Mining the SDSS Database: a Wavelet Approach for Parameterisation and Classification in Support of Novelty Detection in the Astronomy Domain

Marco Grimaldi, Brian Espey, Pádraig Cunningham

04/01/2006

Abstract

In recent years, the advent of large telescopes, efficient linear detectors and high-speed computers has enabled the possibility of observing many astronomical objects simultaneously and obtaining a multi-wavelength description of the sky. By dedicating a telescope to a program of observations in which many images or spectra are acquired simultaneously, vast databases are built up in a matter of a few years.

This new approach to astronomy requires dedicated tools to solve the issues that arise when accessing, analysing and categorising such amounts of data. Data-mining and knowledge discovery techniques find natural application in this scenario.

In this work we discuss the application of machine learning to this domain. We address the task of categorising or labelling the spectra and the associated problem of novelty detection - that is to identify spectra that are significantly different from the known categories. The objective is to train classifiers from a small set of labelled examples and to use those classifiers to automatically categorise the unlabelled examples and highlight novel objects. Thus the key objective is to produce a classifier that generalises well to unlabelled examples.

We present a system that is able to generalise almost perfectly and that provides a novelty detection mechanism that can discover unknown - never seen before - objects. This system has two main parts, a signal processing component that extracts features from the spectra and a machine learning component that uses these features to classify the astronomical objects they represent. This classifier can also highlight novel objects that are promising candidates for further analysis.

A large part of this report is dedicated to the description of a number of signal processing techniques based on the wavelet transform. In fact, the development of the machine learning system is based on the definition of a new set of descriptors that capture key information embedded in the emission spectra of astronomical objects. In order to minimise error due to the intrinsic noise in the spectra *de-spiking*, *de-noising* and *smoothing* procedures are applied automatically to the raw spectra before the descriptors are encoded. Moreover, the use of a multi-resolution approximation such as the wavelet transform enables us to extract parameters related to the emission and absorption lines and easily approximate the continuum component of the spectra.

This work demonstrates that the proposed descriptors (encoded from the signal and embodying physical aspects of the emission spectrum) well represent the domain and that different classifiers based on the k-NN and trained using such features can learn to classify objects correctly. The predictive power of the parameterisation proposed is assessed using three different classifiers: the standard k-Nearest Neighbour (k-NN), a round-robin ensemble and a feature sub-space based ensemble. In order to improve the generalisation performance of the classifiers, a wrapper-based feature selection approach is applied. Our experiments show that feature selection works well in the case of simple k-NN, while results for the feature sub-space ensemble are not conclusive. However, our analysis shows clearly that the proposed descriptors provide a generalisation accuracy in classification that compares well with that obtained using either the original spectrum or a (Principle Components Analysis) PCA based approach. Our work demonstrates that the overhead caused by the signal processing techniques and the parametrisation provides a set of descriptors highly capable of modelling the problem at hand and enabling the possibility of

providing an explanation of the classification. The descriptors correspond to the the main characteristics of the spectrum, thus they directly reflect physical phenomena: the position of a peak is a direct consequence of the typology of chemical elements composing the celestial object; the shape of the continuum may provide information about the presence of a black body radiator and some insight about the intervening material between the source and the observer.

In this work we propose a modification of the k-NN classifier to enhance its sensitivity to the local properties of the feature space. The use of the k-NN is motivated by its ability to learn a local solution of the problem and thus enabling the evaluation of a threshold for the detection of new - never seen before - objects through an analysis of the local properties of the feature space. Our analysis demonstrates that the proposed modification further enhance the sensitivity of the k-NN to local regularities of the feature space: it provides better a recognition-rate of unknown spectra. The experiments show that both the approaches tested (the one based on the proposed descriptors and the second one on PCA) benefit from the proposed modification. Interestingly, the adoption of the proposed features together with a feature sub-space based ensemble and the k-NN modification appears as the best technique for novelty detection, showing results that are open to further refinement and development.

Contents

1 Introduction

In recent years, the advent of large telescopes, efficient linear detectors and high-speed computers has enabled the possibility of observing many astronomical objects simultaneously and obtaining a multi-wavelength description of the sky. By dedicating a telescope to a program of observations in which many images or spectra are acquired simultaneously, vast databases are built up in a matter of a few years. Moreover, large-scale programmes (2dF QSO survey, Sloan Digital Sky Survey, etc.) have been undertaken worldwide in order to provide data for the entire astronomical and astrophysical research community.

The data used in this work are extracted from the Sloan Digital Sky Survey (SDSS). The SDSS project [1], when completed, will have mapped systematically one-quarter of the entire sky, producing a detailed image (and spectrum) of it and will determine the positions and magnitudes of more than 100 million celestial objects. The latest data release (still incomplete) incorporates images that amount to 7.5TB of information, and SQL catalogues that amount to 3TB in total.

This new approach to astronomy requires dedicated tools to solve the issues that arise when accessing, analysing and categorising such amounts of data. Data-mining and knowledge discovery techniques find natural application in this scenario. While there is a role for unsupervised machine learning techniques to organise and cluster the data, we are concerned with supervised machine learning techniques that will categorise the data and will highlight novel objects.

The work presented here has two main contributions. The first is to provide a machine learning system able to almost perfectly model the problem at hand (and automatically label unseen data) and second to provide a system able to use the stored knowledge to discover unknown - never seen before - objects.

This system is based on the definition of a new set of descriptors that capture key information embedded in the emission spectra of astronomical objects. It is our belief that the use of descriptors embodying the semantics of the problem domain will enhance the comprehensibility of the system. Thus, it will facilitate its adoption by experts of different fields (e.g. astronomers and astrophysicists). In fact, the output of the system needs to be easily understood in terms of semantically meaningful differences between already labelled spectra and the spectrum of the unknown object submitted as query. The ability to explain the output of the system in terms of emission/absorption features and differences in the behaviour of the continuum (section 2) is a key property of the proposed system. This property strongly differentiates the proposed approach from the ones based on principle component analysis (PCA). PCA is a dimensionality reduction technique used to transform the raw input spectrum in a new smaller set of features. The new set of descriptors (principal components) is a linear combination of the raw parameters. It has been demonstrated in many studies in the literature that this approach is extremely powerfull and allows the representation of an input spectrum through a few principal components. However, the main drawback is the lack of interpretability as the principle components do not have any meaning for a domain expert.

The feature extraction process we present here is based on a wavelet transform (WT) of the signal. The key feature of the WT is the ability to provide a multi-resolution approximation of a given input signal. Given that different frequencies need different time support in order to be analysed, the wavelet

transform is able to decompose recursively an input signal into different sub-bands (or scales). The multi-resolution approximation (section 3 and 4) enables us to extract parameters related to the emission and absorption lines (4.4) and easily approximate the continuum component of the spectra (4.3). Moreover, the WT is used to remove imperfections from the signal: the spectrum of a celestial object is generally affected by noise and by presence of spikes that may influence its parameterisation. In this context, we make use of the WT (section 4.1 and 4.2) in order to clean the spectra an remove imperfections that may compromise the performace of a classifier.

In order to asses the predictive power of the parameterisation proposed, three different classifiers based on the k-NN are evaluated (section 5). The standard k-NN, a round-robin ensemble and a feature sub-space based ensemble are trained using both the raw spectra and the features proposed. Our experiments (section 8) demonstrate that the features provide a generalisation accuracy that compares well with against that obtained using either the whole spectrum or a PCA based approach. The adoption of the proposed features provide an almost perfectly learnable description of the problem. Given their relative small number (compared to the adoption of the full spectrum as descriptors - about 3800 features), the compact description enables us to test feature selection (section 5.2) on the different classifiers to boost their accuracy. Moreover, the adoption of the proposed features together with classifiers based on the k-NN provides the ability to provide a straightforward explanation of the prediction. The output of the system can be described as function of the neighbourhood of the query and hence as a function of meaningful parameters.

The k-NN technique is enhanced (section 6) and used to implement a novelty detection strategy. The use of the k-NN is motivated by its ability to learn a local solution of the problem and thus enabling the evaluation of a threshold for detection of new - never seen before - objects through an analysis of the local properties of the feature space. The proposed modification is aimed to further enhance the sensitivity of the k-NN to local regularities of the feature space and hence obtain better recognition-rate of unknown spectra.

Starting with the next section, we describe the problem domain, the signal processing techniques adopted and the machine learning technique used. In section 7 we present in detail the dataset used in our experiments, while section 8 describes in detail the results of our experiments. Section 8 provides also a comparison of the results obtained with the proposed spectrum parameterisation and a feature transformation technique based on PCA.

2 Characterisation of astronomical spectra

Astronomical spectra consist of two main components: a continuum, that varies smoothly over many channels, and emission and absorption lines, that appear as abrupt variations over a small number of channels. Continuum features and peaks corresponding to absorption and emission lines are a direct expression of different physical phenomena.

The continuum part of the spectrum is due to *black body* radiation. A black body radiator is a theoretical object that is totally absorbent to all thermal energy that falls on it. As it absorbs energy it heats up and re-radiates the energy as electromagnetic radiation. In outline, the continuum emission of the objects

presented here can be divided into two main types: thermal emission from hot, dense plasma (stellar and galaxy continua), and non-thermal emission from QSOs. As can be seen from Figure 1, the continuum emission from stars has a characteristic curved shape (crudely following a blackbody energy distribution) with a peak which depends on the temperature of the source: peaking at low wavelength (towards the blue end of the spectrum) for hot stars, and towards long wavelengths (the red end of the spectrum) for cooler objects. This smooth continuum is modified by the presence of material in the outer layers of the star, generally resulting in a series of absorption features as is seen in Figure 1. Since a galaxy consists of a combination of stars and gas, the spectrum is a composite of the emission from these components, with a continuum which is the average of stars of different temperatures modified by absorption due to cooler gas between the stars. The combination of stellar and gaseous emission introduces a roll-off to the shorter wavelengths, and sometimes also emission lines when a sufficient amount of heated gas is present (Figure 1).

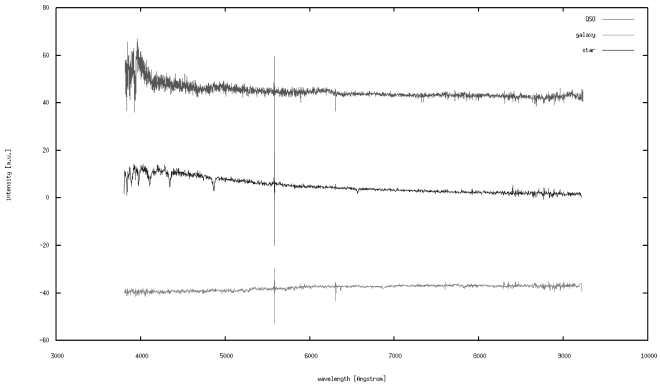


Figure 1: Examples of; QSO spectrum (top), galaxy (center) and star (bottom). These spectra a plotted in arbitrary units on the y-axis so that they can be shown in a single graph.

Although QSOs are a form of active galaxy, the emission from material close to the central black hole in these objects is intense, being equal to, or larger than, the sum total of the stellar component. As a result the continuum (Figure 1) looks very different to that of a normal galaxy, with a shape that is largely determined by energetic processes. The resulting continuum has an energy distribution which is a power-law when plotted against wavelength. The intense continuum radiation in these objects ionises a considerable amount of the galactic gas, generating strong emission lines which are superimposed on the power-law continuum.

Absorption lines may be due to intrinsic absorption of the source itself or to the presence of clouds of gases surrounding the astronomical object. In the latter case absorption lines appear because of the quantum mechanical interactions between electrons orbiting around the gas atoms and photons emitted from the astronomical source.

Emission lines, as in the case of absorption lines, are due to the interaction between light photons and electrons orbiting around atoms. When the intervening material between the astronomical source and the observer absorbs energy,

it re-emits the energy in form of photons of specific frequency and wavelength.

Astronomical objects produce different types of spectra depending on their typology, thus their spectrum is one means of identifying its class.

