HMXB	167
YSO	125
Star	92
QSO	77
SNR	54
Orion_V*	47
AGN	47
GlCl	45

(D) Cluster 3.

main_type	size
NaN	410
Star	256
YSO	254
Orion_V*	197
X	164
TTau*	102
HMXB	78
QSO	70
PM*	57
GlCl	54

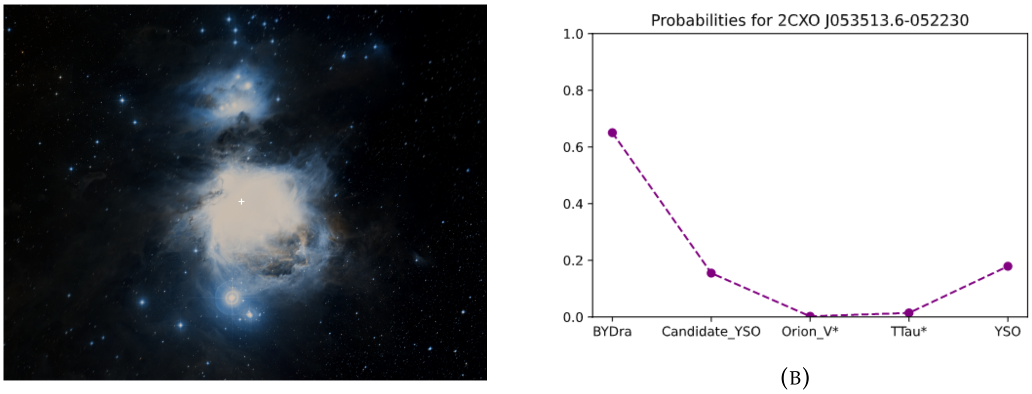
(E) Cluster 4.

main_type	size
NaN	124
Seyfert_2	98
X	89
PartofG	45
Pulsar	40
Radio	40
YSO	39
QSO	35
HMXB	35
Star	33

(F) Cluster 5.

TABLE 4.1: *The 10 most predominant classes in each of the clusters. Recall that we also have ambiguous and non matching sources, the last labeled as "NaN".*

,



(A)

(B)

FIGURE 4.4: *2CXO J053513.6-052230*, *obs_id* 6419, BYDra candidate found by our model in Cluster 0. (A) DSS2 color image of the location of the source in the Orion Nebula. (B) Classification probabilities for the source, showing the 5 predominant classes for its cluster.

example, *X* type says that we have an X-ray source, but that does not tells us much).

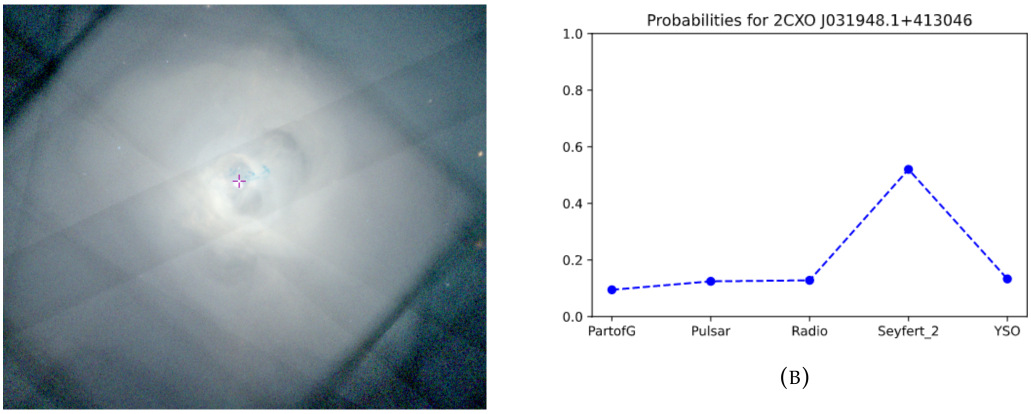
As output, we obtained for each cluster a list of sources with a probability of belonging to the five predominant classes of its cluster. We added another column `main_type` which represents the class that had greater probability for a particular source observation. You can find all the tables and the code available at `https://github.com/BogoCoder/astrox`.

