3rd International Workshop on Spectral Stellar Libraries
ASI Conference Series, 2017, Vol. 14, pp 31–36
Editors: P. Coelho, L. Martins & E. Griffin

Classification of LAMOST DR3 FGK Spectra

Ranjan Gupta[1]*, H.P. Singh[2] and Yue Wu[3]
[1]*IUCAA, Post Bag 4, Ganeshkhind, Pune-411007,India*
[2]*University of Delhi, New Delhi-110007, India*
[3]*NAOC,Chinese Academy of Sciences, 20A Datun Road, Beijing-100012, China*

Received July 1st, 2017 ; accepted Oct 15th, 2017

Abstract. The Large Sky Area Multi-Object Fibre Spectroscopic Telescope (LAMOST) is producing a very large database which consists of spectra from the Chinese 4-meter telescope. LAMOST Data Release 3 (DR3 `http://dr3.lamost.org/`) contains over $5{\times}10^7$ spectra, of which $5.4{\times}10^6$ are stellar spectra. In this preliminary work we have used only the F, G and K spectral types of stars with S/N > 30. Our data set consists of 286,283 spectra that have been classified using Automated Supervised Neural Network (ANNs); this forms the test set, while the training set consists of spectra from the MILES spectral library. Of the three ANN tools used (Tree, Forest and Neural), the best performance was seen for the Tree based classifier; it returned a classification accuracy of 71.4% correct spectral types and an error of 4.61% spectral sub-types. The luminosity classes are not known for the vast majority of LAMOST spectra, but our automated schemes provide this information along with a confidence estimate (with corresponding classification probabilities) for each spectrum. Future work will involve (i) using more classification tools, (ii) improving the classification accuracies, and (iii) applying these upgrades to future LAMOST data releases.

Keywords : stellar spectra; classification; automated schemes; neural networks

1. Introduction

The Large Sky Area Multi-Object Fibre Spectroscopic Telescope (LAMOST) has recently produced a very large spectral data base; the third data release (DR3), allowing

*email: `rag@iucaa.in`

domestic access from December 2016 and international access scheduled to start in July 2017, contains more than 5×10^6 stellar spectra. Further releases (e.g., DR4, for which international access will begin in July 2018) are planned. This paper describes is a preliminary effort to classify part of the stellar spectra by using Supervised Automated Neural Network (ANNs). For more details on LAMOST see: Cui et al. (2015) and Deng et al. (2015). The classification accuracies obtained are at the level of ~71%: 71.4% of the sample of 286,283 were classified correctly for spectral types, with an error of about 4.6% in sub-types. The following sections describe a brief introduction to the ANNs, LAMOST data and its pre-processing, sample spectra and the three ANN tools used, followed by results and conclusion.

2. ANN description

ANNs have been used for more than 5 decades, but only in the past 2 decades they have found numerous applications in several areas, in particular in classifying astronomical data. A supervised ANN is comprised of three basic layers:

Input Layer: A layer where raw information is fed into the network.

Hidden Layer: The activity here is determined by the activities of the input layer and the weights of the connections between the input and the hidden layers.

Output Layer: This depends on the activity of the hidden layers and the weights between the hidden and output layers.

Figure 1 (left panel) shows the plan for a typical ANN architecture used for classification of data.

A supervised ANN includes the following stages: (i) Training (learning) and (ii) Testing session. The training session requires a training data set with examples of inputs and corresponding (desired) outputs. The process of training involves passing the data through layers on interconnecting nodes in several iterations. The errors in each iteration (in terms of deviation from the desired outputs) are noted and propagated back to the input layer by modifying the interconnection weights each time. (iii) The learning curve representing the rms errors versus number of iterations should ideally converge to zero, but the iterations are normally stopped after a certain pre-defined error level is reached. At this stage the ANN is considered to be trained and ready for the testing session. (iv) The testing session uses the trained interconnection weights on test data to classify in terms of the train set classes.

To check the quality of the training, one runs a test session on the test data and checking the resulting classification scatter plot. This step must be carried out prior to an actual testing session. The right panel in Figure 1 shows a typical scatter plot of such a ‘checking’ test, using an input set of 158 spectra. It shows that the training

