Baltic Astronomy, vol. 15, 69–76, 2006.

HOT SUBDWARFS FROM LOW-DISPERSION SURVEYS – METHODS OF EXTRACTION AND AUTOMATIC ANALYSIS

C. Winter C. S. Jeffery A. Ahmad and D. R. Morgan
Armagh Observatory, College Hill, Armagh BT61 9DG, Northern Ireland

Received 2005 August 1

Abstract. We present our first attempts to construct an automatic system for extracting and analysing hot subdwarf spectra from a large set of unknown, low-resolution spectra. The first element in this system is a filtering technique, based on Principal Components Analysis (PCA), which is used to extract hot subdwarf candidates from the unknown data. We field-test the PCA filter on a test data set obtained from the SDSS, and initial results are illuminating. The second element in the system is an automatic pipeline for providing spectral classification and parameterisation of hot subdwarf candidates. Classification is carried out by an artificial neural network, and parameterisation by a χ^2 minimiser over a set of LTE model atmospheres. We combine both elements to extract and analyse a set of 282 hot subdwarf candidates obtained from the SDSS. As such, this work is the first step toward constructing a fully automated tool for analysing large data sets.

Key words: methods: data analysis – methods: numerical – stars: fundamental parameters – stars: subdwarfs – stars: Hertzsprung-Russell (HR) and C-M diagrams

1. INTRODUCTION

Modern digital sky surveys, such as the Sloan Digital Sky Survey (SDSS), are able to produce extremely large data sets which contain observations of millions of objects. To extract and process data of relevance to a particular line of research in a time-efficient manner requires the astronomer to employ automated methods to help sift through the soup of objects not necessarily of interest and then analyse any resulting candidate objects.

As such, we are attempting to construct an automated filtering tool to assist in locating hot subdwarf candidates from any set of low-dispersion spectra. These will then be submitted to an automated analysis pipeline from which classifications and physical parameterisations are derived.

The filtering tool is based around the multivariate statistical technique called Principal Components Analysis (PCA) which can be used to identify patterns in an N-dimensional data set. Our intention is to employ PCA as a coarse spectral classifier to distinguish between hot subdwarf-like objects, and objects/data of little interest, e.g., white dwarfs, QSOs, incomplete or low SNR exemplars. As a test, we apply the tool to a sample of spectra obtained from the SDSS.

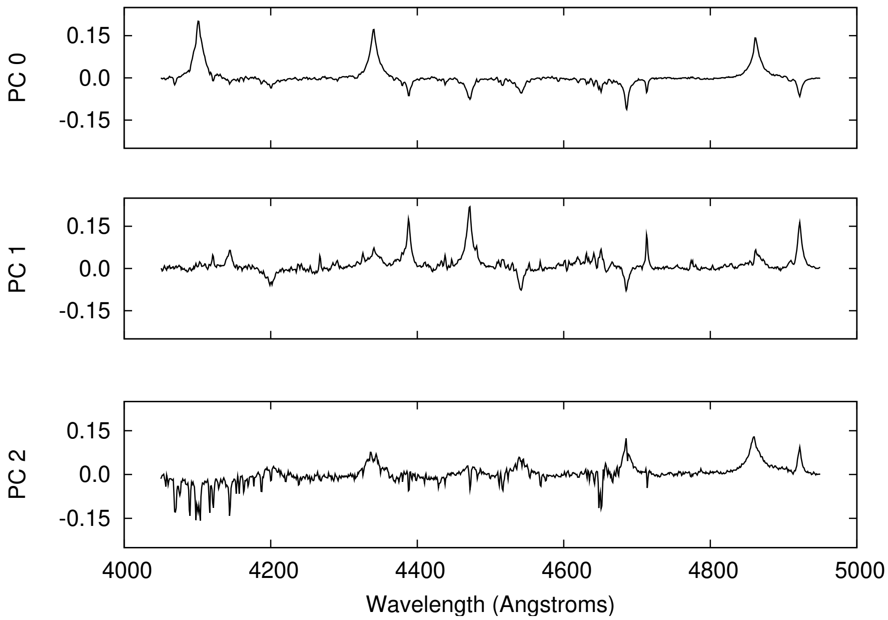

Fig. 2. First three principal components. Eigenfluxes are plotted against wavelength.