In order to verify the classification output of our algorithm, we have explored manually some of the classified sources and based on astrophysical knowledge, identify if the classification was reliable. This process comes as a first exploratory analysis for measuring the performance of the algorithm, but further research could be done in order to find a more suitable approach to quantify the performance of the algorithm, such as constructing a test set with recent contributions to classification of astrophysical X-ray sources.

We present here some examples of selected sources with their classification based on our algorithm, and an interpretation of these results.

An example of a `BYDra` variable candidate is shown in figure 4.4. The source 2CXO J053513.6-052230 is located in the Orion Nebula, which is a known star-forming region, and particularly, it is a large collection of variable stars (Muench et al., 2009). The algorithm has been able to identify this source as a variable star, which given its location, is a candidate for a young star. Additionally, a particular kind of variable star is identified, BY Draconis variable, which could be related to a particular quasiperiodic behaviour in the light curve and a bounded brightness fluctuation (López-Morales et al., 2006). These cases are of interest for further study, in order to extract the predominant properties that made the source fall in this class.

Another example of a `Seyfert_2`, a Type II Seyfert galaxy candidate, is shown in figure 4.5. The source 2CXO J031948.1+413046 is located in the Perseus Cluster, which is a known cluster of galaxies in the constellation Perseus. It contains thousands of galaxies in a huge cloud of gas and dust. As we could observe, the source



(A)

(B)

FIGURE 4.5: *2CXO J031948.1+413046*, both in `obs_id` 11713 and 12036, Seyfert Type 2 candidate found by our model in Cluster 5. (A) Chandra X-ray image of the location of the source in the Perseus Cluster. (B) Classification probabilities for the source, showing the 5 predominant classes for its cluster.

