Star Cluster: The Gaia Revolution (5-7 October 2021)

The use of Deep Learning in stellar classification

Connick K.[1], Gebran M.[1], and F. Paletou[2]

[1]Department of Chemistry and Physics, Saint Mary's College, Notre Dame, IN 46556, USA

[2]Université de Toulouse, Observatoire Midi–Pyrénées, Irap, Cnrs, Cnes, 14 av. E. Belin, F–31400 Toulouse, France

Abstract

Due to an increasing improvement in technological capabilities and processes, as well as a connected broadening of of astronomical data, it has been shown that Deep Learning and Machine Learning have developed as methods to be the most efficient methods of analyzing this new data. In this case, we are applying various ML/DL techniques for the purpose of stellar spectroscopy. Having already ran tests with Principal Component Analysis (PCA) and Sliced Inverse Regression (SIR), we now turn our focus to Convolution Neural Network (CNN), among other techniques, in order to find the most accurate derivations for stellar parameters: effective temperature $T_{\rm eff}$, surface gravity $\log g$, projected equatorial rotational velocity $v_e \sin i$, microturbulence velocity ξ_t and metallicity $[\frac{M}{H}]$.

Introduction

Over the last decade, Machine Learning (ML) applications have been used within astronomy extensively due to a large amount of data that has been recovered from observatories, including both space-born and ground-based locations (Baron, 2019). This increased availability of astronomical data can be accredited largely to the implementation of extensive surveys like SDSS, RAVE, GALAH, and Gaia, among many others, and will continue this expansion for the foreseeable future. It is necessary then, to process this new data in an automated way. There are many ways in which to accomplish this including statistical approaches, dimensionality reduction, wavelet decomposition, as well as Machine Learning (ML) and Deep Learning (DL). All of these processes are examples of preformed attempts to achieve a more accurate derivation of stellar parameters including $T_{\rm eff}$, $\log g$, $v \sin i$, ξ_t, and $[M/H]$. Using techniques like PCA, SIR, and CNN allows the data to be processed efficiently and cleanly so that automated processing can be conducted on a larger scale, as to accommodate for the increase in information resulting from Machine Learning (ML) techniques.

Data preparation and PCA

Our current training databases (TDB) have been constructed from a synthetic set of stars with effective temperature ranging from 7000K to 10000K and within a wavelength range of 4450 Å to 5000 ÅṪhis range of wavelength and effective temperature was chosen to fit the sensitivity of our stellar parameters (($T_{\rm eff}$, $\log g$, $v \sin i$, $[M/H]$), as the ranges fall within the visible domain and contain both Balmer and metallic lines that are necessary for the calculations involving the spectral types selected within these projects. Additionally, the selected region has an insensitivity to microturbulent velocity which can be understood to be ξ_t =2 km/s based on the work of Gebran et al. (2014,2016). Further parameters for this study include: surface gravity, $\log g$, taken to be within the range of 2.0–5.0 dex, projected rotational velocity, $v \sin i$, calculated to be falling between 0–300 km s^{-1}, and metallicity, [M/H], ranging from -1.5 to +1.5 dex. Table 1 (below) depicts these ranges of the varying stellar parameters. In many cases with these calculations, a set of synthetic spectra with added noise, falling within the pre-established parameters and with SNR values ranging from 0 to 300, are used more often than the gathered set of observations.

Table: (1) Ranges of the parameters used for the calculation of synthetic spectra.

Parms	Range
$T_{\rm eff}$ (K)	[7 000,11 000]
$\log g$ (dex)	$[2.0, 5.0]$
$[M/H]$(dex)	$[-1.5, 1.5]$
$v \sin i$ (km s^{-1})	$[0, 300]$
$\lambda/\Delta\lambda$	60 000

Line-blanketed ATLAS9 model atmospheres (Kurucz 1992) were calculated for the purpose of these projects. ATLAS9 models are LTE plane parallel and assume radiative and hydrostatic equilibrium. This ATLAS9 version uses the opacity distribution function (ODF) of Castelli & Kurucz (2003). We included convection in the atmospheres of stars cooler than 8500 K. Convection was treated using a mixing length parameter of 0.5 for 7000 K$\leq T_{\rm eff} \leq$ 8500 K, and 1.25 for $T_{\rm eff} \leq$ 7000 K, following Smalley's (2004) prescriptions.
The grid of synthetic spectra was computed using SYNSPEC48 (Hubeny & Lanz 1992). The metallicity was scaled from -1.5 dex up to +1.5 dex with respect to the Grevesse & Sauval (1998) solar value with a step of 0.1 dex. The line list used in the synthetic spectra calculation was constructed from Kurucz gfhyperall.dat and modified with more recent and accurate atomic data retrieved from the VALD and the NIST databases.

