SDSS Celestial Objects Classification

Andrea Sessa
Mat. 850082

June 23, 2016

Abstract

Given the high dimensionality of the data provided by the Spectra instrumentation; main goal of the project is to provide a classfication of the celestial object observed by Spectra focusing on the data visualization and features selection phase.
Various different feautures selection tecnique are evaluated(Principal Component Anaysis(PCA), Kernel PCA, forward features selection, backward features selection). The different approaches are compared over a 'state-of-the-art' classifier(Support Vector Machine)
Final results shows that (TODO)

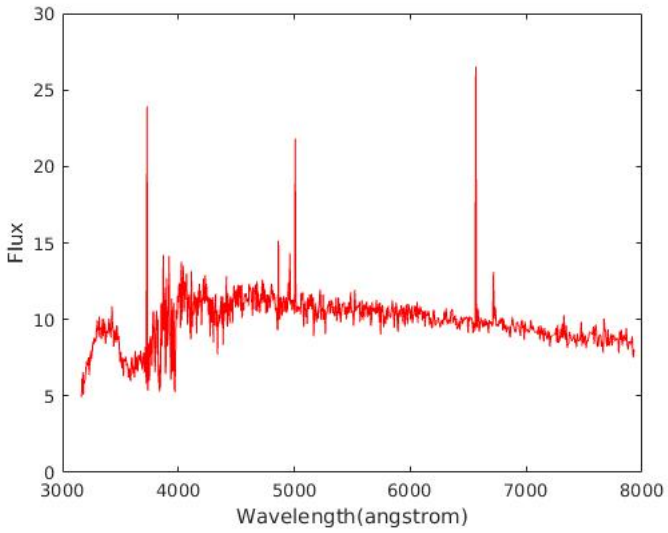
Contents

List of Figures

1 Introduction

Modern astronomy is concerned with the study of very distant celestial objects ie quasars, galaxies, stars, etc.
Often this type of classfication is performed by analyzing the spectrum emitted by such objects.
In general the emission spectrum of a chemical element or of a chemical compound is defined as the eletromagnetic radiation emitted when an atom or a molecule, of the object that we are observing, perform a transition from an high energy states to a low energy state. During the decadiment the atom or molecule the elettromagnetic is iradiated under the form of photon, the associated photon energy (also called flux) is proportional is equal to the energy difference between the two energy states involved in the decadiment.
The important element is that for a given atom there are many possible electron transition, and each of these transition has a specific flux associated.
The Sloan Digital Sky Survey(henceforth referred as SDSS) is a major imaging and spectroscopic survey using a dedicated 2.5-m wide-angle optical telescope at Apache Point Observatory in New Mexico, United States.
The collection of data started in 2000 and continues up to nowdays(latest data realease in June 2013). The dataset comprises almost 2 millions of spectra coming from diffrent objects.
Machine learning plays an important role in the task of classifying these objects, given the fact that many times samples include more than 1000 features.
Features in general includes information about the flux associated with a specific wavelenght.

Figure 1: Typical emission spectrum for an emission galaxy



The problem of feautures selection become very important in this scenario; this project will try to cast some light on the problem by evaluating different approaches over a reducted version of the SDSS dataset.

2 Problem Formulation

2.1 Dataset Description

The dataset that has been used for this project is a reduced version of the SDSS spectroscopic dataset: it consiste of 4000 spectroscopic samples, each of this sample formed by 1000 features that describes the spectrum over the different wavelength.
Objects in the dataset belong to 10 different categories:

1. **UNKNOW**: Unknown celestial objects
2. **STAR**: Generic stellar objects
3. **ABSORTION GALAXY**: Galaxies that show relatively homogeneous spectra, dominated by absorption features from cool giant stars
4. **GALAXY**: Generic galaxy(neither absortion nor emission)
5. **EMISSION GALAXY**: Galaxies whose spectrum is dominated by emission features
6. **NARROW-LINE QSO**: Narrow emission quasars
7. **BROAD-LINE QSO**: Broad emission quasars
8. **SKY**: No better specified
9. **HI-Z QSO**: Quasars with an high red-shift coefficient
10. **LATE-TYPE STAR**: Cooled starts(type K or type M)

2.2 Preprocessing

This project do not use the original SDSS dataset(which contained more than 4000 features per sample!).
The samples present in the original dataset had two main problems:

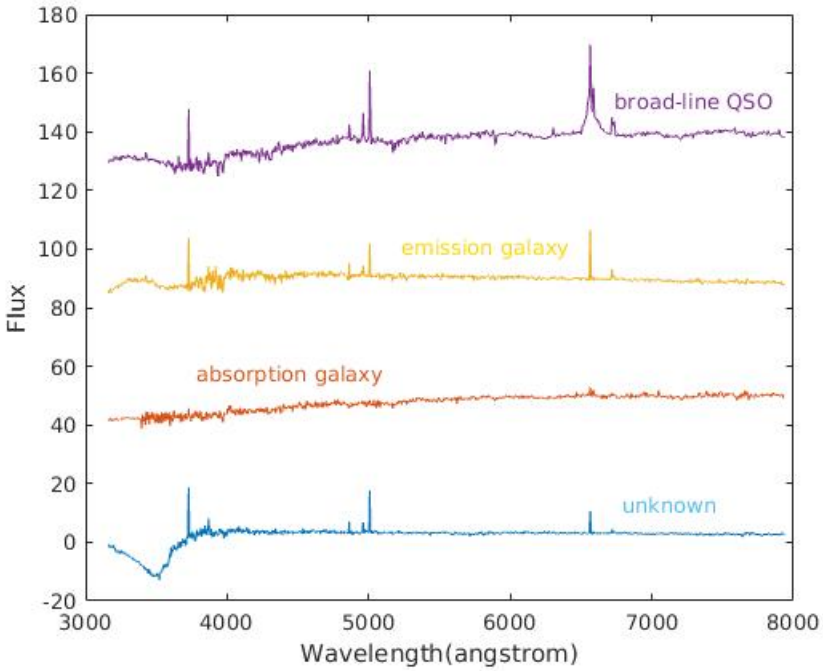
- Each individual spectra present a different redshift factor(z). This phenomena is due to the fact that the object that was being observed, was moving during while emitting light, this cause the frequency of light to 'shift' toward lower energy wavelenght ie. toward 'red' wavelenght.
- Missing data: There are several reasons for gaps to exist: The removal of skylines, bad pixels on the CCD chips all leave gaps at different frame wavelengths for each spectrum, general tecnical problem, etc. All these factor can contribute to incomplete spectra.
  To overcome this problem a Principal Component Anaysis(henceforth referred as PCA) is used to reproject the data on their principal dimension and then used to fill the missing gap. The entire procedure is described in detail in [3]

In addition to the two above stated procedure the dataset has been furtherly downsample till to reduce the number of significant features to 1000(which is still quite high!).
Finally, given the fact that the spectra belong to very different celestial object at very different distance(light years order) from the observation point, a final step of normalization is applied that is Z-Score and mean centering(making them suitable for a subsequent PCA analysis).

2.3 Visualization

In figure 2 are shown some random spectra for different type of objected extracted from the dataset:

Figure 2: Sample spectrum for different objects



3 Methodology

4 Experiments

5 Conclusions

References

[1] Wikipedia, *Emission Spectrum*

[2] Wikipedia, *The Sloan Digital Sky Survey*

[3] C.W. Yip et al, *Spectral Classification of Quasars in the Sloan Digital Sky Survey: Eigenspectra, Redshift, and Luminosity Effects*, Astronomical Journal, 2004