*RevMexAA (Serie de Conferencias)*, **10**, 209–212 (2001)

The Seventh Texas-México Conference on Astrophysics: Flows, Blows and Glows (Austin, Texas, 6-8 April 2000)
Editors: William H. Lee and Silvia Torres-Peimbert

# PREDICTION OF STELLAR ATMOSPHERIC PARAMETERS FROM SPECTRA, SPECTRAL INDICES AND SPECTRAL LINES USING MACHINE LEARNING

Olac Fuentes and Ravi K. Gulati

Instituto Nacional de Astrofísica, Óptica y Electrónica

RESUMEN

En este trabajo presentamos un estudio experimental sobre la aplicación de un algoritmo simple de aprendizaje automático a la predicción de los parámetros atmósfericos estelares $T_{eff}, \log g$ y $[Fe/H]$ usando como datos de entrada tres conjuntos diferentes de características espectrales. Comparamos el funcionamiento del algoritmo de los 3 vecinos más cercanos ponderado por distancia usando como entrada los espectros, los índices espectrales obtenidos de la misma región de longitud de onda y las líneas de absorción obtenidas eliminando del espectro la contribución del contínuo, que es calculada por medio de un algoritmo de envolvente convexo. Los resultados experimentales muestran que las predicciones obtenidas usando índices espectrales y líneas espectrales tienen niveles de precisión muy similares, y que ambas son superiores a las obtenidas usando espectros.

ABSTRACT

In this paper we present an experimental study of the performance of a simple machine learning algorithm applied to the prediction of the stellar atmospheric parameters $T_{eff}, \log g$ and $[Fe/H]$ using as input three different sets of spectral features. We compare the performance of the distance-weighted 3-nearest-neighbor algorithm using as input spectra, a set of spectral indices taken from the same wavelength region, and absorption lines obtained by removing from the spectra the contribution of the continuum, which is computed by means of a linear time convex hull algorithm. Our experiments show that the predictions obtained using spectral indices and spectral lines have very similar accuracy levels, and that both are superior to those obtained using spectra.

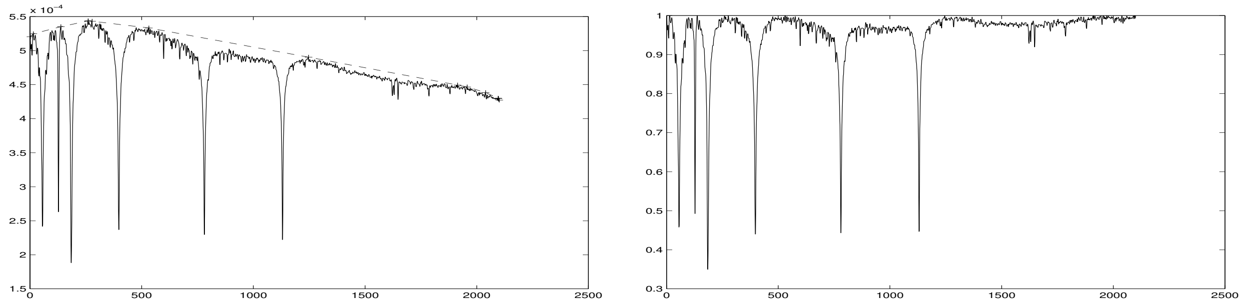
*Key Words:* **METHODS: DATA ANALYSIS — METHODS: NUMERICAL — STARS: ATMOSPHERES**

## 1. INTRODUCTION

With the new class of small telescopes equipped with modern detectors, it is feasible to perform large and deep surveys which will produce a huge and homogeneous amount of data. For example, the Sloan Digital Sky Survey is expected to produce spectra of one million galaxies, 100,000 stars and 100,000 quasars (Szalay 1999). A new paradigm is emerging from such large surveys and it poses the problem of archiving, querying and analyzing data for useful information. Computer technology has reached the point where archiving the data is not a major problem; the main problem now resides in developing tools for automated analysis of the data, since manual analysis has become unfeasible. From the analysis of the data, there will be several areas of astronomy that will be benefited, including the field of stellar population studies, since it is based on the determination of basic observational and physical properties of a large number of stars, which in turn are used to understand evolutionary phenomena in the solar neighborhood and other stellar systems.

