

Automatic Classification of Subdwarf Spectra using a Neural Network

C. Winter¹, C.S. Jeffery¹ and J.S. Drilling²



¹Armagh Observatory, College Hill, Armagh BT61 9DG, N. Ireland

²Dept. of Physics and Astronomy, Louisiana State University, Baton Rouge, LA 70803
cwr@arm.ac.uk, csj@arm.ac.uk, drilling@rouge.phys.lsu.edu

Abstract

We apply a multilayer feed-forward back propagation artificial neural network to a sample of 380 subdwarf spectra classified by Drilling et al. (2002), showing that it is possible to use this technique on large sets of spectra and obtain classifications in good agreement with the standard. We briefly investigate the impact of training set size, showing that large training sets do not necessarily perform significantly better than small sets. Plans for future work in this area are also outlined.

Automated Classification

Stellar spectra require the experience and judgement of a trained expert in order to be classified. However, current and future digital sky survey projects, like the SDSS, along with space-based missions, such as GAIA, will collect huge amounts of spectra – quantities human experts will be unable to cope with. In light of this, investigation into automated classification schemes as supplementary tools is becoming more urgently necessary if we are to stay ahead of the data wave.

Following past examples (Gulati et al. 1994, von Hippel et al. 1994, Bailer-Jones 1996), we aim to establish whether an artificial neural network (ANN) is capable of providing agreeable classifications for a set of subdwarf spectra previously classified by Drilling et al. (2002). Additionally, we briefly investigate how the size and content of the ANN's training data (analogous to spectral classification standards) affects its ability to provide agreeable classifications.

Data Pre-Processing

Our samples of subdwarf spectra were taken from the collection compiled by Drilling et al. (2002) from data provided by Moehler et al. (1990a, 1990b), Dreizler et al. (1990), and Theissen et al. (1993). It comprises a more-or-less representative sample of 174 PG subdwarfs and blue horizontal branch stars, plus a few other stars not included in the PG catalog.

The Drilling classification system uses a spectral type running from sdO1 to sdA (1 – 20), analogous to MK spectral classes. It introduces a helium class (0 – 40) based on H, HeI and HeII line strengths, and uses luminosity classes IV – VIII, where most subdwarfs have luminosity class ~VII. The mapping between Drilling classes and those used elsewhere, e.g. the PG survey (Green et al. 1986), is illustrated in figure 1.

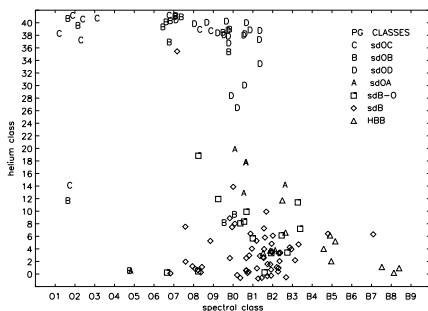


Figure 1. Comparison of Drilling spectral and helium classes with the PG classes (from Drilling et al. 2002)

Our data has been coarsely classified on the helium scale defined by Drilling et al. (2002), with a grain size of 4 helium classes.

Before applying the ANN, data must be in a homogeneous form. Simple investigations revealed dissimilarities in wavelength range, and bin size. A common wavelength range of 4300 – 4850Å was established, along with a common bin size of 0.6Å. Any spectra

unconformable to these were removed from the data set.

Crudely rectifying large cosmic spikes, and instrumental end-effects, spectra were then velocity corrected by way of a cross-correlation function. Further elimination of those spectra with no corresponding spectral classification resulted in a final collection of 380 spectra. These were then resampled onto a common wavelength range of 4200 – 4900Å, with a bin size of 0.6Å, yielding 1167 data points per spectrum.

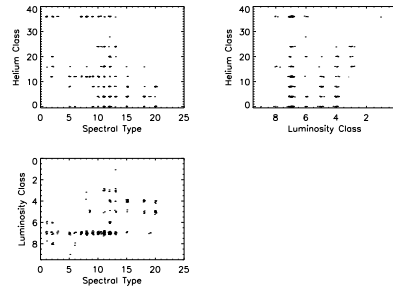


Figure 2. Jittered scatter plots showing the distribution of our final 380 spectra across the three classification dimensions. Note the concentration of spectra at Luminosity Class VII.

The Neural Network

An ANN is a statistical pattern recognition algorithm, able to perform a non-linear, parameterised mapping between two domains. Originally inspired by the structure of neuronal cells in the brain, a typical ANN consists of an interlinked, hierarchical structure of processing nodes. The interested reader should refer to Bishop (1995) for more detailed instruction.