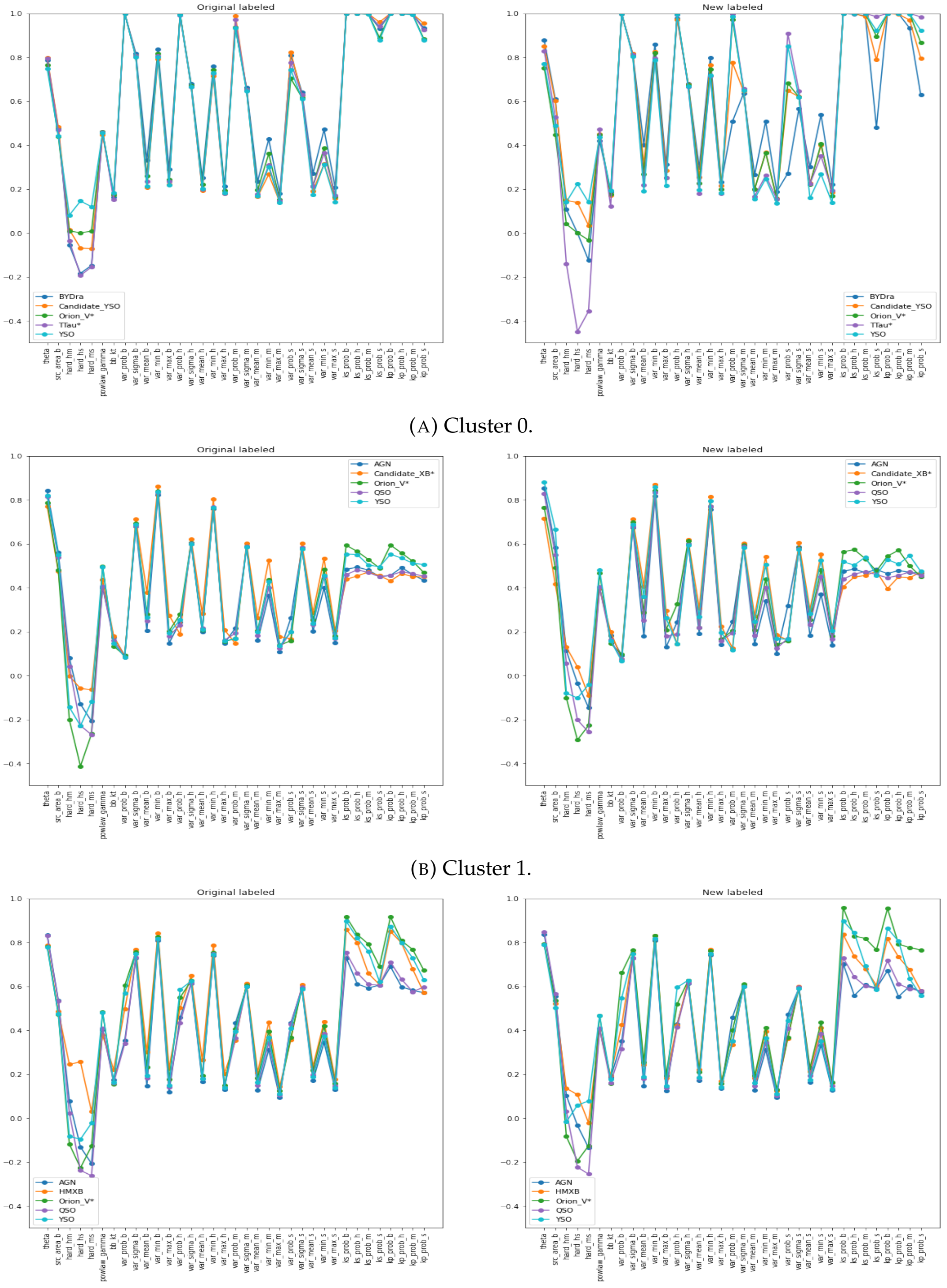
is located at the center of the cluster. Other Seyfert Type 1.5 or 2 have been studied in the center of the Perseus Cluster, such as 3C 84 (Rani et al., 2018) (Véron-Cetty, M.-P. and Véron, P., 2006). These particular galaxies have a very bright core, and are specifically bright in the infrared wavelengths. The number of candidates to be a Seyfert Type 2 galaxy is limited, having NGC 3147 as the best candidate for a "true" case (Matt, G. et al., 2012). Thus, specifying the invariant features that made this source to fall as a `Seyfert_2` galaxy could give us an opportunity for a proposal of a new candidate.

As we could assume, there are limitations in our model, and even though we had a reasonable classification output based on a manual revision, there are some misclassified sources that are of interest for a possible improvement in the model or identification of anomalies. For example, the source *2CXO J232327.9+584842* is classified in an observation as a YSO, but it is a neutron star in the center of the supernova remnant Cas A (Potekhin et al., 2020). Another case is the source *2CXO J053747.4-691019* is classified as an X-ray Binary, but it is a pulsar located in the Large Magellanic Cloud (Lin, Webb, and Barret, 2012). We have just provided some examples, but further analysis could be performed in order to identify errors in the classification. Having this analysis at hand, an investigation over the causes of those misclassifications would be suggested. Let's recall that the we used per-observation data for this work, and thus, observations of the same source could fall in very different classes. Those cases are of interest in order to find hints for a more robust model or a possible anomaly candidate.

In order to summarize the classification results presented here, we could visualize the distributions of the mean values over the properties, for the source observations that originally have a class (those that found a match in SIMBAD) and the new classified sources (with the model). We present this visualization in Figure 4.6. We can see here that most of the new labeled sources follow a similar distribution over the properties compared to the original source observations. We could as well observe that for particular classes we have clear trends over some properties in each cluster, telling to us that predominant classes have particular distributions in their cluster space, confirming again that taking a more general measurement metric in order to consider wider cases of distributions was a reasonable choice. For example, in Cluster 0 we have that *TTau** and *BYDra* have similar distributions over the hardness ratios, and as it could be seen, these distributions are different compared to the other classes in the cluster. Observe that for the new labeled observations, we have very similar distributions over the hardness ratios for the *TTau** type. However, for *BYDra* we have some dissimilarity in the distribution, which in the original case was very similar to the *TTau** case. These results could give us important information about the classification performance, since for the particular case of *BYDra* type, we could say that the algorithm found some ambiguity between the different source observations available. Similar trends of the property distributions could be observed for other cases such as Cluster 1, Cluster 2 and Cluster 4, the last as well dominated by YSOs.

