Universitá degli Studi di Napoli
Federico II

Facoltà di Scienze Matematiche, Fisiche e Naturali

Tesi Sperimentale di Laurea Magistrale in Astrofisica e Scienze dello Spazio

Anno Accademico 2013/2014

CLASSIFICATION OF ASTRONOMICAL TRANSIENTS WITH MACHINE LEARNING METHODS

Candidate
Antonio D'Isanto
matr. N91/16

Supervisors
Ch.mo Prof. Giuseppe Longo
Dr. Massimo Brescia
Dr. Stefano Cavuoti

To my Parents,
who aimed me to the stars!

Contents

Chapter 1

Introduction

The rising era of synoptic imaging surveys has opened the exciting chapter of time-domain astrophysics: one of the fastest growing areas of astrophysical research. A number of important phenomena can, in fact, be studied only in this domain, while new and previously unknown phenomena expect to be discovered. The purpose of this thesis is the classification of astrophysical transients in synoptic surveys, using data mining techniques and methods. The exploration of the temporal domain in search of variable objects and transients has known a constant expansion during the last few years, impacting on all branches of astrophysical research. With the term variable we refer to sky objects whose luminosity presents a more or less accentuate variation in time. As we shall see in what follows, the understanding of the underlying physical mechanisms responsible for the variability represents a crucial aspect in explaining a great variety of phenomena, from Supernovae (SN), to variable stars and Active Galactic Nuclei (AGN), including some of the most energetic events in the Universe, and the produced data volumes have begun to overcome what is possible to visually inspect even for large teams of astronomers, and also crowds of "citizen scientists" are not sufficient to the task. So, an increasingly central role of software and hardware frameworks is needed in order to supply the traditional roles of humans in the real-time loop. In this not so futuristic scenario, data need to be automatically transported, processed, calibrated, and ingested into databases without human intervention.

Each step of such data flow presents many challenges: from the discovery to the detection, to classification and, possibly, to the automatic setup of follow-ups for the most interesting and peculiar variable objects. This requires to employ significant resources, in particular for what concerns the observing time and the technologies used. This will become even more cogent in the near future when a new generation of instruments (such as LSST - Large Syn-

optic Survey Telescope[1], SKA - Square Kilometer Array[2], etc.) will produce an increasingly large amounts of complex data every night. For these instruments a massive application of intelligent and automatic multi-disciplinary methods, enclosed under the umbrella of Astroinformatics (that can be considered as a new scientific matter, standing in the more general family called *X-informatics*), will be an absolute must and in fact, Astroinformatics in particular and X-informatics in general, configure as the *"fourth paradigm"* of scientific research (the others are experimentation, theory and simulation - [27]). In other words, Information Technologies (IT), Data Mining (DM) and Machine Learning (ML) methods need to become an indispensable part of the game.

A central role, in this sense, has been acquired by the Virtual Observatory infrastructure. It is a project that has the aim to create a new way to construct astrophysical research. It is developed in an international framework from national research agencies and expanded collaborations. The major aim of the project is to make possible to researchers and students a simple access to data archives, resources and applications through the web. Programs that are needed for data analysis are available in pre-compiled packets (for example viewing instruments, statistical analysis, regression and all that can be useful to extract knowledge from astronomic data). Therefore, the Virtual Observatory is the result of convergence of research interests and informatics and information technologies.

The application of these methodologies to the discovery and classification of transients (which is the main target of the present work) can be approached from two different points of view: (i) online treatment of data and (ii) offline data analysis. In fact, in some cases it is important to quickly recognize the transient candidates and to perform a rapid follow-up almost in real time, while, in other cases, offline processing may be required to achieve a deeper understanding of the data.

In this work we shall focus on offline classification of variable objects, making use of machine learning approaches, in particular the MLPQNA method ([9] - [8]), and analyzing alternative ones like the random forest method ([21] - [20] - [42]). We will use intensively some statistical methods like the Lomb-Scargle [39], and we shall make extensive use of the Caltech Time Series Characterization Service, a web service devoted to the derivation of photometric features associated with light curves. Most of the work will be performed using the DAMEWARE (Data Mining & Exploration Web Application REsource) infrastructure. The final purpose is to perform a step for a more precise classification based on several methods that in the next future will allow a fully automatized classification of variable objects and transients. In this way, as it has been said before, it shall be possible to

[1] http://www.lsst.org/lsst/

[2] https://www.skatelescope.org/

reach a better comprehension of the known phenomena and to discover new ones yet unknown.

1.1 Time Domain Astronomy: the past

Since the early days, time domain astronomy (hereafter TDA) has enormously grown, including all wavelength ranges and many different parts of astrophysics. In fact, in the history of Astronomy, studies of transient phenomena have always played a key role. In this paragraph we shall outline just a few among the most relevant facts that helped to develop modern TDA. First of all let us introduce the distinction between photometric and astrometric transients.
It is known that astrometry is the branch of Astronomy that involves precise measurements of positions and movements of stars and other celestial bodies. Photometry, instead, concerns with measuring the flux, or intensity of an astronomical object electromagnetic radiation, particularly refering over different wavelength bands of radiation. Therefore, we can define astrometric transients those objects whose variability is due to changes in their positions on the sky. This is the case, for example, of asteroids, comets, etc. Conversely, photometric transients can be defined as those objects whose variability is due to variations in the luminosity of the object caused either by intrinsic or extrinsic phenomena. To the first family belong objects in which the variability is caused by physical variations in its structure which modify also the luminosity flux. It could be the case of supernovae, AGN, cataclismic variables, and so on. Extrinsic variables are instead objects where the variability is induced by other phenomena, such as for instance eclipsing variables.
Modern Astrophysics was born with the first systematic study of a transient. In fact, in 1782 the English amateur astronomer John Goodricke observed the variable star Algol (Beta Persei). We have to recall that, in the ancient era, the static sidereal universe was outside scientific investigation, because it was considered unchangeable. Goodricke noticed the strange variability of Algol[3] and proposed several mechanisms to explain it, as the presence of shape effects (non spherical simmetry) or the passage of a dark body in front of the star. This fact brought to the attention of the scientific community the variability of the universe and we can say that it gave life for the first time to Astrophysics, as a discipline that studies the physical mechanisms and the causes of astronomical phenomena.
Meanwhile, in 1774 Charles Messier had published the *Catalogue des Nebuleuses et des Amas d'Etoiles* nowadays known as the *Messier Catalog*, which

[3]We wish to stress that Algol is a quite bright object clearly visible by naked eye and presents a strong variability. The fact that none before Goodricke seems to have explicitly noted the phenomenon tells a lot about the strenght of the Aristotelic dogma.

can be considered as an involuntary by-product of transient astronomy. In fact, he was a "comet hunter" and he compiled his catalog of nebulae with the aim to better disentangle new comets (astrometric transients) from nebulae (stationary objects).
At the beginning of the 20th century, Henrietta Swan Levitt, one of the *human computers* hired by Edward Charles Pickering at the Harvard College Observatory, by studying variable stars, discovered the Cepheid Period-Luminosity relation. This constituted a key result which enabled the measurement of galactic distances. We will return on this fundamental discovery in the following.
In 1936 Fritz Zwicky and Walter Baade had access to what we now consider the first example of dedicated hardware for transient astronomy: the 18" Schmidt Telescope at the Palomar Observatory (Fig. 1.1). Using this instrument, Zwicky began to workout the first supernovae surveys, and together with Baade, they coined the term "supernova" itself, considered as transitions from normal stars into neutron stars [1]. So they started hunting for supernovae, founding a total of 120 objects. Moreover, Baade proposed the use of supernovae as standard candles, to estimate distance in space. The instrument was also used to discover nearly 50 comets, the most famous of which was the Shoemaker-Levy 9 comet, discovered in 1993, which collided with Jupiter in 1994.
In more recent years the Calan/Tololo Survey was performed, a supernova survey ran from 1989 to 1995 at the University of Chile and the Cerro Tololo Inter-American Observatory to measure a Hubble diagram out to redshifts of 0.1. It led to the discovery of 32 Ia supernovae, which were used as accurate standard candles for measuring distances, bringing to precise measurements of the Hubble Constant H_0 and to the evidence of the accelerated expansion of the Universe and the hypothesis of the presence of *dark energy* or of a cosmological constant dominating the mass/energy of the Universe itself.
Modern transient surveys can offer information only on phenomena which vary significantly on time scales between 1 days and $\approx$ 10 years (ideal for supernovae, but a large portion of the Universe operates at a much slower rate, so we could strongly expand our knowledge if we could extend the time range of our available data) the so called *DASCH project* (Digital Access to a Sky Century @ Harvard) was performed. The aim of this quest is to digitize over 100 years of historical photographic plates at Harvard [24].
Harvard College Observatory was founded in 1839 and soon moved to the forefront of astronomy research, housing the 15-inch "Great Refractor", which resulted to be the largest telescope in the U.S. between 1847 and 1867. In the late 1800s, the observatory began imaging large portions of the sky with telescopes positioned all around the world, and these photographic plates were examined by the already mentioned *human computers* as we previously said when we spoke about the period-luminosity relation for Cepheid. So, Harvard's collection of photographic plates continued to

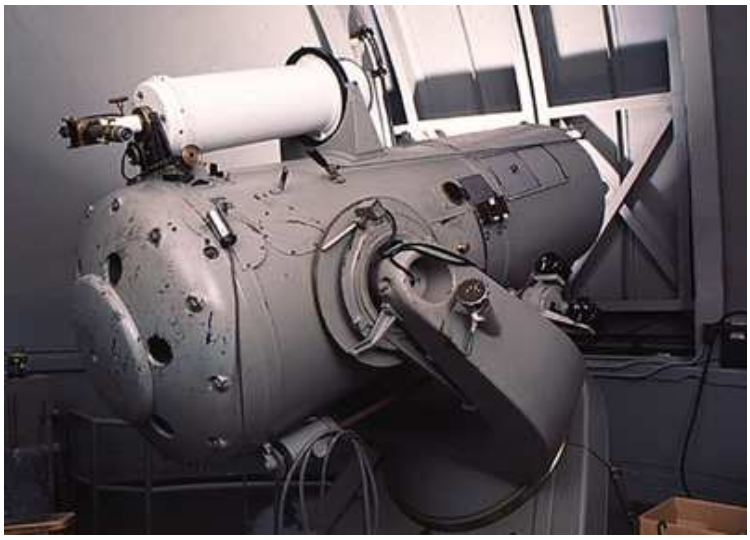


Figure 1.1: The 18" Schmidt telescope at the Palomar Observatory.

grow until the early 1990s, i.e. until when most telescopes had replaced photographic plates with CCDs. Nowadays the archive contains about 500.000 photographic plates, obtained between 1885 and 1993, covering, with different frequency and sampling, the entire sky. Most locations were imaged from hundreds to thousands of times in a 100 year window. Therefore, the project mainly consists in digitizing the plates, detecting sources and measuring their magnitudes, and finally producing the 100-year light curve for every object. The 100-year temporal coverage, compared with < 10 years of coverage by PTF (Palomar Transient Factory[4]) and CRTS (Catalina Real-Time Transient Survey[5]) and the several epochs of SDSS (Sloan Digital Sky Survey[6]), and many other surveys, will enable new studies of long-time scale phenomena, as it can be seen by the comparison in Fig. 1.2.

The overall conclusion is that by expanding TDA surveys to time-scales that are 1 o 2 orders of magnitude longer than those reached by current or planned modern surveys, a range of fundamental classes of objects can be studied as individual objects in well-defined samples[7].

[4] http://www.ptf.caltech.edu/iptf

[5] http://crts.caltech.edu/

[6] http://www.sdss.org/

[7] One of the purposes is to create a historical knowledge (*Historical TDA*), taking a step back and looking to the past, also in the optics of the incoming new projects previously presented. This also gives the idea of the always increasing interest of the astronomical community in the wide field of TDA.

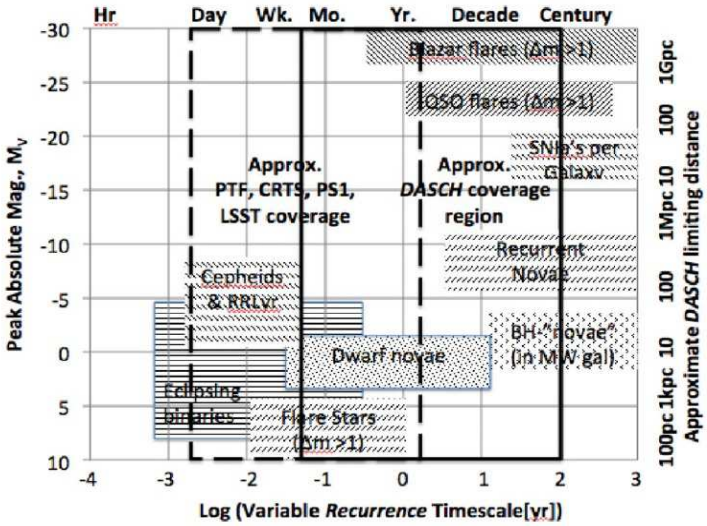


Figure 1.2: Representative classes of variables and transients vs their recurrence time that can be measured for a complete sample with DASCH (right) vs PTF, CRTS, Pan-STARRS-1 and LSST (dashed box, left) or jointly (overlap region).

1.2 Time Domain Astronomy: the present

TDA is opening a totally new discovery space, extending to the time axis the Observable Parameter Space (or OPS). In general the parameter space is defined as the set of all possible combinations of values for all the different parameters contained in a particular mathematical or physical model. So different configurations of the parameters space produce different behaviors of the model. In astrophysics, the set of the parameters is usually obtained from photometric or spectroscopic observables, and from statistical patterns. It is known from the history of science and from literature that every time a technology enables us to open a new portion of the OPS, new types of objects and phenomena are usually discovered. Therefore, adding the temporal dimension to the parameter space has allowed and will allow the discovery of new phenomena and a better characterization of the old ones, with a major comprehension of some physical phenomena (Fig. 1.3).

At the present time, the overall description that emerges is the one depicted by the semantic tree of Fig. 1.4, from which a first classification of variable objects in extrinsic and intrinsic ones can be deduced, as previously explained. As we shall see in more details in the next chapter, extrinsic objects can be asteroids or eclipsed, microlensed and rotating stars, while intrinsic objects are eruptive, cataclysmic, pulsating and secular stars, or AGN.

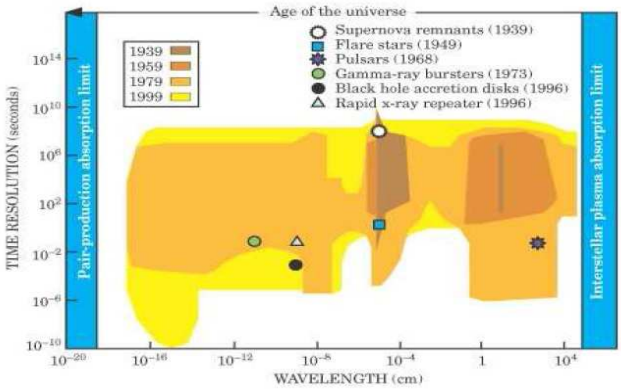


Figure 1.3: The plot, from Harwit [25], shows how our knowledge of the parameter space has increased trough the years, both in wavelength and time resolution of the phenomena. It is interesting to notice that new phenomena (marked with different symbols) are always at the edges of the colored areas, making clear that they were a result of a new technology, opening new windows in the OPS.

Therefore, TDA allows to tackle a broad range of different physical phenomena. In fact, we have to consider that some phenomena can be studied only in the time domain, for example various cosmic explosions, accretion and relativistic phenomena. We can safely state that TDA regards essentially every field of astronomy, from the Solar System to cosmology, and from stellar structure and evolution to extreme relativistic phenomena.

It is needed to emphasize that the data and event discovery rates are expected to increase dramatically, from 0.1 TB and $\approx 10-10^2$ events per night now, to 30 TB and $10^5 - 10^7$events per night in the LSST era, and that the available follow-up facilities would be simply overwhelmed, and will result absolutely unable to react to all potentially interesting events. The traditional manual approach will simply not scale to the next generation of surveys, especially if we are interested in finding the rarer transients. So, the main challenge is to achieve the dynamical, real-time characterization and classification of transient events, and the subsequent optimal decision for their follow-up.

In Fig. 1.5 it is possible to see an example of how such coordination works, for a single event which was observed in the Crab Nebula. This episode illustrates brilliantly how the availability of instruments that survey large areas of the sky, combined with the ability to process the data in real time, has opened new perspectives in TDA.

Moreover, not only electromagnetic signals are involved, if we consider that neutrino and cosmic ray astronomy are ready to explode and gravitational wave astronomy is at its first steps.The community is growing toward this

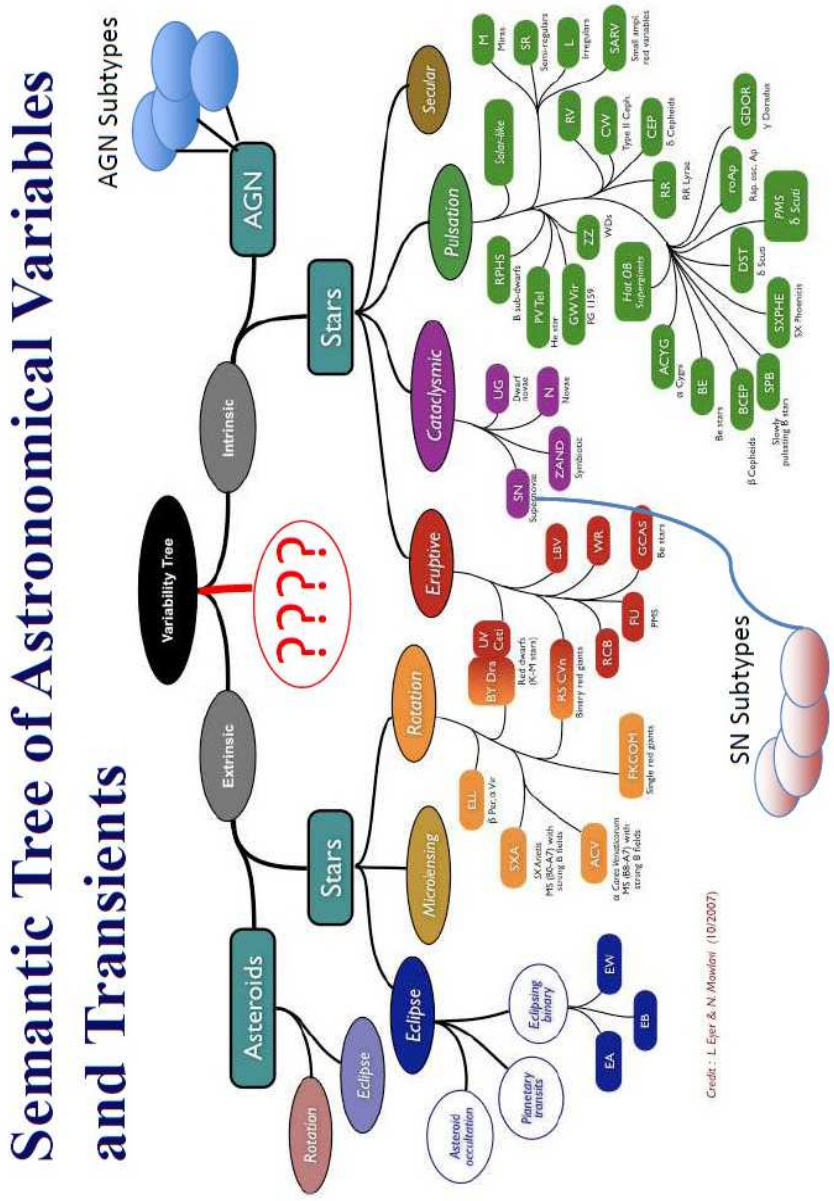


Figure 1.4: Semantic tree of astronomical variables and transients (see Eyer [23]).

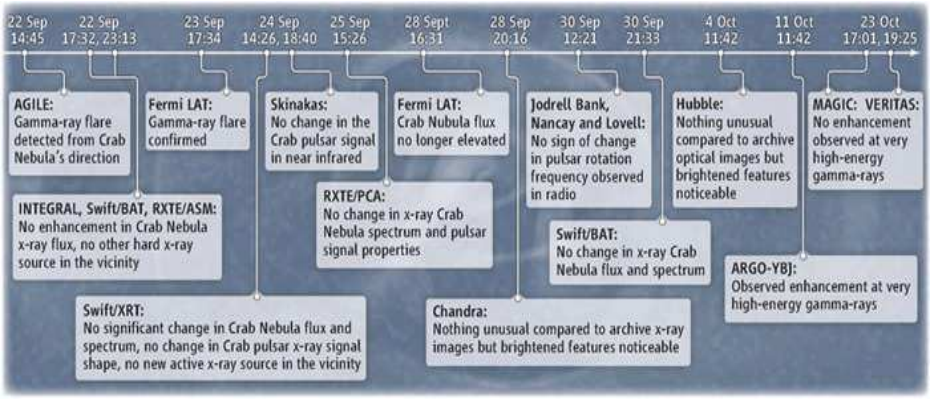


Figure 1.5: Timetable of the Astronomer's Telegram releases on a Crab Nebula flare in universal time, within 1 month after discovery on 22 September 2010 (see [2]).

kind of "multimessenger astronomy". But now it is clear that the huge volume of data to be searched for transients and the multitude of possible decisions to be taken will soon make it impossible to rely on human capabilities to rapidly collect and discriminate time-critical information. Efforts are therefore being put into developing common standards for the implementation of fully automated near real-time systems.
The study of the presented phenomenology implies two different operational modes:

- Offline TDA: understanding of the variable universe from the huge amounts of light curves produced by modern surveys and stored in the digital archives.
- Online TDA: detecting and characterizing in real time photometric transients.