It should also be noted, that as a step of data preparation, the TDB of the synthetic set of stars was put through a process of Principal Component Analysis (PCA) to reduce its dimension and consolidate the data prior to calculations. The synthetic spectra was collected into a matrix S of a size of N$_{spectra}$ $\times$ N$_{\lambda}$ and $\bar{S}$ being the average spectrum, along the N$_{spectra}$-axis.

$$S = \overbrace{\begin{bmatrix} \text{spectrum 1} \\ \text{spectrum 2} \\ \text{spectrum 3} \\ \vdots \\ \text{spectrum N}_{\text{spectra}} \end{bmatrix}}^{N_\lambda} \Bigg| \overbrace{\begin{bmatrix} \text{T}_{\text{eff}_1} & \log \text{g}_1 & \text{v}_\text{e} \sin \text{i}_1 & [\text{Fe/H}]_1 & \xi_{\text{t}1} \\ \text{T}_{\text{eff}_2} & \log \text{g}_2 & \text{v}_\text{e} \sin \text{i}_2 & [\text{Fe/H}]_2 & \xi_{\text{t}2} \\ \text{T}_{\text{eff}_3} & \log \text{g}_3 & \text{v}_\text{e} \sin \text{i}_3 & [\text{Fe/H}]_3 & \xi_{\text{t}3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \multicolumn{5}{c}{\text{Labels for s}_{\text{N}_{\text{spectra}}}} \end{bmatrix}}^{N_{labels}}$$

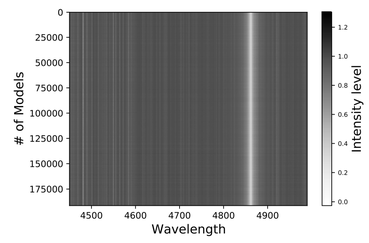

Figure: (1) Color map representing the intensities for a sample of the training databases

From which we then calculate a variance-covariance matrix C is an $N_\lambda \times N_\lambda$ matrix, as well as the eigenvectors of this matrix C. So that we can represent the original set of data in a smaller set that is represented by the first **12** coefficients. Theses coefficients are calculated by projecting each spectrum on the first **12** Principal Components. This reduction of dimensionality allows the calculations to be ran more efficiently and maintain the integrity of the larger original data set.

CNN

The utilization of a Convolutional Neural Network (CNN) enables us to further shrink a large set of data into something more manageable for calculations. Similar to other Machine Learning techniques, we must prepare the training database (TDB) for processing. Having already used PCA to reduce the size of the overall data set of synthetic stellar spectra, CNN further flattens this information through various restraints, allowing us to work with the data in a lower dimension. In this case, PCA has already been used to take down the dimension of the TDB but in many cases, layers of a technique called pooling can be used to initialize this flattening and reduction of data size by grouping together data points of similar parameters. After a combination of reduction processes, the database with which we are working can be taken down to a size containing only a few parameters per spectra. From this point, the parameters can be passed through a network like the one shown in Figure 2, containing both convolutional and fully connected layers, which allows the system to learn the data and predict the resultant stellar parameters.

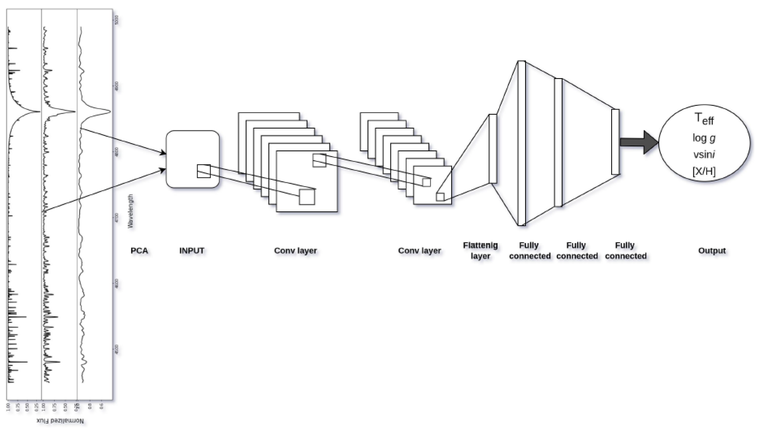

Figure: (2) CNN architecture. A PCA dimension reduction transforms the spectra into a matrix of input coefficient. This input passes through several convolutional layers and fully connected layers in order to train the data predict the stellar parameters.

Within our data set, each stellar parameter utilizes a different combination of network parameter to achieve a certain degree of accuracy. The combination of these parameters, the Signal to Noise ratio within the observations and the architectural structuring of the network all play an important role in insuring the accuracy of the stellar parameter outcome.