The automated analysis pipeline will allow us to gain insight into large volumes of spectra in a short amount of time. At present, the pipeline accepts a quantity of spectra, prepares them for analysis, and simultaneously feeds them to an artificial neural network (ANN), which has been trained to perform classifications using the Drilling et al. (2005) system, and a spectral optimisation toolkit which determines physical parameters by fitting each spectrum to a grid of LTE model atmospheres. Herein, we present the first application of the automated pipeline to a collection of 282 hot subdwarf candidates extracted from the SDSS with the aid of the PCA filter.

2. PRINCIPAL COMPONENTS ANALYSIS FILTER

2.1. Background

The basic idea behind PCA is to describe the variation of an N-dimensional data set in terms of a set of uncorrelated variables, each of which is a mutually-orthogonal linear combination of the original variables. Geometrically speaking, PCA is a rotation and scaling of a data set onto a new set of orthonormal vectors. The vectors (principal components) represent the directions of greatest variance in the data, and are a natural means for classifying the data set. If the first few vectors account for most of the variance in the original data, then they can be used to summarise the data with little loss of information – thus a reduction in dimensionality is achieved.

Technically, the principal components (PCs) are found by calculating the eigenvectors of the covariance matrix of an N-dimensional data set. A complete derivation of the technique can be found in, e.g., Murtagh & Heck (1987).

PCA has been applied on several occasions to the classification of astronomical spectra. Deeming (1964) applied the technique to stellar spectral classification, Connolly et al. (1995) used it to classify galaxy spectra, and Francis et al. (1992) applied it to the classification of quasar spectra. The dimensionality reduction capabilities of PCA have been used in conjunction with artificial neural networks, for example, Storrie-Lombardi et al. (1994).

2.2. Filter construction

We construct our PCA-based filter from the set of hot standards used in the classification system of Drilling et al. (2005). The spectra, which are at a resolution of 2.5 Å, were velocity corrected and resampled onto a uniform wavelength grid of 4050–4950 Å at a dispersion of 1 Å per pixel before performing the PCA. The first three of the derived PCs are shown in Figure 1. One can immediately notice that the first PC (PC 0) distinguishes between H and He lines. The second PC (PC 1) differentiates between He I and He II. This reification makes sense as it is the hydrogen and helium lines which vary most across all of the Drilling et al. (2005) spectra. Weaker lines, and uncorrelated features, such as noise, are distributed across the remaining less significant PCs. The first three derived PCs account for 69% of the variance present in the set of hot standards, and we use them to summarise the data set and to construct our filter.

If we consider the $(3 \times n)$ matrix, $\mathbf{E}$, of these eigenvectors, we can find the 3-dimensional projection vector, $\mathbf{p}$, of a spectrum projected onto the PCs by,

$$\mathbf{p} = \mathbf{x} \cdot \mathbf{E}, \tag{1}$$

where $\mathbf{x}$ is a vector of fluxes of dimension n.

Using the projection vector, we can then derive a reconstructed version of the original spectrum, $\mathbf{x_{rec}}$ by,

$$\mathbf{x_{rec}} = \sum_{k=1}^{k=3} p_k \mathbf{e}_k, \tag{2}$$

where p_k is the k^{th} element of the projection vector, $\mathbf{p}$, and $\mathbf{e}_k$ is the k^{th} eigenvector stored in matrix $\mathbf{E}$. The RMS error between $\mathbf{x}$ and $\mathbf{x_{rec}}$ can be used to differentiate between unknown stars which are most like the Drilling et al. (2005) hot standards, and those objects which we wish to discard.

2.3. Test application

We have applied the PCA filter to a test sample of ∼4600 spectra from the SDSS Data Release 3 database. Our selection criteria naively rely upon the classifications assigned automatically by the SDSS reduction pipeline. The SQL query used is as follows:

```
SELECT s.plate, s.mjd,s.fiberid
FROM BESTDR3..SpecPhotoAll as s
WHERE s.specClass = dbo.fSpecClass('STAR')
AND (s.primTarget & (dbo.fPrimTarget('TARGET_STAR_BHB')
     + dbo.fPrimTarget('TARGET_STAR_SUB_DWARF')) > 0)
AND (s.objType = 2)
```

Each spectrum was velocity corrected using the redshift as derived by the SDSS spectroscopic processing pipeline and stored in the FITS file header, and subsequently resampled onto the same wavelength grid as the Drilling et al. (2005) hot standards before being reconstructed by the PCA filter.