Chapter 2

Phenomenology of transients

In this chapter we shall provide the reader with more details about the classification based on the semantic tree of Fig. 1.4, presented in the previous chapter and we shall furnish a description of the main phenomena and physical process regarding the objects mainly involved in the development of this thesis work.

2.1 The semantic-tree based classification

First of all one must say that it is possible to distinguish transients from simple variable objects using the definition taken from CRTS: transient objects are those which show a magnitude variability of $\Delta m > 2$ mag.
Looking at the semantic tree (Fig. 1.4), the first obvious division is, as it was already said previously, between astrometric and extrinsic ones and photometric or intrinsic ones. We recall that astrometric transients are those phenomena that show a variability induced by variations of their position in the sky with time, so it is not connected to physical properties of the objects. Instead photometric transients consist in those phenomena that owe their variability to a real change of the luminosity of the object itself, caused by intrinsic variations of its physical state and/or parameters.

2.1.1 Astrometric and extrinsic transients

Astrometric and extrinsic transients can be then divided mainly in two categories.

- Asteroids: the causes of their extrinsic variability can be identified in rotational or eclipsing processes.
- Eclipsing, rotating and microlensed stars: In an eclipsing system a star can change its brightness due to an asteroid occultation, to a planetary transit or to the interaction with another star. In the latter

case we speak of Eclipsing Binaries. These systems are formed by physically bound stars, having an orbital plane which lies near the line-of-sight of the observer. The components periodically eclipse each other, causing a decrease in the apparent brightness of the system, with the period of the eclipse that can range from minutes to years. In particular, the case of *planetary transit* underlies for the search of extrasolar planets. This is one of the most active and intriguing field of the modern astrophysical research, and it is performed mainly with the methods of TDA. Rotating stars, instead, show small changes in light that may be due to dark or bright spots on the stellar surfaces. Finally, microlensing is a phenomenon due to the gravitational lens effect, that can be used to detect objects ranging from the mass of a planet to the mass of a star, if obscured by another massive objects, as in the usual lensing phenomenon for galaxies. Microlensing phenomena can be monitored over time through the detection of their light curves.

2.1.2 Intrinsic transients

Intrinsic transients can again be divided in the two major subclasses.

- Variable stars: for what concern stars, we can consider the subcategories of eruptive, cataclysmic and pulsating variables, depending on which phenomenon is at the origin of their variability, and stars displaying a secular evolution, which are usually stars in the post-AGB (Asymptotical Giant Branch) of the H-R diagram (Hertzsprung-Russell [26] - [37]). The entire work described in this thesis in entirely based on intrinsic transients, so in the following paragraphs we will describe these classes in much more detail.

- Galaxies: in the specific, galaxies that show marked variability phenomena are classified as AGN (Active Galactic Nuclei).

Specifically, for what concern stars:

- Eruptive variables: these stars suffer very large variations in brightness due to violent processes and flares occurring in their chromospheres and coronae. The light changes are often accompanied by shell events or mass outflow in the form of stellar winds of variable intensity and by interaction with the surrounding interstellar medium. We recall, as example of eruptive variables, the Wolf-Rayet and the R Coronae Borealis stars. R Coronae Borealis variables are luminous, hydrogen-poor, carbon-rich, supergiant star which spend most of their life time at maximum light, occasionally fading even by nine magnitudes at irregular intervals. Wolf-Rayet stars are very luminous hot Population I stars with effective temperatures between 30000 and 50000 K. They

are characterized by very high mass-loss rate ($\approx 10^{-5}\ M_{\odot} yr^{-1}$). They show light variations with amplitudes of several hundredths of a magnitude and time scales from milliseconds to years. Therefore, eruptive stars are substantially evolved stars that have left the main sequence and are proceeding step by step towards the last phases of their life.

- Cataclysmic variables: are usually close binary systems in which the most massive component is a white dwarf and the companion a main sequence star. In most cases mass is transferred from the companion to the white dwarf through a surrounding accretion disk. This accreted material feeds various types of phenomena, including occasional eruptions and jets. Components of this class of objects are:

  - Novae: these systems are constituted by a white dwarf and a main-sequence low mass star that has filled his Roche lobe. A classical nova can show an increase of brightness from 7 to 15 magnitude in a range of 1 to several hundred days.
  - Dwarf Novae: these systems are constituted by a white dwarf and a red dwarf star cooler of our Sun. They experience semi-regular outbursts with a typical timescale ranging from weeks to years and a range of 4-5 magnitudes.
  - Symbiotic Stars: these are interacting binary systems composed of an evolved red giant and a hot companion star that could be a main sequence star, a white dwarf, or a neutron star. Most symbiotics have orbital periods of a few years while other orbit over several decades.

  But the most famous type of cataclysmic variables of course remain the Supernovae, to which we shall dedicate the next paragraph, due to their importance in our work.

- Pulsating variables: stars characterized by periodic variations of its luminosity. Stellar pulsations can be radial, if the expansion has spherical symmetry, or non-radial, and in this case the shapes of the stars can be asymmetrically distorted. Pulsations can occur at various frequencies, with the lowest allowed frequency called fundamental mode, and the higher frequencies called overtones. For each oscillation mode, these waves have at least one node, where the matter remains steady, at the center of the star and an antinode, where the velocity of the gases is maximum, at the surface.

  The principal categories of pulsating stars are observed to lay in the so called Instability Strip (see Fig. 2.3), a nearly vertical region of the H-R diagram, which defines a range of luminosities, colors and periods, over which pulsation is a stable mode for the star.

We shall analyze the theory of pulsation in more detail in a subsequent paragraph.

An important thing to be noticed is that, in this schema, there are some points of contact between the two great categories of intrinsic and extrinsic transients. In fact, some types of stars that show eruptive phenomena, could have also an extrinsic variability due to rotational effects.

2.2 Supernovae

With the term supernova it is intended the catastrophic explosion occurring in the last stages of the life of a massive star, which is capable to eject a mass of $\approx 10 - 100\ M_{\odot}$, with velocities of about $0.01 - 0.1$ c. The explosion commonly feeds the external environment and the interstellar medium with the heavy elements that were produced in the interior of the star. The burst of radiation in a supernova often briefly outshines the luminosity of the entire host galaxy, before fading from view over several weeks or months.
Supernovae are, without any doubt, among the most spectacular celestial objects ever observed by humans and for sure one of the most energetic phenomena in the Universe. The oldest known supernova was the one observed in 185 AD. Supernovae in 386 and 393 AD are recorded only in Chinese reports with no precise information about their positions. The brightest Supernova ever seen was the one exploded in 1006 AD, which reached a visual magnitude of -7.5 mag. It was described by observers in China, Egypt, Iraq, Japan, Switzerland. However, the most famous supernova is probably the one seen in 1054, which produced an expanding shell of gas and dust today known as the Crab Nebula (see Fig. 2.1). This SN shone brighter than Venus and remained visible for 23 days also during daylight. Another supernova was observed in the 1181 AD by Chinese and Japanese astronomers in the constellation of Cassiopeia. In the same constellation, another famous supernova was observed by the Danish astronomer Tycho Brahe in the 1572 AD, constituting the basis for most of his successive research. Finally, the last confirmed supernova exploded in the Milky Way was the one observed by Kepler in 1604.
For what concern the previous listed supernovae, all of them have left behind the so-called Supernova remnants, and because no supernova has been observed in our Galaxy during the telescopic era, everything we know about these phenomena comes from these remnant and from supernovae in other galaxies.
The features of the optical spectra at maximum light and the characteristics of light curves define the various categories of supernovae. The first division was performed by R. Minkowski in 1941 [32], who defined two main categories, type I and type II. The former differ from the latter for the lack of hydrogen emission line, H, in type I. Type I Supernovae have then been

Figure 2.1: The Crab Nebula resulting from the explosion of the Supernova 1054. In its center there is the so-called Crab Pulsar, a neutron star of about 10 km of diameter.

subdivided in three further classes: type Ia, Ib and Ic, depending on their spectral characteristics. The first one shows the absorption line of the Si IIλ6355 (however we shall see next that Ia Supernovae are originated by a completely different process). Ib show the absorption line of He Iλ5876 together with Calcium and Oxygen emission lines, while Ic do not show any of the previous absorption lines. Type II Supernovae are also divided in type II-L (linear) and type II-P (plateau), depending from the shape of the resulting light curve after the explosion, which can respectively present a steady decline or a slower decline followed by a normal decay.
Type Ia Supernovae were found in all kind of galaxies, ellipticals, spirals and irregulars, and this is an evidence of the fact that their progenitors must be long-lived stars, because in ellipticals there is no ongoing stellar formation. They show the presence of characteristic elements in their spectrum, such as magnesium, silicon, sulphur, calcium and iron. Type Ib, Ic and II instead, seem to explode respectively in stellar formation zones of the arms of spiral galaxies and in H II region of spiral discs or in irregular galaxies, thus indicating that their progenitors must be short-lived, hence massive, stars.

2.2.1 Core Collapse Supernovae

According to what has been said in the previous paragraph, we can now understand that type Ib, Ic and II have a common origin as Core Collapse Supernovae, while type Ia must be considered as completely different phe-

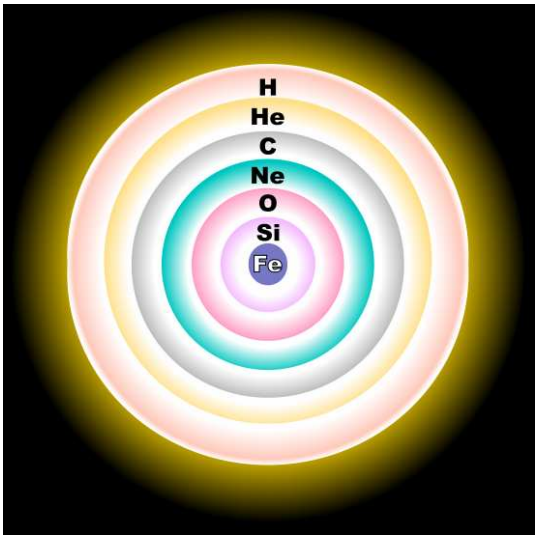


Figure 2.2: Shell structure of the interior of an evolved star that will undergo a supernova explosion.

nomena. In fact, the former ones are originated from the collapse of a massive, evolved stellar core. In particular, type II Supernovae must be stars with masses between $8 - 40\ M_{\odot}$. Instead, stars with a bigger initial mass, like Wolf-Rayet, loose their envelopes bringing to Ib and Ic Supernovae.
These stars pass through the burning phases of hydrogen, helium, carbon, neon, oxygen, and silicon, finally producing an iron core (Fig. 2.2). At this point, because the nuclear binding energy per nucleon has its maximum for iron, no energy can be released by nuclear fusion of this element. Due to the process known as *photodisintegration*, photons at the very high temperatures present in the iron core are capable to destroy heavy nuclei. Meanwhile, the free electrons that contribute to support the star through the electron degeneracy pressure, in these critical conditions, are captured by heavy elements and by protons produced through photodisintegration. Then the core starts to collapse.
The collapse is halted by the repulsive component of the strong nuclear force, when the core has reached about twice the density of atomic nuclei, $\approx 4 - 5 \times 10^{14}\ g/cm^3$. But the sudden halt of the collapsing core produce a rebound mechanism, and shock waves form, directed toward the surface of the star. The shock waves, together with the enormous force generated by neutrinos, which at the opacity caused by the impressive pressure cannot escape as usually, propagate through the still collapsing layers of the star, leading to the supernova explosion. A huge amount of energy is released and the outer layers, containing heavy metals, together with the remaining outer envelope of hydrogen, are expelled.

2.2.2 Type Ia Supernovae

Regarding type Ia Supernovae, there are still uncertainties about the process that originates these kind of phenomena. The most accepted hypothesis is that the formation of these supernovae happens in binary systems constituted by a carbon-oxygen white dwarf and an evolved star. The cause of the explosion can be found in the accreting material on the white dwarf from the companion star, during its red giant phase, until the white dwarf itself reaches its Chandrasekhar limit. At this point, in the most accredited models, the degeneracy pressure is no longer able to support the star against gravity, and the star starts to contract, soon reaching pressure and temperature conditions sufficient to ignite carbon fusion. What happens next is not well understood, but probably the shock waves produced by the explosion ignites a deflagration that completely disrupt the star, without leaving any remnant. Part of the material becomes ^{56}Ni and the remaining lighter elements like Si and C.
The typical light curve can be divided in four phases, all explainable considering the energy released in the decay from ^{56}Ni to ^{56}Fe. We can identify:

- rise time: the period in which the supernova rises very fast to its maximum;
- maximum phase;
- second maximum: a pronounced second maximum has been observed in redder light curves about from 20 to 40 days after the first maximum;
- late decline: about after 50 days the light curves reaches a steady decline phase, exponential in luminosity.

Ia Supernovae reach their maximum about 2 or 3 weeks after the explosion, are brighter of one magnitude then the type II and all of them have the same peak luminosity. For these reasons they can be good standard candles, and if it would possible to measure the absolute magnitude of the supernova, regardless its distance, we could obtain a measure of the Hubble constant. In fact, in 1993, Phillips [34] discovered a linear relationship between the decline rate parameter of the light curve, Δm_{15} (the difference between the magnitude at maximum light and the magnitude after fifteen days), and the absolute peak magnitude of the supernova. This correlation makes possible to greatly improve the precision of distance estimation of Ia Supernovae, using them for the determination of cosmological parameters.

2.3 Pulsating variables and theory of pulsation

The theory of radial stellar pulsation is based on the assumption that this is generated by small perturbations around the hydrodynamical equilibrium

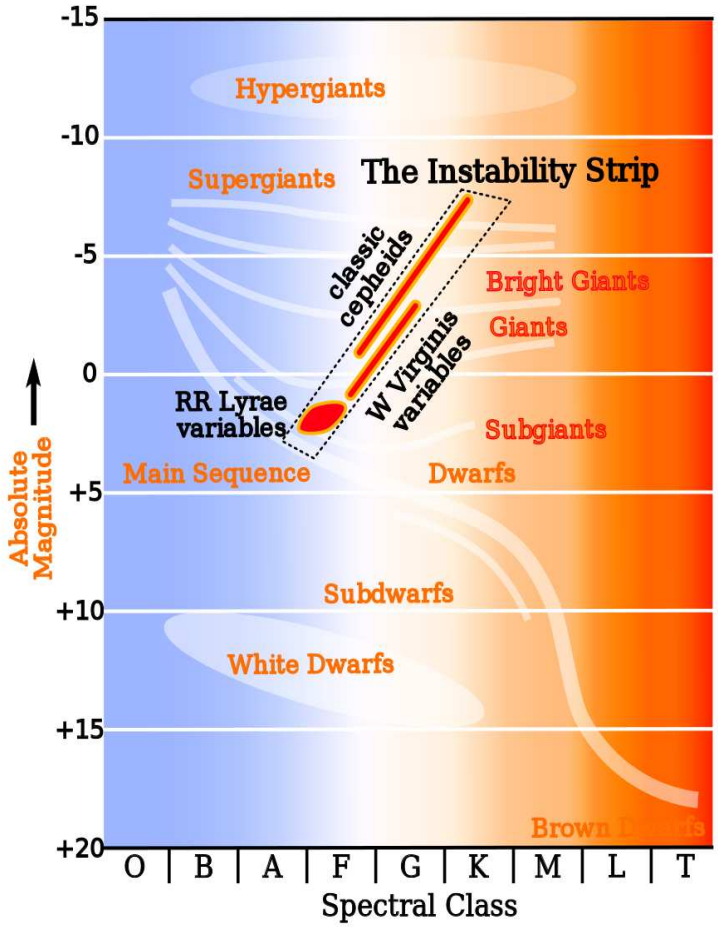


Figure 2.3: Position of some pulsating variables and of the Instability Strip in the H-R diagram

state (during this phase the star places on the Instability Strip, see Fig. 2.3), which can grow to observed amplitudes (linear stability analysis of stellar structure). The first step is to consider the stellar structure equations:

$$\frac{\partial^2 r}{\partial t^2} = -\frac{GM_r}{r^2} - \frac{\partial P}{\partial r} \tag{2.1}$$

$$\frac{\partial r}{\partial M_r} = \frac{1}{4\pi r^2 \rho} \tag{2.2}$$

$$\frac{\partial E}{\partial t} - \frac{P}{\rho^2}\frac{\partial \rho}{\partial t} = \epsilon - \frac{\partial L}{\partial M_r} \tag{2.3}$$

$$L_r = -4\pi r^2 \frac{4ac}{3}\frac{T^3}{\kappa\rho}\frac{\partial T}{\partial r} = -\frac{64\pi^2 ac}{3} r^4 \frac{T^3}{\kappa}\frac{\partial T}{\partial M_r} \tag{2.4}$$

In these equations r is the distance from the center of the star, M_r represents the mass at radius r, L is the luminosity, T the temperature and P the pressure. The energy density ϵ and the opacity κ are functions of the density ρ and the temperature T, and if we consider an equilibrium state, the previous equations become:

$$\frac{\partial P_0}{\partial r_0} = \frac{GM_r}{r_0^4} \tag{2.5}$$

$$\frac{\partial r_0}{\partial M_r} = \frac{1}{4\pi r_0^2 \rho_0} \tag{2.6}$$

$$\frac{\partial L_{r0}}{\partial M_r} = \epsilon_0 \tag{2.7}$$

Therefore, to solve the problem of stellar pulsation, the variables considered can be expressed in terms of an equilibrium quantity and a small perturbation: $r \to r_0 + \delta r$, $P \to P_0 + \delta P$, $\rho \to \rho_0 + \delta\rho$, $L \to L_0 + \delta L$. Putting $\zeta = \delta r / r_0$, so that $r = r_0(1+\zeta)$, we can furthermore write a generic Lagrangian quantity f, as $f = f_0(1 + \delta f / f_0)$. We will proceed assuming that, in case of small perturbations, $|\zeta| << 1$ and $|\delta f / f_0| << 1$ and neglecting all the terms of from second order. With these assumptions, the equations 2.1 - 2.4 can be reduced to a single equation in ζ:

$$\begin{aligned}\frac{\partial^2 \zeta}{\partial t^2} = -\frac{1}{r\rho}\frac{\partial \zeta}{\partial t}\frac{d}{dr}\left[(3\Gamma_1 - 4)P\right] - \left(\frac{1}{\rho r^4}\right)\frac{\partial}{\partial r}\left(\Gamma_1 P r^4 \frac{\partial \zeta}{\partial r}\right) = \\ = \frac{1}{r\rho}\frac{\partial}{\partial r}\left[\rho(\Gamma_3 - 1)\delta\left(\epsilon - \frac{\partial L_r}{\partial M_r}\right)\right]\end{aligned} \tag{2.8}$$

where

$$\Gamma_1 = (d\ln P/d\ln\rho)_{ad} \quad \text{and} \quad \Gamma_3 = (d\ln T/d\ln\rho)_{ad}$$

are the adiabatic exponents of pressure and temperature. Considering only solutions of the form:

$$\zeta(r,t) = \xi(r)e^{i\omega t} \tag{2.9}$$

where $\xi(r)$ is a complex function of the only spatial variable and ω is a frequency. Therefore, in the case of adiabatic oscillations, from Eq. 2.8 we obtain:

$$-\frac{1}{r^4\rho}\frac{d}{dr}\left[\Gamma_1 P r^4 \frac{d\xi}{dr}\right] - \frac{1}{r\rho}\left\{\frac{d}{dr}\left[(3\Gamma_1 - 4)P\right]\right\}\xi = \omega^2\xi \tag{2.10}$$

This is an eigenvalue equation which admits discrete solution characterized by eigenfunctions ζ_k, where every k is a node with $\zeta_k = 0$, and eigenvalues ω_k. ω_0 is the fundamental mode, while the other frequencies are the overtones. Obviously, the solution of such equation requires special conditions at the center and at the surface of the star.
The driving mechanism which sustains the pulsation must be found in the opacity of the star. It was suggested by Eddington [22] that certain layers of the star, during the compression phase due to pulsation, might become quite opaque to radiation. But the increase of the opacity generates an accumulation of heat under these layers, which brings to an increase of pressure and an expansion of the star. At this point, there is a new decrease of opacity and pressure, the star contracts again and a new cycle begins. In 1980, J.P. Cox [14] found that the mechanism proposed by Eddington can successfully operate in the partially ionization zones of the pulsating star.

2.3.1 Types of pulsating variables

The parameters that permits to distinguish between the various types of pulsating variables are the pulsation period, mass and evolutionary status of the star, besides the characteristics of the pulsation itself.

- RR Lyrae stars: short period (1 hour to 30 hours), pulsating, blue giant stars, usually of spectral class A. The amplitude of variation is usually from 0.3 to 2 magnitudes.
- δ Scuti: their variations in luminosity are due to both radial and non radial pulsations of their surface. Fluctuations in brightness are comprised between 0.003 and 0.9 magnitudes in V, over a period of a few hours. They can be A0 to F5 type giant or main sequence stars.