In the case of Cluster 2, we recall that it is one of the cases in which it is not clear what particular class of sources is representing, since, for example, the most predominant classes are *QSO* and *YSO*. Observe, however, that we can find a similar distribution over the properties for this classes, suggesting that they share similar X-ray properties, which is of interest, since this could be explained by abnormal phenomena. However, this could encourage an addition of properties in other wavelengths, mainly the optical, where descriptors such as the *redshift* would be valuable for separating these kind of sources.

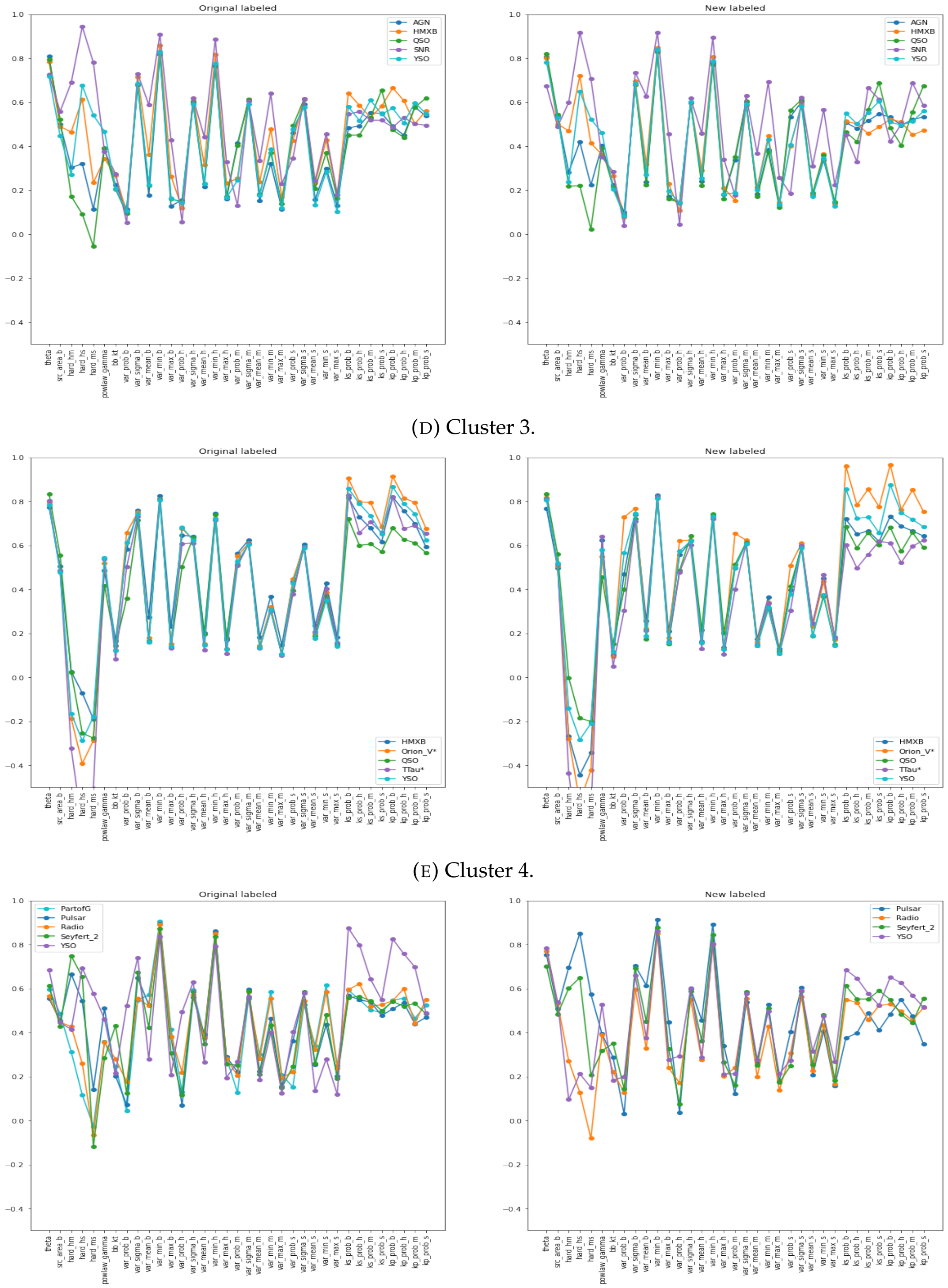
For the Cluster 3, which is another case of a mixed cluster with no clear representation, we can observe big differences between the predominants distribution over the properties. For example, we can see that *QSO* and *SNR* are very different in the hardness ratios, one tending to the lower values, and the other tending to higher, respectively. Thus, a pattern of the distribution over the properties is not really clear, but this gives us evidence that we have sources with similar properties that tend to group them in one cluster, and at the same time those sources can diverge in large scales within the cluster. A similar behaviour can be encountered in the Cluster 5, but in this case the one type of source diverging the most is YSO, a type of source that does not suit very much in the predominants of *Seyfert_2* or *PartofG*. Even though that we have very different classes (*YSOs* and extra-galactic objects), properties are similar between them and could be of particular interest for a future work.