The main constraint of focus to insure this accuracy include: **Data augmentation**, **Initialisers: Kernel and bias**, **Optimizer**, **Learning Rate**, **Dropout**, **Pooling**, **Activation functions**, **Epochs**, **Batches**, **Loss Functions**

- Data augmentation is the technique in which the TDB increases in diversity through the application of various transformations to the previously existing TDB.
- Intializers allow us to set an initial weight to the set being ran. While there are many excellent initalizers capable of doing just this, many end up ultimately having similar accuracy, so it is important to keep this in mind while running tests with altered parameters.
- Similar to data augmentation, dropout is a regularization technique used within neural networks to decrease the effect of over fitting on the overall model. This technique randomly selects neurons from each pass of the training stage to remove them from the further stages, gathering an average of these sub-networks.
- Epochs and batches tend of have influence over one another. Altering the number of epochs within a system has an effect on the number of times that the weight of each respective neuron can be updated. This has an inverse relationship with that of batches. Batches help to avoid the over-saturation of computer memory and help to increase the speed of iterations. Finding a balance between these two can help to preserve the integrity of the TDB, allow the computer to run correctly, and maintain a reasonable speed of processing the data.

Through a number of trials concerning the current training database (TDB) we should be able to determine the optimal set of parameters for running the CNN in the most efficient way. Having used other techniques prior to CNN, like PCA, we are able to condense a large database, like the one we are using with synthetic spectral data, into a manageable set with which calculations can be ran in the most accurate and efficient of ways. Utilizing the aforementioned parameters as points for adjustments, we can find the best combination of factors for the CNN for our input data.

Preliminary results for $v \sin i$

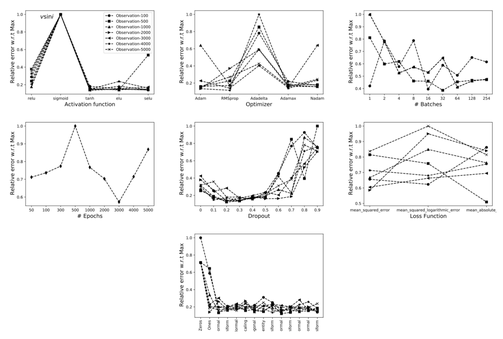

Figure: Effect of varying the CNN parameters on the accuracy of $v \sin i$for different epoch numbers. The results are displayed by dividing the observation standard deviation by their maximum value in each test.

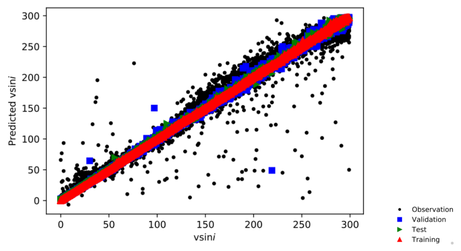

Figure: Predicted stellar parameters using the optimal CNN configurations for $v \sin i$ as a function of the input ones for the training, validation and test databases as well as for the noise added observations.

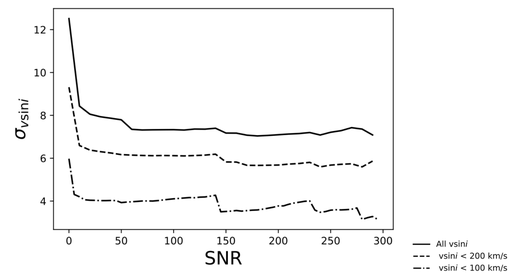

Figure: Average error bars for the observation predicted $v \sin i$ as a function of the signal to noise ratio and for different ranges of stellar rotation.

Future work:

- Find the best combination of hyper-parameters for the CNN for $T_{\rm eff}$, $\log g$, and $[M/H]$.
- Find the best architecture for the CNN.
- Introduce a correction on normalization.
- Apply auto-encoding to remove the noise of the spectra during the learning or validation.

References

Baron, D., 2019, Machine Learning in Astronomy: a practical overview. arXiv e-prints, page arXiv:1904.07248.
Castelli, F., & Kurucz, R. L. 2003, in Modelling of Stellar Atmospheres, Proc. IAU Symp., 210, 20
Gebran, M., Farah, W., Paletou, F., Monier, R., and Watson, V. (2016). A new method for the inversion of atmospheric parameters of A/Am stars. A & A, 589:A83.
Gebran, M., Monier, R., Royer, F., Lobel, A., and Blomme, R. (2014). Microturbulence in A/F Am/Fm stars. In Mathys, G., Griffin, E. R., Kochukhov, O., Monier, R., and Wahlgren, G. M., editors, Putting A Stars into Context: Evolution, Environment, and Related Stars, pages 193–198.
Grevesse, N., & Sauval, A. J. 1998, Space Science Reviews, 85, 161
Hubeny, I., & Lanz, T. 1992, Bulletin of the American Astronomical Society, 24, 1152
Kurucz, R.L. 1992, Rev. Mexicana AyA 23, 45
Smalley B., Proceedings IAU Symposium n224, 2004