- RV Tauri: yellow supergiants with characteristic light variation which alternates deep and shallow minima. The period between two deep minima ranges usually between 30 to 150 days and the variation in magnitude can be up to 3. Some of these stars show also long-term cyclic variations from hundreds to thousands of days. The spectral class often ranges from G to K.

- Pulsating white dwarf: their luminosity variations are due to non radial gravity wave pulsations. The variations are small (1% - 30%) and the periods are comprised from hundreds to thousands of seconds.

- Long period variables: pulsating red giants or supergiants in which variations occur over long timescales of months or years. We can distinguish the two major subclasses of Mira and Semiregular variables.

- Irregular variable stars: red supergiants white little or no periodicity at all.

But the most famous example of pulsating stars remain Cepheid variables. These are massive stars, with spectral type that can change during pulsation, from F at maximum luminosity to G or K at minimum. Pulsation is mainly radial. It is possible to identify four classes of Cepheid variables:

- Classical Cepheids: also called type I Cepheids, fundamental mode pulsators with periods that vary from 1 to 70 days.

- Beat Cepheids: they display the presence of two or more simultaneously operating pulsation modes, generally the fundamental and the first overtone, with periods between 2 and 7 days.

- S Cepheids: probably first-overtone pulsators, with periods in the same range of Beat Cepheids.

- W Virginis: population II Cepheids, they are fundamental mode pulsators with periods between 1 and 30 days.

Cepheids exhibit strong correlations between their periods, luminosity and colors, but not for amplitudes, which do not seem to correlate with other observables. In the next paragraph we will analyze this in more detail.

2.3.2 Period-Luminosity relation

In 1912 Henrietta Swan Leavitt, an American astronomer and *human computer* of Edward Pickering at the Harvard Observatory, discovered a linear correlation between the apparent magnitude and the logarithm of the period for a sample of stars, in the specific Classical Cepheids, in the Large

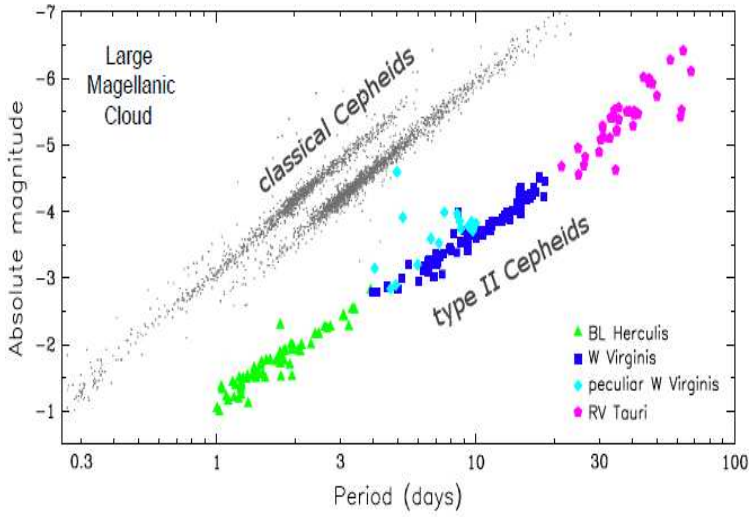


Figure 2.4: Period-luminosity relations of classical (grey points) and type II Cepheids (color symbols) in the Large Magellanic Cloud, as taken by OGLE. As the "luminosity" the reddening-free Wesenheit index [13] was used, defined as $WI = I - 1.55(V - I) - DM$, where I and V are mean luminosities of Cepheids in these passbands, and $DM = 18.5$ mag is the distance modulus of the Large Magellanic Cloud.

Magellanic Cloud (LMC). However, the relation is valid also for the absolute magnitude, because all the stars of the LMC can be considered at the same distance. The relation discovered by Leavitt was called the "Period-Luminosity relation", and can be expressed as:

$$M = a + b * \log_{10} P \tag{2.11}$$

An example of the Period-Luminosity relation is reported in Fig. 2.4. Once it has been properly calibrated, this relation allows us to derive, from the measured period of a Cepheid, its absolute magnitude and so its distance module. Obviously, one has to take into account the effects of interstellar reddening, which will produce systematic errors that could be reported into the distance scale.

2.4 Active Galactic Nuclei

Galaxies hosting Active Galactic Nuclei (Fig. 2.5), that contain all AGN subclasses such as Blazars, Seyfert Galaxies, Quasars and so on, are also

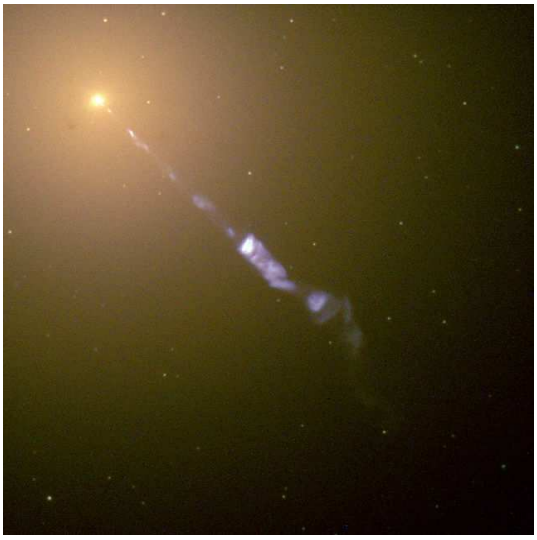


Figure 2.5: The active radiogalaxy M87 as seen by Hubble Space Telescope, with its characteristic 5000-light-year-long jet. It is thought to be produced by the synchrotron radiation of the particles accelerated from the central engine.

usually variable. AGN, however, are very particular variables. In fact they emit strongly over a wide range of wavelengths, from X-ray to radio. Many AGN vary in brightness by huge amounts over relatively short timescales, such as months, days, or even hours. AGN are conveniently divided in two main classes, radio-loud and radio-quiet, depending on whether or not they emit in the radio portion of the electromagnetic spectrum respectively.
Nowadays, the different types of AGN and their physical properties have found explanation in a unified model that bases the activity of these objects on a central engine constituted by a supermassive black hole on which the dynamical e termodynamical properties of the entire galaxy are based. It results evident that the strong emission coming from AGN could be explained only considering accretion onto a supermassive black hole (in the range of $10^6 - 10^{10}\ M_\odot$). In fact, we must remember that gravitational accretion is the most efficient known way of using mass to get energy, much more efficient than nuclear fusion.
The unified model proposes that different types of AGN are a single type of physical object observed under different conditions, as showed in Fig. 2.6. The currently accepted idea is that this models are "orientation-based unified models", meaning that the apparent differences between the various types of objects arise simply because of their different orientations to the observer. Moreover, it has been proposed that, confirmed the presence of a supermassive black hole in the nucleus of almost all galaxies, the AGN phase

is just a step in the evolutionary history of a galaxy.
However, once fixed the division in radio-quiet and radio-loud AGN, it is possible to identify the following subcategories.

Radio-quiet AGN

- Low-ionization nuclear emission-line regions (LINERs): weak nuclear emission-line regions. It is still debated if the are truly AGN.
- Seyfert galaxies: these objects show optical range nuclear continuum emission, narrow and occasionally broad emission lines, occasionally strong nuclear X-ray emission and sometimes a weak small-scale radio jet. They are divided into two types known as Seyfert 1 and 2: Seyfert 1 show strong broad emission lines while Seyfert 2 do not, and Seyfert 1 are more likely to show strong low-energy X-ray emission. The host galaxies of Seyferts are usually spiral or irregular galaxies.
- Radio-quiet quasars/QSOs: characterized by a very high redshift, quasars were originally "quasi-stellar" in optical images as they had optical luminosities that were greater than that of their host galaxy. They show strong optical continuum emission, broad and narrow emission lines, and strong X-ray continuum emission. The host galaxies of quasars can be spirals, irregulars or ellipticals.

Radio-loud AGN

- Radio-loud quasars: they behave exactly like radio-quiet quasars, with the addition of emission from a jet. Thus, they show strong optical continuum emission, broad and narrow emission lines, and strong X-ray emission, together with nuclear and often extended radio emission.
- Blazars, i.e. BL Lac objects and OVV (optical violent variable) quasars: their variable emission is believed to originate in a relativistic jet oriented close to the line of sight. Both classes are distinguished by rapidly variable, polarized optical, radio and X-ray emission. BL Lac objects show no optical emission lines, broad or narrow, so that their redshifts can only be determined from features in the spectra of their host galaxies. The emission-line features may be intrinsically absent or simply swamped by the additional variable component. OVV quasars behave more like standard radio-loud quasars with the addition of a rapidly variable component.
- Radio galaxies: these objects show nuclear and extended radio emission. Their other AGN properties are heterogeneous, but their host galaxies, whatever their emission-line type, are essentially always ellipticals.

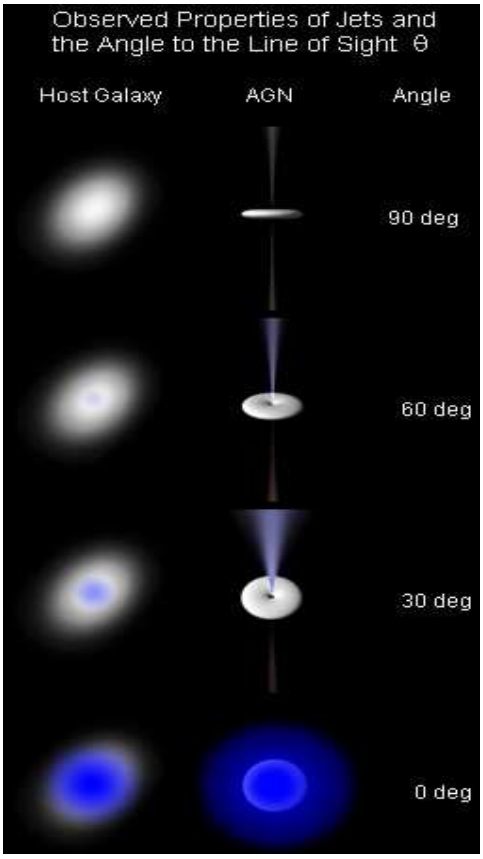


Figure 2.6: Unification by viewing angle. From bottom to top: down the jet - Blazar, at an angle to the jet - Quasar/Seyfert 1 Galaxy, at 90 degrees from the jet - Radio galaxy / Seyfert 2 Galaxy.

Chapter 3

Automated classification of transients

For what classification is concerned, the main aspect that must be taken into account is that nowadays data volumes have begun to surpass what is possible to visually inspect by even large teams of astronomers and volunteer *citizen scientists*. This implies an increasingly more central role of software and hardware frameworks to substitute humans in the real-time loop. Data need to be automatically transported, processed, photometered and inserted into databases almost without human intervention.
Of course, autonomous discovery of transients and variables is a big challenge. Threshold cuts in photometric quality, changes in apparent magnitudes, matched filtering, etc., can be very effective tools to discover new events, but other types of variables and transients could be not easily recovered from these kinds of approaches. Furthermore, previous machine-learning based discovery have been optimized on domain-specific discovery, leaving apart the multitude of other variables not of direct interest for a particular project.
The challenge is to conflate the process of discovery with classification, using different machineries and methods working on the same problem with various approaches. In this view, the advantages of a computational approach, rather than human-centric, become clear:

- machines, properly trained, are faster than humans both in discovery and classification of candidates/events; at least in theory they allow for operations at arbitrarily high data rates;

- more efficient use of follow-up (e.g. spectroscopic, photometric, etc.) facilities;

- experimentations with new discovery and classification schema require little more than re-running new codes on existing data;

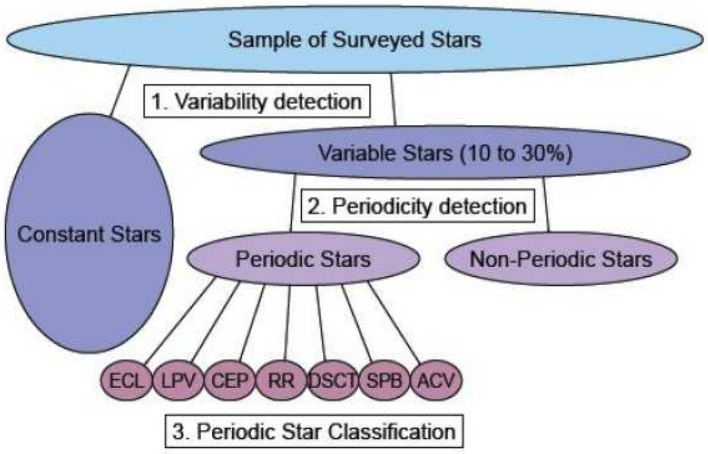


Figure 3.1: Scheme of classification for variable and periodic stars by Dubath [21]

- machine-learned classification is reproducible and very often deterministic;
- the reproducibility allows for calibration of the uncertainties of classification probability statements, giving assurances that classifications are robust as the survey proceeds.

In this framework, there may still be a vital role for humans in the real-time loop, in the case of ambiguous classifications or uncertain follow-up paths for a particular source, but the main idea is that the whole process must not be guided by humans.

3.1 Periodic objects classification

As it has already been mentioned, this thesis focuses on offline classification and therefore real-time issues are not crucial. The study of their periodicity represents the baseline for a deeper analysis of transients. The traditional and most logic approach consists in three main steps (see Dubath [21]). During the first one it tries to separate variable objects from the ones that do not show variability. Then the second part the method considers the periodicity of the objects and measures their period. Finally, in the third part, one can proceed to the classification of the periodic objects (see Fig. 3.1 - 3.2 for the scheme).

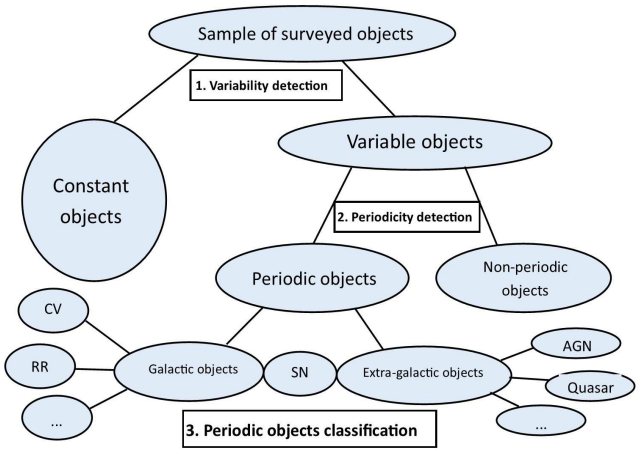


Figure 3.2: Modified scheme for a most general classification of variable and periodic objects.

To determine the variability in a certain data sample, it is possible to use many criteria. We analyze the one due to Stetson [42]. This criterion employs an index used to determine the probability that a given object presents a certain variability degree, so determining the *p-value* of the distribution. We recall that the p-value is defined as the probability, under the assumption of a certain hypothesis, of obtaining a result equal or more extreme than what was actually observed (Fig. 3.3). The index provides the principal measure of confidence that the variability is real, and not due to noise. In fact noise can be confused for a variable source if not correctly dealt with.
These are two expressions of the Stetson index:

$$I = \sqrt{\frac{1}{n(n-1)}} \sum \left(\frac{b_i - \overline{b}}{\sigma_{b,i}}\right)\left(\frac{v_i - \overline{v}}{\sigma_{v,i}}\right) \tag{3.1}$$

$$J = \frac{\sum w_k sgn(P_k)\sqrt{|P_k|}}{\sum w_k} \tag{3.2}$$

where:

$$P_k = \delta_{i(k)}\delta_{j(k)} \tag{3.3}$$

$$\delta = \sqrt{\frac{n}{n-1}}\frac{v - \overline{v}}{\sigma_v} \tag{3.4}$$

In this expressions b_i and v_i are the apparent magnitudes obtained for the candidate object in two observations closely spaced in time, $\sigma_{b,i}$ and $\sigma_{v,i}$

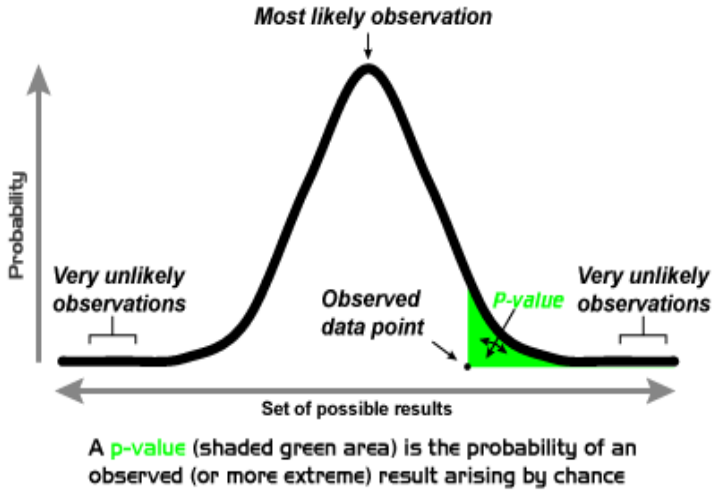


Figure 3.3: Example of a p-value computation. The p-value is the area in green under the curve, past the observed data point.

are the standard errors of those magnitudes, $\overline{b}$ and $\overline{v}$ are the weighted mean magnitudes in the two filters, and n is the number of observation pairs. w_k instead is a weight and δ is a magnitude residual of a given observation from the average of all observations in the same bandpass, scaled by the standard error.
The J value is a more robust version of the same index, which, combined with the distribution kurtosis, gives:

$$K = \frac{1/N \sum |\delta_i|}{\sqrt{1/N \sum {\delta_i}^2}} \tag{3.5}$$

where the index i runs over all N observations available without regard to pairing. Then one can show that, in the limit where the total range of variation is vastly larger than the σ's of the individual observations, and for a Gaussian magnitude distribution $K \to \sqrt{2/\pi} = 0.798$. Hence the final version of the index can be written as:

$$L = (\frac{JK}{0.798})(\frac{\sum w}{w_{all}}) \tag{3.6}$$

where the factor $\sum w/w_{all}$, with w_{all} being the total weights an object would have if successfully measured in all frame pairs, takes into account possible problems of detection of the same object if it results to be absent from one or more frames. In this way, those candidates that were successfully measured the most times will also be the first to be followed up. A value of L can be

determined for every object in the field having some minimum total weight, and stars exceeding some threshold value of L may be subjected to period searches and light curve fits.
After the potentially variable objects have been identified, the second step requires to disentangle periodic from non periodic objects. One possibility is to evaluate the periodogram function through the Lomb-Scargle method [39]:

$$P_x(\omega) = \frac{1}{2}\left(\frac{[\sum X_j \cos\omega(t_j-\tau)]^2}{\sum \cos^2\omega(t_j-\tau)}\right) + \left(\frac{[\sum X_j \sin\omega(t_j-\tau)]^2}{\sum \sin^2\omega(t_j-\tau)}\right) \tag{3.7}$$

This function is a discrete expression of the power spectrum of the signal, and the periods are taken as the peak frequency of the distribution.

3.2 An automated classification method

There are many automated methods that can be used to achieve the final classification. In this paragraph, we want to describe briefly one of them, the Random Forest method, that has also been used by Donalek [17], which we adopted as a template for comparison. Then, in the next chapter, we will focus on classification based on neural networks.
Firstly developed by Leo Breiman and Adele Cutler ([3] - [20]), a random forest is a classifier consisting of a collection of tree-structured classifiers $\{h(\mathbf{x}, \Theta_k), k = 1, ...\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree cast a unit vote for the most popular class at input $\mathbf{x}$. The algorithm is defined as follows (see Fig. 3.4):

1. A bootstrap object sample is obtained, by building it substituting objects from the training set, with the same size as the original set, but with some objects represented multiple times, while others left out (Out of Bag stars, OOB from now - the same number of the objected used multiple times are omitted and will be used to estimate the prediction error).

2. The tree is recursively grown by partitioning the bootstrap sample into subgroups having always more and more homogeneous type contents. At each node, m_{try} divisions into two groups are considered, each using one attribute from a randomly selected set of m_{try} attributes. The best split is selected and the process is then repeated for the child nodes.

3. Finally, a maximum tree is constructed, i.e., a tree with terminal nodes containing only a single type of objects.

Large numbers of trees are built and each tree provides a predicted type for an object. The most probable type is the most frequent one in the sample of

predictions from the different trees. The error rate and confusion matrix can be built by comparing the predicted with the actual types. The attribute importance is given by the difference in classification error averaged over all trees obtained by the OOB sample, permuting the attributes to infer about their importance.
The procedure to build a list of the most important, not too correlated, attributes is as follows (Fig. 3.5):

1. A ranked list of attributes, from the most to the least important, is built using a 20000-tree random forest with the full attribute set.
2. The most important attribute is selected and all other attributes with a Spearman correlation coefficient (Spearman [41]) above 80% are discarded.

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3.8}$$

3. A new ranked attribute list is built by re-running a random forest with the selected and the remaining attributes.
4. The second most important attribute is selected and all other attributes highly correlated with any of the first two are discarded, repeating the same procedure.
5. The process is iterated, obtaining a full ranked list of not too correlated attributes.