(A) Cluster 0.

(B) Cluster 1.

(C) Cluster 2.



(D) Cluster 3.

(E) Cluster 4.

(F) Cluster 5.

FIGURE 4.6: Mean values of the properties for original and new labeled sources in each cluster. This visualization intends to show the similarities between the distributions over the properties, product of the model based on the Mahalanobis distance.

Chapter 5

Conclusions and future work

In this thesis we developed a procedure to probabilistically classify sources of the Chandra Source Catalog. This procedure is based on unsupervised learning and distance metrics, particularly the Mahalanobis distance. First of all, in Chapter 2, we describe how we made a query over CSC2, in order to bound the amount of registers and have enough significant sources with spectra that could be fitted by a model. After that, we preprocessed the obtained data, based on normalization and log transformation, in order to have data distributed in similar scales that could be compared and determined by our machine learning model. In Chapter 3 we described the methods used, first the unsupervised learning algorithms, at the end focusing in Gaussian Mixtures Model. After using the unsupervised learning technique, we crossmatched the data in order to extract available classes, and finally we used an algorithm based on Mahalanobis distances in order to provide probabilistic classification to all the sources. We described and discussed the results in Chapter 3, having as a main output for this first approach 10090 source observations of the Chandra Source Catalog with an assigned probabilistic classification. These source observations are distributed around the clusters, and each register have an assigned class `main_type` and its respective probabilities of belonging to the 5 most predominant classes in its cluster. Codes and tables are available at `https://github.com/BogoCoder/astrox`.

Preliminary results of this thesis were presented at the Chandra Data Science Workshop 2021, organized by the Chandra X-ray Center. In the future we aim to extend this research and publish a paper summarizing the most important findings.

From the main findings already presented in this work, we want to highlight the following ideas:

- Through this research we have seen that it is possible to assign a probabilistic classification to the Chandra sources using our approach.
- We have seen that this approach works well for particular objects such as *YSOs*, or extra-galactic sources, but is ambiguous in other cases.
- Those cases presenting ambiguity are of importance too, and a possible improvement to the pipeline that could separate those kind of sources is adding optical information.
- This thesis presents an early approach for the final goal of the research, which is to present a probabilistic classification to all the sources in the Chandra Source Catalog that lacks of a class or have not been studied in detail.

As a future work, there are some important points that we have not explored yet in this work, but that are a potential opportunity of improvement or complement to the methods and results presented.