At present, the reconstruction filter is successful in filtering out objects with spectroscopic features that differ significantly from those of the hot subdwarfs, e.g., the white dwarfs in our test sample, or incomplete spectra.

The filter also allows a SNR threshold to be determined beyond which all samples can be discarded on the principle that they are too noisy for further analysis. Unfortunately, the filter does not readily allow the extraction of objects that are spectroscopically similar to subdwarfs but which contain small, yet important, differences that set them apart somehow. For instance, the test data sample, because it was selected based on the automatic classifications given by the SDSS (which are not very precise), contained many A- and F-type BHB stars which were clearly not subdwarfs because of very strong Balmer lines and the appearance of metal lines, and also other objects with weaker Balmer lines but containing the molecular G-band.

In calculating the reconstruction error of such spectra, the simple RMS error measure is not sensitive enough to allow small differences to influence the final error value in any meaningful way. The error calculation could be enhanced perhaps by introducing a weighting scheme, however, although these stars may have much in common with subdwarfs spectroscopically, they are very dissimilar photometrically and can be avoided easily by searching the SDSS by color.

3. AUTOMATED ANALYSIS PIPELINE

3.1. Outline

The input for our analysis pipeline comes from the PCA filter described in the previous section. The filter does still admit a number of false positives at this stage, so a manual inspection of the data is necessary.

Candidate spectra are subjected to a data preparation stage. The SDSS continuum fitting procedure does not do a great job in most cases, so we renormalise the calibrated spectra using a cubic spline fitting algorithm. The normalised spectra are then resampled onto a uniform wavelength grid of 4050–4950 Å, at a dispersion of 1 Å per pixel, ready for further analysis.

Classification is performed by an artificial neural network (ANN) which has been trained to classify onto the system defined by Drilling et al. (2005). This is an MK-like system which extends and refines the earlier work of Drilling (1996) and Jeffery et al. (1997) – it defines a spectral type (analogous to MK spectral classes), luminosity class and a helium class based on H, He I and He II line strengths. The ANN can classify onto this system with approximate errors (1σ) of ± 2 subtypes for spectral type, ± 1 subclass for luminosity class, and ± 4 subclasses for the helium class.

Physical parameters ($T_{\rm eff}$, $\log g$, $\log(n{\rm He}/n{\rm H})$) are derived from candidate spectra by χ^2 fitting to a grid of LTE model atmospheres. This is performed automatically by our spectrum optimisation toolkit, SFIT2 (Jeffery et al. 2001). To accommodate the physical diversity of the hot subdwarfs, our grid of model atmospheres coarsely covers the parameter space $T_{\rm eff}$: 15 000–50 000; $\log g$: 3–6; nHe:

0.001–0.999.

3.2. SDSS data samples

We selected two samples of spectra from the SDSS Data Release 3 database. The first sample consists of all the hot standard stars used for SDSS spectrophotometric calibrations.

Our second sample was selected based on photometric colors. Although the SDSS spectroscopic processing pipeline does make an attempt to classify stellar spectra, we learned in section 2.3 that it does not do so in any great detail, so we cannot rely upon the SDSS classifications to accurately extract a complete sample of subdwarf candidates from the database. Our selection criteria are summarised in the following SQL query:

```
SELECT s.fiberID,s.mjd,s.plate
FROM BESTDR3..SpecPhotoAll as s
WHERE s.psfMag_u < 21
AND (s.psfMag_u - s.psfMag_g) < 0.7
AND (s.psfMag_g - s.psfMag_r) < -0.1
AND s.specClass <> dbo.fSpecClass( QSO )
```

After passing all the retrieved spectra through our PCA filter, manually extracting any false positives and removing samples occurring in both data sets, we obtained a final sample of 282 hot subdwarf candidates. A summary of data set statistics is given in Table 1.