The importance value decreases in the list, but never reaches zero, so it is important to understand where to cut the list.
In order to reduce the number of attributes, it is used the following algorithm:

1. The data sample is partitioned for a 10-fold cross validation (CVAL by now).
2. On each CVAL training set, a ranked list of attributes is established using the random forest importance measures.
3. On each CVAL training set, a model is trained on all attributes and used to predict types for the CVAL test set. The CVAL error rate is recorded and the process is repeated after removing the least important attribute. Iterating by removing one attribute at a time and stopping when only 2 attributes are left, a vector of CVAL error rates is obtained.
4. A mean error vector is computed by taking the mean of the 10 values obtained for each attribute subset.

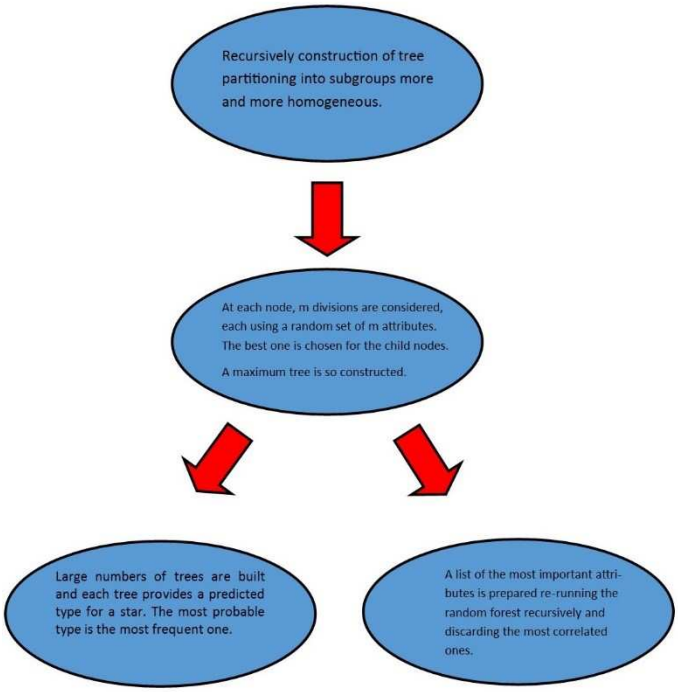


Figure 3.4: Simplified scheme that summarizes the principles of the random forest method.

5. Steps 1 to 4 are repeated 20 times. The mean value and the standard deviation of the 20 CVAL mean errors are computed for each attribute number, combining the results of the classification experiments achieved with a specific attribute number.

The optimum number of attributes can then be inferred by the plot resulting from this procedure. Finally it is possible to proceed with classification and determine the confusion matrix.

3.3 Photometric features

The process of classification relies upon the ability to recognize and quantify the differences between light curves. To build a supervised machine-learning classifier, many instances of light curves are required for each class of interest. These labeled instances are used in the training and testing processes. Since the data are, in general, not sampled at regular intervals, nor are all instances of a certain class observed with the same number of epochs and S/N ratio, the identification of the differences directly from the time-series data is both

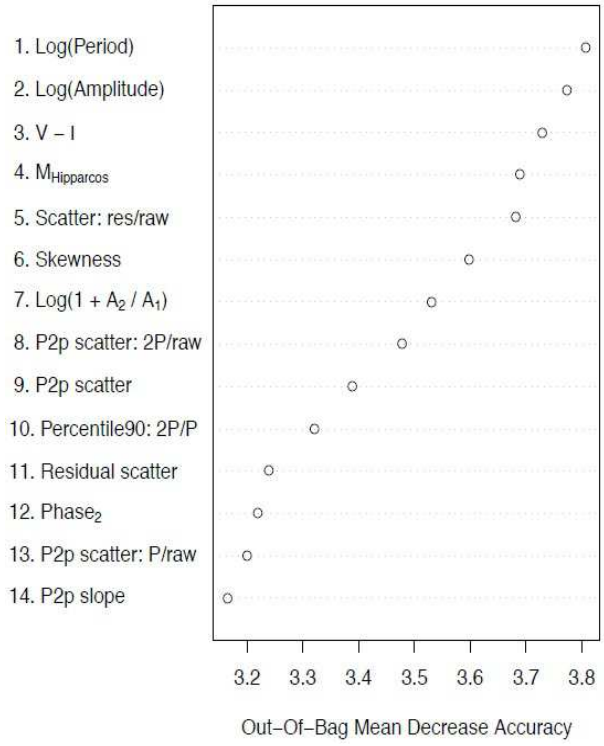


Figure 3.5: An example of a ranked list of 14 most important, not too correlated attributes, from Dubath [20]. The Spearman correlation coefficient of any of the above attributes pairs is smaller than the 80%. The attribute importance is measured with the random forest OOB mean decrease accuracy.

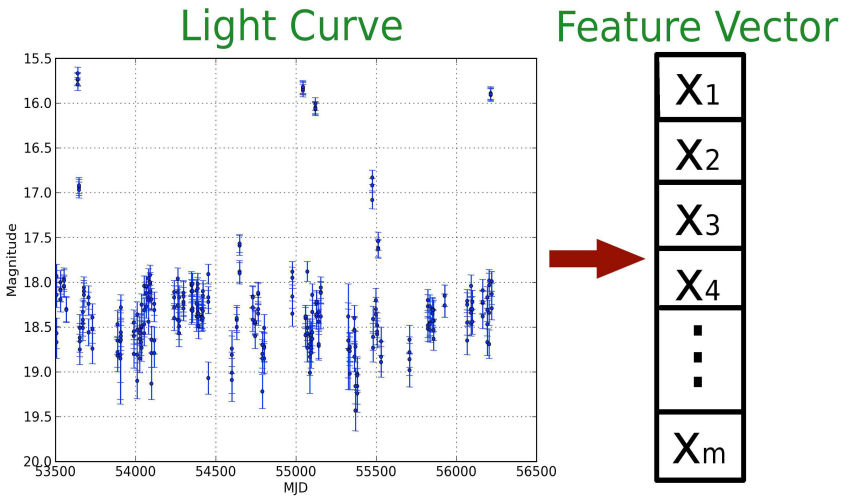


Figure 3.6: Using the CTSCS a set of photometric features is extracted from each light curve forming a feature vector. Here a light curve of a Cataclysmic Variable from CRTS is shown.

conceptually and computationally challenging.
Instead, we homogenize the data by transforming each light curve into a set of real-number line features using statistical and model-specific fitting procedures (Fig. 3.6). These features can be identified, for example, with the attributes used for the random forest method, as said in the previous paragraph.
The features needed for our purpose were calculated by raw light curves using the web service "Caltech Time Series Characterization Service"[1] (CTSCS). With the help of this web service it was possible to determine the 31 non-periodic features ([35] - [15]) for a data sample taken from the Catalina Real-Time Transient Survey (CRTS). Moreover, it is possible to upload also a user defined catalog. Furthermore there exist a number of features that can be determined from the Lomb-Scargle method. In the next paragraph we list the above features, with a brief description for each of them.

3.3.1 Description of the features

As said before, it is possible to divide features in periodic and non-periodic ones. The formers are extracted using the Lomb-Scargle method, while the latter are statistical parameters derived from the light curve analysis.

- **Amplitude**: arithmetic average between maximum and minimum

[1] http://nirgun.caltech.edu:8000/scripts/description.html#data_input

magnitude.

$$amplitude = \frac{mag_{max} - mag_{min}}{2} \tag{3.9}$$

- **Beyond1std**: fraction of photometric magnitudes ($\leq$ 1) that are above or under a certain standard deviation from the weighted average (by photometric errors).

$$beyond1std = P(|mag - \overline{mag}| > \sigma) \tag{3.10}$$

- **Flux Percentage Ratio**: The percentile is the value of a variable under which there is a certain percentage of observations. The flux percentile $F_{n,m}$ was defined to be the difference between the flux values at percentiles n and m, and the following flux percentile ratios were used:

$$fpr\_mid20 = F_{40,60}/F_{5,95}$$
$$fpr\_mid35 = F_{32.5,67.5}/F_{5,95}$$
$$fpr\_mid50 = F_{25,75}/F_{5,95}$$
$$fpr\_mid65 = F_{17.5,82.5}/F_{5,95}$$
$$fpr\_mid80 = F_{10,90}/F_{5,95}$$

- **Linear Trend**: slope of the light curve in the linear fit, that is to say the b parameter in the following linear relation.

$$mag = a * t + b \tag{3.11}$$

$$linear\_trend = b \tag{3.12}$$

- **Maximum Slope**: the maximum difference obtained measuring magnitudes at successive instants.

$$maximum\_slope = max(|\frac{(mag_{i+1} - mag_i)}{(t_{i+1} - t_i)}|) = \frac{\Delta mag}{\Delta t} \tag{3.13}$$

- **Median Absolute Deviation**: median of the deviation of fluxes from the median flux.

$$med\_abs\_dev = median_i(|x_i - median_j(x_j)|) \tag{3.14}$$

- **Median Buffer Range Percentage**: fraction of observations that are within 10% of the median flux.

$$med\_buf\_range\_per = P(|x_i - median_j(x_j)| < 0.1 * median_j(x_j)) \tag{3.15}$$

- **Pair Slope Trend**: percentage of the last 30 couples of consecutive measures of fluxes that show positive slope.

$$pair\_slope\_trend = P(x_{i+1} - x_i > 0, i = n - 30, ..., n) \tag{3.16}$$

- **Percent Amplitude**: maximum percentage difference between maximum or minimum flux and the median.

$$percent\_amplitude = max(|x_{max} - median(x)|, |x_{min} - median(x)|) \tag{3.17}$$

- **Percent Difference Flux Percentile**: Difference between the second and the 98th percentile flux, converted in magnitudes. It is calculated by the ratio $F_{5,95}$ on median flux.

$$pdfp = \frac{(mag_{95} - mag_5)}{median(mag)} \tag{3.18}$$

- **QSO - NOT QSO**: the χ^2/qso and χ^2/non-qso statistics and their significance levels from the quasar variability metric of Butler and Bloom [11]. These parameters, obtained from a function of time modeled using a covariance matrix, make possible to determine a probability distribution for an object to be or not to be a quasar.

- **Skew**: the skewness is an index of the asymmetry of a distribution. It is given by the ratio between the 3rd order momentum and the variance cube.

$$skew = \frac{\mu_3}{\sigma^3} \tag{3.19}$$

- **Small Kurtosis**: the kurtosis represents the departure of a distribution by normality and it is given by the ratio between the 4th order momentum and the variance square. For small kurtosis it is intended the reliable kurtosis on a small number of epochs.

$$kurtosis = \frac{\mu_4}{\sigma^2} \tag{3.20}$$

- **Standard deviation**: standard deviation of the fluxes.

- **Stetson J-K**: the Stetson variability index, which describes variability for Cepheids by p-value determination, as described in Chapter 3.

- **Lomb-Scargle Periodogram**: the period obtained by the peak frequency of the Lomb-Scargle periodogram (Scargle [39]), as described in Chapter 3. There are also a faster version of the algorithm, that determines the top five periods and their false-peak probabilities, and the Generalized Lomb-Scargle Periodogram (see Zechmeister [43]), that instead determines the first five periods obtaining them from a generalization of the Lomb-Scargle method, using appropriate weights.

- **Self Correlation**: the correlation function expresses the statistical correlation between random variables in different points of space and time. If correlation functions between variables representing the same quantity measured in two different points are considered, we speak about an autocorrelation function.

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{3.21}$$

$$C(s,t) = corr(X(s), X(t)) \tag{3.22}$$

- **Structure Function**: the first order structure function obtained using the square differential.

$$sf = f(|mag_i - mag_j|^2, |t_i - t_j|) \tag{3.23}$$

- **Debosscher Frequency Statistics**: the frequency statistic analysis described by Debosscher et al. [15], that is to say the slope of the linear trend, the first three frequencies and their first four harmonics (amplitude and phase for each of them) and the ratio between the variance of the light curve before and after the subtraction of a harmonic fit with the first frequency.

- **R Cor Bor**: the fraction of magnitudes that is below 1.5 magnitudes respect to the median.

$$rcorbor = P(mag > (median(mag) + 1.5)) \tag{3.24}$$

- **AOV**: the period according to the analysis of variance method of Schwarzenberg-Czerny [40].

- **Magnitude Ratio**: an index used to estimate if the object spends most of the time above or below the median.

$$mag\_ratio = P(mag > median(mag)) \tag{3.25}$$

- **Phase Dispersion Minimization**: the period obtained by the minimization of the variance of data with respect to the medium light curve.

- **Fast** χ: this technique uses Fourier series truncated at the H harmonic to model the periodic function. The quality of data is obtained from the Fourier coefficients' χ^2 together with the frequency f.

- **Periodic features**: these are a series of features obtained by light curves using the generalized Lomb-Scargle method. The light curves are modeled as follows:

$$y_i(t|f_i) = a_i \sin(2\pi f_i t) + b_i \cos(2\pi f_i t) + b_{i,0} \tag{3.26}$$

  To determine periodic variations then it is possible to do a minimization of the square sum:

$$\chi^2 = \sum [d_k - y_i(t_k)]^2 / \sigma_k^2 \tag{3.27}$$

So one can define the generalized periodogram:

$$P_f(f) = \frac{(N-1)}{2} \frac{\chi_0^2 - \chi_m^2(f)}{\chi_0^2} \tag{3.28}$$

where:

$$\chi_0^2 = \sum [d_k - \mu]^2 / \sigma_k^2 \tag{3.29}$$

$$\mu = \sum [d_k / \sigma_k^2] / \sum 1/\sigma_k^2 \tag{3.30}$$

Then a fit of light curves is done using the sum of a linear term plus a sum of sinusoids:

$$y(t) = ct + \sum \sum y_i(t|if_i) \tag{3.31}$$

The features used are so obtained:

$$A_{i,j} = \sqrt{a_{i,j}^2 + b_{i,j}^2} \tag{3.32}$$

$$PH_{i,j} = \tan^{-1}(b_{i,j}, a_{i,j}) \tag{3.33}$$

$$f_i \tag{3.34}$$

Finally, other four features are used, obtained by the ratio of the previous features and the offset c.

Chapter 4

Machine learning with Neural Networks

As already mentioned in the introduction, this thesis work tries to classify transients using a machine learning approach based on the use of neural networks.

4.1 Neural networks

A neural network is an analysis instrument modeled on the human brain structure, inserted in an informatics device. It can be constituted both by software and/or by dedicated hardware. Its purpose is to simulate a heavily interconnected computational structure, consisting of many relatively simple individual process elements, the neurons, which make simple calculations on the input signal, then passing the output one to another neuron. These elementary objects are usually organized in groups or layers. Layers can generally receive input signals (input layers), emit output signals (output layers), or be inaccessible to both types of signals, having only connections with other layers (hidden layers).

4.1.1 Biological foundations

In almost all living organisms there are complex organizations of neural cells, with configurations defined by external environment, memorization and reaction to stimuli. Human brain represents the most extraordinary product of biological evolution, due to his capacity to elaborate information. With the aim to do these operations, biological networks use a massive number of simple computational elements, neurons, highly interconnected so as to vary their configuration in response to external stimuli: in this sense we can speak about learning and artificial models trying to catch this distinctive feature of biology.

Generally a neuron is constituted from three principle parts: soma (cell body), axon (the unique output neuron line, branching off in thousands of lines) and dendrite (input neuron line, receiving input signals by other axons through synapses). The cell body makes a weighted sum (integration) of input signals. If the result exceeds a certain threshold value, then the neuron is activated and an action potential is produced and sent to the axon. If the result does not exceed the threshold value, the neuron remains in the rest state. An artificial neural network receives external signals on an input nodes' layer (elaboration units), each one connected with numerous internal nodes, organized in more layers. Every node elaborates the received signals and transmits the result to the nodes in the subsequent nodes layer.

4.1.2 History and utilization

The wide variety of neural networks models cannot leave aside from its basic constituent, the artificial neuron proposed by W.S. McCulloch and W. Pitts in 1943 [31], which outlines a linear threshold combiner, with multiple input binary data and a single output binary data. An appropriate number of these elements, connected to form a network, is capable to calculate simple boolean functions.
In 1958, F. Rosenblatt [36] introduces the first neural network schema, called *perceptron*, which is the precursor of current neural networks, for identification and classification of shapes, with the aim to furnish an interpretation of biological systems general organization. So, the probabilistic model of Rosenblatt looks at the analysis, in mathematical sense, of functions such as information storing and their influence on models' identification. It constitutes a crucial improvement with respect to the binary model of McCulloch and Pitts, because the synaptic weights are variable and therefore the perceptron is capable to *learn*.
Rosenblatt's work stimulate a great number of studies and researches and causes strong interest and expectations on scientific community, which underwent a stop in 1969, when Marvin Minsky and Seymour A. Papert [33] show the operative limits of simple two layers networks based on perceptron, demonstrating the impossibility to resolve many classes of problems: in fact, this type of neural network is not quite powerful for calculating the XOR (exclusive or) function.
The mathematical context to train *Multilayer Perceptron* networks (MLP) was established by the American mathematician Paul Werbos in his doctorate thesis in 1974. One of the best known and efficient methods for neural networks training is the so called *error backpropagation* algorithm, proposed in 1986 by Rumelhart, Hinton and Williams, that systematically modifies weights of connections between nodes, bringing the network response always nearer to the one desired. The backpropagation (BP) algorithm is a learning technique by examples, constituting a generalization of the perceptron learn-

ing algorithm developed by Rosenblatt in the Sixties. Through this technique it was possible, as it has already been said, treating just applications characterized as linearly separable boolean functions. The new algorithm, which allowed to overcome perceptron limitations and to resolve the problem of non linear separability (so calculating the XOR function), marked the definitive revival of neural networks, as showed also by the great variety of commercial applications.
Neural networks are usually used in contexts where data could be partially wrong or where does not exist analytical models to face the problem. Typical utilizations are in optical character recognition software (OCR), in facial recognition systems, and more generally in systems that treat data subjected to errors or rumor. Neural networks are also one of the most used instrument in *Data Mining* analysis. They are also used as predictive instrument in financial or weather analysis. In last years their importance has enormously grown also in bioinformatic and astrophysics, in which they are used for researching functional and structural models in proteins and nucleic acids in the first case and, as previously said, in regression and classification problems for what concerns the astrophysical aspects. Giving properly a series of input (training or learning phase), the network can give the most probable output.

4.1.3 Structure

A neural network is characterized by three fundamental elements:

- The architecture or network topology, that is the particular way in which layers are interconnected and through which they receive input and output; the connection between two generic neurons occurs through a link called weight.
- The activation or transfer function chosen for the neurons, which, in analogy with biological neuron, represents the answer modality to external stimuli. Generally the same function is chosen for all neurons of the layers composing the network, but this is no a strict bond, but an architectural strategy.
- The algorithm used during the learning phase of the network.

These three characteristics can be thought as the highest level of vision of a neural network model. It is important to say that the method, or the methods, must be defined unequivocally, because by this process depends the ability whereby the network learns and progressively improves the response. In the neural networks context, the *learning process* can be seen as the problem to update network architecture and connection weights, so that the network itself can efficiently perform its specific task. In general, during the learning phase, the network fixes the weights values that the input

configurations connections must have. Its performances improve progressively by updating the weights over time, by the repeated presentation of configurations belonging to the same class.
It is necessary to distinguish at least three different learning typologies (the most important ones, but there exist also other ones). In particular, one can have:

- Supervised learning: based on a training set including typical input examples with the corresponding outputs. The network is trained with a proper algorithm, which uses the a priori knowledge to modify weights and other parameters of the network itself, so as to minimize the prevision error related to the sample used for training. If the training phase is successful, the network learns how to recognize the unknown relation that connects the input variables with the output ones, and so it becomes capable to make previsions also where the output is not known a priori. In other words, the final target of supervised learning is the prevision of the output value for every valid input value, basing just on a relatively small number of correspondence examples (that is to say, input-output couples).

- Unsupervised learning: based on training algorithms that modify network weights referring exclusively to a set of data that includes just input variables. These algorithms try to group input data and to find proper classes that result to be representative of the data themselves, making use of topological or probabilistic methods. Unsupervised learning is also used to develop compression data techniques.

- Reinforcement learning: in this case an algorithm aims to find a certain modus operandi, starting from an observational process on external environment; every action has a consequence on environment, and it produces a feedback that guides the algorithm itself in the learning process. This class of problems postulates an agent, endowed with perception power, which explores an environment in which it undertakes a series of actions. The environment itself furnishes an incentive or disincentive as response, as appropriate. Algorithms for reinforcement learning ultimately try to determine a policy inclined to maximize incentives received by the agent during its exploration of the problem. Reinforcement learning differs from supervised one because there were not presented input-output couples of known examples, and one does not proceed to the explicit correction of suboptimal actions. Furthermore, the algorithm is focused on real time performance, that implies a balance between the exploration of unkwown situations and exploitation of current knowledge.

In the present work we shall use only supervised methods.