- Include CSC2-SDSS (Sloan Digital Survery Summary) unambiguous matches, adding optical information to our pipeline that would be potential differentials for the source observations entered into the model.
- An exploration of direct semi-supervised techniques including this optical information. Although the work done in this thesis could be considered as well as semi-supervised, it may be a good idea considered methods that are specifically designed for these kind of problems.
- Further analysis in atypical observations that were misclassified is required. These ambiguities could be mitigated by adding multi-wavelength data to the model. Nevertheless, these particular cases are of interest in order to find possible anomalies or to understand better chances of improvement.

Bibliography

Alpaydin, Ethem (2014). *Introduction to Machine Learning*. 3rd ed. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press. ISBN: 978-0-262-02818-9.

Ansari, Zoe, Agnello, Adriano, and Gall, Christa (2021). "Mixture models for photometric redshifts". In: *A&A* 650, A90. DOI: 10.1051/0004-6361/202039675. URL: https://doi.org/10.1051/0004-6361/202039675.

Astropy Collaboration et al. (Oct. 2013). "Astropy: A community Python package for astronomy". In: *Astronomy & Astrophysics* 558, A33, A33. DOI: 10.1051/0004-6361/201322068. arXiv: 1307.6212 [astro-ph.IM].

Astropy Collaboration et al. (Sept. 2018). "The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package". In: *The Astronomical Journal* 156.3, 123, p. 123. DOI: 10.3847/1538-3881/aabc4f. arXiv: 1801.02634 [astro-ph.IM].

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag. ISBN: 0387310738.

Deisenroth, Marc Peter, A. Aldo Faisal, and Cheng Soon Ong (2020). *Mathematics for Machine Learning*. Cambridge University Press.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1, pp. 1–38. ISSN: 00359246. URL: http://www.jstor.org/stable/2984875.

Evans, Ian N. et al. (2010). "THE CHANDRA SOURCE CATALOG". In: *The Astrophysical Journal Supplement Series* 189.1, pp. 37–82. DOI: 10.1088/0067-0049/189/1/37. URL: https://doi.org/10.1088/0067-0049/189/1/37.

Event Horizon Telescope Collaboration et al. (Apr. 2019). "First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole". In: *The Astrophysical Journal Letters* 875.1, L1, p. L1. DOI: 10.3847/2041-8213/ab0ec7. arXiv: 1906.11238 [astro-ph.GA].

Farrell, Sean A, Tara Murphy, and Kitty K Lo (2015). "Autoclassification of the variable 3xmm sources using the random forest machine learning algorithm". In: *The Astrophysical Journal* 813.1, p. 28.

Harris, Charles R. et al. (Sept. 2020). "Array programming with NumPy". In: *Nature* 585.7825, pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

Hunter, J. D. (2007). "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

Lin, Dacheng, Natalie A. Webb, and Didier Barret (2012). "CLASSIFICATION OF X-RAY SOURCES IN THEXMM-NEWTONSERENDIPITOUS SOURCE CATALOG". In: 756.1, p. 27. DOI: 10.1088/0004-637x/756/1/27. URL: https://doi.org/10.1088/0004-637x/756/1/27.

Lo, Kitty K et al. (2014). "Automatic classification of time-variable X-ray sources". In: *The Astrophysical Journal* 786.1, p. 20.

Logan, C. H. A. and Fotopoulou, S. (2020). “Unsupervised star, galaxy, QSO classification - Application of HDBSCAN”. In: *A&A* 633, A154. DOI: 10.1051/0004-6361/201936648. URL: https://doi.org/10.1051/0004-6361/201936648.

López-Morales, Mercedes et al. (2006). “Limits to Transits of the Neptune-Mass Planet Orbiting GJ 5811”. In: 118.849, pp. 1506–1509. DOI: 10.1086/508904. URL: https://doi.org/10.1086/508904.

Mahalanobis, Prasanta Chandra (1936). “On the generalized distance in statistics”. In: *Proceedings of the National Institute of Sciences (Calcutta)* 2, pp. 49–55.