Table 1. With the combination of our PCA filter and visual selection, we were able to extract 282 unique subdwarf candidates from a total of 8171 spectra retrieved from the SDSS.

Data set	Spectra retrieved	Candidates extracted
Hot standards	1417	152
Color-color	6754	249
Total	8171	401 (282 Unqiue)

3.3. Results

Classification and parameterisation results for the 282 SDSS spectra are presented in Figure 2. The analogues between the classification parameters and physical parameters can be clearly seen by comparing the left-hand plots with the right-hand plots in Figure 2.

We note the patterns visible in the $\log g$ vs. $T_{\rm eff}$ plot: the gap at approximately $T_{\rm eff} \approx 20,000$K, $\log g \approx 4.5$; and the line extending leftwards at $\log g \approx 5.0$. We are unsure as to their nature, but the work of Green et al. (2006) discovered similar patterns from an analysis of an independent set of subdwarf stars.

An example of the model atmosphere fits and corresponding ANN classification of four stars from the sample are presented in Figure 3.

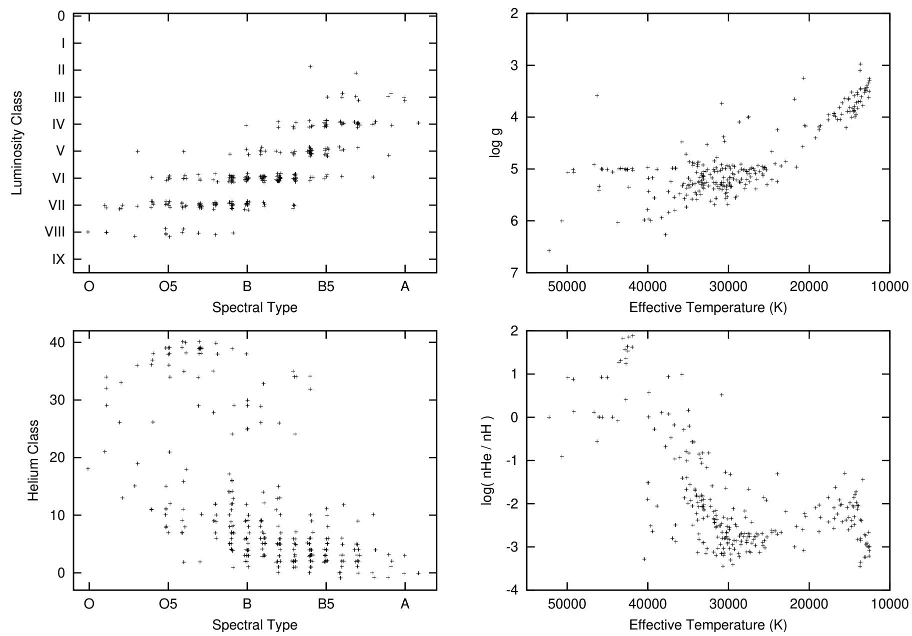

Fig. 2. Results of classifying and parameterising the 282 hot subdwarf candidates extracted from the SDSS. Classifications are presented in the two left-hand plots, with parameterisations presented in the two right-hand plots. The analogues between classification parameters and physical parameters is clearly evident.

4. FUTURE DIRECTIONS

The automatic filtering and analysis pipeline presented here is a part of what will be a more extensive system for automatic spectral analysis (see Jeffery 2003).

Given the nature of the hot subdwarfs, our use of LTE model atmospheres represents a problem for the analysis of very hot stars wherein NLTE effects become more apparent. However, our analysis pipeline can be easily extended to incorporate third-party model atmosphere codes that will allow us to address this issue in future work. We have used this study to help us refine our analysis pipeline and uncover the main hazards in dealing with large quantities of data. As we work to further improve the system, we must also tackle a number of issues, such as:

• **Data management.** Keeping track of different data sets, selection criteria, extracted candidates, their classifications and parameterisations, and maintaining general order when working with large quantities of spectra is an important task. Doing it manually is time-consuming, prone to user-forgetfulness, and can result in unfortunate industrial accidents (`rm -fr *`)[1].

