Age determination of the nuclear stellar population of Active Galactic Nuclei using Locally Weighted Regression

Trilce Estrada-Piedra, Juan Pablo Torres-Papaqui, Roberto Terlevich, Olac Fuentes, Elena Terlevich

Instituto Nacional de Astrofísica Óptica y Electrónica, Puebla, 72840, Mexico

Abstract. We present a new technique to segregate old and young stellar populations in galactic spectra using machine learning methods. We used an ensemble of classifiers, each classifier in the ensemble specializes in young or old populations and was trained with locally weighted regression and tested using ten-fold cross-validation. Since the relevant information concentrates in certain regions of the spectra we used the method of sequential floating backward selection offline for feature selection. The application to Seyfert galaxies proved that this technique is very insensitive to the dilution by the Active Galactic Nucleus (AGN) continuum. Comparing with exhaustive search we concluded that both methods are similar in terms of accuracy but the machine learning method is faster by about two orders of magnitude.

1. Introduction

Recent spectroscopic surveys of nearby AGN have proven that a large fraction show high-order hydrogen Balmer absorption lines in the near-UV (González-Delgado et al 1999) (Joguet et al 2001). These features are characteristic of young stars and therefore represent strong evidence of recent star formation in these galaxies. From a theoretical point of view, it is very important to determine the age of these starbursts, in order to understand the nature of the starburst-AGN connection and galaxy formation and evolution. The characterization of the nuclear star forming region (its age and mass) is very difficult to achieve in AGN, due to the contamination of the nuclear stellar absorption lines by the AGN component itself. The recent release of high-resolution spectra of large number of galaxies by the Sloan Digital Sky Survey (SDSS) consortium allows spectroscopic studies to be performed now on thousands of galaxies with active nuclei. In this work we propose a Machine Learning (ML) method to determine the age of stellar populations in synthetic galactic spectra. The experimental results obtained here show the efficiency of the automatic learning method applied to astronomy.

2.   Background

2.1.   Sequential Floating Backward Selection (SFBS)

SFBS is a feature selection algorithm that allows to work with non-monotonic data. It constructs in parallel the feature sets of all dimensionalities up to a specified threshold and consists of applying after each feature exclusion a number of features inclusion as long as the resulting subsets are better than those previously evaluated at that level. It makes a dynamically controlled number of iterations and achieves good results without static parameters (Pudil 1994).

2.2.   Locally Weighted Regression (LWR)

LWR is an instance based learning method; it assumes instances can be represented as points in an Euclidean space (Moore 2001). Its training consists of explicitly retaining the training data and using them each time a prediction needs to be made. LWR performs a regression around a point of interest using only a local region around that point. Locally weighted regression can fit complex functions in an accurate way and data modifications have little impact on the training.

2.3.   Ensembles of Classifiers

An ensemble of classifiers is a group of classifiers trained independently whose outputs are combined in some way, usually by voting (Mitchell 1997). They are normally more accurate than the individual classifiers that make it up.

3.   Data

The data is composed by 14 high resolution synthetic spectra combined in pairs considering 10 levels of dilution and including Gaussian noise. The experiments were calibrated using two population synthesis models with different ages (A. Bressan, private communication):

- A young population with ages between $10^{7.0}$ and $10^{8.6}$ years, representing a starburst component,
- An old population with ages between $10^{8.0}$ and $10^{9.6}$ years, representing the bulge component.

The spectra consist of 5655 points, from the optical to the near UV with wavelengths between 3600 and 5300Å and 0.2Å sampling. A subsampling process was performed to a resolution of 1Å in order to make the data compatible with the resolution of the observed spectra, in our case the SDSS spectral data. The features selected are the Balmer and Calcium II lines that are characteristic of the young and old populations respectively. The points of interest are selected for each component and population using the SFBS algorithm.