4.1.4 Multilayer Perceptron

The *Multilayer Perceptron* (MLP) is the most commonly used architecture for practical applications of neural networks. Generally a MLP is constituted by an input neuron layer, one or more hidden layers, each one composed by a certain number of neurons, and an output layer, constituted by as many neurons as the response variables are. The different neurons are interconnected by weights, that is to say parameters which are estimated during the training phase using the so called learning set. Practically MLP networks with just one hidden layer are often used, because they furnish satisfactory results and are computationally less expensive than networks with more layers.
The MLP realizes a complex non linear mapping between input and output of the network. Denote with $\mathbf{x} = \{x_1, x_2, ..., x_d\}$ the N input values. The first layer generates a series of linear combinations of the input values, with the aim to obtain a set of intermediate activation variables $a_j^{(1)}$ such that:

$$a_j^{(1)} = \sum_{i=1}^{d} w_{ji}^{(1)} x_i + b_j^{(1)}, j = 1, ..., M \tag{4.1}$$

where every $a_j^{(1)}$ variable is associated to a single neuron of the M units of the hidden layer. The $w_{ji}^{(1)}$ values represent the elements of the weight matrix of the first layer, while the $b_j^{(1)}$ are the bias parameters (which consider a systematical error or a selection effect) associated to the hidden layer units. So the $a_j^{(1)}$ variables are transformed into the non linear activation function of the hidden layer. For example, if the used function is the hyperbolic tangent, the output values from the hidden neurons are:

$$z_j = \tanh(a_j^{(1)}), j = 1, ..., M \tag{4.2}$$

Then the z_j values are transformed again by the second layer of weights and biases to obtain a second layer of activation values $a_k^{(2)}$, given by the formula:

$$a_k^{(2)} = \sum_{j=1}^{M} w_{kj}^{(1)} z_j + b_k^{(2)}, k = 1, ..., c \tag{4.3}$$

where c is the number of the output units. Finally, the output activation function is applied to these values, through which the final values y_k, where $k = 1, ..., c$, are obtained. Depending on the nature of the considered problem, one can have:

- regression problems, with a linear activation function, i.e. $y_k = a_k^{(2)}$;
- classification problems, with a gaussian activation function, independently applied to everyone of the output neurons, i.e.:

$$y_k = \frac{1}{1 + \exp\left(-a_k^{(2)}\right)} \tag{4.4}$$

The main training algorithm of a MLP is the *backpropagation*, based on a correction error rule. Essentially, backpropagation consists in two steps, forward and backward respectively, through the network layers. In the first step, the forward one, an input vector is applied to the network inputs, propagating layer by layer. Finally, an output is generated, corresponding to the actual response of the network itself. Contrarily, in the successive step, backward, weights are adjusted through the error correction law. In a more specific manner, the network answer is subtracted to the values of a correct-known values' sample, denoted with $\mathbf{t} = \{t_1, t_{2,\ldots,t_c}\}$, so that an error signal is producted and propagated through the network. Obviously the signal error form can be defined in many different ways, depending from the problem we are considering. In particular, one can have two fundamental cases:

- for regression problems, a quadratic sum error function is adopted:

$$E = \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{c} \left\{ y_k(\mathbf{x}^n; \mathbf{w}) - t_k^n \right\}^2 ; \tag{4.5}$$

- for classification problems, a *cross-entropy* error function is often preferred:

$$E = -\sum_{n} \sum_{k=1}^{c} \left\{ t_k^n \ln y_k^n + (1 - t_k^n) \ln \left(1 - y_k^n\right) \right\} . \tag{4.6}$$

Weights are then adjusted in such a manner that the output of the network approaches to the desired values in a statistical sense and the procedure is repeated until the result varies only in a negligible way. In our case we will use a more efficient variant of the backpropagation algorithm, called *quasi-newtonian method*.

4.1.5 MLPQNA

MLPQNA stands for the traditional neural network MLP model implemented with a Quasi Newton Approximation (QNA) as learning rule. The network used for our experiments is offered by the DAMEWARE infrastructure [10] - [4] - [5] - [6] - [7]. In the case of the QNA learning rule implementation, the algorithm used is an adapted version of the classical Newton method for optimization problems. The Newton method is the general basis for a whole family of so called Quasi-Newtonian methods. The QNA is an

optimization of the learning rule, also because the implementation is based on a statistical approximation of the Hessian matrix of the error, through a cyclic gradient calculation. The learning rule of our MLP is the Quasi Newton Approximation, which differs from the Newton Algorithm in terms of the calculation of the Hessian of the error function. In fact Newtonian models are variable metric methods used to find local maxima and minima of functions and, in the case of MLPs they can be used to find the stationary (i.e. the zero gradient) point of the learning function.
We know that the classical Newton method uses the Hessian of a function in the following way. The step of the method is defined as a product of an inverse Hessian matrix and a function gradient. If the function is a positive definite quadratic form, we can reach the function minimum in one step. In case of an indefinite quadratic form (which has no minimum), we will reach the maximum or saddle point. In short, the method finds the stationary point of a quadratic form. In practice, we usually have functions which are not quadratic forms and, however, the Newton method can converge both to a minimum and a maximum. More generally, the Hessian of a function is not always available and in many cases it is far too complex to be computed. More often we can only calculate the function gradient which can be used to derive the Hessian via N consequent gradient calculations. The gradient in every point w is in fact given by:

$$\nabla E = H \times (w - w^*) \tag{4.7}$$

where w corresponds to the minimum of the error function, which satisfies the condition:

$$w^* = w - H^{-1} \times \nabla E \tag{4.8}$$

The vector $H^{-1}\nabla E$ is known as Newton direction.
Quasi Newton methods solve this problem as follows: they use a positive definite approximation instead of a Hessian. If the Hessian is positive definite, we make the step using the Newton method. If the Hessian is indefinite, we modify it to make it positive definite, and then perform a step using the Newton method. In practice, it QNA is an optimization of the learning rule based on a statistical approximation of the Hessian by cyclic gradient calculation which, as already mentioned, is at the base of the classical Back Propagation method.
The QNA instead of calculating the H matrix and then its inverse, uses a series of intermediate steps of lower computational cost to generate a sequence of matrices which result more and more accurate approximations of H^{-1}. During the exploration of the parameter space, in order to find the minimum error direction, QNA starts in the wrong direction. This direction is chosen because at the first step the method has to follow the error gradient and so it takes the direction of steepest descent. However, in subsequent

steps, it incorporates information from the gradient. By using the second derivatives, QNA is able to avoid local minima and to follow more precisely the error function trend, revealing a "natural" capability to find the absolute minimum error of the optimization problem.
The following features are implemented in the MLPQNA present in DAMEWARE, and, during this thesis work we will widely use them:

- only batch learning mode is available (i.e. the network error is calculated at the end of the submission of the complete training dataset);
- strict separation between classification and regression functionality modes;
- for classification mode, the Cross Entropy method is available to compare output and target network values. It is possible to alternatively use standard MSE rule, that is mandatory for regression mode;
- K-fold cross validation method to improve training performances and to avoid overfitting problems;
- resume training from past experiments, by using the weights stored in an external file at the end of the training phase;
- confusion matrix calculated and stored in an external file for both classification and regression modes (in the last case an adapted version is provided). It is useful after training and test sessions to evaluate model performances.

The MLP network topology parameters and QNA training rule parameters are the following:

- input neurons: the number of neurons assigned to the input layer. It will correspond to the number of feature selected for the experiment, as will be described in Chapter 5;
- hidden: the number of hidden layers selected and the correspondent number of hidden neurons;
- output: the number of output neurons. In all our experiments it will be always fixed to 1, because we will only do binary classifications;
- W-step: one of the two stopping criteria. The algorithm stops if approximation error step size is less than this value. A step value equal to zero means to use the parameter MaxIts as unique stopping criterion;
- Restarts: number of restarts of hessian approximation from random positions, performed at each iteration;

- Decay: regularization factor for weight decay. The term $dec*||net\_w||^2$ is added to the error function, where net_w is the total number of weights in the network. When properly chosen, the generalization error of the network is highly improved. This is a fundamental parameter;
- MaxIts: max number of iterations of hessian approximation. If zero the step parameter is used as stopping criterion;
- CVAL: the k parameter for Cross Validation.

In particular, for what concern this last point, we already mentioned Cross Validation in the case of the attribute selection for the random forest method, in Chapter 3. More generally it is an automatized process used to avoid overfitting on the training set. It occurs when a statistical model describes random error or noise instead of the underlying relationship. Generally, overfitting appears when a model is excessively complex, such as having too many parameters relative to the number of observations. A model which has been overfit will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data. In particular, overfitting occurs when a model begins to memorize training data rather than learning to generalize from trend.
Therefore, as we said, Cross Validation is one of the techniques that can be used to avoid this phenomenon. In our case, it was done by performing 10 different training runs with the following procedure:

1. The training set is split into 10 random subsets, each one composed by 10% of the dataset.
2. At each training run the 90% of the dataset is applied for training and the excluded 10% for validation. The number of runs is fixed to the k value.

The analysis of the results of our experiments are based on the so called confusion matrix. From a confusion matrix defined as showed in Tab. 4.1 it is possible to derive the following parameters (in capital letters there is the extended name of the parameter, in brackets the label that will be used in tables):

- TOTAL EFFICIENCY (Eff): ratio between the number of correctly classified objects and the total number of objects in the data set.

$$Eff = \frac{N_{11} + N_{22}}{N_{11} + N_{12} + N_{21} + N_{22}} \tag{4.9}$$

- PURITY OF A CLASS (Pur1 and Pur2): ratio between the number of correctly classified objects of a class and the number of objects

		OUTPUT	
	-	CLASS 1	CLASS 2
TARGET	CLASS 1	N_{11}	N_{12}
	CLASS 2	N_{21}	N_{22}

Table 4.1: Structure of the confusion matrix as obtained from the MLPQNA.

classified in that class, also known as efficiency of a class.

$$Pur1 = \frac{N_{11}}{N_{11} + N_{21}} \tag{4.10}$$

$$Pur2 = \frac{N_{22}}{N_{12} + N_{22}} \tag{4.11}$$

- COMPLETENESS OF A CLASS (Comp1 and Comp2): ratio between the number of correctly classified objects in that class and the total number of objects of that class in the data set.

$$Comp1 = \frac{N_{11}}{N_{11} + N_{12}} \tag{4.12}$$

$$Comp2 = \frac{N_{22}}{N_{21} + N_{22}} \tag{4.13}$$

- CONTAMINATION OF A CLASS (Cont1 and Cont2): it is the dual of the purity, namely it is the ratio of misclassified object in a class and the number of objects classified in that class.

$$Cont1 = 1 - Pur1 = \frac{N_{21}}{N_{11} + N_{21}} \tag{4.14}$$

$$Cont2 = 1 - Pur2 = \frac{N_{12}}{N_{12} + N_{22}} \tag{4.15}$$

These parameters make possible to describe completely the distribution of the patterns after the process of classification training and test.

Chapter 5

The DAMEWARE infrastructure

The data burst that in the recent years is changing the way to perform astrophysical research, requires a new generation of software tools, largely automatic, scalable and highly reliable. A great importance has been acquired from the

The DAMEWARE (Data Mining & Exploration Web Application REsource) infrastructure is born with the aim to perform the so called *Knowledge Discovery in Databases* (KDD), enabling a learning paradigm to treat massive data sets by the development of new algorithms of lower computational complexity. In this way it is possible to infer knowledge from data and validate the obtained results. It was an innovative, general purpose, Web-based, VO (Virtual Observatory) compliant, and distributed data mining infrastructure specialized in massive data sets exploration with machine learning methods. Nowadays it has evolved to become a general purpose platform able to find applications also in other domains of human knowledge and research.

One of the main features of DAME is its usability and scalability, considering the fact that KDD is a complex process. In fact, we must consider that one can find good results only on a trial and error base by comparing outputs of different methods and different experiments with the same method, with a lengthy fine tuning phase that could result hard to a not experienced user, requiring a good knowledge of the mathematics underlying the methods, of the computing infrastructures and of the complex workflows which need to be implemented. For these reasons, through the use of the Web application paradigm and of an extensive and user friendly documentation, DAME represents the first attempt to bring the KDD models to users hiding most of their complexity behind an hybrid distributed well designed computing infrastructure.

Obviously it is important to remember that by making an intensive use of background knowledge it is possible to reduce the amount of data that are

required by a specific problem during the learning phase.
Using a simple browser, DAME offers several tools for data analysis, such as clustering, classification, regression, feature extraction etc., together with models and algorithms. No software needs to be installed on the local machine of the user, configuring and executing experiments on a virtualized computing infrastructure. Moreover, it is possible to extend the original library of available tools, by adding plug-in and executing code through a simple guided procedure, without any restriction about the native programming language.

5.1 Design and architecture

DAME was conceived to provide the scientific community with an extensible, integrated environment for data mining and exploration. With this aim, it had to:

- support the VO standards and formats, in particular for data interoperability;
- to abstract the application deployment and execution, so to provide the VO with a general purpose computing platform exploiting modern technologies.

An important aspect that must be considered is the a-synchronous access. In fact, most available web based data mining services run synchronously, so executing jobs during a single HTTP transaction. This is obviously simpler, but it does not fit well with long-run tasks, because all the entities in the chain of command must remain up for the duration of the activity, losing it if anyone stops.
For what concern the main structure of the web service (see Fig. 5.1), in the DAME data mining infrastructure the choice of any machine learning mode, a supervised or unsupervised one, is always accompanied by the functionality domain, that is to say the mode to explore the available data (regression, classification, clustering, etc.).
The combination of the chosen data mining model and functionality makes possible to do experiments, for which a use case must be selected: one may have training, test, validation and run use cases, in order to perform, respectively, learning, verification, validation and execution phases. Most models provide also a full use case, that executes all listed cases automatically as a sequence.
From the technological point of view, DAMEWARE consists of five main components: Front End (FE), Framework (FW), Registry and Data Base (REDB), Driver (DR) and Data Mining Models (DMM).

Model	Name	Category	Functionality
MLPBP	Multi Layer Perceptron with Back Propagation	Supervised	Classification, regression
FMLPGA	Fast MLP trained by Genetic Algorithm	Supervised	Classification, regression
MLPQNA	MLP with Quasi Newton Approximation	Supervised	Classification, regression
MLPLEMON	MLP with Levenberg-Marquardt Optimization Network	Supervised	Classification, regression
SVM	Support Vector Machine	Supervised	Classification, regression
RandomForest	Random Forest Algorithm	Supervised	Classification, regression
ESOM	Evolving Self Organizing Maps	Unsupervised	Clustering
K-Means		Unsupervised	Clustering
SOFM	Self Organizing Feature Maps	Unsupervised	Clustering
SOM	Self Organizing Maps	Unsupervised	Clustering
PPS	Probabilistic Principal Surfaces	Unsupervised	Feature Extraction

Table 5.1: Data mining models and functionalities available in the DAMEWARE framework. Column 1: acronym; column 2: extended name; column 3: category; column 4: functionality.

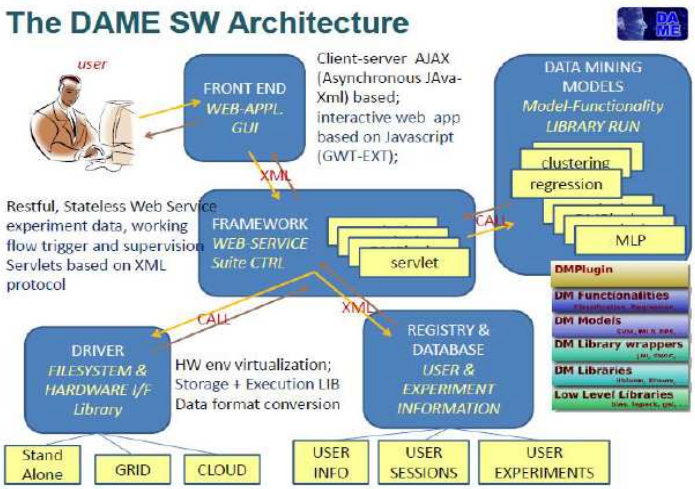


Figure 5.1: The general software architecture of DAMEWARE (Cavuoti [12]).

The DAME design architecture is implemented following the standard LAR (Layered Application Architecture) strategy, which leads to a software system based on a layered logical structure, where different layers communicate with each other via simple and well-defined rules:

- Data Access Layer (DAL): the persistent data management layer, responsible of the data archiving system, including consistency and reliability maintenance.
- Business Logic Layer (BLL): the core of the system, responsible of the management of all services and applications implemented in the infrastructure, including information flow control and supervision.
- User Interface (UI): responsible of the interaction mechanisms between the BLL and the users including data and command I/O and views rendering.

The main concepts that lay behind the distributed data mining applications implemented in the DAME Suite are based on three issues:

- virtual organization of data: this is an extension of a basic feature of the VO;
- hardware resource-oriented: this is obtained by using computing infrastructure, like grid, which enable parallel processing of tasks, using idle capacity, with the aim to obtain large number of instances running for short periods of time;
- software service-oriented: this is the base of usual cloud computing paradigm. The data mining applications implemented runs on top of virtual machines seen at the user level as services (specifically web services), standardized in terms of data management and working flow.

The hardware infrastructure of DAMEWARE, instead, is based on two subnetworks addressable from an unique access point, the website, which provides an embedded access to the user to all DAME web applications and services. The integrity of the system is guaranteed by a registration procedure, which gives the possibility to access all facilities from just one account. Depending on the computing and storage power requested by the job and by the processing load currently running on the network, an internal mechanism redirects the jobs to a job-queue in a pre-emptive scheduling scheme. The interaction with the infrastructure is completely asynchronous and a specialized software component has the responsibility to store off-line job results in the user storage workspaces, that can be retrieved and downloaded in subsequent accesses. This hybrid architecture makes possible to execute simultaneous experiments that gathered all together bring the best results.

Instead, from the software point of view, DAME is based on the following features:

- modularity: software components with standard interfacing, easy to be replaced;
- standardization: basically, in terms of information I/O between user and infrastructure as well as between software components;
- hardware virtualization: i.e. independent from the hardware deployment platform (single or multi processor, grid etc.);
- interoperability: by following VO requirements;
- expandability: because many parts of the infrastructure require to be increased and updated along its lifetime;
- asynchronous interaction: there is not a synchronous interaction between the end user and the client server mechanisms, so the user is not constrained to remain connected after launching an experiment in order to wait for the end of execution;
- language-independent programming: this basically concerns the API (Application Programming Interface) forming the data mining model libraries and packages. Although most of the available models and algorithms were internally implemented, this is not considered as mandatory. The suite provided a Java based standard wrapping system to achieve the standard interface with multi-language APIs;
- distributed computing: the components can be deployed on the same machine as well as on different networked computers;
- pluggable: with the new plugin procedure users can extend the data mining model library integrated into the web app, by simply download and run a Java application, which through a driven procedure generate source code to be integrated into the web app software infrastructure.

Chapter 6

Data

In this thesis we used data from a single synoptic survey, the CRTS. This choice came after comparing several data sets from different surveys and the fundamental parameters characterizing them, as is possible to see in Tab. 6.1 and 6.2.
The survey resulted as the most suitable for our purpose was the CRTS, so, as said before, the catalogs have been created from this survey. Relatively small synoptic surveys like CRTS, today, can be considered both as scientific and technological precursors and testbeds for the biggest surveys of the next future, such as LSST or SKA.

6.1 Catalina Real Time Transient Survey

The Catalina Real Time Transient Survey (Tab. 6.1) makes use of existing synoptic telescopes and image data resources from the Catalina Sky Survey (CSS). CSS uses three wide-field telescopes: the 0.68 m Catalina Schmidt at Catalina Station, AZ, the 0.5 m Uppsala Schmidt (Siding Spring Survey - SSS), at Siding Spring Observatory, NSW, Australia, and the Mt. Lemmon Survey (MLS), a 1.5 m reflector located at Mt. Lemmon, AZ. For each telescope, a camera with a single, cooled, 4k x 4k back-illuminated, unfiltered CCD is used. The combined CSS+SSS+MLS data streams can cover up to $\approx 2000\ deg^2$ per night to a limiting magnitude of V $\approx 19 - 20$ mag, plus a smaller area ($\approx 200\ deg^2$ per night) to a limiting magnitude of V ≈ 21.5 mag.
The CRTS covers the total area of $\approx 33000\ deg^2$, excluding the Galactic plane within $|b| < 10 - 15$, down to $\approx 19 - 21$ mag per exposure, with increasing time baselines from 10 min to 8 years; there are now typically $\approx 300 - 400$ exposures per pointing, and coadded images deeper than ≈ 23 mag.
The CRTS has detected astrophysical transients and variable objects outside the Solar System performing searches in the catalog domain and by the use

Survey	CRTS	PQ
Coverage	33000 deg^2	150000 deg^2
Coverage per night	2200 $deg^2/night$	500 $deg^2/night$
Field of View	8 deg^2	9.4 deg^2
Declination	$-75 < dec < 70$	$-25 < dec < 25$
RA	/	/
Galactic latitude	$\|b\| > 15$	/
Nr of passes/field/night	4	From 5 to 25
f_open	0.7	/
Effective collecting area	2.326 m^2	1 m^2
t_exp	$20 - 30$ sec	150 $sec/cos\delta$
Overall instrument efficiency	0.7	0.4
Full Width Half Maximum	3	2
Merit figure	5470	/
Limiting magnitude	21.5 (V)	21.5 (r)
Transient detected	7500 (CSDR2)	4800 (15% confirmed)
Time baseline	From 10 min to 6 yrs	From hours to years
Public data release	CSDR2	Public data release 1.0
Number of objects	500 million	/
Magnitude interval	$11.5 < V < 21.5$	/
Reference for data	[19]	[16]
Included surveys	CSS, MLS, SSS	/

Table 6.1: Some useful parameters of CRTS and Palomar Quest (PQ) surveys.