Matt, G. et al. (2012). “The Suzaku X-ray spectrum of NGC 3147 - Further insights on the best "true" Seyfert 2 galaxy candidate”. In: *A&A* 540, A111. DOI: 10.1051/0004-6361/201118729. URL: https://doi.org/10.1051/0004-6361/201118729.

McLachlan, Geoffrey J (2005). *Discriminant analysis and statistical pattern recognition*. Vol. 583. John Wiley & Sons.

Merloni, A. et al. (2012). *eROSITA Science Book: Mapping the Structure of the Energetic Universe*. arXiv: 1209.3114 [astro-ph.HE].

Muench, August et al. (Jan. 2009). “Star Formation in the Orion Nebula I: Stellar Content”. In:

Neal, Radford M. and Geoffrey E. Hinton (1998). “A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants”. In: *Learning in Graphical Models*. Ed. by Michael I. Jordan. Dordrecht: Springer Netherlands, pp. 355–368. ISBN: 978-94-011-5014-9. DOI: 10.1007/978-94-011-5014-9_12. URL: https://doi.org/10.1007/978-94-011-5014-9_12.

pandas development team, The (Feb. 2020). *pandas-dev/pandas: Pandas*. Version latest. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

Pedregosa, F. et al. (2011). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Perlman, Eric S. and Andrew S. Wilson (July 2005). “The X-Ray Emissions from the M87 Jet: Diagnostics and Physical Interpretation”. In: *The Astrophysical Journal* 627.1, pp. 140–155. DOI: 10.1086/430340. arXiv: astro-ph/0503024 [astro-ph].

Pineau, F-X et al. (2010). “Comparison of classification methods for XMM sources”. In: *Astronomical Data Analysis Software and Systems XIX*. Vol. 434, p. 369.

Potekhin, A Y et al. (June 2020). “Thermal luminosities of cooling neutron stars”. In: *Monthly Notices of the Royal Astronomical Society* 496.4, pp. 5052–5071. ISSN: 0035-8711. DOI: 10.1093/mnras/staa1871. eprint: https://academic.oup.com/mnras/article-pdf/496/4/5052/33507099/staa1871.pdf. URL: https://doi.org/10.1093/mnras/staa1871.

Rani, B. et al. (2018). “NuStar View of the Central Region of the Perseus Cluster”. In: 866.1, p. L13. DOI: 10.3847/2041-8213/aae48f. URL: https://doi.org/10.3847/2041-8213/aae48f.

Rostami Osanloo, Mehrdad et al. (2019). “Classification of Extragalactic X-Ray Sources Using Machin Learning Methods”. In: *AAS* 233, pp. 457–03.

Taylor, M. B. (Dec. 2005). “TOPCAT & STIL: Starlink Table/VOTable Processing Software”. In: *Astronomical Data Analysis Software and Systems XIV*. Ed. by P. Shopbell, M. Britton, and R. Ebert. Vol. 347. Astronomical Society of the Pacific Conference Series, p. 29.

Véron-Cetty, M.-P. and Véron, P. (2006). “A catalogue of quasars and active nuclei: 12th edition”. In: *A&A* 455.2, pp. 773–777. DOI: 10.1051/0004-6361:20065177. URL: https://doi.org/10.1051/0004-6361:20065177.

Waskom, Michael L. (2021). “seaborn: statistical data visualization”. In: *Journal of Open Source Software* 6.60, p. 3021. DOI: 10.21105/joss.03021. URL: https://doi.org/10.21105/joss.03021.

Wenger, M. et al. (Apr. 2000). “The SIMBAD astronomical database. The CDS reference database for astronomical objects”. In: *AAPS* 143, pp. 9–22. DOI: 10.1051/aas:2000332. arXiv: astro-ph/0002110 [astro-ph].

Wilkes, Belinda and Wallace Tucker, eds. (2019). *The Chandra X-ray Observatory*. 2514-3433. IOP Publishing. ISBN: 978-0-7503-2163-1. DOI: 10.1088/2514-3433/ab43dc. URL: http://dx.doi.org/10.1088/2514-3433/ab43dc.