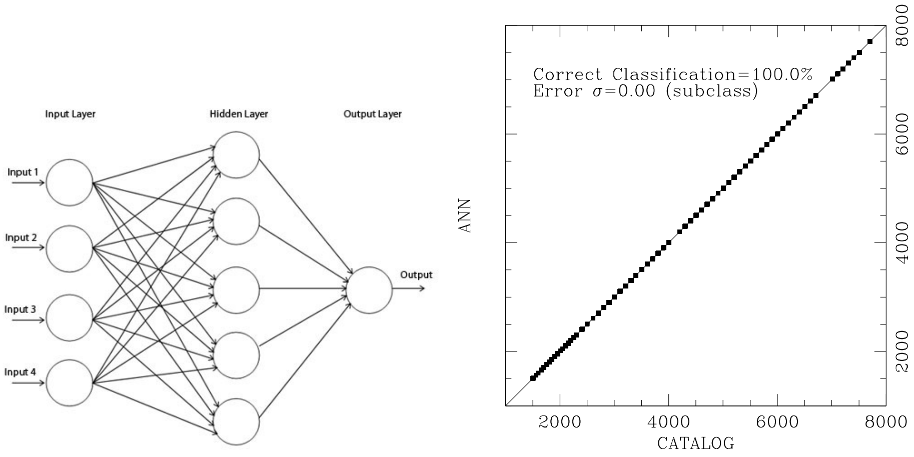

Figure 1. The left panel shows a sketch of a typical ANN Interconnecting Architecture; the right panel shows a classification scatter plot for trainer v/s trained test session

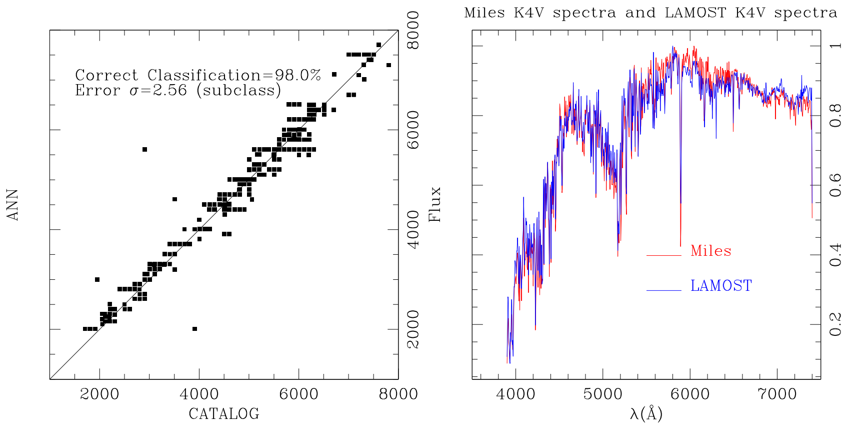

Figure 2. The left panel shows a shows a classification scatter plot for a test session. The right panel shows a LAMOST K4V spectra being matched with a MILES spectrum

has been successful, as there were 100% correct classifications and zero classification errors. The units on the X- and Y-axes represent spectral type and luminosity class. For example, a value of 3555.5 means a spectral type of A5.5 spectral type, and a luminosity class on III (first three digits stand for O, B, A, F, G, K, M and range from 100 to 700, and the last three numbers, including the decimal, represent luminosity classes I, II, II, IV, V and range from 1.5 to 9.5 in steps of 1.0). In these 2-D scatter plots the luminosity class is not represented, so a separate scatter plot is required for checking the errors in luminosity classifications.

The left panel in the Figure 2 shows a scatter plot from a typical test session that used 850 test spectra. One can notice that the scatter in the classifications indicate the classification errors (shown there as as σ = 2.56, which means 2.56 sub-type errors and classifications that were 98% correct).

3. LAMOST DR3 data and pre-processing

We used FGK spectra from LAMOST DR3 (Data Release for the years 2011–2015), in our preliminary attempt to classify the large data base. The spectral type information for this data base is provided by the LAMOST 1D pipeline (Luo et al. 2015). We selected 286,283 spectra where the S/N are > 30. Original LAMOST spectra have a resolution of 0.305 nm, and the set selected was pre-processed to the wavelength range 390–740 nm in 0.5-nm bins, with a degraded resolution of 0.45 nm (i.e. 701 spectral points). That formed the ANN test set. A training set of 63 different FGK spectral types and covering the same wavelength band was selected from the MILES spectral library (Sanchez-Blazquez et al. 2006) and pre-processed from the original 0.25-nm resolution to match that of the LAMOST spectra. The right panel of Figure 2 shows a comparison between one of the pre-processed MILES spectra and a corresponding LAMOST one.

We used our own ANN code (Gulati et al. 1994) which is a MBPN-based architecture. We also used two more classification tools from MATLAB, based on Decision Tree and Random Forest. We also attempted to reduce the size of the data set by applying Principle Component Analysis (PCA) (Singh et al. 1998); we used 15 PCAs, which covered almost 98% of the information content in the spectral data. It should be noted that since the output from the PCA did not contain the original data in full, its use will result in larger classification errors.

4. Results and Conclusions

The classification results are divided in two parts: histogram plots, and 3-D plots for the trials of ANN/Tree/Forest and the corresponding PCA results. The errors are given in spectral sub-types, e.g. a 4.0 sub-type error denotes that G5V star could be either G1V or G9V. Figures 3 and 4, respectively, show just the best classification results (i.e., the Tree classifier histogram and 3-D plot, and – in the case of the PCA output – the corresponding Forest PCA plots).