In recent years, various techniques developed in the field of artificial intelligence have been applied to the analysis of astronomical data, in an attempt to cope with the information overload problem. By far the most commonly used approach has been artificial neural networks. Artificial neural networks have been used for

The Seventh Texas-México Conference on Astrophysics: Flows, Blows and Glows (Austin, Texas, 6-8 April 2000)
Editors: William H. Lee and Silvia Torres-Peimbert




(a) (b)

Fig. 1. (a) Normalized spectrum (solid line), convex hull approximation to the continuum (dotted lines), and convex hull points (plus signs). (b) Spectral lines obtained by dividing the original spectra by the continuum.

spectral classification of stars (Gulati et al. 1994; Bailer-Jones et al. 1997), for spectral classification of galaxies (Sodré & Cuevas 1994), for morphological classification of galaxies (Storrie-Lombardi et al. 1992), and for discriminating stars and galaxies in deep-field photographs (Odewahn & Nielsen 1994). Recent work on the application of other machine learning methods to automated classification and prediction in astronomy includes the exploration of instance-based machine learning methods and genetic algorithms (Fuentes & Gulati 2000; Ramírez, Fuentes & Gulati 2001, in preparation).

While a great amount of effort has been invested in developing automated methods for prediction and classification in astronomy, the problem of selecting the features that are most reliable as inputs to these systems has received little attention. In this paper we aim to determine which set of spectral features is most useful as an input to a learning algorithm to predict the stellar atmospheric parameters $T_{eff}, \log g$ and $[Fe/H]$.

## 2. DATA

Jones (1996) has produced a homogeneous library of spectra for 684 stars observed at KPNO with the coudé feed instrument. A set of spectral indices were measured from the spectra in the wavelength regions $3820-4500$ Å and $4780-5450$ Å by following the definition of the Lick indices (Worthey et al. 1994), the Rose indices (Rose 1994) and new Lick-type Balmer indices (Jones & Worthey 1995). For our experimental work, we used both the spectra and the indices in conjunction with the physical atmospheric parameters given in the catalog.

## 3. METHODOLOGY FOR AUTOMATED CLASSIFICATION

One of the simplest and most commonly used machine learning methods is the k-nearest-neighbor algorithm. In this method, we simply store all of the available training data, and, when a query is presented, we find the training examples that are most similar to it (its k nearest neighbors), and assign to it an output parameter that corresponds to the weighted average of the parameters of its neighbors, where the weight given to an example is inversely proportional to its distance from the query point. Thus, the output parameters of a query point are given by

$$f(x_q) = \frac{\sum_{i=1}^{k} w_i f(x_i)}{\sum_{i=1}^{k} w_i},$$

where $x_i$ is a vector containing the attributes (indices, spectrum or lines) of star $i$, $f(x_i)$ is the vector of atmospheric parameters of star $i$, and $w_i$ is the weight assigned to star $i$, given by the inverse of its Euclidean distance to star $q$

$$w_i = \frac{1}{d(x_q, x_i)}.$$

The Seventh Texas-México Conference on Astrophysics: Flows, Blows and Glows (Austin, Texas, 6-8 April 2000)
Editors: William H. Lee and Silvia Torres-Peimbert

