Classification of Stars using Stellar Spectra collected by the Sloan Digital Sky Survey

Michael Brice and Răzvan Andonie

Computer Science Department: Central Washington University

July 16, 2019

Outline

Introduction

.

- Avoids complex transformation and statistical analysis of the spectra space.
- Use spectra without redshift corrections.
- Use Machine Learning classifiers.
- Uses Standard Feature Selection Methods to reduce the number of flux measurements.

Astronomy Background

Stellar Spectra

- A *spectrum* (plural *spectra*) is defined as the way in which light is distributed with wavelength [6].
- Incandescent light bulbs, when viewed with the unaided eye, appears to emit white light, but when viewed through a prism or a diffraction grating, the bulb is actually emitting a rainbow or spectrum of light. Stellar spectra work the same way.
- Unlike the example with the incandescent light bulb, stellar spectra have interesting properties.

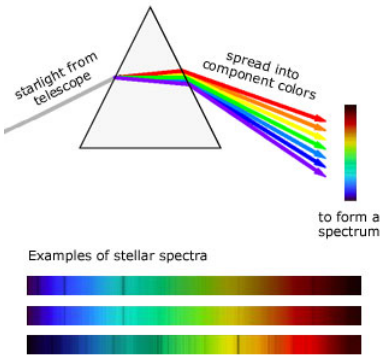

Figure 1: A spectrum [1].

Stellar Spectra

- Stellar spectra contain wavelength measurements and flux measurements.
- *Wavelengths* are discrete values that represent a range of wavelengths of light. For example, a wavelength value can be 6,000$\pm$5 *Ångströms* (*Å*).
- *Flux measurements* are the number of photons that pass through an area per second per measured wavelength.
- Fig. 2 shows an example spectrum of a flux scaled G2 (Sun like) star. The downward spikes seen in Fig. 2 are absorption lines.

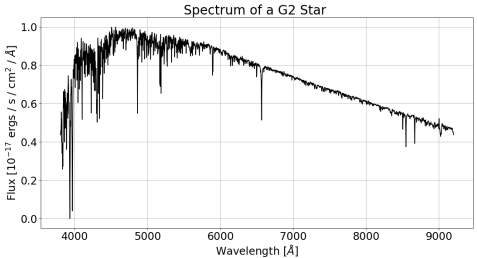

Figure 2: Example: Stellar Spectrum of a flux scaled G2 star collected by the SDSS.

Stellar Classification Types

- Two types of stellar classification schemes: The Harvard spectral classification and the Morgan - Keenan Luminosity Classes (MK) [5].
- The Harvard spectral classification is a surface temperature classification utilizing absorption lines
  - O, B, A, F, G, K, and M
  - Subclasses 0 - 9
- MK Luminosity Classes extends on the Harvard classes by adding a luminosity class and using the widths of the absorption lines.
  - I, II, III, IV, V, VI
  - Super Giants, Bright Giants, Sub-Giants, Main Sequence (Dwarfs), and Sub-Dwarfs [5].
- For example, Betelgeuse is a red super giant and a stellar class of M1 (Harvard) Ia (MK) star and Proxima Centauri is a main sequence red dwarf and a stellar class of M6 V star.

Importance of Stellar Classes

- Astronomers use stellar classes, in conjunction with other data, to learn many details about stars.
- The Hertzsprung-Russell (H-R) diagram, shown in Fig. 3, is a plot of stars where the horizontal axis is the spectral class or surface temperature and the vertical axis is the luminosity or absolute magnitude.
- The location of a star on the H-R diagram tells astronomers information about that star.[5].

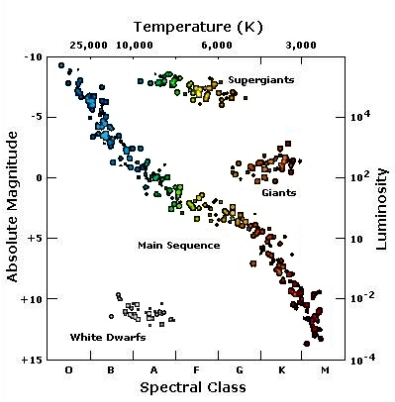

Figure 3: Hertzsprung Russell Diagram [13]

Redshift I

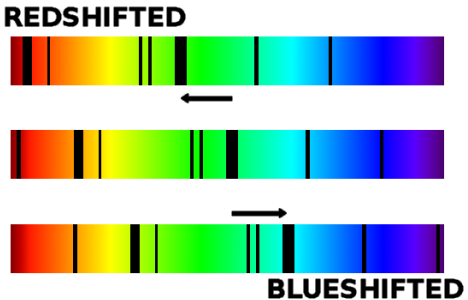

Figure 4: Example of Redshift [8]

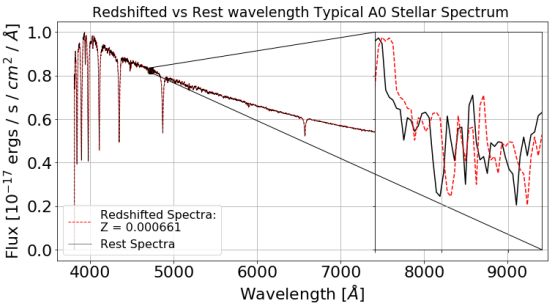

Figure 5: Another example of Redshift.