Of particular relevance in characterising real world signals is the presence of noise. Emission spectra have different signal to noise ratios depending on various factors, e.g. the apparatus setup, sky conditions and apparent brightness of the source. The effect of noise is to hide some properties of the spectrum, and thus to make its parameterisation and categorisation problematic. Moreover the spectra may show the presence of spikes and glitches. Presence of spikes in emission spectra may be due to the interaction of cosmic rays with the acquisition apparatus, while glitches are due to particular condition of the sky during the acquisition of the spectra or depending on the apparatus setup itself. Topically, spikes appear in the spectrum as sharp positive peaks with an intensity exceeding in value any other spectral feature. Glitches (as in the case of SDSS spectra) appear in the spectrum as a sequence of sharp negative/positive peaks exceeding in value any other spectral feature.

The parameterisation of a spectrum requires features able to describe with sufficient accuracy its peculiar characteristics: emission lines, absorption lines, continuum behaviour. This requirements can be met by performing time-frequency analysis of the spectrum.

While astronomical spectra are expressed as a function of wavelength, time-series analysis considers input signals expressed as a function of time. However, without any loss of generality and in perfect analogy with time-series analysis, we can consider a spectrum as *an ordered sequence of values of a variable at equally spaced time* (wavelength) *intervals*. Thus, given the above stated change of variables ($time = wavelength$), a frequency analysis of the input signal (spectrum) can provide different information: low frequencies bring information about the continuum; mid frequencies about emissions and absorption lines; high frequencies about the noise in the signal [5, 6]. In the following, if not otherwise specified, we will refer to the spectrum (or input signal) as a *function of time*. Thus, the term "frequency" will reflect its meaning in the time-series analysis context.

In this work we address the problem of de-spiking, de-noising and parameterisation of the input signal using a wavelet decomposition. In section 3 we briefly introduce the wavelet transform and the algorithm applied to perform the decomposition.

3 The wavelet transform

The wavelet transform (WT) is a well known signal analysis technique [18, 3] that has been applied in various research fields, such as compression [5, 6, 26], de-noising, statistics, time series analysis, image processing, signal analysis and data mining (e.g.: [5, 6, 28, 17, 7]). The key feature of the WT is the ability to provide a *multi-resolution* approximation of a given input signal. Given the fact that different frequencies need different time support in order to be analysed, the wavelet transform recursively decomposes an input signal into different sub-bands (or scales). The time resolution of each sub-band matches the frequency constraints: low time resolution (big time support) for low frequencies and high time resolution (small time support - pixel spacing) for high frequencies.

The continuous wavelet transform projects a real function $f(t)$ over a prototype function $\psi(t)$ (the mother wavelet) according to the following equation:

$$W(s,r) = \int f(t)\frac{1}{\sqrt{s}}\psi(\frac{t-r}{s})dt \tag{1}$$

where s characterises the scale and r the translation factor. The continuous wavelet transform is a convolution of the data sequence with a scaled and translated version of the mother wavelet. In analogy with the windowed Fourier transform, the mother wavelet is a function (continuous in both time and frequency) that serves as the analysing window. Typically, the form of the mother wavelet is selected based on the signal processing problem domain.

In the discrete domain, two approaches can be adopted in order to compute the wavelet transform: the octave-band and the wavelet packet decomposition. In this work we focused on the octave band decomposition. This is achieved using a pyramidal algorithm [18]: the original signal ($f[t]$) is decomposed using a pair of *quadrature mirror filters* into two different approximations containing respectively the high and the low frequency components of the signal. The two approximations are respectively indicated as $D_1[t]$ (containing the details of the spectrum - namely the high frequency components) and $A_1[t]$ (containing the low frequency component). At each step in the pyramid the low component of the signal is further decomposed. Thus the decomposition of $A_1[t]$ provides two other approximations, $A_2[t]$ and $D_2[t]$. The octave band decomposition provides a series of signal approximations where the frequency resolution is *a priori* defined by the decomposition rule. The number of different approximations is ruled by the chosen number (j_0) of decompositions performed.

Two different approaches [18] have been proposed to calculate the octave-band transform of a given discrete signal: the *dyadic* deconvolution and the *à trous* algorithm. In this work we adopt a discrete wavelet decomposition using the *à trous* algorithm. It offers a number of advantages over the *dyadic* approach [27];

1. the transform is invariant under translation,
2. the transform is carried out in direct space: artefacts due to the periodization do not occur,
3. the evolution of the transform from one scale to the next can be easily followed.

Adopting the *à trous* algorithm and the octave-band decomposition schema, the original signal ($f[t]$) can be expressed as the sum of the wavelet coefficients at the different decomposition levels ($D_j[t]$) and the smoothed array ($A_{j_0}[t]$) containing the lowest frequencies of the signal:

$$f[t] = A_{j_0}[t] + \Sigma_j D_j[t] \tag{2}$$

where j varies between (inclusive) 1 and j_0 - the selected maximum number of decompositions applied.

3.1 Spectral analysis

The work of Stark et al [27] presents an interesting signal processing technique to analyse spectra. Their method is based on an octave-band decomposition employing the *à trous* algorithm and a cubic-spline mother wavelet. The authors show clearly that the wavelet decomposition can be applied to the analysis of continuum sources that show superposed interstellar/circumstellar absorption and emission band that are shallow and broad [27].

In this work we apply the same approach to characterise the emission spectra of QSOs, stars and galaxies. The key idea is to apply the octave-band decomposition and obtain different signal approximations centred at different frequencies. This multi-resolution analysis allows us to selectively extract information about peaks with different widths: the signal approximation containing high frequencies (e.g. $D_1[t]$ and $D_2[t]$) of the input signal provides information about the presence of sharp peaks in the spectrum. By selectively analysing less detailed approximations (e.g.$D_4[t]$, $D_5[t]$, ...), information about the presence of broad peaks can be encoded. Moreover, it is possible to extract the continuum component of the spectrum directly from the wavelet transform as the smoothed array $A_{j_0}[t]$ containing the lowest frequencies of the signal.

Applying a multi-resolution decomposition of the signal guarantees a characterisation of the signal more complete than a simple: "pick the N most intensive peaks in the spectrum". However, the wavelet decomposition and its parameterisation has some drawbacks, the most noticeable one is the need of identify a suitable level of decomposition j_0. The value for j_0 can not be *a priori* determined and it is domain dependent. In section 4 we discuss the methodology applied to estimate the value of j_0 for astronomical spectra characterisation.

Moreover, it is worth noting that a multi-resolution parameterisation of the input signal is somewhat redundant, the presence a broad and intense peak in the spectrum will appear on different sub-bands. This fact is not necessarily an issue. As we will discuss in section 5, the presence of redundancy in a set of descriptors can be used to build an efficient ensemble of classifiers.

A third draw-back in using a wavelet transform for analysis is related to the existence of secondary peaks in the different sub-bands. As discussed in [27], the wavelet analysis of a spectrum that contains a strong emission (absorption) band may indicate the presence of two weaker absorption (emission) bands symmetrically displaced form the location of the original band. If this behaviour is not desirable for analysis of certain physical phenomena, in the context of classification it can be considered less problematic. The detection of "fake" absorption lines will generate descriptors that can be considered as irrelevant or redundant for classification. In both cases, as we will discuss in section 5, some machine learning methods - such as feature selection and/or ensembles of classifiers - may be applied to resolve this issue.

4 Signal processing and feature extraction

In this section we discuss how the wavelet transform is applied to solve the problem of de-noising and de-spiking astronomical spectra. Moreover we present the features that are extracted from the input signal and used for classification.

4.1 De-noising

In signal processing, the application of the wavelet transform for de-noising is a well known strategy [5, 6, 16, 22, 26, 9, 25, 8]. The general idea is to apply the wavelet decomposition to the input signal and filter the wavelet coefficients according to a pre-defined policy. Once the wavelet coefficients have been filtered, the inverse transform is applied and the de-noised signal is obtained. This simple process guarantees a significant reduction of the noise, preserving the edge sharpness in the signal [29]. Donoho et al [5, 6] demonstrated that applying the *wavelet shrinkage* (or *soft-thresholding*) method for de-noising has a number of excellent properties, such as being near optimal in a min-max sense and having a better rate of convergence with respect of the *hard-thresholding* method [29]. The hard-thresholding method chooses all the wavelet coefficients that are greater than a given threshold τ and sets the others to zero:

$$w_j(t) = \begin{cases} t & if\ |t| \geq \tau \\ 0 & otherwise \end{cases} \tag{3}$$

The soft-thresholding methods shrinks by τ the wavelet coefficients toward zero:

$$w_j(t) = \begin{cases} t - \tau & if\ t \geq \tau \\ 0 & if |t| < \tau \\ t + \tau & if\ t \leq -\tau \end{cases} \tag{4}$$

Regardless of the method applied, the threshold τ is usually inferred from the input signal, as a function of the signal energy or signal to noise ratio and/or noise variance. Donoho et al. [5, 6] argued that in the case of Gaussian noise with unit standard deviation the threshold at each scale can be determined directly from the wavelet coefficients as follows:

$$\sigma_j = \frac{median(abs(w_j))}{0.6745} \tag{5}$$

where σ_j is the estimated noise level for the scale j. The value for τ can be calculated as follows [5, 6]:

$$\tau_{j,n} = \sigma_j \cdot \sqrt{2log(n)} \tag{6}$$

where n is the length of the input signal. A number of research papers have been published on the topic of hard and soft thresholding demonstrating the usefulness of these approaches and their variants for de-noising of images and spectra (e.g. [26, 9, 25, 8, 29, 16, 22, 5, 6]).
In this work the de-noising procedure is applied estimating σ (equation 5) and τ (equation 6) from the signal approximation $D_1[t]$. Hence, the signal is de-noised by performing soft-thresholding on the coefficients of $D_1[t]$.

4.2 De-spiking

If the presence of noise in a spectrum can be addressed by soft-thresholding as above mentioned and discussed in [8], the presence of spikes is not solvable using the same approach. The soft-thresholding method relies on the fact that noise equally influences all the channels in the spectrum. A spike obviously affects

only a small region of it by multiplying few pixels by a huge value. The presence of a spike is "seen" by the wavelet transform as a sharp edge, therefore it is not removed by the de-noising procedure [5]. However, as discussed by Ehrentreich et al. [8], the wavelet transform can be still useful in the identification of spikes. As reported by the authors, the first-level details - the signal approximation obtained at the first decomposition level and containing the high frequency details - can be used as discriminant. Once the spike location is identified in the original spectrum, spike removal can be performed by conventional methods such as interpolation or regression [8].
In this work, the process of de-spiking astronomical spectra relies on a experimentally determined threshold evaluated from the first-level details ($D_1[t]$ - section 3). The signal is decomposed using the octave-band approach and the *à trous* algorithm. Using a small sample of spectra with and without spikes, we experimentally evaluated how the presence of spikes is reflected in $D_1[t]$. The threshold for spike detection (τ_s) is calculated as two times the standard deviation of the wavelet coefficients at $D_1[t]$:

$$\tau_s = 2 \cdot stdev(w(t)_{D1}) \tag{7}$$

Every peak in the $D_1[t]$ approximation that exceeds such a threshold is marked as a possible spike. In order to make the algorithm robust with respect to a false detection, the width at the base of the spike in the signal is checked. A candidate spike is removed by linear interpolation if the following two rules apply:

1. $w(t_0)_{D1} > \tau_s$
2. $width(P(t_0)) < \omega$

where $width(P(t_0))$ is the width of the peak containing the point t_0 and ω is a parameter evaluated experimentally and depending on the resolution of the spectrum. In this work ω assumes the value of about 70 Ångstroms (15 pixels on the x-axis). This value is calculated based on an average pixel resolution of about 4.5 Angstroms, a value estimated on the basis of the SDSS datarelease sheet.

This simple procedure allows us to automatically de-spike a number of spectra sharing the same wavelength resolution but different signal to noise ratio. It is important to note that the de-spiking procedure produces the side effect of losing information about some spectral features within the resolution (ω) chosen. Glitches are removed by performing the de-spiking procedure on the original signal and on the inverted signal (obtained by multiplying by -1 each channel value).