• **Visualisation and interaction.** Analysing large data sets presents the problem of how to visualise thousands of spectra in a meaningful way. Also, when results have been obtained, the user must be able to produce plots and interact

---

[1] This is the Unix shell incantation which wipes clean the files in a user's current working directory and all its sub-directories.

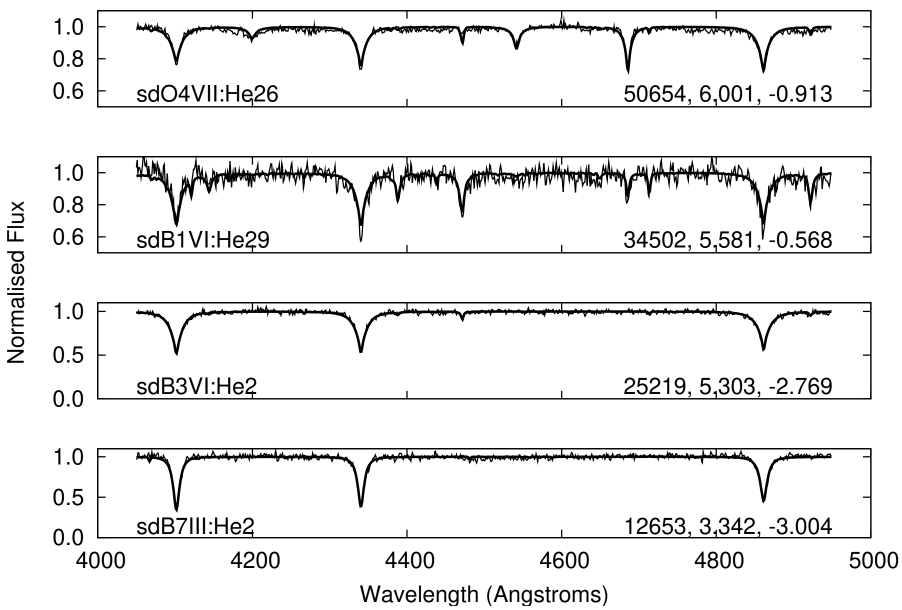

Fig. 3. A selection of stars from the set of 282 subdwarf candidates. The bottom-left of each plot details that star's spectral type as determined by the ANN, and the derived physical parameters are given at the bottom-right ($T_{\rm eff}$, $\log g$, $\log(n{\rm He}/n{\rm H})$). The thicker line in each plot is the best-fit model determined by SFIT2.

with them such that objects of interest can be displayed, allowing the data to be explored and connections to be made.

ACKNOWLEDGEMENTS. This work was carried out as part of the CosmoGrid project, funded under the Programme for Research in Third Level Institutions (PRTLI) administered by the Irish Higher Education Authority under the National Development Plan and with partial support from the European Regional Development Fund.

REFERENCES

Connolly A. J., Szalay A. S., Bershady M. A., Kinney A. L., Calzetti D. 1995, AJ, 110, 1071

Deeming T. J. 1964, MNRAS, 127, 493

Drilling J. S. 1996, in *Hydrogen Deficient Stars*, eds. C. S. Jeffery & U. Heber, ASP Conf. Ser., 96, 461

Drilling J. S., Jeffery C. S., Moehler S., Heber U., Napiwotzki R. 2005, in preparation

Francis P. J., Hewett P. C., Foltz C. B., Chaffee F. H. 1992, ApJ, 398, 476

Green E. M., Fontaine G., Hyde E. A., Charpinet S., Chayer P. 2006, Baltic

Astronomy, 15, 167

Jeffery C. S. 2003, in *Stellar Atmosphere Modeling*, eds. I. Hubeny, D. Mihalas & K. Werner, ASP Conf. Ser., 288, 137

Jeffery C. S., Drilling J. S., Harrison P. M., Heber U., Moehler S. 1997, A&AS, 125, 501

Jeffery C. S., Woolf V. M., Pollacco D. L. 2001, A&A, 376, 497

Murtagh F., Heck A., 1987, *Multivariate Data Analysis*, Reidel, Dordrecht

Storrie-Lombardi M. C., Irwin M. J., von Hippel T., Storrie-Lombardi L. J. 1994, Vistas in Astronomy, 38, 331