Redshift II

- Redshift is defined by (1), where Z is the redshift value and λ is the wavelength [5, 6].
- Equation (2) defines how the flux density changes due to redshift, where f is the flux density at wavelength λ (see [6]).
- However, for the purpose of this thesis, since Z for stars from the SDSS database have $Z \ll 1$, (3) is used.

$$Z = \frac{\lambda_{observed} - \lambda_{rest}}{\lambda_{rest}} \quad (1)$$

$$f(\lambda_{rest}) = (1 + Z)^2 f(\lambda_{observed}) \quad (2)$$

$$f(\lambda_{rest}) \approx f(\lambda_{observed}) \quad (3)$$

Redshift III

Human Redshift Extraction

1. Obtain the spectrum of a star which shows Absorption Lines.
2. From the pattern of lines, identify which line corresponds to which atom, ion, or molecule.
   - This pattern of lines are used to classify into the Harvard Spectral and Morgan Keenan Luminosity Classes.
3. Measure the shift of any one of those lines with respect to its expected wavelength, as measured in a laboratory on Earth.
4. Apply a formula that relates the observed shift to velocity along the line-of-sight (eq. (1))

Data Source

Sloan Digital Sky Survey (SDSS)

.

- The SDSS has created the most detailed three-dimensional maps of the Universe ever made, with deep multi-color images of one third of the sky, and spectra for more than three million astronomical objects [10].

Bolton *et al.* [4]

1. Uses templates.
2. Overall best fit (in terms of lowest reduced chi-squared) is used as classification.

SDSS Data Pre-processing

- The data is pre-processed by SDSS scientists through the methods presented by Dawson *et al.* [7] and Stoughton *et al.* [12].
- Two spectrographs used by the SDSS to collect the stellar spectra: the Baryon Oscillation Spectroscopic Survey (BOSS) spectrograph and the SDSS spectrograph described in Smee *et al.* [11].

Classification into the Harvard Spectral Classification Scheme

Introduction

- Avoid complex transformation and statistical analysis of the spectra space.
- Spectra without redshift corrections.
- Use feature selection to reduce the number of flux measurements. Feature selection may destroy the shape and the structure of each stellar spectrum by only using the most relevant flux measurements.
- Classifies only into the Harvard Spectral Classification Scheme.

Related Work I

Bailer-Jones, Irwin, and Hippel [2]

- Used a neural network with 50 nodes in the input layer, 5 nodes in the hidden layer and a single output node
- The average classification error is 1.07 SpT (spectra type)

Schierscher and Paunzen [9]

- Used a committee of Artificial Neural Networks, 10 neural networks are used
- Classes are generated from the effective temperature of the star and not from its Harvard class
- Reduces 2,400 dimensions to 435 using interesting areas in the spectra
- 85% Accuracy in relation to the SEGUE Stellar Parameter Pipeline

Related Work II

Zhang, Luo, and Tu [16]

- Used a non-parameter regression method
- Standard Deviation of $\sigma = 0.7994$

Xing and Guo [14]

- Used a Support Vector Machine (SVM)
- Utilized Principle Component Analysis (PCA) and Wavelet reduction
- 81.66% for SVM, 81.30% for PCA + SVM, and 93.26% for Wavelet + SVM

Dataset

- SDSS data run 14 using the SDSS spectrograph, which collected 600,967 stellar spectra.
- Some spectra were rejected.
- The usable dataset contains 578,346 stellar spectra and 22 of the 70 classes.
- The data is balanced using Undersampling (12,584 samples) and Hybrid (367,004 samples) balancing techniques.

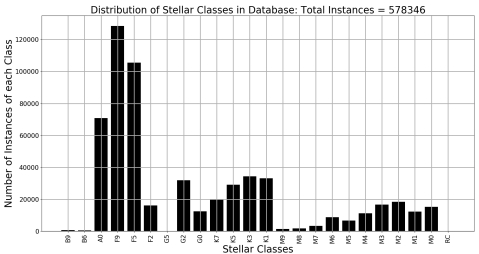

Figure 6: Distribution of data.