Figures 2, 3 and 4 show 3 sample spectra and the effect of the de-spiking procedure. It appears clear that the simple procedure presented above is quite effective in removing spikes and glitches. In Figure 2 the spike/glitch between 5000 and 6000 Ångstroms has been almost totally removed. The positive part has been completely deleted, while the negative part has been reduced. Moreover many "small" spikes at high wavelength ($>$ 7000 Ångstroms) have been successfully removed.

In Figure 3 the spike/glitch visible between 5000 and 6000 Ångstroms has been effectively removed, as is the case of the spike/glitch between 6000 and

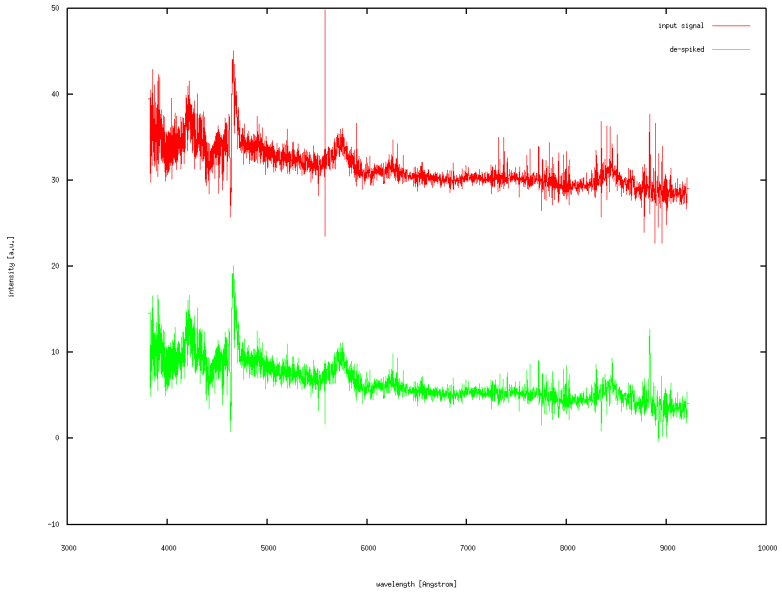


Figure 2: A QSO spectrum original (top) and after the de-spiking procedure(bottom)

7000 Ångstroms. On the other hand the small glitch between 7000 and 8000 Ångstroms has not been corrected by the algorithm. This behaviour may be ascribed to the presence of the other two imperfections in the same spectrum: the mean of the D1 coefficients is affected by their presence. A possible solution would be to iteratively process the spectrum until no *correction* can be made.

Figure 4 shows that the proposed algorithm does not influence the absorption lines typical of star spectra (in the range 4000-5000 Ångstroms), while it has been able to partially correct the spike/glitch visible between 5000 and 6000 Ångstroms.

Once spikes and glitches are removed by applying the above described strategy, the spectrum is de-noised using a standard wavelet shrinkage method (equation 4). The shrinkage threshold τ is estimated directly from the input signal as described in section 4.1. As suggested in [8], we estimate the noise level from the $D_1[t]$ approximation according to equation 5. The soft-thresholding is applied using equations 4 and 6 on the $D_1[t]$ approximation of the de-spiked spectrum. The "noise-free" signal is obtained by reconstructing the original signal according to equation 2. Finally, the spectrum is smoothed using a cubic-spline interpolation of width 4 pixels (about 19 Ångstroms). This step is applied in order to facilitate the peak-detection algorithms to parameterise the peaks in the spectrum. Given the fact that the width of the smoothing filter is less than the value of ω chosen in the de-spiking procedures, no further loss of information is expected.

Figure 5 shows a sample QSO spectrum, and the effects of the described procedures. It worth noting that the spike/glitch between 5000 and 6000 Ångstroms is partially removed by the de-spiking procedure (second spectrum from top),

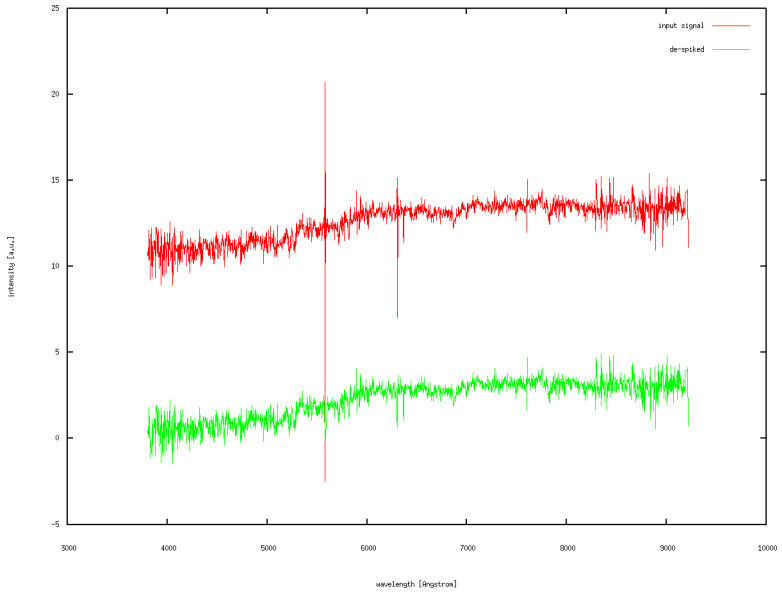


Figure 3: A galaxy spectrum original (top) and after the de-spiking procedure

not deleted by the de-noising algorithm and finally eliminated by the cubic-spline smoothing. Moreover the spectral feature at about 8900 Ånstroms has been enhanced by the de-spiking/de-noising/smoothing procedure.

The effect of the de-spiking and the de-noising procedure is evaluated in terms of the gain in classification accuracy. In section 8.2 we present and discuss how the accuracy of a classifier improves by applying the presented algorithms.

Once the de-spiking, de-noising and smoothing procedures are applied, the input signal is normalised to its integrated value and the emission/absorption lines and the continuum characterised as we present in the following sections.

4.3 Continuum

In order to extract the shape of the local continuum in noisy astronomical spectra, different strategies are often used. Polynomials of low degree are fit to the spectrum in order to determine the shape of the continuum. Broad band-pass filters are often applied in order to obtain a smoothed representation of the input signal. As presented and discussed in [27], the wavelet transform provides a simple way to determine the shape of the continuum of an input spectrum: the smoothed array $A_{j_0}[t]$ (equation 2) can be considered as a good representation of the continuum - it contains all the information at a very low spectral resolution. The only parameter that has to be determined is the decomposition level j_0: the maximum width of the lines we need to characterise. We estimate the best value for j_0 using the accuracy of a classifier trained in recognising the class of input spectra belonging to QSOs, galaxies and stars (section 8.3). Each spectrum is parameterised using the information extracted from the smoothed array $A_{j_0}[t]$ (section 4.3.1). Thus, the best j_0 is estimated in terms of prediction power.

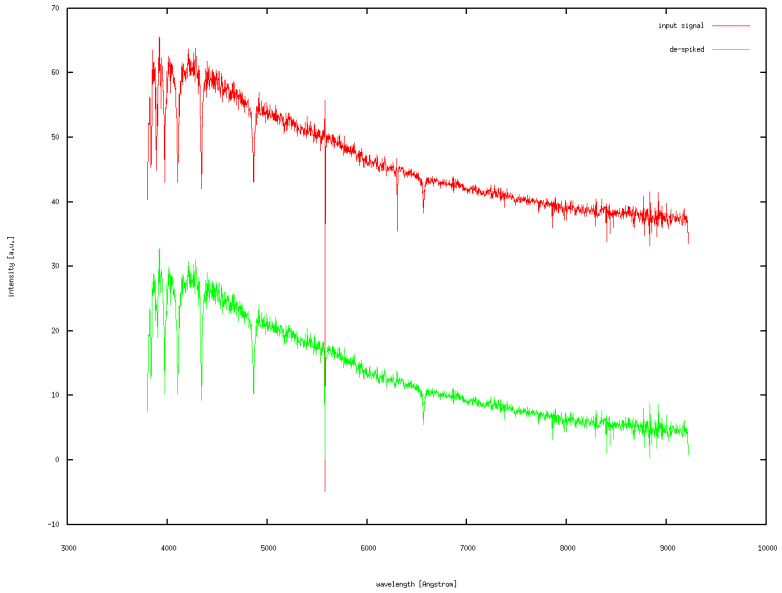


Figure 4: A star spectrum original (top) and after the de-spiking procedure

4.3.1 Continuum parameterisation

The continuum of a spectrum is not characterisable by max/min concepts as in the case of emission or absorption lines. As introduced in section 2, the continuum is produced by black body radiation and its spectrum is modified by intervening material and physical phenomenon occurring between the source and the observer. Since a comprehensive model of all the different phenomena is not feasible, the parameterisation we propose is aimed to simply record the shape of the continuum in different regions of the spectrum using a cubic spline interpolation. This is done under the assumptions that similar objects will show similar continuum behaviour. The parameterisation is as follows: firstly, the continuum is extracted as the coarsest approximation ($A_{j_0}[t]$) of the wavelet transform and divided in a discrete number of intervals (N_I). Secondly the continuum is interpolated (cubic-spline interpolation) using as knots the points delimiting each interval. Finally the 4 parameters of the 3rd degree polynomials fitting each interval and the upper and lower bound of the interval are recorded as features.

The total number of features (N_f^c) used to parameterise the continuum can be calculated as follows:

$$N_f^c = 6 \cdot N_I \tag{8}$$

As in the case of the level of decomposition j_0, the value of N_I is experimentally evaluated in order to maximise the classification accuracy. Section 8.3 presents an extensive evaluation of the accuracy of a classifier varying j_0 and N_I.

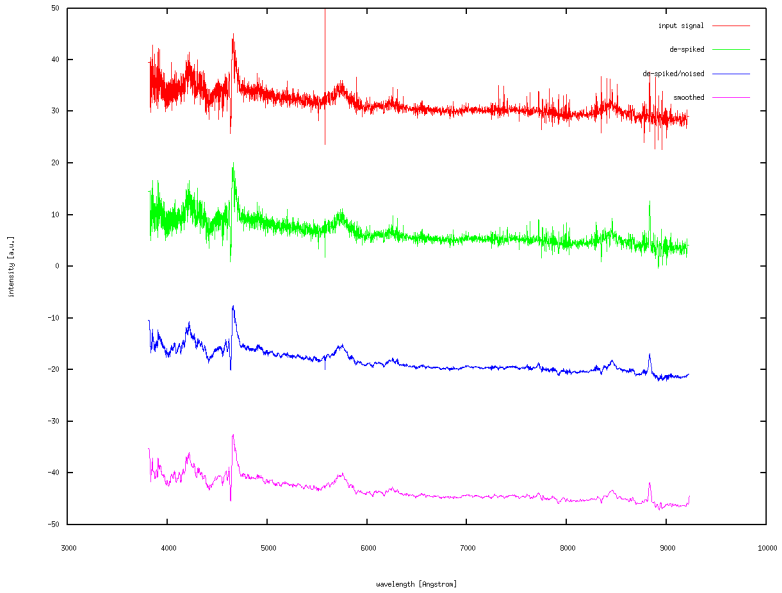


Figure 5: QSO processed: de-spiking, de-noising and cubic-spline smoothing

4.4 Emission and absorption lines

Once the maximum decomposition level j_0 has been evaluated as discussed in section 4.3, the emission and absorption lines are straightforwardly parameterised from the signal approximations $D_1[t]$, $D_2[t]$, ..., $D_{j_0}[t]$. In the next section we present the features extracted from the different approximations.