4.   Implementation

We made specialized classifiers using LWR trained in the high-order Balmer and Ca II absorption lines to determine the young and the old populations
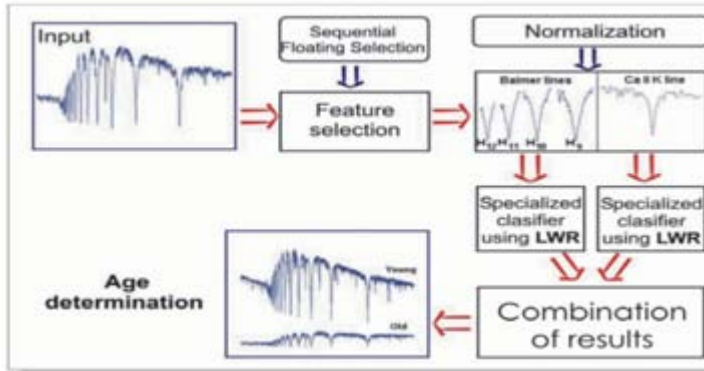
Figure 1.   Process of age identification

respectively; at the end of the process the results were combined by an ensemble of classifiers. For each classifier we selected a subset of features maximizing the probability of correct classification, this goal was achieved applying the SFBS method offline. The general process of age identification is shown in Figure 1 and is as follows:

1. Select features in the relevant regions. This is done online using the information retrieved by SFBS in a previous step.
2. Extract the information in the Ca II K line, which is typical of the older bulge population.
3. Identify the old population using specialized classifiers.
4. For each spectrum with no classification: extract the information in the Balmer lines that is characteristic of recent star formation.
5. Identify the young population using specialized classifiers.
6. Combine results in an ensemble.

5.   Experimental Results

Table 1.   Summary of results.

Algorithm	Noise	Young		Old		Time sec.
		Sample Error	True Error	Sample Error	True Error	
Exhaustive Search	0-11	-	0	-	0	12
Machine Learning	0	0.005	[0.003,0.01]	0.001	[0.003,0.01]	0.014
0.2Å	3-11%	0.007	[0.001,0.01]	0.025	[0.01,0.04]	0.092
Machine Learning	0	0.039	[0.02,0.05]	0.049	[0.03,0.06]	0.16
1Å	3-11%	0.059	[0.04,0.077]	0.069	[0.05,0.08]	0.35

We experimented using data with different resolutions and adding different noise levels. First we used 14 high spectral resolution synthetic models at 0.2Å,

after that we sub-sampled the spectra to 1Åto evaluate the performance of the ML method and to decide if it is possible to make an extension to handle the SDSS spectra. The ensemble was tested using ten fold cross-validation; in ten fold cross validation we divide the data into 10 subsets of equal size; we train the classifier 10 times, each time leaving out one of the subsets and using it for testing the algorithm. The sample error is calculated each time and is averaged to obtain the true error. An accuracy of about 0.3 *dex* in logarithmic age was achieved. The main results are summarized in Table 1. The time of prediction using LWR is linear in the number of examples and the experiments show that the time is reduced drastically with respect to the technique that does not involve ML. The method was then applied to the optical/near UV spectra of nuclear regions of nearby Seyfert galaxies covering the wavelength region $3600 - 5300$Å and it was found to be rather insensitive to the emission line and continuum contamination.

6. Conclusions and Future Work

The results obtained by ML are compared with those produced by exhaustive search in terms of time and precision. The machine learning method could find the correct ages, and was faster than exhaustive search, and has the additional advantage of the generalization capabilities inherent to this kind of algorithms. We conclude that the ML method can be extended to work with real spectra if we use a realistic noise model. The two-dimensional classification used for age identification in galactic nuclear spectra is similar in many ways to other problems and can be taken as a guideline in different problems (for example classification of binary stars and search for supernovae in galactic spectra). An extension of the method to handle observational spectra more reliably is to be published elsewhere. A second goal will be to implement a ML algorithm to find the AGN dilution and modify the synthetic spectra to predict three ages of stellar populations instead of just two.

References

González-Delgado R., Leitherer C. & Heckman T. 1999, ApJ, 125, 489

Joguet B. et al. 2001, A&A, 380, 19

Mitchell T.M 1997, Machine Learning, 230-247

Moore A. et al. 2001, Tutorial, 34-35

Pudil P. et al. 1994, ICPR94, 279-283.