The feed-forward back propagation neural network code STATNET, by Dr. Coryn Bailer-Jones (<http://www.mpia-hd.mpg.de/homes/calj/statnet.html>), was used in this study.

Our main objective is to show whether ANNs are able to perform the task of spectral classification, hence the ANN architecture was kept reasonably simple. A committee of 5 networks was used, with each network consisting of 1 input layer with 1167 input nodes, 1 hidden layer of 5 nodes, and 1 output node.

To test the effect of training set size on ANN performance, two training sets were created for each parameter space we wanted to classify in. One set comprised 100 spectra for training, with the resulting ANN being tested on the remaining 280 spectra. Similarly, the second training set contained 280 spectra, with the remaining 100 spectra used to test the ANN. In each case, training set samples were chosen stochastically from the main data set such that the parameter space was represented evenly. An uneven representation limits the ANN's ability to generalise, and would thus reduce performance.

Results

• ANN classification was limited to spectral type and helium class only. ~64% of our spectra reside in luminosity class VII, thereby making this class over-represented, restricting the ability of the ANN to make accurate classifications in other areas of the parameter space.

Table 1: Summary of results.

	SpT		HeC	
	100	280	100	280
RMS	2.09	1.99	4.79	4.55
r	0.89	0.90	0.92	0.94

• Beginning with spectral type, the training set of 100 spectra allowed the ANN to provide classifications to within 2.09 subtypes, with a correlation coefficient of 0.89. Training with a set of 280 spectra, classifications were accurate to within 1.99 subtypes, with a correlation coefficient of 0.90.

• In terms of helium class, a training set of 100 spectra enabled the ANN to classify to within 4.79 classes,

with a correlation coefficient of 0.92. Using the training set of 280 spectra, classifications were within 4.55 classes, with a correlation coefficient of 0.94. The rather large errors in helium classifications are due to the coarse grain of the original classification scale.

• In each case, we see that the ANN trained on a larger sample of 280 spectra yields a classification error not significantly smaller than the ANN trained using the smaller sample of 100, suggesting a large training set is not necessarily required for good performance.

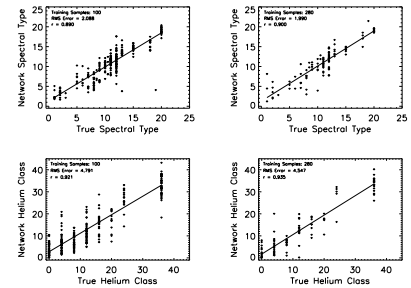


Figure 2. The scatter plots show true classifications against network classifications. Also plotted for each case is a least-squares best fit line.

Conclusions

• We have established that ANNs are capable of providing spectral classifications agreeable with those made according to the classification standards.

• In addition, a large training set is not necessarily required for the ANN to yield good results.

• Future work will allow us to determine if ANNs can yield even better performance. We plan to investigate a number of possibilities:

- Attempting to locate an optimal training set, and whether such a set should contain the classification standard spectra;
- Restricting the ANN's attention to the same spectral lines as used in defining the classification standards;
- Network structure is an important factor in ANN performance. If we vary ANN architecture, by adding a second hidden layer, adjusting the number of processing nodes, etc., what is the corresponding effect on performance?
- In many cases, our data set contains several spectra from the same star. Is the ANN giving each spectrum the same classification?
- Pre-processing spectra with Principal Components Analysis to remove noisy features and compress the number of ANN inputs;
- Increasing the size of the data set to provide a richer representation of the parameter space, allowing further studies into hot subdwarfs.

References

Bailer-Jones C. A. L. 1996, PhD thesis, University of Cambridge
 Bishop C. M. 1995, Neural Networks for Pattern Recognition (Oxford: Oxford University Press)
 Dreizler S., Heber U., Werner K., Moehler S., de Boer K. S. 1990, A&A, 235, 234
 Drilling, JS, Moehler, S, Jeffery, CS, Heber, U, and Napiewotzki, R 2002, Probing the Personalities of Stars and Galaxies, ed. Richard Gray, in press.
 Green R. F., Schmidt M., Liebert J. 1986, ApJS, 61, 305
 Gulati R. K., Gupta R., Gothoskar P., & Khobragade S. 1994, ApJ, 426, 340
 Moehler S., Richter T., de Boer K. S., Dettmar R. J., Heber U. 1990, A&AS, 86, 53
 Moehler S., Heber U., de Boer K. S. 1990, A&A, 239, 265
 Theissen A., Moehler S., Heber U., de Boer K. S. 1993, A&A, 273, 524
 von Hippel T., Storrie-Lombardi L. J., Storrie-Lombardi M. C., & Irwin M. J. 1994, MNRAS, 269, 97