4.4.1 Emission and absorption parameterisation

Emission lines and absorption lines are characterised from their intensity and position in each sub-band $D_j[t]$. From each sub-band, the position and intensity of the most intensive peaks (the most intensive emission line and the most intensive absorption line) are recorded. Hence the total number of features ($N_f^{e/a}$) can be expressed as follows:

$$N_f^{e/a} = 4 \cdot j \cdot N_p \tag{9}$$

where j is the total number of sub-bands (decomposition level) taken into account and N_p the number of peaks extracted form each sub-band. For each subband j, the feature array consits of N_p quadruples of values describing emission/absorption line's intensity and position odered from the strongest to the weakest. The first two values in the quadruple describe the position and intensity of the positive peak in $D_j[t]$, while the last two give the position and intensity of the negative peak in the same subband.
As in the case of j_0 and N_I (section 4.3), N_p is not defined *a priori*, but is instead estimated experimentally in order to maximise the accuracy of the classifier (section 8). The width of the emission/absorption lines is not recorded as it is somewhat expressed by its presence in one (or multiple) sub-band (see Figure 6 for an example).

Figure 6, 7 and 8 show different wavelet approximations of a QSO, galaxy and star respectively. D2 corresponds to the high frequency coefficients extracted at decomposition level 2. Similarly D3 and D5 are the high frequency coefficients extracted at decomposition levels 3 and 5 respectively). C is the continuum as represented by the low frequency coefficients extracted at decomposition level 6.

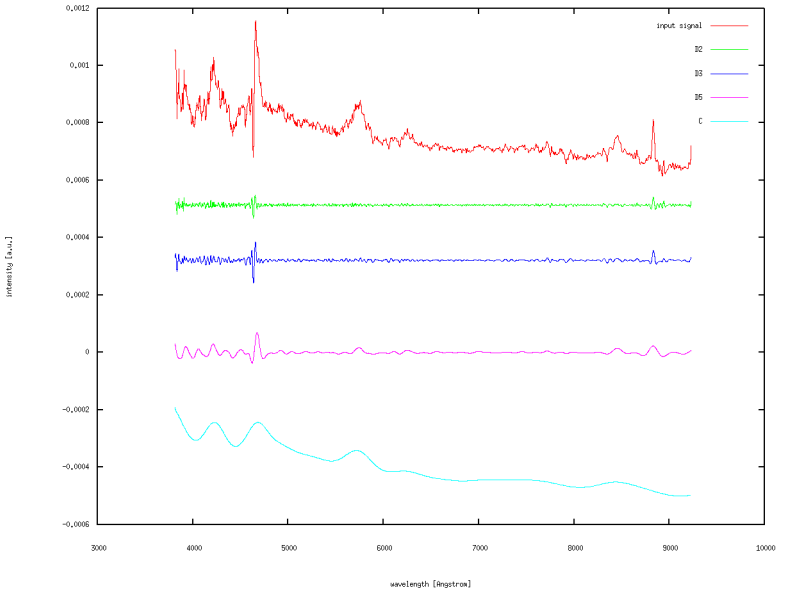


Figure 6: Wavelet approximations of a QSO spectrum. D2 corresponds to the high frequency coeffcients extracted at decomposition level 2. D3 (and similarly D5) are the high frequency coeffcients extracted at decomposition level 3 (and 5). C is the continuum as the low frequency coeffcients extracted at decomposition level 6.

5 Classification of astronomical spectra

5.1 The k-NN classifier

The k-nearest neighbour classifier is one of the simplest and most effective predictors and has been studied for almost four decades. It is an instance-based algorithm taking a conceptually straightforward approach to approximate real or discrete valued target functions [21]. Since the k-NN is a lazy predictor, the learning process consists in simply storing the presented data. All instances correspond to points in an n-dimensional space and the nearest neighbours of a given query are usually defined in terms of the standard Euclidean distance [21]. The predicted class is inferred from the classes of the k nearest cases. In terms of query neighbourhood (k), the 'probability' of the query q belonging to class C can be estimated as follows:

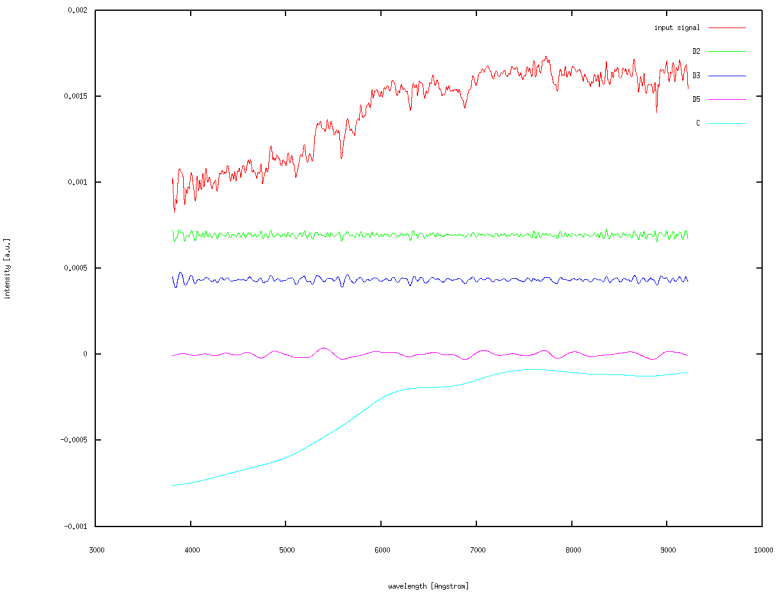


Figure 7: Wavelet approximations of a galaxy spectrum

$$P(C|q) = \frac{\sum_{i=1}^{k} w_i \cdot 1_{K_i=C}}{\sum_{i=1}^{k} w_i} \tag{10}$$

where the weight w_i is as follows:

$$w_i = \frac{1}{d(q, k_i)} \tag{11}$$

and k_i indicates the i-th nearest neighbour, k the total number of neighbours considered, K_i the class of k_i and $d(q, k_i)$ the Euclidean distance of k_i from q.

5.2 Feature selection

A lazy learner, such as k-NN, uses the whole set of features describing the instances in order to predict the class of the unseen query q. The simplicity of this approach has an important drawback. Both relevant and irrelevant features are used for the classification. Thus, the classifier performance can be highly influenced by the ability of a given set of features to describe the problem. In order to overcome this issue we implement a hill-climbing search based on the *wrapper approach* [24]. In the wrapper approach the induction algorithm is used as a black-box: it is iteratively run on a portion of the dataset. The evaluation function (usually an estimation of the generalisation accuracy or the generalisation error) is used to estimate which subset scores the best. The resulting classifier is tested against a separate subset of the data not used during the search.

In this work, the general schema used to evaluate the different feature subsets is based on a *inner* 10-fold cross-validation. The training-set is divided into 10 folds of which nine are used to train the algorithm using the selected

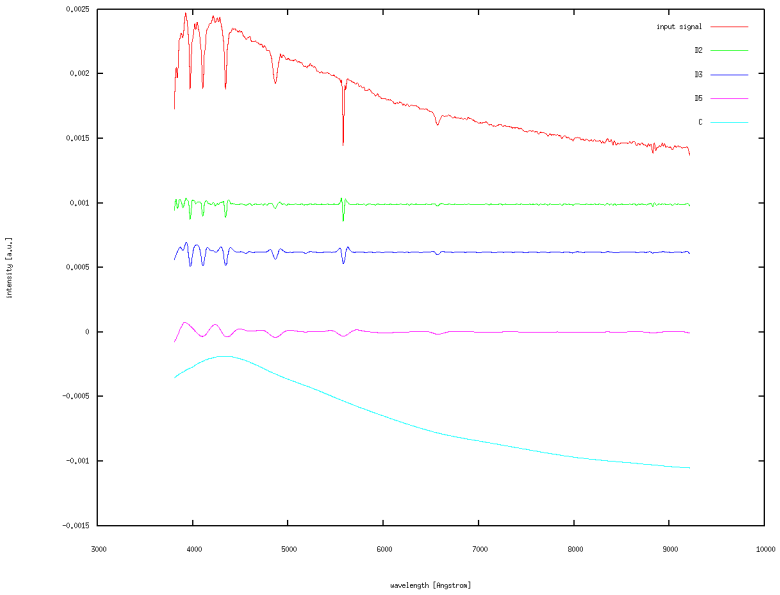


Figure 8: Wavelet approximations of a star spectrum

feature set, and the remaining one to test the accuracy of the classifier. The estimation of the accuracy is achieved by running the training/test process on the 10 different combinations of training-set/test-set.

A key issue in searching the feature space for the best sub-set is the order in which the attributes are tested since the wrapper approach is essentially a greedy search in the feature space for the best feature mask. In order to better drive such a search, *information gain ratio* [23] is used in this work to rank the features. Information gain ratio is a measure based on the notion of entropy that estimates the effectiveness of an attribute (feature) in classifying the training data [21]: it measures the expected reduction in entropy caused by partitioning the examples according to a given feature.

In this work both forward and backward hill-climbing are evaluated. The forward hill-climbing search is implemented as described by the following schema:

- at the beginning of the search the features are ranked according to the information gain ratio measure.
- the accuracy of the classifier is evaluated through a 10-fold cross-validation using only the feature that scores the best according to the given measure (information gain ratio).
- the bit in the feature mask corresponding to the next best feature is flipped (0 to 1) and the new accuracy evaluated through the 10-fold cross-validation.
- if the accuracy has improved the new feature mask is kept, otherwise the feature mask corresponding to the previous state is selected.
- sequentially the process is repeated until all the features are evaluated – from the most important to the less important.

Similarly to the forward case, the backward hill-climbing search is implemented as follows:

- at the beginning of the search the features are ranked according to the information gain ratio measure.
- the accuracy of the classifier is evaluated through a 10-fold cross-validation using all the features.
- the bit in the feature mask corresponding to the worst feature is flipped (1 to 0) and the new accuracy evaluated through the 10-fold cross-validation.
- if the accuracy has improved the new feature mask is kept, otherwise the feature mask corresponding to the previous state is selected.
- sequentially the process is repeated until all the features are evaluated – from the less important to the most important.

5.3 Ensemble methods

The idea behind the definition of an ensemble of predictors comes from a simple observation: it is commonly the case that a group of experts in a given domain can make better decisions than a single one. The obvious fact that a single expert may not posses the wide knowledge necessary to cover all the possible aspects of the domain is the intuition motivating the research of ensemble methods [4]. Different strategies for building a committee of experts have been proposed. These include; the manipulation of the training examples, the manipulation of the input features, the manipulation of the output targets and also the injection randomness [4].

In this work we adopt two different approaches to build an ensemble of classifiers, round-robin binarization and ensembles based on different sub-spaces of the original feature space. In the next sections we introduce the ensemble strategies adopted.

5.3.1 Round-robin ensemble

A round-robin ensemble converts a c-class problem into a series of *two*-class problems by creating one classifier for each pair of classes [13]. New items are classified by submitting them to the $c(c-1)/2$ binary predictors. The final prediction is achieved by weighted majority voting. The weights correspond to the probability estimated by each component classifier for the given query q.

5.3.2 Feature sub-space ensemble

Sub-sampling the feature space and training a simple classifier for each sub-space is an alternative methodology for building an ensemble. This strategy differs completely from the round-robin approach. It does not decompose the decision space based on the classification task. Instead, the strength of feature sub-space ensembles depends on having a variety of simple classifiers trained on different feature sub-sets sampled from the original space. This approach is very similar to a bagging technique [4] where the ensemble is built using different subsets of the instances in the training data. As reported in [4], this kind of ensemble is particularly appropriate in situations where redundant features exist.

Different methodologies can be applied to build a feature sub-space ensemble. For example, each ensemble member can be trained on different feature-subsets of predefined dimension, where each feature-subset may be drawn randomly from the original set.

In this work we build the feature sub-space ensemble reflecting the signal characterisation that has been performed: each ensemble member is built using features encoded from each level of the decomposition (section 4.4.1). This approach is motivated by the fact that we expect redundancy among features encoded from different levels (section 3.1), while features recorded from the same level are expected to carry no redundancy. Thus, given j_0 levels of decomposition, $j_0 + 1$ ensemble members are built: the features obtained from the $D_j[t]$ approximation plus the features encoded from the continuum ($A_{j_0}[t]$).

The final prediction is achieved by weighted majority voting. The weights correspond to the probability estimated by each component classifier for the given query q.

