Astronomical Data Analysis Software and Systems XIII
ASP Conference Series, Vol. 314, 2004
F. Ochsenbein, M. Allen, and D. Egret, eds.

Automated Determination of Stellar Population Parameters in Galaxies Using Active Instance-based Learning

Thamar Solorio, Olac Fuentes, Roberto Terlevich, Elena Terlevich

INAOE, Luis Enrique Erro #1, Tonantzintla, Puebla, 72840, México

Alessandro Bressan

Osservatorio Astronomico di Padova Vicolo dell' Osservatorio 5, 1-35122 Padova, Italy

Abstract. In this work we focus on the determination of the relative distributions of young, intermediate-age and old populations of stars in galaxies. Starting from a grid of theoretical population synthesis models we constructed a set of model galaxies with a distribution of ages, metallicities and intrinsic reddening. Using this set we have explored a new fitting method that presents several advantages over conventional methods. We propose an optimization technique that combines active learning with an instance-based machine learning algorithm. Experimental results show that this method can estimate with high speed and accuracy the physical parameters of the stellar populations.

1. Introduction

The availability for the first time of huge astronomical spectroscopic surveys such as the SDSS, with more than 10^6 spectra, will allow the determination of intrinsic physical parameters of a large number of galaxies, including the age distribution or star formation history and metallicity distribution of their stellar populations.

The importance of the accurate knowledge of these parameters for cosmological studies and for the understanding of galaxy formation and evolution cannot be overestimated. Template fitting has been used to carry out estimates of the distribution of age and metallicity from spectral data. Although this technique achieves good results, it is very expensive in terms of computing time and therefore can be applied only to small samples.

Starting from a grid of theoretical population synthesis models we constructed a set of model galaxies with a distribution of ages, metallicities and intrinsic reddening. Using this set we have explored a new method that maximizes speed and accuracy. Our proposed technique combines standard least-squares fitting with an active instance-based machine learning algorithm. Experimental results show that this method can estimate with high speed and accuracy the physical parameters of the stellar populations. Based on empirical evidence we believe that this method can be applied with equal success to other astronomical problems, reducing the computational cost and thus providing the capability of analyzing larger quantities of astronomical data.

2. Description of the Models

For the spectral synthesis of simple stellar populations the atmospheric models have been folded with the predicted number of stars along isochrones of given age and metal content (Bressan et al. 1994). The atmosphere models have been inserted in low resolution Kurucz models (Kurucz 1993) in order to preserve the complete energy distribution.

The models have the following characteristics:

- Ages are from $10^6 yr$ to $2 \times 10^{10} yr$ in logarithmic steps: $[10^6 yr, 10^8 yr, 10^{8.3} yr, 10^{8.6} yr, 10^9 yr, 10^{9.6} yr, 10^{9.78} yr, 10^{10} yr, 10^{10.2} yr]$
- Metallicity has the values Z=[0.0004, 0.004, 0.008, 0.02, 0.05] in Solar units
- The resolution is smoothed at the desired value.

For the present experiments we used solar metallicity (0.02) and a resolution of 20 Å.

3. The Proposed Solution

Given an observed galaxy spectrum we would like to determine the relative distribution of ages and their intrinsic reddening. We restricted the problem to finding three contribution of ages: starbursts of age 1Myr, an intermediate age population with age between 100Myr and 1000Myr and an old population with age greater than 1000Myr. Each of the three populations is affected by the same reddening law which is defined as follows:

$$R(c_i, \lambda) = 1 - e^{\lambda \times c_i} \tag{1}$$

where c_i is the free parameter of each stellar population and λ is the wavelength, in this case going from 890Å to 2.301 μm. In order to determine the free parameters of reddening and the relative contributions we pose the problem as an optimization problem, where a modified version of a machine learning algorithm is trained to estimate the reddening parameters of the three populations. Once we have an estimate of the reddening we can compute the relative contribution of ages, $\overrightarrow{A}$, with a pseudo inverse matrix as follows:
Let $M = [\overrightarrow{m_1}, ..., \overrightarrow{m_9}]$ be the grid of our nine theoretical models described earlier. $\overrightarrow{o}$ is the observed spectrum, and $\overrightarrow{r} = c_1, c_2, c_3$ is the vector of the free reddening parameters predicted by the learning algorithm for $\overrightarrow{o}$. We can compute $S = [\overrightarrow{F_{1r}}, ..., \overrightarrow{F_{9r}}]$, by applying to the theoretical models the reddening function as defined in equation 2.

$$\overrightarrow{F_{ir}}(\lambda) = \overrightarrow{m_i}(\lambda) \times R(c_i, \lambda) \tag{2}$$

We know that the observed spectrum $\overrightarrow{o}$ is the product of S and the unknown relative contributions $\overrightarrow{A}$,

$$\overrightarrow{o} = \overrightarrow{A} \times S \tag{3}$$

$$\overrightarrow{A} = S^* \times \overrightarrow{o} \tag{4}$$

then by computing S^*, the pseudo-inverse of S, we can determine the relative contribution of ages, as equation 4 shows. The following section introduces the optimization procedure used in this work.