Data pre-processing

- The flux is scaled from 0 to 1.
- Each stellar spectrum in the dataset collected different amounts of flux measurements.
- Problems for Feature Matrix (next slide)
  - An average number of flux measurements is computed using the first 5,000 spectra.
  - Used to fit each spectrum to a standardized number of flux measurements.
  - Average number of flux measurements is 3,834.
  - Wavelength Fitting
  - Redshift

Feature Matrix

Table 1: Example of the feature matrix

	Wavelength: 3,807.15 Å	Wavelength: 3,808.03 Å	Wavelength: 3,808.90 Å	Wavelength: 3,810.66 Å
Spectrum 1	Flux	Flux	Flux	Flux
Spectrum 2	Flux	Flux	Flux	Flux

Table 2: Example of the feature matrix that has redshifted data

Star Class	Wavelength: 6,560.8 Å	Wavelength: 6,561.8 Å	Wavelength: 6,562.8 Å	Wavelength: 6,563.8 Å	Wavelength: 6,564.8 Å
A0		$H\alpha$			
A0	$H\alpha$				
A0			$H\alpha$		
A0				$H\alpha$	
A0					$H\alpha$

Classifier and Feature Selection Methods

Random Forest (RF)

- 10 Trees
- Other Authors report good accuracy [15]
- Weka demonstrated potential for good accuracy

Support Vector Machine (SVM)

- Gaussian Kernel / RBF Kernel
- Authors report good accuracy [14]
- Weka demonstrated potential for good accuracy

Chi Squared

- Chosen because of simplicity and low computational time.
- Commonly used and is reliable [3].

Fisher Score

- Chosen because of simplicity.
- Performed well on other large astronomical datasets [17].

Results I

Accuracy of Undersampled (12,584 Samples) Stellar Spectra				
K Best Features	Redshift	Rest	Redshift	Rest
	Chi Squared + RF		Chi Squared + SVM	
500	83.59%	88.24%	40.41%	41.27%
250	80.92%	86.72%	38.73%	39.59%
100	80.81%	86.17%	33.60%	35.69%
	Fisher Score + RF		Fisher Score + SVM	
500	84.95%	87.40%	64.66%	63.81%
250	83.45%	86.27%	64.44%	61.02%
100	82.00%	83.46%	66.51%	56.06%

Results II

Accuracy of Hybrid (367,004 Samples) Stellar Spectra				
K Best Features	Chi Squared + RF		Fisher Score + RF	
	Redshift	Rest	Redshift	Rest
500	96.55%	98.21%	96.87%	97.32%
250	95.46%	97.31%	96.49%	97.58%
100	95.11%	97.02%	95.89%	97.22%

Results III

Execution Time (Seconds) for Hybrid (367,004 Samples) Rest Spectra with 500 Best Features		
	Chi Squared + RF	Fisher Score + RF
Feature Selection	17.48	7.94×10^5
Train	109.06	121.24
Test	0.74	0.79

Conclusions

Conclusions

- The entire spectrum is not necessary to obtain high accuracy.
- Despite feature matrix complications, redshifted spectra can be accurately classified.

Reference I

[1] Looking at starlight.
Spectrum Image Reference.

[2] C. A. L. Bailer-Jones, M. Irwin, and T. von Hippel.
Automated classification of stellar spectra - II. Two-dimensional classification with neural networks and principal components analysis.
298:361–377, aug 1998.

[3] Veronica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos.
Feature Selection for High-Dimensional Data.
Springer International Publishing, 2015.

Reference II

[4] Adam S. Bolton et al.
Spectral classification and redshift measurement for the SDSS-III baryon oscillation spectroscopic survey.
The Astronomical Journal, 144(144):1–20, 2012.

[5] Bradley W. Carroll and Dale A. Ostlie.
An Introduction to Modern Astrophysics.
Cambridge University Press, 2nd edition, 2017.

[6] Fredrick R. Chromey.
To Measure The Sky: An Introduction to Observational Astronomy.
Cambridge University Press, 2010.

Reference III

[7] Kyle S. Dawson et al.
The baryon oscillation spectroscopic survey of SDSS-III.
The Astronomical Journal, 145(10):1–41, 2013.

[8] Samihahplanet.
One shift, two shift, redshift, blueshift.
Redshift Image Reference.

[9] F. Schierscher and E. Paunzen.
An artificial neural network approach to classify sdss stellar spectra.
Astronomische Nachrichten, 332(6):597–601, 2011.

[10] Sloan Digital Sky Survey.
Sloan Digital Sky Survey.

Reference IV

[11] Stephen A. Smee et al.
The multi-object, fiber-fed spectrographs for SDSS and the baryon oscillation spectroscopic survey.
The Astronomical Journal, 146(32):1–40, 2013.

[12] Chris Stoughton et al.
Sloan digital sky survey: Early data release.
The Astronomical Journal, 123:485–548, 2002.

[13] University of Iowa Department of Physics and Astronomy.
Imaging the universe.

Reference V

[14] Fei Xing and Ping Guo.
Classification of stellar spectral data using svm.
In Fu-Liang Yin, Jun Wang, and Chengan Guo, editors, *Advances in Neural Networks (ISNN 2004)*, pages 616–621, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[15] Zhenping Yi and Jingchang Pan.
Application of random forest to stellar spectra classification.
978-1-4244-6516-3/10/26.00 2010 IEEE, pages 3129–3132, 2010.

[16] J. N. Zhang, A. L. Luo, and L. P. Tu.
A stratified approach for automatic stellar spectra classification.
In *2008 International Congress on Image and Signal Processing (CISP 2008)*, pages 249–252, 2008.

Reference VI

[17] Hongwen Zheng and Yanxia Zhang.
Feature selection for high-dimensional data in astronomy.
Advances in Space Research, 41:1960–1964, 09 2007.