5.3.3 Ensembles and feature selection

The effectiveness of the two ensemble strategies in conjunction with k-NN classifiers is evaluated with and without the application of the feature selection strategies proposed in section 5.2. Feature selection is applied to each member of the ensemble. Partitioning the solution space according to a given rule, or sub-sampling the feature space, causes the gain ratio to provide different feature rankings; hence the feature selection process increases the diversity among the constituent classifiers. However, we will not take into account an explicit measure for the diversity, as proposed in [30]. Diversity arises naturally in the different ensembles.

6 Novelty detection

Novelty detection is one of the fundamental requirements of a good classification system in the domain of astronomy and spectra identification. Given an almost perfectly learnable problem, a machine learning system may not only be used to automatically label the vast amount of data. The knowledge recorded can be used as ground truth to discover new - never seen before - objects. Novelty detection is a fundamental strategy in recognising test data containing information about objects that were not known at the time the model was trained [19].

The literature presents a considerable number of different strategies upon which a novelty detection algorithm can be modeled. As presented by Markou et al. [19], these strategies can be subdivided into the following categories:

1. statistical approaches;
2. parametric approaches;
3. probabilistic/GMM approaches;
4. non-parametric approaches.

It is generally recognised [19] that there is no single best model for novelty detection. In fact a number of different studies (e.g. [2, 10, 12, 11, 14, 15, 20]) suggests that success depends on different variables such as the method adopted and the statistical property of the data at hand. However, the different strategies generally share the same logic, firstly the data are modelled according to the given approach and secondly a distance function and a threshold are used to identify abnormality.

In the context of parametric approaches, Chow [2] and Hansen at al. [15] show that it is possible to calculate the error and reject trade-off for a pattern recognition system in the context of binary classification. Fumera et al. [12], extended the concept to multi-class problems and linearly combined multiple classifiers. The authors demonstrated that the probability estimated from a given classifier can provide a better error-reject trade-off than one based on pure parametric assumptions [2]; when the posterior probability distributions are not exact and present an unknown degree of error the optimal threshold [2] is not a valid strategy.

In this work, the novelty detection schema is based on multiple thresholds evaluated from the data using instance based classifiers. Given a class C and query q, the threshold is evaluated as a function of the probability associated with the prediction (equation 10) of q belonging to C. Each instance of C in the training set is tested using a sample neighbourhood and, if correctly classified, the probability associated with the prediction recorded. The thereshold for class C is hence evaluated as the mean of all the probabilities stored. The other thresholds are similarly estimated by cycling through the different classes. The general schema adopted to estimate the probabilities is similar to the wrapper approach (section 5.2): the training set is partitioned and a 10-fold cross-validation applied.

The novelty detection approach proposed here allows us to define a threshold value directly inferred from the local properties of the training set and refelecting the bias of the classifier. The k-NN is used to model the typical (average) probability boundary associated with each class and defined by the data. Moreover, any dependency from outliers [14] is removed since the queries not correctly classified are not considered in the estimation of the thresholds.

A direct advantage of this approach is the ability to provide a confidence measure associated with each threshold. The confidence in predicting a new object is inferred from the data and is strongly related to the threshold. Given all the items correctly classified as class C, the confidence is defined as *the ratio of instances that have a probability of belonging to class C less than the threshold*. It expresses the confidence in predicting a new object based on the properties of the training set.

The confidence on the threshold can be used as a training parameter: the system can be tuned to provide novelty detection with a given confidence threshold. Given a confidence value, the thresholds (class dependent) are adjusted to control the probability with which the given ratio of objects of class C is correctly classified. In fact we take advantage of this property: in section 8.7 we present the evaluation of the performance of the novelty detection algorithm varying the confidence level as a learning parameter.

A major drawback of this approach is apparent in how the probability associated to the prediction is estimated by the k-NN. In the next section we discuss in detail this issue and propose a strategy to overcome it.

6.1 Novelty detection and the *k*-NN classifier

Adopting the k-NN classifier as learner, the probability of a query q being classified as class C is locally inferred from the k nearest neighbours of the query. The following equation gives a formal definition:.

$$P(C|q) = \frac{\sum_{i=1}^{k} w_i \cdot 1_{K_i=C}}{\sum_{i=1}^{k} w_i} \tag{12}$$

where the weight w is as follows:

$$w_i = \frac{1}{1 + d(q, k_i)} \tag{13}$$

and k_i indicates the i-th nearest neighbour, k the total number of neighbours considered, K_i the class of k_i and $d(q, k_i)$ the Euclidean distance of k_i from q. Generally, the weight w_i is calculated as the straight inverse of the Euclidean distance (11). However, we prefer the proposed form (equation 13) because the maximum degree of similarity (weight) is equal to 1 when the query q and the neighbour k coincide. For increasing distance, the weight decreases smoothly, asymptotically matching the behaviour of the standard expression.
Equation 12 expresses the probability of q belonging to C given the local properties of the feature space: the probability is a function of the class of the nearest neighbours of the query. The denominator in equation 13 is a constant that nomalizes the contributions of the weights. This normalisation procedure has the side effect of hiding the real distance of the query from its neighbourhood. This following example makes this clear.
Figure 9 shows an example of k-NN classification: the queries $q1$ and $q2$ are classified according to the classes of their neighbours $a1$, $a2$, $a3$ ($k = 3$). The regions marked as A and B are areas where instances belonging to the respective classes are concentrated. According to equation 12 both $q1$ and $q2$ are classified as examples of class A with a probability of 100%.

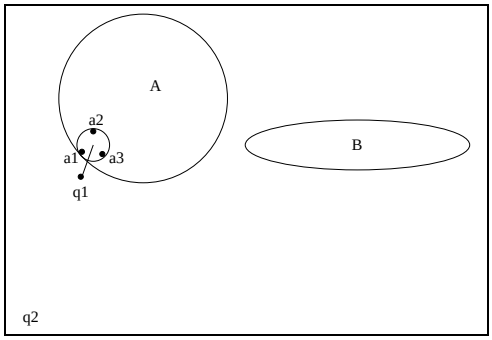
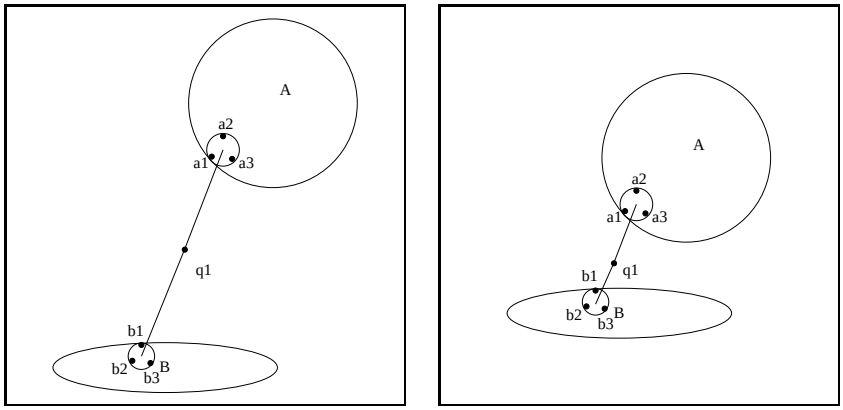


Figure 9: Example of classification using a k-NN ($k = 3$). $q1$ is clearly an example of A, while $q2$ may be an example of an unknown class.

It is evident from the figure that while $q1$ is likely to be an example of class A, $q2$ can be an example of a class not known at training time given its real distance from the the region A.

Another two cases are illustrated in Figure 10. In this scenario six neighbours ($k = 6$) are used to classify the query $q1$. According to equation 12 in both the cases (a) and (b), the query $q1$ is classified as an example of class B with a probability slightly over 50%. However, while in case (b) the query q is likely to be an outlier of class B, in case (a) the query may be an example of an unknown class, given the bigger distance from both the class bonduaries.



(a) (b)

Figure 10: Two examples of classification using a k-NN ($k = 6$).

In order to avoid these errors and to be sensitive to the real distance of the query from its neighbours, we rewrite equation 12 as follows:

$$P(C|q) = \frac{\sum_{i=1}^{k} w_i \cdot 1_{K_i=C}}{\sum_{i=1}^{k} max(w_i)} \tag{14}$$

where $max(w_i)$ is the maximum value that equation 13 can assume for any q. The maximum value is equal to 1 when the query q is the same as its neighbour k_i (equation 13). Thus, equation 14 can be written as follows:

$$P(C|q) = \frac{\sum_{i=1}^{k} w_i \cdot 1_{K_i=C}}{k} \tag{15}$$

The proposed equation provides a class probability that is strongly dependent on the similarity (distance) among instances in the feature space. The probability of q belonging to a given class C decreases smoothly from the regions where the neighbourhood is clustered. Given a properly defined threshold, the modified k-NN will be able to solve the examples in Figure 9 and 10. It is worth noting that the new normalisation procedure does not change the ordering of the neighbours of the query.

6.2 Novelty detection and ensembles of classifiers

The novelty detection strategy presented in the above section can be easily extended to an ensemble of classifiers. As presented in section 5.3, an ensemble is a classifier that often offers a better stability and better generealization accuracy than its component members. An ensemble takes advantage of the

linear combination of the outputs of its component classifiers to better model the data. Regardless of the method adopted to generate the new classifier and without any loss of generality, the probability associated with the prediction of an ensemble can be formulated as follows:

$$P(C|q) = \frac{\sum_{i=1}^{m} P_i(C|q)}{m} \tag{16}$$

where m is the number of component classifiers (defined by the typology of ensemble adopted, e.g. round robin or feature sub-space) and $P_i(C|q)$ is the probability given by the i-th member (equation 15).

The introduction of the aggregation function (generalised in equation 16) adds an extra layer in the classification procedure. This extra layer has the role of smoothing the decision surface by averaging the outputs of the ensemble members. Novelty detection based on ensembles relies on their generalisation capability to better assess the probability threshold. On the other hand, the extra layer of complexity has the side effect of weakening the relation between probability and distance. However, given equation equation 15, the probability is still a function of the distance in the feature space.

In section 8.7 we evaluate the impact of using ensembles of classifiers as well as the adoption of equation 15 in novelty detection.

7 Datasets

As previously mentioned, the datasets used in this work are extracted from the Sloan Digital Sky Survey (SDSS) [1].

The dataset labelled as *5class* is a balanced set of objects belonging to 5 different classes. The elements (600) are picked randomly, but with the constraint that they all have been catalogued by the survey whit a confidence of 95% (or more). The *5class* dataset includes the following classes:

1. *galaxy*
2. *star*
3. *star late*
4. *qso 0.15-0.30*
5. *qso 0.50-0.75*

The label *qso 0.15-0.30* indicates QSOs (quasars) at a redshift (z) between 0.15 and 0.30. Similarly *qso 0.50-0.75* indicates QSOs at z between 0.50 and 0.75.

The other five datasets are drawn form the same source [1]. Each of these datasets include elements of one single class. As in the case above, the elements are randomly selected and have been catalogued by the survey with a confidence of 95%. Table 1 illustrates the main characteristics of the datasets; the names are self explanatory.

The dataset labelled *unknown* contains objects of unknown class. The policy adopted by the survey is to apply such label to Objects not identifiable from their spectra. This includes spectra with bad artifacts, low signal to noise ratio

Table 1: Performance of the *fsse** in recognising QSOs at higher redshift

name	***n. of classes***	***elements per class***	***total elements***
5class	5	600	3000
qso 0.75-1.00	1	118	118
qso 1.00-1.25	1	48	48
qso 1.50-1.75	1	36	36
qso 1.75-2.00	1	36	36
unknown	1	600	600

and objects truly not identifiable. Given the nature of the dataset *unknown*, an accuracy of 100% in identifying its element as new - never seen before - objects is not expected. However, our evaluation (section 8.7) will be oriented toward the maximisation of the recognition ratio - together with the minimisation of the ratio of known objects identified as new.