Survey	SDSS II	PTF
Coverage	300 deg^2	1/2 of the entire sky
Coverage per night	/	1000 $deg^2/night$
Field of View	$3 * 3$ deg^2	7.78 deg^2
Declination	$-1.25 < dec < 1.25$	/
RA	$-60 < ra < 60$	/
Galactic latitude	$b < 0$	/
Nr of passes/field/night	/	2
f_open	/	0.7
Effective collecting area	4 m^2	1.131 m^2
t_exp	54 sec	60 sec
Overall instrument efficiency	0.4	0.7
Full Width Half Maximum	1.5	2
Merit figure	/	4820
Limiting magnitude	22.5	21
Transient detected	580	1860
Time baseline	/	From 1 min to 5 days
Public data release	DRSN1	/
Number of objects	230 million	/
Magnitude interval	/	/
Reference for data	[38]	[29] - [28]
Included surveys	/	/

Table 6.2: Some useful parameters of SDSS II and Palomar Transient Factory (PTF) surveys.

of image subtraction. Sources that show significant changes in brightness, or which appear for the first time where previously no sources where detected, are identified. The contrast threshold is set high (flux changes of at least ≈ 1 mag and $\approx 5\sigma$), with the aim to find the most dramatic, and also most interesting, transients.
The survey has detected ≈ 7500 unique, high-amplitude, transients, including at least 1800 SN, at least 1000 CVs (the majority of them previously uncatalogued), over 2500 of blazars/OVV AGN, hundreds of flare stars, etc. It was recently made available for download the second data release (CSDR2[1]), containing about 500 million light curves. Photometric data are obtained using SExtractor.
It is possible to choose different search options into the database[2]. It is possible to perform searches around a single location (giving Ra and Dec, the name of the selected object or the ID) or around multiple locations, loading a data file containing ID, Ra, and Dec (with a limit of 100 locations). Moreover it is possible to perform a search for period. For every catalog extracted, one can select the table and the data formats. The database is organized in different catalogs, described in the following.
Master objects are the sources detected in coadds (Master frames) from 20 CSS images. Objects detected in individual images are linked to these objects based on their position. The matching radius is a function of seeing and telescope resolution. Information about all master images is placed in the MasterFrame, whereas the photometry of the objects is in MastercatCSS (MastercatMLS and MastercatSSS for MLS and SSS coadd sources).
The individual object catalogs include all the detections from the North and South grid fields of CSS. Each detection is linked to a master source and placed in the photometry catalog (Photcat). Sources with no match to master objects are put in the separate Orphan object catalog (OrphancatCSS). Orphan objects include real sources such as asteroids as well as other transients. Other spurious single detections are also included. However, some objects have been removed based on quality flags. Information common to an image is placed in the frame catalog (FramecatCSS).
A fundamental feature of the CRTS is its fully open policy: in fact all detected transients are immediately published, with no proprietary period at all, bringing enormous benefits to the entire astronomical community and maximizing the scientific returns by encouraging follow-up by other groups.

6.2 Final catalog

We prepared a catalog with data from the CRTS database, from which the photometric features described in the paragraph 3.3 were extracted using

[1] http://nesssi.cacr.caltech.edu/DataRelease/
[2] http://nesssi.cacr.caltech.edu/DataRelease/schema.html

the CTSCS web-service, with 1619 patterns and 29 columns (name, ra, dec, 25 photometric features and class). These data will be used to train the MLPQNA. The catalog is composed by the following classes (on the right there is the label used in the catalog and in brackets the number of patterns for each class is reported):

- Cataclismic Variables - CV (461);
- Supernovae - SN (536);
- Blazar - Bl (124);
- Active Galactic Nuclei - AGN (140);
- Flare Stars - Fl (66);
- RR Lyrae - RRL (292).

Chapter 7

Classification experiments

This chapter describes the central part of this thesis. The experimental strategy followed in the development of the work is presented and the description of the experiments, with the tables presenting the detailed results obtained.

7.1 Experimental strategy

Before describing in details the experiments, we will summarize the adopted strategy and the various steps in the preparation of the catalogs. First of all our work focused only on two-class classification, since it is an exploratory work and this is the simplest type of classification. Only some of the available photometric features were selected for the training in the different cases. In fact, for each experiment, a new catalog was prepared, by selecting only a subset of features plus the binary flags representing the target classification values. Then, in each experiment the catalog was randomly split in two parts, one for training and one for test, containing 80% and 20% of the objects respectively, using the random row shuffle function available within DAMEWARE, in order to ensure a proper coverage of the parameter space. In Tab. 7.1 an example of such a catalog is presented. All the catalogs in the following will be based on the same criteria.
We used the following strategy to proceed with the experiments:

1. A first series of experiments was performed in order to find the best initial configuration of the MLPQNA classifier. The classification is between the two classes CV and ALL (AGN + SN + Fl + Bl - RRL were removed to be added again in late experiments).

2. To understand what were the best working group of features for the classification of the previous classes, we performed a pruning of the features. We started from a nucleus of features, chosen heuristically, and then we added the other features, one by one, recursively, selecting the one that gave the best results. Then we repeated the same operation

amplitude	beyond1std	fpr_mid50	fpr_mid65	std	target
1.5	0.571429	0.61319	0.806252	1.082759	1
0.47817	0.33333	0.2247	0.41031	0.29104	0
1.33274	0.45454	0.71078	0.83724	0.95925	0
1.36	0.310954	0.279479	0.42138	0.50272	1

Table 7.1: Example of few records in one of the catalogs used for the experiment, with some features selected and the target parameter indicating the right classification for all training and test pattern.

starting from a different nucleus, chosen by the inspection of features histograms, finally comparing the results. The main purpose of this method was to see what could be the ideal combination of features for this classification.

3. Using the two nuclei and the best setup previously obtained, we changed the type of classification, by performing experiments that we shall call for EXTRA-GALACTIC vs GALACTIC (AGN + Bl vs CV + SN + Fl), to see if there are improvements with respect to the previous separation. This definition comes from the fact that we grouped together AGN with BL Lac objects (i.e. extra-galactic objects), against CV+SN+Fl. The inclusion of SN in the latter being due that even though they are mainly observed in external galaxies, they still are stars and therefore of a completely different category with respect to active galactic nuclei. The idea, in fact, is to explore different types of classification to test the method in different situations.

4. Finally we did experiments adopting the same groups of features used in Donalek [17], to compare the results obtained with different classifiers. In this case the catalog used was the same of the article [17], with the classification of the two classes SN vs ALL (AGN + Bl + CV + Fl + RRL).

Following this strategy we aimed of exploring the performances of the method, identifying its strength and its weakness, by considering also its results within the context of a wider framework of transient classification, starting from raw data up to the final classification.
Therefore: for each experiment (in the cases of the pruning operations, only for the best ones), we repeated the experiment with the same configuration, but on a new catalog, in which the two classes have been preventively balanced one each other. We noticed that there is always a great lack of balance between the classes considered (it will be clear when we will report the number of the patterns composing the different classes). So we properly

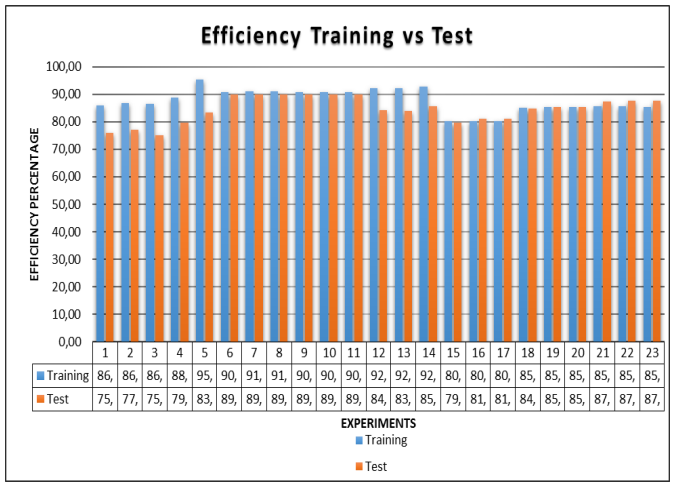


Figure 7.1: Histogram showing a comparison between the efficiency values during the two phases of training and test in some of the experiments done. Only experiments 6 and 7 uses CVAL.

(randomly) cut the catalogs, by leaving the larger class with about the 5% of patterns more than the smaller one. In this way, we could study the dependence of the results from the number of patterns in the two classes, and from the group of features used.
Finally, before focusing our attention on the experiments, we shall anticipate some of the results obtained, in the subsequent histograms, where we report the values of the various parameters obtained from the confusion matrix for training and test phases of more than 20 experiments.
The aim is to verify if there is compatibility between the two values and if the classifier is working properly.
The histograms are reported from Fig. 7.1 to Fig. 7.7 in the next pages. We can see that, except for some cases, there is a good agreement between the results of the training and test phases. For the cases where this is not true, with a variation from 10% to 20%, we have noticed that this happens mainly in the first experiments, based on CV classification (the most ambiguous one), and done with groups of features and MLPQNA structure not yet fixed at all. So we think this could be a good explanation for these strong variations, that in the last experiments practically disappeared. Notice that, among these experiments, only number 6 and 7 use the k-fold Cross Validation.

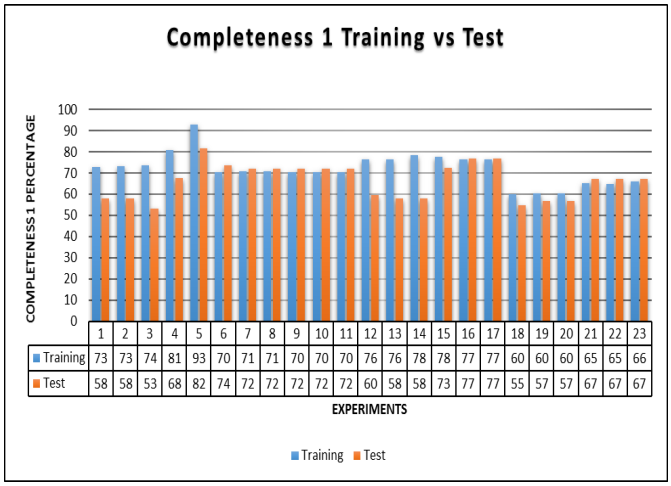


Figure 7.2: Histogram showing a comparison between the completeness values of the first class (Cataclismic Variables, AGN + Blazar, or Supernovae, depending from the experiments that will be presented in the next paragraph and that were chosen to cover all the experimental phase) during the two phases of training and test in some of the experiments done. Only experiments 6 and 7 uses CVAL.

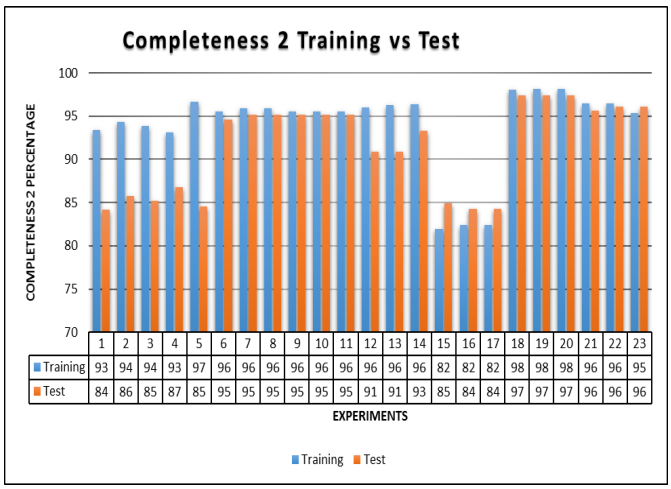


Figure 7.3: Histogram showing a comparison between the completeness values of the second class (ALL the other classes opposite to the first classes listed in Fig 7.2, depending from the experiments that will be presented in the next paragraph and that were chosen to cover all the experimental phase) during the two phases of training and test in some of the experiments done. Only experiments 6 and 7 uses CVAL.

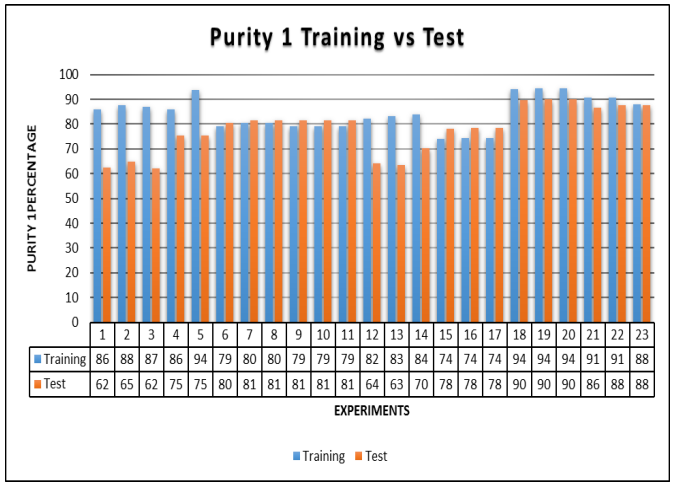


Figure 7.4: Histogram showing a comparison between the purity values of the first class (Cataclismic Variables, AGN + Blazar, or Supernovae, depending from the experiments that will be presented in the next paragraph and that were chosen to cover all the experimental phase) during the two phases of training and test in some of the experiments done. Only experiments 6 and 7 uses CVAL.

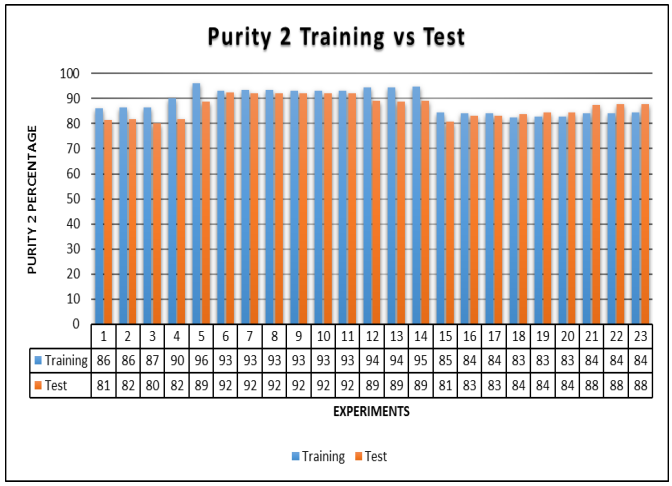


Figure 7.5: Histogram showing a comparison between the purity values of the second class (ALL the other classes opposite to the first classes listed in Fig. 7.4, depending from the experiments that will be presented in the next paragraph and that were chosen to cover all the experimental phase) during the two phases of training and test in some of the experiments done. Only experiments 6 and 7 uses CVAL.

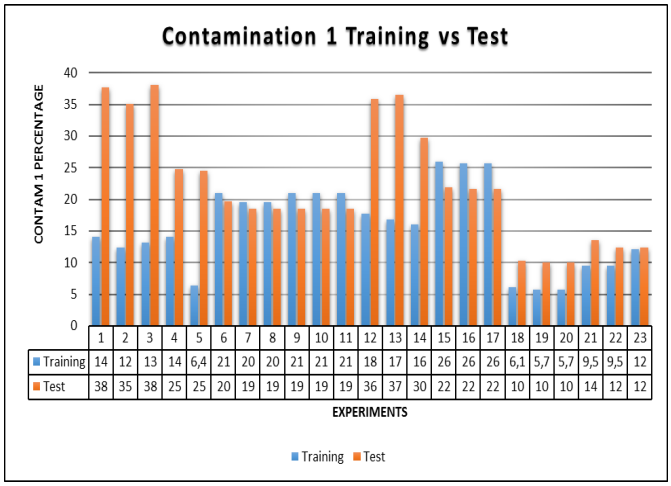


Figure 7.6: Histogram showing a comparison between the contamination values of the first class (Cataclismic Variables, AGN + Blazar, or Supernovae, depending from the experiments that will be presented in the next paragraph and that were chosen to cover all the experimental phase) during the two phases of training and test in some of the experiments done. Only experiments 6 and 7 uses CVAL.

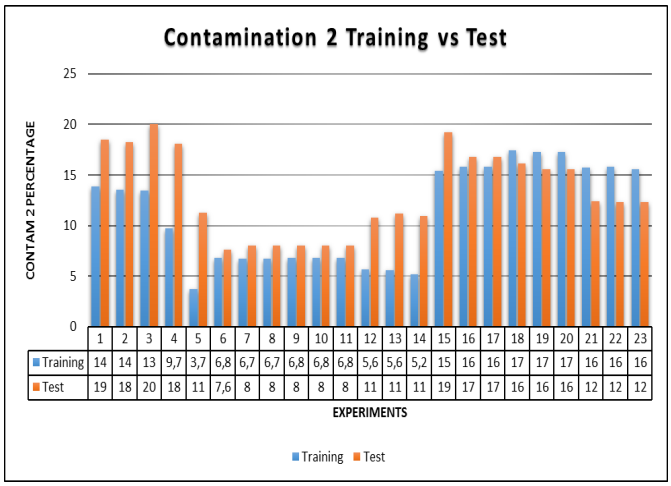


Figure 7.7: Histogram showing a comparison between the contamination values of the second class (ALL the other classes opposite to the first classes listed in Fig. 7.6, depending from the experiments that will be presented in the next paragraph and that were chosen to cover all the experimental phase) during the two phases of training and test in some of the experiments done. Only experiments 6 and 7 uses CVAL.

	Hidden layers	Decay	Wstep	Eff (%)	Pur1 (%)	Pur2 (%)	Comp1 (%)	Comp2 (%)
Test1	1	0.01	0.001	78	75	80	65	86
Test2	1	0.001	0.001	73	68	75	56	83
Test3	2	0.01	0.001	74	70	75	55	85
Test4	2	0.001	0.001	73	69	75	54	85

Table 7.2: Table showing the different settings of the MLPQNA and the results obtained in percentage of objects. Class 1 is referred to CV (461 patterns), while class 2 to is referred to ALL the others (866 patterns). The values of Restart and MaxIts parameters are fixed respectively to 60 and 10000 and the number of input neurons is fixed to 5.

7.2 Experiments

In this paragraph we shall report the detailed description of the experiments done, following the strategy previously depicted in paragraph 7.1. We did three different classifications: CV vs ALL, EXTRA-GALACTIC vs GALACTIC and SN vs ALL.

7.2.1 Feature space identification

In this paragraph we describe the realization of the first point of our strategy. After the first test experiments, that we will not report here, we focused on the goal to find the best MLPQNA configuration, by selecting a nucleus of five features (Nucleus 1 from now) and working only on classification of Cataclismic Variables class (class 1) versus ALL the others (class 2 - SN + Bl + AGN + Fl, with RRL class removed).
We had 461 CV and 866 ALL the other patterns. The selected features are:

- amplitude;
- beyond1std;
- percent_amplitude;
- skew;
- kurtosis.

These features were selected by the heuristic criteria discussed before. We obtained the results showed in Tab. 7.2.

N	Feature	Eff	Comp1	Comp2	Pur1	Pur2	Cont1	Cont2
6	**linear_trend**	79.70	67.68	86.83	75.28	81.92	24.72	18.08
7	**med_buf_range**	80.83	69.77	86.11	70.59	85.63	29.41	14.36
8	**pair_slope_trend**	83.46	81.63	84.52	75.47	88.75	24.53	11.25

Table 7.3: Table showing the results of the pruning operation for the 1 Hidden Layer configuration. All the values are in percentage of objects. Only the best feature added is reported for every value of N (the number of input features). Class 1 refers to CV (461 patterns), class 2 refers to ALL (866 patterns).

N	Feature	Eff	Comp1	Comp2	Pur1	Pur2	Cont1	Cont2
6	**linear_trend**	78.19	66.67	85.03	72.53	81.14	27.47	18.86
7	**med_buf_range**	82.33	70.93	87.78	73.49	86.34	26.51	13.66
8	**fpr_mid65**	81.95	80.64	82.66	71.43	88.82	28.57	11.18

Table 7.4: Table showing the results of the pruning operation for the 2 Hidden Layers configuration. All the values are in percentage of objects. Only the best feature added is reported for every value of N (the number of input features). Class 1 refers to CV (461 patterns), class 2 refers to ALL (866 patterns).

7.2.2 Cataclismic Variables vs ALL classification

To perform the second point of the strategy previously explained, in the following we used the MLPQNA with the first and third configuration, shown in Tab. 7.2, because they clearly obtained the best results. Then, we decided to perform a pruning of the remaining features, with the aim to minimize effects of correlation and to identify the best group of features for CV classification (CV: class 1 - ALL the others: class 2). We recursively added all the other features to the initial nucleus, one by one, and selected the most significant. In Tab. 7.3 - 7.4 we report the results of the experiments only for the best feature.

At this point, we repeated the experiments for the two best setups (Test 1 and 3) and for the pruning, using the best features selected, as shown in Tab. 7.3 - 7.4, with the prescriptions indicated in paragraph 7.1, using a balanced catalog. In fact, we noticed that in most cases there is a large difference between the values of completeness for the two classes, and this problem could be resolved only by balancing the two classes. In Tab. 7.5 we report the results of these experiments.