The scatter-plot correlation coefficients (in terms of correctly classified percentage) and the classification errors (in terms of sub-types) are summarized in Table 1.

Most of the classification errors seem to be shifted by about 1-2 sub-types towards higher values. That could be explained by the original LAMOST having wrong designated spectral types.

Table 1. Classification Results

ANN Tool	Correct Classification	Classification Error
ANN	66.0%	4.78
ANN-PCA	55.0%	5.60
Tree	71.4%	4.61
Tree-PCA	53.0%	5.68
Forest	67.9%	4.88
Forest-PCA	61.0%	5.20

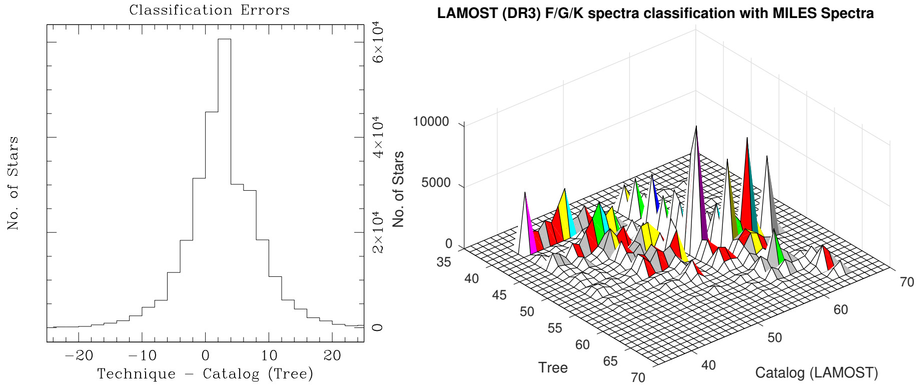

Figure 3. Left panel shows a histogram for classification errors by the Tree classifier (histogram bins are in sub-spectral types) and the right panel shows the corresponding 3-D plot

The conclusions of this work are as follows:

- We found that the best classification of the full set of spectra is achieved with the Tree Classifier; for the PCA-reduced set the best classifier was the Forest Classifier.
- Using a PCA reduces the size of the data matrix, but classification based on the full spectra will always provide better results.
- Since the LAMOST data base has only estimates of spectral types and no information about the Luminosity Class (for the majority of cases), our ANN tools will now provide the corresponding Luminosity Classes along with overall classification probabilities.
- Future work will (i) provide complete spectral classification for objects in DR3 (and later releases), (ii) provide estimates of atmospheric parameters for stars, and (iii) use more classification tools to improve the classification results.

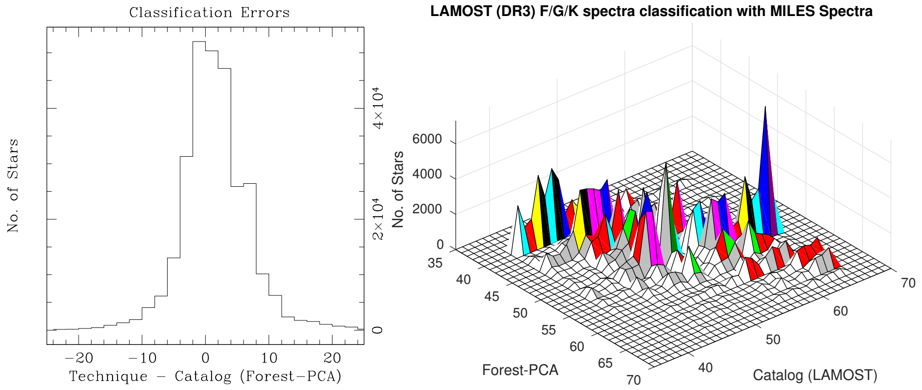

Figure 4. Left panel shows a histogram for classification errors by the PCA Forest classifier (histogram bins are in sub-spectral types) and the right panel shows the corresponding 3-D plot

5. Acknowledgements

H.P. Singh's participation at the conference was made possible through FAPESP project 2016/13479-0. Y. Wu acknowledges the National Science Foundation of China (NSFC) under grant No. 11403056 and the fund supplied by the Guangdong Provincial Engineering Technology Research Centre for Data Science.

References

[1] Cui X. Q., et al., 2012, RAA, 12, 1197

[2] Deng L. C., et al., 2012, RAA, 12, 735

[3] Gulati R. K., Gupta R., Gothoskar P., Khobragade S., 1994, ApJ, 426, 340

[4] Luo A-Li., et al, 2015, RAA, 15(8), 1095

[5] Sánchez-Blázquez P., Peletier R. F., Jiménez-Vicente J., Cardiel N., Cenarro A. J., Falcón-Barroso J., Gorgas J., Selam S., Vazdekis A., 2006, MNRAS, 371, 703

[6] Singh H. P., Gulati R. K., Gupta R., 1998, MNRAS, 295, 312