8 Evaluation and discussion

8.1 Evaluation methodology

The evaluation methodology is based on a randomised 10-fold cross-validation repeated 10 times. The dataset is divided into 10 folds of which nine are used to train the algorithm and the remaining one to test the accuracy of the classifier. The estimation of the accuracy is achieved by running the training/test process on the 10 different combinations of training-set/test-set. At the beginning of each 10-fold cross-validation the instances are shuffled randomly. This is done in order to assess the generalisation accuracy of the predictor. The generalisation score is calculated as the mean of 10 independent 10-fold cross-validations. The error of the measure is calculated as twice the standard deviation. Thus, the confidence in the estimation of the generalisation accuracy is about 95%. As described in section 5.2, the feature selection process is based on a second (*inner*) 10-fold cross-validation: the training-set is divided into 10 folds of which nine are used to train the algorithm using the selected feature set, and the remaining one to test the accuracy of the classifier. The estimation of the accuracy is achieved by running the training/test process on the 10 different combinations of training-set/test-set.

8.2 The effect of de-noising de-spiking and smoothing the raw spectra

In this section we show the effect of normalising, de-spiking, de-noising and smoothing the input spectrum. In order to asses the benefit of these procedures, we perform a number of experiments taking as dataset the *5class* problem (section 7). The classifier is a simple k-NN ($k = 5$), while no feature encoding has been performed on the spectra. Each spectrum is parameterised by its flux value per channel. A total of 3810 channels have been considered. In the case of spectra having more than 3810 channels, the signal is cropped and the excess data points eliminated from further consideration. The reason why we consider the

whole spectrum is simple, the accuracy obtained can be considered a *straw man* against which we can compare the results obtained using the features proposed as descriptors.

Table 2 shows how the generalisation accuracy changes by applying the different methodologies.

Table 2: Accuracy of a k-NN with; normalising, de-spiking, de-noising and smoothing the input spectra

procedure	***accuracy [%]***	***error [%]***
none	95.0	0.2
norm	97.6	0.2
ds.norm	97.4	0.2
ds.dn.norm	98.8	0.2
ds.dn.ss.norm	98.9	0.2

The label *none* indicates that the rough spectrum is used. Label *norm* indicates that the spectrum has been normalised to its integral value and *ds* that the de-spiking procedure is applied. Similarly, the label *dn* indicates that the spectra have been de-noised and *ss* that the cubic-spline smoothing has been performmed. The label *ds.dn.ss.norm* indicates that first the spectra has been de-spiked, then de-noised, smoothed and finally normalised.

Table 2 shows clearly that the normalisation procedure adopted is helpful in terms of classification: it removes dependencies of the spectra from contingent variables (e.g. apparatus set-up, sky conditions, ...) at acquisition time. Moreover, it is clear that the de-noising methodology is effective in terms of classification accuracy: an increase of about 1% is noticeable comparing the results obtained with *ds.norm* and *ds.dn.norm*. Finally, it confirms that no information is lost by applying the smoothing strategy. However, Table 2 suggests that if anything the de-spiking strategy does not help accuracy, even if the score obtained by *norm* and *ds.norm* are equivalent within the error. This equivalence is not surprising if we consider the fact that a spike influences few channels (about 15) out of the many (3810) considered as features in these experiments. While the de-spiking strategy does not help a classifier working on the raw data, it is of considerable importance when working with representation coming from the wavelet decomposition.

In order to illustrate this a new set of experiments on the same dataset has been performed. A simple k-NN classifier is trained using features extracted from different wavelet decomposition levels. Specifically, from a given level j the position and intensity of the most intensive peak (intensity and position) is used as feature, the classifier trained and the generalisation accuracy evaluated. This process is repeated varying the decomposition level, applying the procedures *norm* and *norm.ds*. Table 3 shows the results obtained.

Table 3 shows clearly that considering one peak the classification accuracy benefits greatly from the proposed de-spiking procedure. The presence of spikes has a bigger impact when features are extracted from approximations containing high frequency components of the signal: the gain in accuracy at less detailed approximations is less prominent. Table 3 suggests that the presence of spikes is less problematic if features are selected from level 5 onwards.

Table 3: The effect of de-spiking on the representation based on the wavelet decomposition.

procedure	***level j***	***accuracy [%]***	***error [%]***
norm	1	41.0	0.8
ds.norm	1	54.4	0.6
norm	2	53.6	0.8
ds.norm	2	65.1	0.4
norm	3	65.5	0.6
ds.norm	3	72.7	0.6
norm	4	76.2	0.4
ds.norm	4	79.4	0.4
norm	5	84.9	0.6
ds.norm	5	84.7	0.4

In the following we will extract features considering normalised, de-spiked, de-noised and smoothed spectra as input signal.

8.3 Parameterisation of the continuum

In this section we present an evaluation of the number of decomposition levels j_0 required to maximise the accuracy of a predictor trained on features encoded from the coarse approximation $A_{j_0}[t]$ (section 4.3). As presented in section 4.3.1, the number of intervals (N_I) used for cubic spline interpolation is also evaluated. The selected classifier is a standard k-NN with $k = 5$.

Figure 11 shows a three-dimensional plot of the generalisation accuracy obtained varying j_0 and N_I. The error of each measure is about 1%.

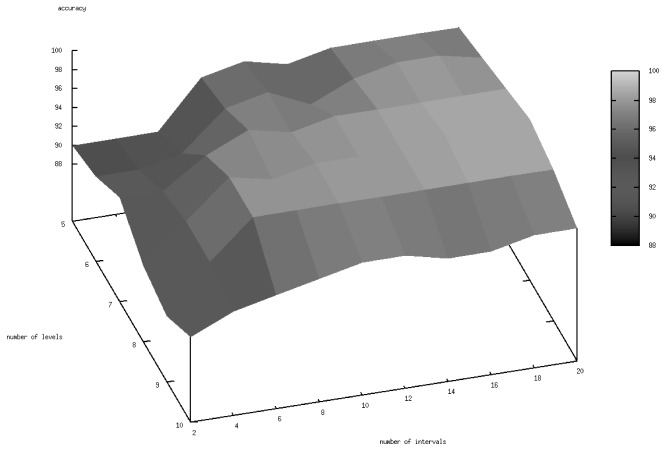


Figure 11: Accuracy of a k-NN varying j and N_I

The figure shows that a characterisation by cubic spline interpolation of the continuum is a valid strategy for spectra classification. By extracting features from the coarse approximation $A_8[t]$ and with $N_I = 6$, the classifier scores 98%.

In this case, as discussed in section 4.3.1, the number of features used is 36. By dividing the spectrum into 16 intervals (96 features) the classifier scores 99%.

Table 4 summaries the accuracy scores obtained at level $j_0 = 8$ varying N_I.

Table 4: Accuracy of a k-NN using 8 levels of decomposition varying N_I

N_I	***number of features***	***accuracy [%]***	***error [%]***
2	12	90	1
4	24	94	1
6	36	98	1
8	48	97	1
10	60	98	1
12	72	98	1
14	84	98	1
16	96	99	1
18	108	99	1
20	120	99	1

In order to minimise the complexity of the description while maximising the accuracy of the classifier (*Occam's Razor*) we propose to parameterise the continuum of the spectrum obtained at $j_0 = 8$ with 36 features ($N_I = 6$). A direct comparison between Table 3 and Table 4 shows that the suggested parameterisation guarantees an accuracy score equivalent (within the error bars) to the one achieved considering the whole spectrum (3810 features) as the feature set.

8.4 Varying the number of peaks for emission/absorption lines characterisation

In this section we present the evaluation of the number of peaks (N_p), section 4.4.1) needed to parameterise the input signal. As in the case of the continuum, we use a k-NN ($k = 5$) to asses the gain in generalisation accuracy varying the number of peaks taken into account (N_p). As discussed in section 8.3, the maximum decomposition level j_0 is equal to 8. Thus, 8 different arrays of features are evaluated, one for each $D_j[t]$ obtained from the wavelet transform.

Figure 12 shows the accuracy of the classifier using features extracted from the different signal approximations varying N_p, The error of each measure is approximately 0.4%

The figure shows that varying N_p, the accuracy reaches its maximum at 2 and then degrades smoothly with the adding of more information (increasing the number of peaks).

Table 5 summaries the results presented in Figure 12 for a selected number of decomposition levels (j).

The Table (and Figure 12) clearly tells us that a classifier trained with features encoded at low j hits its accuracy maximum when $N_p = 2$. Further increasing N_p has the effect of decreasing the generalisation accuracy of the k-NN. Augmenting the value of j (e.g. $j = 6$, $j = 8$ in Table 5), the classification accuracy hits its maximum at $N_p = 1$, showing the same accuracy (within the error bars) for $N_p = 2$. Further increasing the value of N_p, the accuracy score

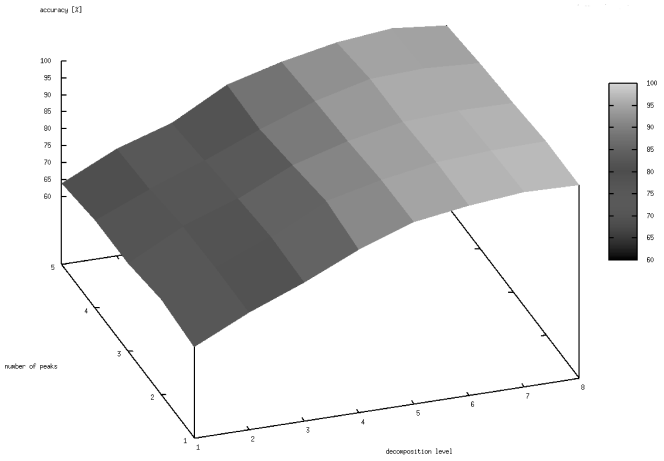


Figure 12: Accuracy of a k-NN at different j, varying N_p

smoothly degrades. In perfect analogy with the parameterisation of the continuum (section 8.3) - maximising the accuracy of each predictor while keeping the parameterisation simple, we parameterise each decomposition level $D_j[t]$ with information about the two most intense peaks (2 emission peaks and 2 absorption peaks).

8.5 Putting things together

In this section we present the results obtained by parameterising the spectra with a total of 100 features. As discussed in section 8.3 and section 8.4, 36 features are obtained by parameterising the continuum, while 64 are obtained by parameterising emission and absorption lines at different scales of the wavelet decomposition. As discussed in section 8.2, the spectra are pre-processed with the proposed strategies before the features are extracted. The dataset is the *5class* problem described in section 7.

Figure 13 shows the accuracy of a simple k-NN through varying the number of k (nearest neighbour).

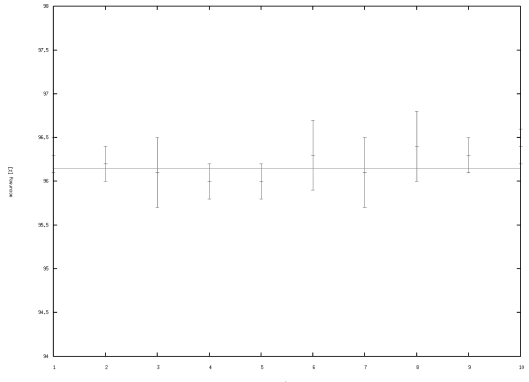


Figure 13: Accuracy of a k-NN varying number of k

Table 5: Accuracy of a k-NN at different levels of decomposition varying N_p

j	N_p	***number of features***	***accuracy [%]***	***error [%]***
1	1	4	66.8	0.4
1	2	8	68.1	0.4
1	3	12	66.0	0.4
1	4	16	65.8	0.4
1	5	20	64.0	0.4
3	1	4	80.8	0.4
3	2	8	81.5	0.4
3	3	12	80.0	0.4
3	4	16	78.5	0.4
3	5	20	76.7	0.4
6	1	4	95.9	0.4
6	2	8	96.0	0.4
6	3	12	95.2	0.4
6	4	16	95.0	0.4
6	5	20	93.0	0.4
8	1	4	97.0	0.4
8	2	8	97.1	0.4
8	3	12	95.6	0.4
8	4	16	94.7	0.4
8	5	20	93.0	0.4

The figure shows that the accuracy of the classifier is mostly independent of the number of neighbours selected (k): the accuracy stays constant at about 96.15% in the ranges of k explored. Thus, we arbitrarily choose to have a number of k equal to the number of classes for the ensemble classifiers evaluated in the following sections. The accuracy of the k-NN with $k = 5$ is equal to (96.0±0.2)% ($2 \cdot \sigma$).
Table 6 summarises the results obtained with a simple k-NN, a round-robin ensemble (*rre*) and a feature sub-space ensemble (*fsse*). As discussed above, the number of k is equal to 5.