We notice that, except in the case of the nucleus 1, there is always a balancing of the completeness values, with respect to the previous experiments, but we did not see any improvement in the total efficiency, probably due to the reduced number of patterns. However, these experiments did not help us to

N	HL	Eff	Comp1	Comp2	Pur1	Pur2	Cont1	Cont2
5 (Nucleus1)	1	72.25	61.96	81.82	76.00	69.83	24.00	30.17
5 (Nucleus1)	2	68.06	58.69	76.77	70.13	66.67	29.87	33.33
6	1	83.77	85.57	81.91	83.00	84.61	17.00	15.38
6	2	80.10	79.38	80.85	81.05	79.17	18.95	20.83
7	1	80.63	82.22	79.21	77.89	83.33	22.10	16.67
7	2	70.68	73.33	68.32	67.35	74.19	32.65	25.81
8	1	79.58	74.04	86.21	86.52	73.53	13.48	26.47
8	2	83.25	84.44	82.18	80.85	85.57	19.15	14.43

Table 7.5: Table showing the results in percentage with the balanced catalog CV vs ALL. All the experiments were done using both the topology with one and two hidden layers. We added the best features selected by the previous pruning for every case (the one showed in Tab. 7.3 - 7.4). Class 1 refers to CV (461 patterns), class 2 refers to ALL (490 patterns).

choose between the 1 or 2 hidden layers topology and, moreover, by repeating some experiments, we noticed a strong variation in the results (of $3-4\%$ in terms of efficiency). We supposed that the reason for this behavior could arise by an ill defined selection of features.
Therefore we tried to select a new nucleus (Nucleus 2) of features, by inspecting their histograms, without dividing the patterns in classes (e.g. by considering the whole catalog), and choosing the most regular ones. In this way we obtained a new nucleus composed by the following features:

- amplitude;
- beyond1std;
- fpr_mid50;
- fpr_mid65;
- std.

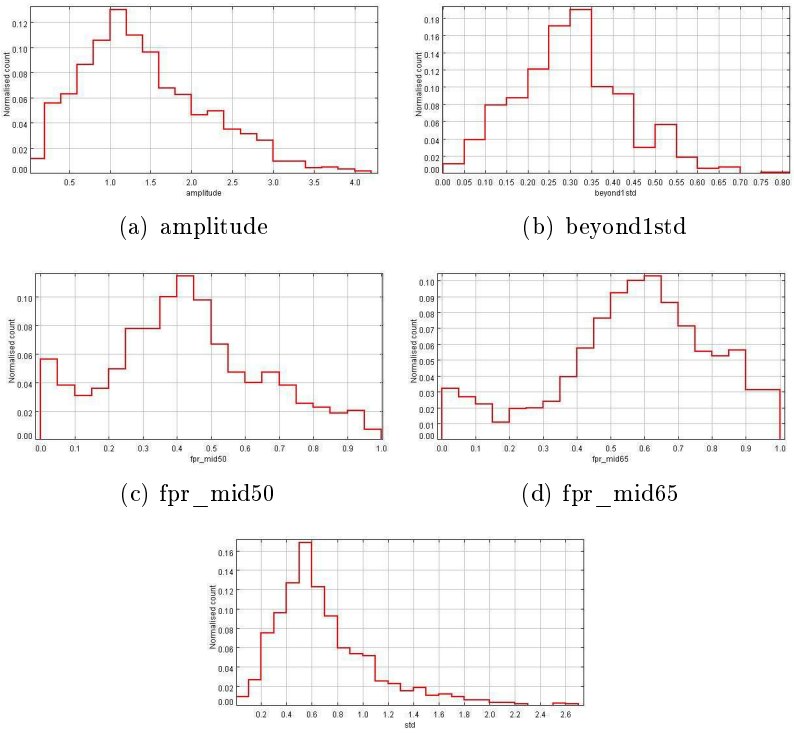
In Fig. 7.8 we show the histograms of the selected features.
Then we performed a new series of experiments using this nucleus. Each experiment was replicated three times, in order to avoid systematic trends. We can observe that the efficiency variation between the experiments is reduced, leading to more stable results, by not using CVAL and by freezing the configuration of the MLPQNA to a single hidden layer.
After that, we performed a pruning with some features (selected for their regularity properties by the histograms), by identifying the best one (for instance, pair_slope_trend). In Tab. 7.6 we report the results.
Also in this case, we decided to repeat the experiments with a balanced catalog. In Tab. 7.7 we report the results obtained.

Nucleus2	**Eff**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	75.94	57.83	84.15	62.34	81.48	37.66	18.52
Exp2	77.07	57.83	85.79	64.86	81.77	35.13	18.23
Exp3	75.19	53.01	85.24	61.97	80.00	38.03	20.00
6 Features	**Eff**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	77.44	64.83	84.00	67.82	82.12	32.18	17.88
Exp2	76.31	61.54	84.00	66.67	80.77	33.33	19.23
Exp3	76.31	70.33	79.43	64.00	83.73	36.00	16.26

Table 7.6: Results in percentage of the experiments on the new nucleus of features, obtained by histogram analysis, and the new pruning to add a sixth feature. Only the best feature is reported (pair_slope_tren). Class 1 refers to CV (461 patterns), class 2 refers to ALL (866 patterns). The configuration adopted is the one of Test 1 of Tab. 7.2.

Nucleus2 bal.	**Eff**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	71.20	67.02	75.26	72.41	70.19	27.59	29.81
Exp2	71.20	56.83	85.57	79.10	66.93	20.89	33.06
Exp3	69.11	65.96	72.16	69.66	68.63	30.34	31.37
6 Features bal.	**Eff**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	67.54	68.82	66.33	65.98	69.15	34.02	30.85
Exp2	70.68	72.04	69.39	69.07	72.34	30.93	27.66
Exp3	70.68	66.67	74.49	71.26	70.19	28.73	29.81

Table 7.7: Table showing the results in percentage with the balanced catalog CV vs ALL for the Nucleus2. In the six features case, we added the best feature selected by the previous pruning (pair_slope_trend). Class 1 refers to CV (461 patterns), class 2 refers to ALL (490 patterns).



(a) amplitude (b) beyond1std

(c) fpr_mid50 (d) fpr_mid65

(e) std

Figure 7.8: Histograms of the features of the Nucleus 2 over all the catalog: (a) amplitude, (b) beyond1std, (c) fpr_mid50, (d) fpr_mid65, (e) std.

In this case there is not a substantial improvement in the results, probably because there is a stronger dependence on this group of features (i.e correlation phenomena between the nucleus used and the classification CV vs ALL). Furthermore, we decided to stop the pruning at this point because we did not notice a substantial improvement by adding the sixth feature, and also CV experiments, trying a new class separation.

7.2.3 EXTRA-GALACTIC vs GALACTIC classification

As fixed in point 3 of the experimental strategy of paragraph 7.1, we decided to proceed with a new classification, in which the classifier had to separate between two new classes: EXTRA-GALACTIC (Bl + AGN - class 1) and GALACTIC (CV + SN + Fl - class 2) objects. From this point we shall use only the configuration with one hidden layer and without using CVAL, which seems to induce instability in the results, probably due to the very limited number of input patterns.
We had 264 galaxies (124 Bl and 140 AGN) and 1063 stars (461 CV, 536

Nucleus1	Nucleus2
amplitude	amplitude
beyond1std	beyond1std
percent_amplitude	fpr_mid50
skew	fpr_mid65
kurtosis	std

Table 7.8: The two main nuclei of features obtained from the previous analysis, that were used in the experiments.

Nucleus1	**Eff**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	83.46	61.40	89.47	61.40	89.47	38.60	10.53
Exp2	84.21	56.14	91.87	65.31	88.48	34.69	11.52
Exp3	84.59	57.89	91.87	66.00	88.89	34.00	11.11
Nucleus2	**Eff**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	85.34	65.57	91.22	68.96	89.90	31.03	10.10
Exp2	87.22	67.21	93.17	74.54	90.52	25.45	9.48
Exp3	89.47	70.49	95.12	81.13	91.55	18.87	8.45

Table 7.9: Results in percentage of the EXTRA-GALACTIC vs GALACTIC experiments for the two nuclei of Tab. 7.8. The experiments are repeated for three times with the same network configuration (the one used for Test 1 in Tab. 7.2, without cross validation). Class 1 refers to EXTRA-GALACTIC (264 patterns) objects, Class 2 refers to GALACTIC (1063 patterns) objects.

SN and 66 Fl). As in the previous experiments, we used the two established nuclei of features and each experiment was replicated three times. We recapitulate the composition of the two nuclei in Tab. 7.8.
The results obtained are reported in Tab. 7.9. We noticed a great improvement in the classification process with this method, so we were encouraged to proceed first with the experiments using a balanced catalog (the results for completeness however show a great difference per class also in this case), and then with a pruning operation on the Decay parameter.
We recall that this is one of the internal model parameters, indicating the weight regularization decay. If accurately chosen, there could be an important improvement of the generalization error of the trained neural network, with also an acceleration of training.
In fact there is a strong dependence of the Decay from the specific case we are considering, the number and type of features, and so on.
For what concerns the experiments with balanced catalog, we obtained the results reported in Tab. 7.10. These results show a balancing in both cases, but in the first case there is a global worsening in the results, while in the second case there is a smaller effect which however still leads to a good

Nucleus1 bal.	Eff	Comp1	Comp2	Pur1	Pur2	Cont1	Cont2
Exp1	74.31	79.49	71.43	60.78	86.21	39.21	13.79
Exp2	77.98	71.79	81.43	68.29	83.82	31.71	16.18
Exp3	73.39	76.92	71.43	60.00	84.74	40.00	15.25
Nucleus2 bal.	**Eff**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	83.49	78.57	88.68	88.00	79.66	12.00	20.34
Exp2	80.73	78.57	83.02	83.02	78.57	16.98	21.43
Exp3	81.65	82.14	81.13	82.14	81.13	17.86	18.87

Table 7.10: Results in percentage of the experiments EXTRA-GALACTIC (class 1 - 264 patterns) vs GALACTIC (class 2 - 280 patterns) for the two different Nuclei, with the balanced catalog, using the same network configuration of Tab. 7.9 (in particular the Decay parameter remains fixed to 0.01).

Nucleus1	Eff	Comp1	Comp2	Pur1	Pur2	Cont1	Cont2
Exp1	86.09	52.63	95.21	75.00	88.05	25.00	11.95
Exp2	86.09	52.63	95.21	75.00	88.05	25.00	11.95
Exp3	86.09	52.63	95.21	75.00	88.05	25.00	11.95
Nucleus2	**Eff**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	89.85	73.77	94.63	80.36	92.38	19.64	7.62
Exp2	89.85	72.13	95.12	81.48	91.98	18.52	8.02
Exp3	89.85	72.13	95.12	81.48	91.98	18.52	8.02

Table 7.11: Results in percentage of the EXTRA-GALACTIC vs GALACTIC experiments for the two nuclei of Tab. 7.8, after the pruning operation on the Decay parameter. Only the best results, with the selected Decay value of 0.5, are reported. The experiments are repeated for three times with the same network configuration. Class 1 refers to EXTRA-GALACTIC (264 patterns) objects, Class 2 refers to GALACTIC (1063 patterns) objects.

result. Again we can identify this behavior in the dependence from the group of features used. Concerning instead the pruning of the Decay parameter, we obtained that the best value is 0.5, with the results for the two nuclei, respectively, reported in Tab. 7.11.

We proceeded again by balancing the classes, with the goal to obtain results with less difference in the completeness values. The results are showed in Tab. 7.12. This shows the improvement, together with the great balancing between classes, for the second nucleus. Therefore in the subsequent experiments we will use the second nucleus, and a 0.5 Decay value, because this is the configuration that gives the best results.

Nucleus1 bal.	Eff	Comp1	Comp2	Pur1	Pur2	Cont1	Cont2
Exp1	76.15	79.49	74.28	63.26	86.87	36.73	13.33
Exp2	77.06	79.49	75.71	64.58	86.88	35.42	13.11
Exp3	76.15	79.49	74.28	63.26	86.87	36.73	13.33
Nucleus2 bal.	**Eff**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	89.91	87.50	92.45	92.45	87.50	7.55	12.50
Exp2	90.82	89.28	92.45	92.59	89.09	7.41	10.91
Exp3	90.82	89.28	92.45	92.59	89.09	7.41	10.91

Table 7.12: Results in percentage of the experiments EXTRA-GALACTIC (class 1 - 264 patterns) vs GALACTIC (class 2 - 280 patterns) for the two different nuclei, with the balanced catalog and the Decay value set to 0.5

Nucleus2	Eff	Comp1	Comp2	Pur1	Pur2	Cont1	Cont2
Exp1	79.70	72.75	84.97	78.09	80.74	21.90	19.25
Exp2	81.20	76.99	84.31	78.38	83.22	21.62	16.77
Exp3	81.20	76.99	84.31	78.38	83.22	21.62	16.77

Table 7.13: Results in percentage of the SN vs ALL experiments for the Nucleus 2 of Tab. 7.8. The Decay parameter is fixed to the value of 0.5. Class 1 refers to SN (536 patterns), Class 2 refers to ALL the others (791 patterns).

7.2.4 Supernovae experiments

Finally, as stated in the fourth point of paragraph 7.1, we continued our work by performing experiments for Supernovae (Class 1), versus ALL other classes (Class 2). In this series of experiments we used the best MLPQNA structure previously fixed. For the first group of experiments we used the nucleus that gave us the best results (Nucleus2 - Tab. 7.8) obtaining the results reported in Tab. 7.13 (we had 536 SN and 791 ALL the other patterns). Proceeding instead with the usual balancing of the catalog, we obtained the results showed in Tab. 7.14.

Nucleus2 bal.	Eff	Comp1	Comp2	Pur1	Pur2	Cont1	Cont2
Exp1	77.37	76.41	78.26	76.41	78.26	23.58	21.74
Exp2	77.37	76.41	78.26	76.41	78.26	23.58	21.74
Exp3	78.28	79.24	77.39	76.36	80.18	23.64	19.82

Table 7.14: Results in percentage of the SN vs ALL experiments for the Nucleus 2 of Tab. 7.8 with a balanced catalog. The Decay parameter is fixed to the value of 0.5. Class 1 refers to SN (536 patterns), Class 2 refers to ALL the others (566 patterns).

This leads to a better balancing of the completeness values but also to the worsening of the overall results. After that, we decided to make a comparison with the results of Donalek [17], which worked on the same types of classifications and with the same catalog, but using different classifiers, i.e. K-nearest-neighbor (KNN) and Decision Trees (DT). Therefore we added again the sixth class containing RR Lyrae (536 SN and 1083 ALL the other patterns) to our catalog, in order to obtain the same catalog as in the article considered. We then performed the experiments with our MLPQNA model, but using some groups of features chosen by various feature selection automated methods, from Donalek [17]. In Donalek [17] the feature selection method are:

- Fast Relief Algorithm (reliefF): simple but efficient procedure to estimate the quality of attributes according to how well their values distinguish between istances;
- Fisher Discriminant Ratio (fdr): it can be used to rank a number of features with respect to their class-discriminatory power;
- Correlation-based Feature Selection (cfs): it is a method which selects features that have low redundancy and results strongly predictive of a single class, considering that features strongly predictive of a class are highly correlated with that class and uncorrelated with each other;
- Fast Correlation Based Filter (fcbf): it is a supervised filter based feature selection algorithm, similar to cfs;
- Multi Class Feature Selection (mcfs): it is an unsupervised feature selection method based on the spectral analysis of the data.

In Tab. 7.15 we report the groups of features used by Donalek [17] for the various method. By assuming the same groups of features as in Donalek [17], we obtained the results reported in Tab. 7.16. We indicated also the column with the misclassified objects, to make a comparison with the results obtained in the article previously indicated (Tab. 7.17).
As it can be seen our results are comparable to those obtained by Donalek [17], especially in the case of the cfs/fcbf and fdr methods (we report just these detailed results here, in the following we will furnish results for all the different methods of feature selection in a comparing table - Tab. 8.1), but we must however note a rather large difference between the completeness values.
Therefore we repeated the usual procedure, obtaining the results reported in Tab. 7.18. We obtained good results in the last case only (fdr group of features). Then we decided to proceed with a new pruning on the Decay parameter for the two best groups previously indicated (cbs/fcbf and fdr methods), with the goal to improve our results. The pruning resulted in an

reliefF	cfs/fcbf	mcfs	fdr
amplitude	beyond1std	max_slope	fpr_mid50
beyond1std	linear_trend	percent_amplitude	linear_trend
fpr_mid80	percent_amplitude	pdfp	pdfp
skew		kurtosis	skew
std			kurtosis
magratio			std

Table 7.15: The different groups of features used for the experiments as in Donalek [17]. These groups were determined via different automated methods for feature selection.

ReliefF	Eff	Misclass	Comp1	Comp2	Pur1	Pur2	Cont1	Cont2
Exp1	77.16	22.84	60.91	85.51	68.37	80.97	31.63	19.03
Exp2	77.47	22.53	66.36	83.18	66.97	82.79	33.03	17.21
Exp3	77.47	22.53	58.18	87.38	70.33	80.26	29.67	19.74
cfs/fcbf	**Eff**	**Misclass**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	84.88	15.12	54.74	97.38	89.65	83.83	10.34	16.16
Exp2	85.49	14.51	56.84	97.38	90.00	84.47	10.00	15.53
Exp3	85.49	14.51	56.84	97.38	90.00	84.47	10.00	15.53
mcfs	**Eff**	**Misclass**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	64.51	35.49	29.46	83.02	47.83	69.02	52.17	30.98
Exp2	64.51	35.49	29.46	83.02	47.83	69.02	52.17	30.98
Exp3	64.51	35.49	29.46	83.02	47.83	69.02	52.17	30.98
fdr	**Eff**	**Misclass**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	84.88	15.12	71.17	92.02	82.29	85.96	17.71	14.03
Exp2	85.80	14.20	71.17	93.43	84.95	86.15	15.05	13.85
Exp3	86.11	13.89	70.27	94.37	86.67	85.90	13.33	14.10

Table 7.16: Results in percentage of the experiments with the different groups of features of Tab. 7.15. The column Misclass indicates the percentage of misclassified object, as the complement of the efficiency parameter. Class 1 refers to SN (536 patterns), Class 2 refers to ALL the others (1083 patterns). The Decay parameter is fixed to the value of 0.5 for all the experiments.

Feature Selection Strategy	KNN Loss	DT Loss
ReliefF	22%	15%
CFS	24%	17%
FCBF	24%	17%
MCFS	32%	19%
FDR	22%	16%

Table 7.17: Percentage of misclassified objects, obtained by Donalek [17], using two different classifiers.

ReliefF bal.	Eff	Misclass	Comp1	Comp2	Pur1	Pur2	Cont1	Cont2
Exp1	76.47	23.53	78.38	74.54	75.65	77.36	24.35	22.64
Exp2	76.47	23.53	78.38	74.54	75.65	77.36	24.35	22.64
Exp3	78.28	21.72	81.08	75.45	76.92	79.81	23.08	20.19
cfs/fcbf bal.	**Eff**	**Misclass**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	78.73	21.27	58.09	97.41	95.31	71.97	4.69	28.02
Exp2	78.73	21.27	59.05	96.55	93.94	72.26	6.06	27.74
Exp3	78.73	21.27	59.05	96.55	93.94	72.26	6.06	27.74
mcfs bal.	**Eff**	**Misclass**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	64.70	35.30	76.58	52.73	62.04	69.05	37.96	30.95
Exp2	64.70	35.30	76.58	52.73	62.04	69.05	37.96	30.95
Exp3	64.70	35.30	76.58	52.73	62.04	69.05	37.96	30.95
fdr bal.	**Eff**	**Misclass**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	83.71	16.29	80.58	86.44	83.84	83.61	16.16	16.39
Exp2	82.35	17.65	80.58	83.89	81.37	83.19	18.63	16.81
Exp3	84.61	15.39	81.55	87.29	84.85	84.43	15.15	15.57

Table 7.18: Results in percentage of the experiments with the different groups of features of Tab. 7.15, after balancing catalogs. The column Misclass indicates the percentage of misclassified object, as the complement of the efficiency parameter. Class 1 refers to SN (536 patterns), Class 2 refers to ALL the others (566 patterns). The Decay parameter is fixed to the value of 0.5 for all the experiments.

overall improvement of the results, by using different best Decay values for the two different group of features, respectively, 0.05 for the cfs/fcbf group and 0.005 for the fdr one. The results are reported in Tab. 7.19 - 7.20.
In the first case, we could notice an improvement of about 2-3%, while in the second case the improvement is smaller. Moreover, by balancing the catalog also in this case and by repeating the experiments, using the respective best values of Decay previously obtained, we found the values reported in Tab. 7.21 - 7.22.
With the aim to understand if the worsening of the overall results is due to the reduction of the number of patterns in the balanced cases, we repeated the experiments for the cfs/fcbf and fdr groups of features with a reduced catalog (randomly cut), containing the same total number of objects for the balanced case (1102 patterns), but without balancing the two classes. So we had 372 SN (class 1) and 730 ALL the others (class 2). We repeated the experiments in the two cases with Decay parameter fixed to 0.5, and with the best respective best values of Decay obtained after the pruning. The results are reported in Tab. 7.23 - 7.24.
We cannot notice a worsening of the results, as in the balanced experiments, but off course the results obtained show the previous strong fluctuation in the values of completeness for the two classes. But we can say that, at least in these cases, there is not a strong dependence from the number of patterns.