4. The Optimization Procedure

We are interested in the problem of finding the parameters of a known analytic function that best match an observation. Let $\overrightarrow{o}$ be the observed galactic spectrum variable, let $f(\overrightarrow{r})$ be a function with the same dimensionality as o. The goal of the optimization procedure is to obtain the value of $f(\overrightarrow{r})$ that minimizes the error $e = |o - f(\overrightarrow{r})|$. In order to solve the problem more efficiently, we pose it as a learning problem, where a learning algorithm learns the reddening parameters $\overrightarrow{r}$, and with a forward model we compute $f(\overrightarrow{r})$. The training set used by the algorithm, $\langle f(\overrightarrow{x_i}), \overrightarrow{t}_i\rangle$, is formed by randomly generated reddening parameters, $\overrightarrow{x_i}$, and their corresponding galactic spectra, $\overrightarrow{t_i}$, where contributions of ages were also generated randomly; its test set consists of the galactic spectra to be analyzed denoted here by $\overrightarrow{o}_1, ..., \overrightarrow{o}_n$ and it outputs an estimate of $\overrightarrow{r_1}, ..., \overrightarrow{r_n}$ that is expected to minimize the errors $e_1, ..., e_n$. When a new set of solutions $\overrightarrow{r_1}, ..., \overrightarrow{r_n}$ is proposed by the algorithm, we compute their corresponding $f(\overrightarrow{r_1}), ..., f(\overrightarrow{r_n})$, using equations 2,3 and 4, and use the new pairs $\langle f(\overrightarrow{r_i}), \overrightarrow{o}'_i\rangle$ to augment the training set, and continue this iterative process until convergence is attained. Since this type of active learning adds to the training set examples that are progressively closer to the points of interest, the errors are guaranteed to decrease in every iteration. The pseudocode of the algorithm is the following:

1. Generate randomly an initial set of vectors $\overrightarrow{x}_i, ...\overrightarrow{x}_m$ and compute their corresponding $f(\overrightarrow{x}_1), ..., f(\overrightarrow{x}_m)$.
2. Let $P = \langle f(\overrightarrow{x}_1), \overrightarrow{t}_1\rangle, ..., \langle f(\overrightarrow{x}_m), \overrightarrow{t}_m\rangle$ be the initial training set.
3. Let $T = \overrightarrow{o}_1, ..., \overrightarrow{o}_n$ be the test set.
4. While T is not empty
    1. Train an approximator A using P as training set
    2. For each $\overrightarrow{o}_i$ in T
        - Use A to predict $\overrightarrow{r}'_i$
        - Generate $\overrightarrow{o}'_i$
        - $P = P \cup \langle f(\overrightarrow{r}'_i), \overrightarrow{o}'_i\rangle$
        - If $|o_i - \overrightarrow{o}'_i| < threshold$ remove o_i from T

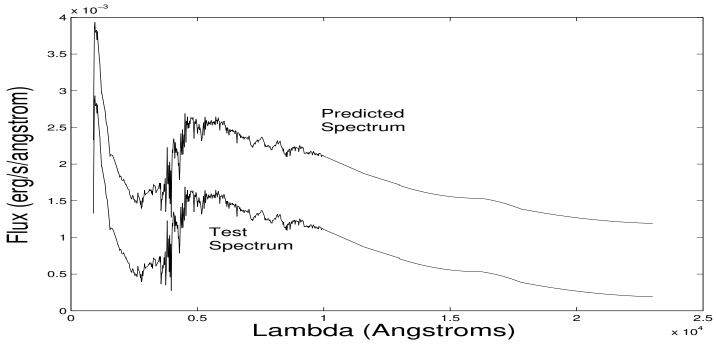
In this problem the approximator mentioned in step 4.1 is Locally Weighted Linear Regression, an instance-based learning algorithm that has shown good results in similar optimization problems (Fuentes & Solorio 2003).

5. Experimental Results

	r_1	r_2	r_3
mae$\times 10^6$	0.0149	0.0482	0.4182

Table 1. Mean absolute errors in reddening parameters

In order to evaluate our proposed solution we experimented generating randomly 500 spectra together with metallicities and intrinsic reddening, we then



A=[0.1132, 0.1763, 0, 0, 0, 0, 0, 0, 0.7105],R=[-0.0004, -0.0011, -0.0008]
A'=[0.1132, 0.1763, 0, 0, 0, 0, 0, 0, 0.7105],R'=[-0.0005, -0.0011, -0.0008]

Figure 1. In this figure we show test and predicted spectra shifted by a constant amount to aid visualization. Vectors A and R are the parameters for the test spectrum, while A' and R' are the corresponding predicted parameters.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9
mae$\times 10^6$	4.58	2.92	1.79	2.78	6.48	2.83	5.79	4.33	1.90

Table 2. Mean absolute errors in predicted population fractions

generated their corresponding spectra. From this set we selected randomly 150 spectra that were used as the test set, the remaining spectra were used as the training set. We repeated this process 10 times, and reported the overall average. Table 2 presents mean absolute errors in estimating age distributions, in Table 1 we show the errors in the reddening parameters. Figure 1 shows a comparison between a test example and the predicted one. On average, it takes 15 seconds to predict the parameters of a single spectrum.

6. Conclusions

We presented in this work an optimization algorithm that can estimate with high accuracy age distributions and reddening of stellar population in galaxies. The algorithm achieves convergence by iteratively creating new data points that lie in the vicinity of the query point. One important feature of this method is its high speed, it takes 15 seconds to estimate the parameters of a single spectrum. This represents a great advantage over other more conventional methods proposed for this problem, which may take several hours to find the solution for a single spectrum.

References

Bressan A. & Chiosi C. & Fagotto F. 1994, ApJS 94,63

Fuentes O. & Solorio T. 2003, AIA2003, Spain

Kurucz R. L. 1993, CD-ROM13: Atlas9, SAO, Harvard, Cambridge