Table 6: Accuracy of different classifiers on the *5class* problem

classifier	***accuracy [%]***	***error [%]***
k-NN	96.0	0.2
rre	96.0	0.2
fsse	98.7	0.1

The table shows that a *fsse* classifier outperforms both the simple k-NN and *rre* scoring $(98.7 \pm 0.1)\%$. The *fsse* takes advantage of the redundancy in the proposed features to better generalise the encoded knowledge. A direct comparison with the results obtained with the *rre* ($(96.0 \pm 0.2)\%$) shows that the number of members in the ensemble does not play an important role in this domain. The *rre* has ten members while the *fss* has nine members. Table 2 shows that using the whole spectrum (*ds.dn.ss.norm*) the classifier (*sknn*) scores

$(98.9 \pm 0.2)\%$. A similar accuracy (within the error bars) is obtained with the *fsse* trained using the propose features.

8.6 Feature selection

As discussed in section 5.2, the feature selection applies the wrapper approach model and uses the ranking provided by information gain ratio to drive the search in the feature space. Both forward and backward search are tested. Table 7 summaries the results.

Table 7: Accuracy scores obtained applying forward and backward search.

classifier	***fs.fw [%]***	***error [%]***	***fs.bk [%]***	***error [%]***
k-NN	99.0	0.2	96.3	0.6
rre	98.6	0.2	95.8	0.2
fsse	98.6	0.2	98.5	0.2

In Table 7, label *fs.fw* indicates that forward search is applied, *fs.bk* indicates backward selection.

A direct comparison between Table 6 and Table 7 shows that, in the case of k-NN and round-robin (*rre*) ensemble, forward search improves the accuracy of the classifiers. In this case, the simple k-NN shows slightly better results than the round-robin counterpart. On the other hand, backward selection does not provide the expected improvement in generalisation accuracy.

In the case of the feature sub-space ensemble (*fsse*), neither forward nor backward search are effective in improving the classifier accuracy. This phenomenon may be ascribed to a lack of redundancy in the descriptors available to each ensemble member. In fact, as discussed in section 6, the redundancy among the features is eliminated by implementing the *fsse* strategy itself. Thus, it is unlikely that overfitting occurs in the ensemble members and hence the lack of improvement in generalisation accuracy. A way to force each member to overfit the problem at hand is to increase the number of descriptors provided, increasing the number of peaks encoded and the number of bins used to characterise the continuum (section 4.3.1 and section 4.4.1). This seems a sound strategy to increase the specialisation of the ensemble member and hence improving the accuracy of the ensemble.

8.6.1 Comparison with PCA

In order to better asses the prediction power of the set of descriptors proposed, in this section we provide a comparison with results obtained using PCA. PCA is applied to the whole spectrum (3810 features) and different numbers of principal components are then used to build the classifier. Thus the PCA is being used to compress the data prior to training the classifier. Table 8 summaries the results obtained.

A comparison among Table 8,Table 6 and Table 7, shows that the results obtained with PCA are very similar to the ones obtained using the parameterisation proposed. In fact, a simple k-NN trained using the proposed features and implementing forward search scores $(99 \pm 0.2)\%$, equivalent to the accuracy obtained with PCA using 8, 10 and 12 components.

Table 8: Accuracy scores obtained applying PCA varying the number of principal components

n of principal components	***accuracy [%]***	***error [%]***
2	96.3	0.2
4	98.5	0.1
6	98.9	0.2
8	99.1	0.2
10	99.0	0.2
12	99.0	0.1

These results prove that the proposed approach is highly effective in parameterising the problem. The accuracy obtained using 100 redundant descriptors matches the accuracy obtained using about 8 features extracted as linear combination of 3810 parameters (the whole spectrum).

The signal processing techniques and the associated parametrisation provides a set of descriptors that is effective for categorising the spectra. This has the advantage of providing an interpretable and explainable classification process. The descriptors correspond to the the main characteristics of the spectrum, thus they directly reflect physical phenomena: the position of a peak is direct consequence of the typology of chemical elements composing the celestial object. The shape of the continuum (i.e. the coefficients of the polynomials) provides information about the presence of a black body radiator and some insight about the intervening material between the source and the observer. The explanation may be provided in terms of difference of the descriptors values between the query and the nearest neighbour(s).

Thus we have two means of compressing the raw data (PCA and the wavelet-based parameterisation) that can be used to build a classifier as good as or better than that built on the raw data. The PCA compresses the training data from 3810 to 12 features, the wavelet-based parameterisation produces 100 features that have a special semantics in this domain.

8.7 Novelty detection

In this section we present the results obtained by applying the novelty detection algorithm described in section 6. We evaluate the ability of a feature subspace ensemble (*fsse*) based on a simple k-NN to discover new objects. As in the case of the evaluation of feature selection (section 8.1), the estimation of the threshold value for novelty discovery is based on an inner 10-fold cross-validation. As discussed in section 6, the accuracy (in terms of *false negatives* and *false positives*) is evaluated varying the confidence level, i.e. the proportion of rightly classified objects. It is important to be careful with the terminology here as a *negative* corresponds to an object that is not considered to belong to one of the known classes, i.e. a novel object is a *negative*. The false negative figure corresponds to the proportion of known objects classified as new objects (never seen before). False positives are the new objects classified as known objects. The proportion of false negatives is evaluated using a 10-fold cross-validation on the *5class* problem as dataset. As discussed in section 8.1, the procedure is repeated 10 times randomly shuffling the dataset. The proportion

of false positives is estimated using *5class* as training set and *unknowns* as test set. As in the previous case, the procedure is repeated 10 times and the error figures we report are the average over the 10 repetitions; the error is estimated as twice the standard deviation. Classifiers using both equation 12 and equation 15 are tested. The symbol $*$ indicates that the classifier makes use of the modified probability function (equation 15). The error is indicated between brackets beside the obtained value.

In Table 9, we present the results obtained for the standard k-NN.

Table 9: False negative and false positive ratio using both k-NN and k-NN*

	confidence				
classifier	***80%***	***85%***	***90%***	***95%***	***100%***
false negative proportion in [%]					
k-NN	14.6 (0.2)	13.2 (0.6)	10.8 (0.4)	7.1 (0.1)	0.7 (0.2)
k-NN*	23.3 (0.4)	18.5 (0.4)	13.4 (0.4)	7.6 (0.6)	0.5 (0.2)
false positive proportion in [%]					
k-NN	23.8 (0.4)	25.1 (0.4)	31 (3)	59 (3)	92 (2)
k-NN*	6.4 (0.6)	12.0 (0.8)	20 (2)	35 (4)	76 (2)

The table shows that the required confidence level has a strong influence on the false negative and false positive figures as would be expected. At the highest confidence level the false negative figure approaches zero and the false positive tends towards 100%. The modified probability function (equation 15), has a significant impact in reducing the proportion of false positives. As expected (and discussed in section 6.1), the modified equation enhances the local properties of the classifier, providing an effective way to discover new objects. On the other hand, the table suggests that the adoption of the proposed probability equation has the effect of slightly worsening the false negative score of the classifier. However, it is worth noting that the difference in score of the two configurations tends to zero as the confidence level is increased.

In Table 10, we present the results obtained using a *fsse* as classifier in the same experiment.

Table 10: False negative and false positive ratio using both *fsse* and a *fsse**

	confidence				
classifier	***80%***	***85%***	***90%***	***95%***	***100%***
false negative proportion in [%]					
fsse	20.7 (0.8)	15.9 (0.6)	10.9 (0.4)	6.0 (0.2)	0.4 (0.1)
fsse*	20.9 (0.6)	16.0 (0.8)	11.3 (0.6)	6.2 (0.4)	0.4 (0.2)
false positive proportion in [%]					
fsse	10.5 (0.1)	14.3 (0.4)	24 (1)	47 (3)	97 (3)
fsse*	0.8 (0.2)	0.8 (0.2)	3.6 (0.8)	21 (3)	83 (2)

The table shows clearly that in the case of the feature sub-space ensemble, the proposed probability equation is highly effective. In terms of false negative, both the approaches provide equivalent results (within the error bars), while in terms of false positive the proposed approach outperforms the standard one. As

argued in section 6.2, the novelty detection based on ensembles relies on their generalisation capability to better asses the probability threshold (see Table 6).

In practice the error figures associated with *fsse** for confidence levels of 90% or 95% would represent the most acceptable trade-off. At the 90% confidence level the *fsse** would miss 3.6% of novel objects and produce 11.3% false alarms.

8.7.1 PCA and novelty detection

In this section we compare the performance of the wavelet-based compression against PCA on novelty detection. An equivalent algorithm to that presented in section 6 is applied. First, PCA is applied to the feature set and a suitable number of principal component selected. Then a confidence threshold for each class is calculated and the performance of novelty detection finally on the test data is assessed. Both versions of the class probability metrics are assessed (equation 12 and equation 15). The experimental methodology is as presented above. As suggested by our earlier evaluation (see Table 8) 8 principal components are used. Table 11 summarises the results obtained.

Table 11: False negative and false positive ratio using both *pca* and *pca**

	confidence				
classifier	***80%***	***85%***	***90%***	***95%***	***100%***
false negative proportion in [%]					
pca	1.9 (0.4)	1.9 (0.2)	1.9 (0.2)	1.9 (0.2)	0.3 (0.1)
*pca**	19.4 (0.6)	14.6 (0.4)	9.9 (0.4)	5.2 (0.4)	0.4 (0.2)
false positive proportion in [%]					
pca	63.3 (0.1)	63.3 (0.1)	63.3 (0.1)	63.4 (0.2)	88 (4)
*pca**	5.5 (0.4)	7.7 (0.4)	13.4 (0.8)	23 (1)	61 (2)

It is clear that in terms of false negatives the standard probability function performs better than the proposed one. Setting a confidence threshold of 80%, the difference in false negative proportion is about 18%. The difference in value gets smaller and smaller as the confidence threshold is increased. With a confidence threshold of 95%, the difference is about 3%. With a confidence threshold of 100%, both the approaches score similarly (within the error bars).

On the other hand, considering the proportion of false positives, the proposed probability function outperforms the standard one. Setting a confidence threshold of 80%, the difference in false negative proportion is more than 50%. This difference value tends to decrease as the confidence level is increased. However, with a confidence threshold of 100%, the difference is still more than 20%.

As in the case of k-NN and *fsse*, the proposed probability function provides a better false positive ratio. However, the huge gain in performance (more than 20% in the case of pca* and confidence threshold equal to 100%) is paid for by a poorer proportion of false negatives (no consequences at confidence threshold equal to 100%, using *pca*, *fsse* and k-NN).

Figure 14, shows the accuracy behaviours of PCA according to the different setups(Table 11).

In Table 12 we direct compare the results obtained with pca* and fsse* both in terms of false negatives and false positives.