Decay 0.5	**Eff**	**Misclass**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	84.88	15.12	54.74	97.38	89.65	83.83	10.34	16.16
Exp2	85.49	14.51	56.84	97.38	90.00	84.47	10.00	15.53
Exp3	85.49	14.51	56.84	97.38	90.00	84.47	10.00	15.53
Decay 0.05	**Eff**	**Misclass**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	87.34	12.66	67.37	95.63	86.49	87.60	13.51	12.40
Exp2	87.65	12.35	67.37	96.07	87.67	87.65	12.33	12.35
Exp3	87.65	12.35	67.37	96.07	87.67	87.65	12.33	12.35

Table 7.19: Best results in percentage of the pruning experiments for the **cfs/fcbf** group of features of Tab. 7.15. The column Misclass indicates the percentage of misclassified object, as the complement of the efficiency parameter. Class 1 refers to SN (536 patterns), Class 2 refers to ALL the others (1083 patterns). Previous results for Decay 0.5 are also reported for convenience of the reader.

Decay 0.5	**Eff**	**Misclass**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	84.88	15.12	71.17	92.02	82.29	85.96	17.71	14.03
Exp2	85.80	14.20	71.17	93.43	84.95	86.15	15.05	13.85
Exp3	86.11	13.89	70.27	94.37	86.67	85.90	13.33	14.10
Decay 0.005	**Eff**	**Misclass**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	86.11	13.89	78.38	90.14	80.55	88.89	19.44	11.11
Exp2	86.11	13.89	78.38	90.14	80.55	88.89	19.44	11.11
Exp3	86.11	13.89	79.28	89.67	80.00	89.25	20.00	10.75

Table 7.20: Best results in percentage of the pruning experiments for the **fdr** group of features of Tab. 7.15. The column Misclass indicates the percentage of misclassified object, as the complement of the efficiency parameter. Class 1 refers to SN (536 patterns), Class 2 refers to ALL the others (1083 patterns). Previous results for Decay 0.5 are also reported for convenience of the reader.

Decay 0.5	**Eff**	**Misclass**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	78.73	21.27	58.09	97.41	95.31	71.97	4.69	28.02
Exp2	78.73	21.27	59.05	96.55	93.94	72.26	6.06	27.74
Exp3	78.73	21.27	59.05	96.55	93.94	72.26	6.06	27.74
Decay 0.05	**Eff**	**Misclass**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	77.37	22.63	64.76	88.79	83.95	73.57	16.05	26.43
Exp2	78.28	21.72	64.76	90.52	86.08	73.94	13.92	26.06
Exp3	79.64	20.36	66.67	91.38	87.50	75.18	12.50	24.82

Table 7.21: Best results in percentage of the pruning experiments for the **cfs/fcbf** group of features of Tab. 7.15, using balanced catalogs. The column Misclass indicates the percentage of misclassified object, as the complement of the efficiency parameter. Class 1 refers to SN (536 patterns), Class 2 refers to ALL the others (566 patterns). Previous results for Decay 0.5 are also reported for convenience of the reader.

Decay 0.5	Eff	Misclass	Comp1	Comp2	Pur1	Pur2	Cont1	Cont2
Exp1	83.71	16.29	80.58	86.44	83.84	83.61	16.16	16.39
Exp2	82.35	17.65	80.58	83.89	81.37	83.19	18.63	16.81
Exp3	84.61	15.39	81.55	87.29	84.85	84.43	15.15	15.57
Decay 0.005	**Eff**	**Misclass**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	82.35	17.65	82.52	82.20	80.19	84.35	19.81	15.65
Exp2	81.45	18.55	75.73	86.44	82.98	80.31	17.02	19.68
Exp3	78.73	21.27	83.49	74.58	74.14	83.81	25.86	16.19

Table 7.22: Best results in percentage of the pruning experiments for the **fdr** group of features of Tab. 7.15, using balanced catalogs. The column Misclass indicates the percentage of misclassified object, as the complement of the efficiency parameter. Class 1 refers to SN (536 patterns), Class 2 refers to ALL the others (566 patterns). Previous results for Decay 0.5 are also reported for convenience of the reader.

Decay 0.5	Eff	Misclass	Comp1	Comp2	Pur1	Pur2	Cont1	Cont2
Exp1	87.78	12.22	70.89	97.18	93.33	85.71	6.67	14.28
Exp2	87.78	12.22	70.89	97.18	93.33	85.71	6.67	14.28
Exp3	87.78	12.22	70.89	97.18	93.33	85.71	6.67	14.28
Decay 0.05	**Eff**	**Misclass**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	87.78	12.22	75.95	94.37	88.23	87.58	11.76	12.42
Exp2	87.33	12.67	77.21	92.56	85.91	88.00	14.08	12.00
Exp3	88.23	11.77	75.95	95.07	89.55	87.66	10.45	12.34

Table 7.23: Results in percentage for the reduced catalog with the **cfs**/**fcbf** group of features of Tab. 7.15. The reduced catalog has the same total number of objects of the balanced case (1102 patterns), but the classes, 372 SN (class 1) and 730 ALL the others (class 2), are not balanced. The column Misclass indicates the percentage of misclassified object, as the complement of the efficiency parameter. The values used for the Decay parameter are the initial one of 0.5 and the best one (0.05) obtained for this group after the pruning operation.

Decay 0.5	**Eff**	**Misclass**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	84.61	15.39	68.42	93.10	83.87	84.90	16.13	15.09
Exp2	82.80	17.20	64.47	92.41	81.67	83.23	18.33	16.77
Exp3	83.71	16.29	67.10	92.41	82.26	84.28	17.74	15.72
Decay 0.005	**Eff**	**Misclass**	**Comp1**	**Comp2**	**Pur1**	**Pur2**	**Cont1**	**Cont2**
Exp1	84.61	15.39	71.05	91.72	81.82	85.81	18.18	14.19
Exp2	83.26	16.74	69.74	90.34	79.10	85.06	20.89	14.93
Exp3	82.35	17.65	71.05	88.27	76.06	85.33	23.94	14.67

Table 7.24: Results in percentage for the reduced catalog with the **fdr** group of features of Tab. 7.15. The reduced catalog has the same total number of objects of the balanced case (1102 patterns), but the classes, 372 SN (class 1) and 730 ALL the others (class 2), are not balanced. The column Misclass indicates the percentage of misclassified object, as the complement of the efficiency parameter. The values used for the Decay parameter are the initial one of 0.5 and the best one (0.005) obtained for this group after the pruning operation.

Chapter 8

Conclusions

As discussed in the introduction, Time Domain Astronomy, or TDA, is among the most challenging and rapidly evolving fields of Astrophysics. TDA, in fact, is crucial both to better understand old phenomena (such as stellar variability, active galactic nuclei, supernovae) and to discover new ones. From a practical point of view, TDA presents formidable technological challenges which have already changed, and even more so will do in the future, the methods, problems and goals of everyday astronomical practice. As we said in the introduction, the future of observational astrophysics will be performed mainly by extracting the useful information from huge datasets produced by a new generation of instruments. As a results of this changing scenario, astronomers will need to automatize as much as possible the procedures for data analysis and for the interpretation of the data. This thesis adopted this new perspective and focused on the use of a neural network to classify transients as a first step towards producing a framework where different classifiers will work in collaborative manner on the same data to obtain a classification of variable objects reliable, accurate and reproducible.

This thesis made use of the DAMEWARE infrastructure, that represents a crucial technological improvement in the construction of an environment where everyone can work on data, with powerful instruments, in a simple, standardized and accessible way.

We performed three types of experiments (all binary classifications): Cataclismic Variables versus all other classes, EXTRA-GALACTIC (AGN + Blazars) versus GALACTIC (Supernovae + Cataclismic Variables + Flare stars), Supernovae versus all other classes.

These experiments were done with the aim to test different types of classifications to verify the behavior of the neural network on the different classes involved. We also varied the groups of features used, to analyze the dependence of the classification performances on them. Finally, with the last series of experiments, Supernovae versus ALL, we compared our results to those obtained by Donalek [17], who worked on the same dataset but using

different classifiers and different automated methods of feature selection.
More in detail, the results obtained during the first set of experiments, (Cataclismic Variables vs ALL), even after the pruning operation, produced results which are the worst ones, probably due to the wrong balancing of the two classes. However these experiments allowed us to define a good topology for the MLPQNA and provided some nuclei of features which became the starting point for the subsequent work. This operation also resolved the problem of the fluctuations between the training and test phases, as it is showed in the histograms from Fig. 7.1 to Fig. 7.7.
In the EXTRA-GALACTIC vs GALACTIC experiments, despite the fact that the patterns were not balanced, the results were much better, very likely because the distinction between the two classes has a deeper physical meaning, which was reflected in different temporal behaviors.
Finally, in the experiments regarding Supernovae vs ALL, using the same groups of features as in Donalek [17], we noticed an improvement with respect to the experiments done with our selection of features. The results obtained with two of these groups of features are comparable and we decided to try to improve them by a pruning on the Decay parameter. The pruning lead to a substantial improvement in the first case, a smaller one in the second case; conflicting results which were caused by a dependence of the best Decay value on the different groups of selected features.
In Tab. 8.1 we compare our results and those obtained in [17]. These encouraging values show that the MLPQNA can be considered a good tool for transient classification, especially in view of the fact that this result must be considered only preliminary and can be largely improved in the future.
Furthermore, we noticed that, by balancing the classes in the training set, in most cases, we obtained a decreasing difference between the completeness values for the two classes, but also a general worsening of all the parameters with respect to the not-balanced case. This is likely due to the reduced number of patterns after the balancing. In some cases, by repeating the same experiments, we noticed also an increase in the fluctuations. We also performed experiments with a reduced, but non balanced, catalog, with the aim to verify whether there was a dependence of the results on the number of patterns in order to try to understand the worsening of the overall results. These experiments, performed only for our best cases (cfs/fcbf and fdr), disclaimed this hypothesis, at least in the case SN vs ALL. This aspect, however, will require further analysis.
Despite of this fact, we can consider as the most robust the results obtained by balancing the training data, which produced better classification in the two classes (as demonstrated by the reduced difference between completeness values). The balance, though not improving our results, allowed us to evaluate the strong dependence of the classification process and of the MLPQNA behavior on the number of patterns and on the groups and number of features used. The results obtained in the balanced case are reported

Feature selection strategy	KNN (%)	DT (%)	MLPQNA (%)	MLPQNA bal. (%)	MLPQNA red. (%)
reliefF	22	15	23	23	/
cfs/fcbf	24	17	12	22	12
mcfs	32	19	35	35	/
fdr	22	16	14	19	16

Table 8.1: Comparison between the results obtained by Donalek [17] with two different classifiers and those obtained by us with the MLPQNA and using the same group of features. The percentage indicates the number of misclassified objects (average on the three experiments) as a complement of the efficiency. We report also the balanced results, but it is clear that these are not comparable with the results of [17], that are not balanced, and the results with the reduced catalog, as reported in Tab. 7.23 - 7.24. We recall also that in the two cases of reliefF (Fast Relief Algorithm) and mcfs (Multi-Class Feature Selection) groups of features, the results are for a Decay value of 0.5, because we did not perform the pruning operation. The reduced case, instead, was done only for cfs/fcbc (Correlation-based Feature Selection/Fast Correlation Based Filter) and fdr (Fisher Discriminat Ratio) cases (our best cases), to verify the existence of a dependence from the number of patterns that could explain the worsening of the overall results in the balanced cases.

in Tab. 8.1, though they cannot be considered comparable with the results obtained in Donalek [17], because these were not performed using a balanced training set.
The final purpose of the classification process with neural networks, together with the comparison with other methods, as we said before, must be seen in the framework depicted by the classification schema of Fig. 8.1. We focused just on a small subset of this classification problem, exemplified in Fig. 8.1, which will require a much more complex hierarchic workflow. Starting from raw data, it aims on achieving a precise classification, using different methods and feature selection algorithms, that, also with the help of external knowledge, in the next future, will make possible to realize a complete automatized classification process.

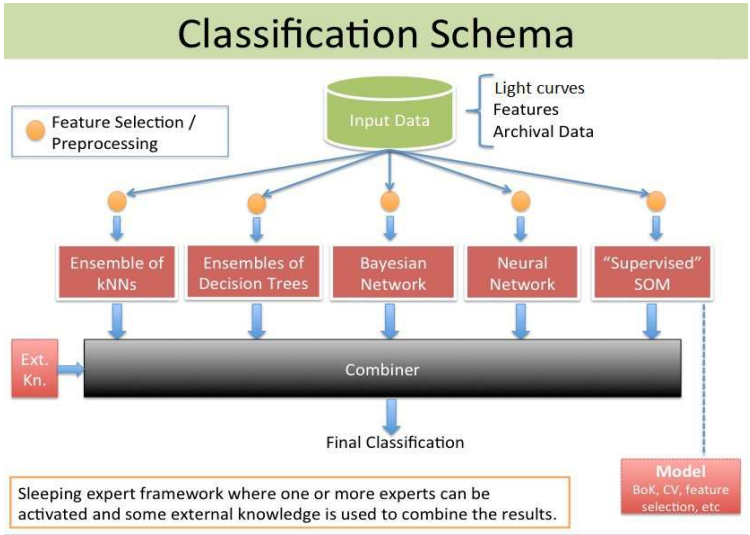


Figure 8.1: Classification schema with different classifiers (Donalek [18]). A weighted average of the results obtained by different methods can represent a simple way to obtain the final classification, with the help of some external knowledge.

Acknowledgments

There are many people that, directly or indirectly, have made possible the realization of this thesis work. First of all, I have to thank Prof. Giuseppe Longo, my teacher and supervisor, who is teaching me not only Astrophysics, but how to become an Astronomer, thus realizing my childhood dream. Then, special thanks are due also to my other two supervisors, Dr. Massimo Brescia and Dr. Stefano Cavuoti, for the enormous help given to me during the work and for their endurance in difficult moments and when I did mistakes, and to Dr. Ciro Donalek, for his availability and clarity in giving me suggestions to improve my work.
Surely, I cannot forget all my colleagues and friends in the Astrophysics Laboratory, Alfonso Nocella, Giuseppe Riccio, Mauro Garofalo, Demetra De Cicco, Civita Vellucci, Virgilio De Stefano, and so on, who were always ready to give me help and precious suggestions, and with their friendship made my days easier.
Finally, the most special thanks are to my parents, to which this thesis is dedicated. It is thanks to their efforts and sacrifices that today I am reaching this important goal. I can only hope to continue to make them proud in the next future.

Bibliography

[1] Baade W., Zwicky F., "Supernovae and Cosmic rays" Physical Review 45, 138, 1934b

[2] Bernardini E., "Astronomy in the Time Domain", Vol. 331, Science, February 2011

[3] Breiman L., Machine Learning, 45(1), 5, 2001

[4] Brescia, M., et al., "DAME: A Distributed Web Based Framework for Knowledge Discovery in Databases", Mem. SAIt Suppl, 13, 56, 2009

[5] Brescia M., Cavuoti S., D'Abrusco R., Laurino O., Longo G., V International Workshop on Distributed Cooperative Laboratories: "Instrumenting the Grid", in Remote Instrumentation for eScience and Related Aspects, F. Davoli et al. (eds.), Springer:NY, 2011

[6] Brescia M., Cavuoti S., Paolillo M., Longo G., Puzia T., "The Detection of Globular Clusters in Galaxies as a data mining problem", MNRAS, 421, issue 2, 1155, 2012a

[7] Brescia M., Longo G., Castellani M., et al., Mem. SAIt Suppl, 19, 324, 2012b

[8] Brescia et al., "The detection of globular clusters in galaxies as a data mining problem", Monthly Notices of the Royal Astronomical Society, 421, 2, 1155-1165, 2012

[9] Brescia et al., "Photometric redshifts for Quasars in multiband Surveys", ApJ, 772, 2, 140, 12 pp., 2013

[10] Brescia M., Cavuoti S., Longo G., et al., "DAMEWARE: A Web Cyberinfrastructure for Astrophysical Data Mining", Publications of the Astronomical Society of the Pacific Vol. 126, No. 942, pp. 783-797, August 2014

[11] Butler N.R., Bloom J.S., "Optimal time-series selection of Quasars", AJ, 141, 93, 2011

[12] Cavuoti S., Brescia M., Longo G., Garofalo M., Nocella A., 2012, "DAME: A Web Oriented Infrastructure for Scientific Data Mining and Exploration", Science - World Scientific Publishing Co. Pte. Ltd., ISBN 9789814383295, pp. 241-247, 2012

[13] Chow-Choong Ngeow, Shashi M. Kanbur, "The linearity of the Wesenheit function for the Large Magellanic Cloud Cepheids", Monthly Notice of the Royal Astronomical Society, 360, S. 1033-1039, 2005

[14] Cox J.P., "Theory of stellar pulsation", Princeton, N.J.: Princeton Univ. Press., 1980

[15] Debosscher J. et al., "Automated supervised classification of variable stars", AA, 475, 1159, 2007

[16] Djorgovski et al., "The Palomar-Quest Digital Synoptic Sky Survey", Astron. Nach., 329, 263, 2008

[17] Donalek et al., "Feature Selection Strategies for Classifying High Dimensional Astronomical Data Sets", ArXiv:1310.1976, 2013

[18] Donalek C., Djorgovski G., Longo G. et al., "Machine Learning Techniques in the Time Domain", Naples, May 2013

[19] Drake A.J. et al., "First Results from the Catalina Real-time Transient Survey", ApJ, 696, 870, 2009

[20] Dubath P., Rimoldini L., Suveges M. et al., "Random forest automated supervised classification of Hipparcos periodic variable stars", Mon. Not. R. Astron. Soc., May, 651, 2011

[21] Dubath P., "Hipparcos Variable Star Detection and Classification Efficiency", Astrostatistics and Data Mining, Springer Series in Astrostatistics, Volume 2. ISBN 978-1-4614-3322-4. Springer Science+Business Media New York, p. 117, 2012

[22] Eddington A.S., "Stars, Gaseous, On the pulsations of a gaseous star", MNRAS, Vol. 79, p.2-22, 1918

[23] Eyer L., Mowlavi N., 2007, "Variable Stars across the HR Diagram", JPhCS 118, 2010

[24] Grindlay J. et al., "Opening the 100-year window for time-domain astronomy - New Horizons in Time-Domain Astronomy", Proceedings of the International Astronomical Union, IAU Symposium, Volume 285, p. 29-34

[25] Harwit M., "The Growth of Astrophysical Understanding", Phys. Today, 56, 38, 2003

[26] Hertzprung E., "Uber die Sterne der Unterabteilung c und ac nach der Spektralklassifikation von Antonia C. Maury", Astronomische Nachrichten 179 (4296): 373-380, 1909

[27] Hey T., Tansley S., Tolle K., "The Fourth Paradigm", Microsoft Research

[28] Law et al., "The Palomar Transient Factory Survey Camera: 1st Year Performance and Results", SPIE 7735, 2010

[29] Law et al., "The Palomar Transient Factory: System Overview, Performance and First Results", PASP 121 1395L

[30] Longo G., Brescia M., "Time Domain Astronomy: a new frontier of astronomy, Interdepartmental Physics-Mathematics lecture, Department of Physics, University Federico II, Napoli, June 7, 2012

[31] McCulloch W.S., Pitts W., Bullettin of Mathematical Biophysics 5:115-133, 1943

[32] Minkowski R., "Spectra of Supernovae", Publications of the Astronomical Society of the Pacific, Vol. 53, No. 314, pp. 224-225, August 1941

[33] Minsky M.L., Papert S.A., "Perceptrons", Cambridge, MA: MIT Press, 1969

[34] Phillips M.M., "The absolute magnitudes of Type IA supernovae", ApJ, Part 2 - Letters (ISSN 0004-637X), vol. 413, no. 2, p. L105-L108, 1993

[35] Richards J.W. et al., "Variable star classification", ApJ, 733, 1; arXiv:1101.1959, 2011

[36] Rosenblatt F., "Psychological Review", Vol 65(6), 386-408, November 1958

[37] Russell H.N., "Relations Between the Spectra and Other Characteristics of the Stars", Popular Astronomy 22: 275-294, 1914

[38] Sako M. et al., "The Sloan Digital Sky Survey-II Supernova Survey: Search Algorithm and Follow-up Observations", AJ, 135:348-373, January 2008

[39] Scargle J.D., "Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data", ApJ, 263, 835, 1982

[40] Schwarzenberg-Czerny A., "On the advantage of using analysis of variance for period search", MNRAS, 241, 153, 1989

[41] Spearman C., "The proof and measurement of association between two things", Amer. J. Psychol. 15: 72âĂŞ101, 1904

[42] Stetson P.B., "On the Automatic Determination of Light-Curve Parameters for Cepheid Variables", Publ. Astron. Soc. Pacific 108, 851, 1996

[43] Zechmeister M., Kurster M., "The generalised Lomb-Scargle periodogram. A new formalism for the floating-mean and Keplerian periodograms", AA, 496, 577, 2009

List of Tables

List of Figures