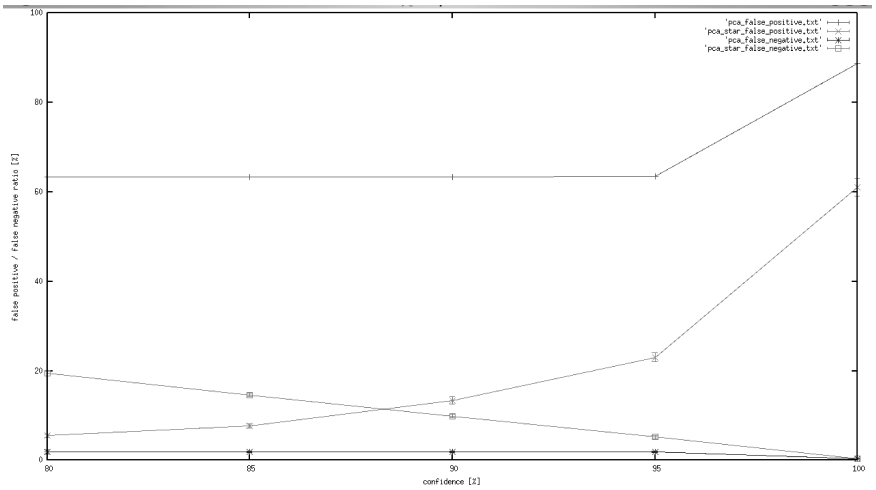


Figure 14: Accuracy graphs as expressed in Table 11

Table 12: False negative and false positive proportions using both *fsse** and *pca**

	confidence				
classifier	***80%***	***85%***	***90%***	***95%***	***100%***
false negative proportion in [%]					
*fsse**	20.9 (0.6)	16.0 (0.8)	11.3 (0.6)	6.2 (0.4)	0.4 (0.2)
*pca**	19.4 (0.6)	14.6 (0.4)	9.9 (0.4)	5.2 (0.4)	0.4 (0.2)
false positive proportion in [%]					
*fsse**	0.8 (0.2)	0.8 (0.2)	3.6 (0.8)	21 (3)	83 (2)
*pca**	5.5 (0.4)	7.7 (0.4)	13.4 (0.8)	23 (1)	61 (2)

The table shows that *pca** performs slightly better than *fsse** in terms of false negatives. However, the difference in score approaches zero as the confidence value is increased. In terms of false positives, *fsse** outperforms *pca** in the interval of confidence [80% : 90%]. Setting the confidence threshold to 95%, both the techniques score the same (within the error bars).

These results suggest that the proposed wavelet-based characterisation of the spectra coupled with a feature sub-space ensemble outperforms a PCA-based approach in recognizing new - never seen before - objects. On the other hand, PCA seems to be slightly more effective with respect to false negatives.

8.8 QSOs and redshift

In this section we present an evaluation of the performance of a *fsse** in recognising QSO spectra at different redshifts. The *5class* data is used as the training set and *qso 0.75-1.00*, *qso 1.00-1.25*, *qso 1.50-1.75* and *qso 1.75-2.00* as test sets. As in the previous case, the procedure is repeated 10 times and the generalisation error is estimated as the mean of the 10 repetitions; the error is estimated as twice the standard deviation. The confidence threshold adopted for novelty detection is equal to 90%. As Table 10 shows, this value provides a

good balance between false positives and false negatives.

Table 13 shows the results obtained. The *redshift* column indicates the redshift of the QSOs used as queries. Column *0.15-0.30* indicates the proportion of query objects classified as QSOs at redshift *0.15-0.30*. Similarly, column *0.50-0.75* indicates the ratio of query objects classified as QSOs at redshift *0.50-0.75*. Column *wrong* indicates the ratio of QSOs wrongly classified (classified as galaxy, star or star-late). Beside the accuracy we indicate between brackets the error in the estimation. In some cases, the error is equal to zero: when no variation in the score appears in the 10 random repetitions.

Table 13: Performance of the *fsse** in recognising QSOs at higher redshift

redshift	**true positive** *[%]*	**0.15-0.30***[%]*	**0.50-0.75***[%]*	**wrong** *[%]*
0.75-1.00	6 (1)	0 (0.0)	94 (1)	0.0 (0.0)
1.00-1.25	67 (2)	5 (1)	27 (2)	2.1 (0.0)
1.50-1.75	67 (3)	2.8 (0.0)	25 (3)	5.6 (0.0)
1.75-2.00	80 (3)	0.0 (0.0)	16 (1)	4 (2)

The table shows that the performance in novelty discovery varies depending on the nature of the query. For objects highly dissimilar to the known spectra - in the case of *qso 1.75-2.00* - the accuracy is about 80%. The figure degrades as the dissimilarity attenuates: 67% in the case of *qso 1.50-1.75* and *qso 1.00-1.25*; only 6% of the *qso 0.75-1.00* are marked as new objects. However, our results prove that the the ensemble is very effective in classifying QSOs at different redshifts. In the case of *qso 0.75-1.00*, the accuracy is about 100%: about 6% as new objects and about 94% as QSOs at lower redshift. For queries at higher redshift, the accuracy degrades: about 98% for *qso 1.00-1.25*, about 94 in the case of *qso 1.50-1.75* and *qso 1.75-2.00*.

9 Conclusions and future work

The objective of this research is to develop machine learning tools that will help in the analysis of astronomical data such as that coming from the Sloan Digital Sky Survey. The objective is to take a relatively small set of hand labelled data and develop classifiers that will generalise from that to identify other examples that can be assigned those labels and also highlight novel examples that warrant attention.

The work presented here has two aspects. There is the wavelet-based techniques for representing the data in a reduced set of features that are more meaningful. There is also the process for *ensembling* nearest neighbour classifiers that offers impressive classification performance on the data using this wavelet-based representation.

It is clear that the wavelet-based signal analysis techniques reduce the impact of noise in the emission spectra. As section 8.2 demonstrates, the application of the proposed strategies for de-spiking, de-noising and smoothing improve on the generalisation accuracy of a classifier trained using the raw data.

Through an empirical evaluation (section 8.3, section 8.3 and section 8.5), we fine tuned the wavelet decomposition adopted for signal characterisation and

defined a minimum number of features able to succesfuly describe the problem at hand.

Section 8.6 shows that feature selection applied on a simple k-NN provides a generalsation accuracy of about 99%, matching the generalisation accuracy provided by PCA (section 8.6.1). Our work demonstrates that the overhead caused by the signal processing techniques and the parametrisation provides a set of descriptors highly capable of generalising the problem at hand. Moreover, they enable the possibility of providing a meaningful explanation of the classification suggested by the system. The descriptors correspond to the the main characteristics of the spectrum, thus they directly reflect physical phenomena: the position of a peak is direct consequence of the typology of chemical elements composing the celestial object. The shape of the continuum provides information about the presence of a black body radiator and some insight about the intervening material between the source and the observer. This explanation may be provided in terms of the difference in the descriptor values between the query and the nearest neighbour(s).

This work demonstrates that the novelty detection approach proposed together with the enhancement to the probability function for the k-NN classifier (6.1) are highly effective (section 8.7). In fact, the proposed modification of the probability function provides a better recognition-rate for unknown objects with all the classifiers tested (Table 9, Table 10, Table 11 and Table 12), while showing comparable results in term of false negative ratio. Interestingly, the adoption of the proposed features together with a feature sub-space based ensemble and the k-NN modification appears as the best technique for novelty detection (better than PCA), showing results that are open to further refinement and development (8.7.1). In fact, using a number of datasets containing QSOs at various redshifts (section 8.8) we demonstrated that using a feature sub-space ensemble the system is highly effective in recognising new objects. The queries most similar to known objects (QSOs at low redshift) are recognised as such, while the most different ones are marked as new objects.

As reported in section 8.6, in the case of feature sub-space ensemble (*fsse*), wrapper-based feature subset selection is not effective in improving the classifier accuracy. We suggest that this is due to a lack in redundancy in the descriptors available to each ensemble member. In the near future we plan to increase the number of descriptors provided for each ensemble member. We expect in this to increase the specialisation of each component classifier and thus increase the degree of overfitting in each of them. This should improve the overall performance of the ensemble. We expect that this procedure will bring advantages in both classification accuracy and novelty detection rates (false negative and false positive).

10 Acknowledgements

Grant funding: We acknowledge the support of Enterprise Ireland BRGS grant SC/2002/370 via EU-funded NDP.

ICHEC: The authors wish to acknowledge the SFI/HEA Irish Centre for High-End Computing (ICHEC) for the provision of computational facilities and support.

SDSS: Funding for the SDSS and SDSS-II has been provided by the Alfred

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

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

References

[1] Sloan digital sky survey. `http://www.sdss.org`.

[2] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):159–184, 1970.

[3] I. Daubechies. *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

[4] T. G. Dietterich. Ensemble methods in machine learning. *Lecture Notes in Computer Science*, 1857:1–15, 2000.

[5] D. Donoho. Nonlinear wavalet methods for recovery of signals, densities, and spectra from indirect and noisy data. In *Proceedings of Symposium in Applied Mathematics*, volume 47, pages 173–204, 1993.

[6] D. Donoho and I. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. of the American Statistical Assoc.*, 90(432):1200–1224, 1994.

[7] F. Ehrentreich. Wavelet transform applications in analytical chemistry. *Analytical and Bioanalytical Chemistry*, 372(1):115–121, January 2002.

[8] F. Ehrentreich and L. Summchen. Spike removal and denoising of raman spectra by wavelet transform methods. *Analytical Chemistry*, 73(17):4364–4373, September 2001.

[9] M. Fligge and S. Solanki. Noise reduction in astronomical spectra using wavelet packets. *Atron. Astrophys. Suppl. Ser.*, 124:579–587, September 1997.

[10] P. Foggia, C. Sansone, F. Tortorella, and M. Vento. Multiclassification: reject criteria for the bayesian combiner, 1999.

[11] G. Fumera and F. Roli. Error rejection in linearly combined multiple classifiers. *Lecture Notes in Computer Science*, 2096:329–338, 2001.

[12] G. Fumera, F. Roli, and G. Giacinto. Reject optin with multiple thresholds. *Patter Recognition*, 33:2099–2101, 2000.

[13] J. Fürnkranz. Round robin rule learning. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 146–153, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[14] S. Guttormsson, R. M. II, M. El-Sharkawi, and I. Kerszenbaum. Elliptical novelty grouping for on-line short-turn detection ofexcited running rotors. *IEEE Transactions on Energy Conversion*, 14(1):16–22, 1999.

[15] L. K. Hansen, C. Liisberg, and P. Salamon. The error-reject tradeoff. *Open Systems & Information Dynamics*, 4(2):159–184, 1997.

[16] M. Lang, H. Guo, J. Odegard, C. Burrus, and R. Wells. Nonlinear processing of a shift invariant dwt for noise reduction. In *SPIE Symp. OE/Aerospace Sensing and Dual Use Photonics, Algorithm for Synthetic Aperture Radar Image*, Orlando, FL, April 1995.

[17] T. Li, Q. Li, S. Zhu, and M. Ogihara. A survey on wavelet applications in data mining. *SIGKDD Explorations*, 4(2):49–68, 2002.

[18] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[19] M. Markou and S. Singh. Novelty detection: a review (part 1): statistical approaches. *Signal Process.*, 83(12):2481–2497, 2003.

[20] M. Markou and S. Singh. Novelty detection: a review (part 2): neural network based approaches. *Signal Process.*, 83(12):2499–2521, 2003.

[21] T. Mitchell. *Machine Learning*. McGraw Hills, 1997.

[22] F. Murtagh, J. Starck, and A. Bijaoui. Image restoration with noise suppression using a multiresolution support. *Astronomy and Astrophysics Supplement Series*, 112:179–189, 1995.

[23] J. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[24] R.Kohavi and G.H.John. Wrappers for feature subset selection. *IEEE Transaction on Neural Networks*, 97(1-2):273–324, 1997.

[25] J. Starck and F. Murtagh. Automatic noise estimation from the multiresolution support. *Publications of the Astronomical Society of the Pacific*, 110(744):193–199, 1998.

[26] J. Starck, F. Murtagh, B. Pirenne, and M. Albrecht. Astronomical image compression based on noise suppression. *Publications of the Astronomical Society of the Pacific*, 108:446–454, May 1996.

[27] J.-L. Stark, R. Siebenmorgen, and R. Gredel. Spectral analysis using the wavelet transform. *The astrophysical journal*, 482:1011–1020, June 1997.

[28] M. Unser and A. Aldroubi. A review of wavelets in biomedical applications. *Proceedings of the IEEE*, 84(4):626–638, April 1996.

[29] J. Weaver, X. Yansun, D. J. Healy, and L. Cromwell. Filtering noise from images with wavelet transforms. *Magn Reson Med.*, 21(2):288–295, 1991.

[30] G. Zenobi and P. Cunningham. Using diversity in preparing ensemble of classifiers based on different subsets to minimize generalization error. In *12th European Conference on Machine Leaming (ECML 2001)*. Springer Verlag, 2001.