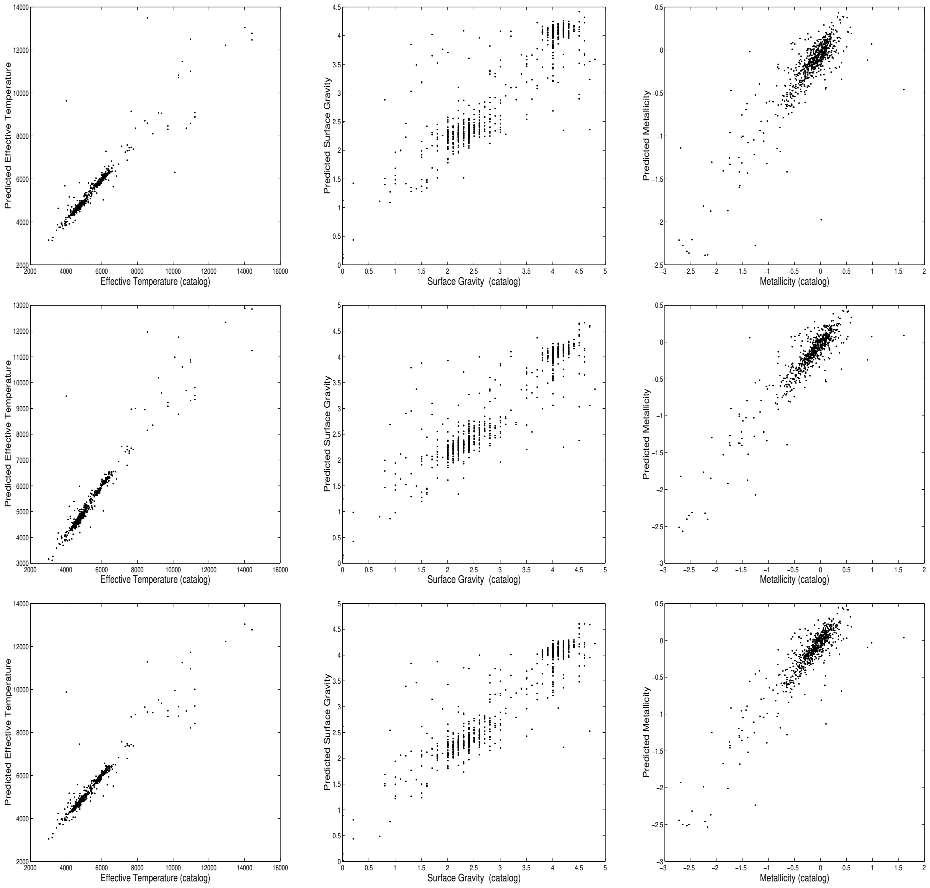



Fig. 2. Catalog versus predicted stellar atmospheric parameters obtained applying the distance-weighted 3-nearest-neighbor algorithm using as input normalized spectra (first row), spectral indices (second row) and spectral lines (third row).

Despite its simplicity, this method is often competitive with other more sophisticated machine learning methods, such as artificial neural networks and decision trees.

To find an approximation to the contribution of the continuum to each spectrum we used the linear time (given a sorted set of points) convex hull algorithm proposed by Graham[1]. Given a sequence of spectral measurements $S = \langle f(\lambda_1), f(\lambda_2), \dots, f(\lambda_n) \rangle$, the algorithm finds $H = \langle f(\lambda_{h_1}),\ f(\lambda_{h_2}),\ \dots,\ f(\lambda_{h_k}) \rangle$, the smallest subsequence of $S$ such that any point in $S$ is either under, or on the line joining a pair of consecutive points $f(\lambda_{h_i}), f(\lambda_{h_{i+1}})$ in $H$. Figure 1 shows an example of a spectrum, the continuum approximation found by the convex hull algorithm, and the absorption lines obtained by dividing the spectrum by the continuum. It should be noted that the ultimate goal of this methodology is to provide accurate and efficient prediction of the atmospheric parameters. Therefore, the accuracy of this approximation, when compared to that obtained by other automatic or semi-automatic methods, is not as important as the time that its required to compute it and the consistency it has over the whole dataset.

## 4. EXPERIMENTAL RESULTS

We applied the distance-weighted 3-nearest-neighbors algorithm described in the previous section to the prediction of $T_{eff}, \log g$ and $[Fe/H]$ using as input the three different feature sets: spectra, spectral indices and

[1] A full description of the algorithm is beyond the scope of this paper but it can be found in (Preparata & Shamos 1985).

TABLE 1

SUMMARY OF MEAN ABSOLUTE ERRORS IN THE PREDICTION OF STELLAR ATMOSPHERIC PARAMETERS USING THE DISTANCE-WEIGHTED 3-NEAREST-NEIGHBOR ALGORITHM WITH THREE DIFFERENT INPUT DATASETS.

| Data | $T_{eff}$ Error [K] | $\log g$ Error [dex] | $[Fe/H]$ [dex] |
|---|---|---|---|
| Normalized Spectra | 157 | 0.27 | 0.15 |
| Spectral Indices | 139 | 0.24 | 0.13 |
| Lines | 143 | 0.23 | 0.13 |

lines. The resulting predictions are shown in Table 2. In all the plots the horizontal axis represents the values of the atmospheric parameters found in the catalog, while the vertical axis represents the values obtained by the algorithm. Figure 2 presents a summary of the mean absolute errors obtained by the algorithm for each of the input sets. It can be seen that the predictions obtained using spectral indices and spectral lines have very similar accuracy levels, and that both are superior to those obtained using spectra. From this we can conclude that processing the spectra to obtain either indices or lines will result in an improved performance of the automatic prediction method, and that the selection of whether to use indices or lines does not have a significant effect on performance.

## 5. CONCLUSIONS

We have presented the application of the 3-nearest-neighbors learning algorithm to the prediction of stellar atmospheric parameters using as input spectra, spectral indices and spectral lines obtained by an automated method based on a convex hull algorithm. Our experimental results have shown that all three forms of information can be used to predict the parameters accurately, but that spectral indices and spectral lines yield smaller errors than spectra. Future work will include similar experiments to determine if these results generalize to other machine learning methods such as decision trees and neural networks.

We are grateful to F. Ramírez and L. Altamirano for helpful discussions. Financial support for this work has been provided by CONACyT (project number J31877).

## REFERENCES

Bailer–Jones C. A. L., Irwin M., Gilmore G., von Hippel T. 1997, MNRAS, 292, 157
Fuentes, O. & Gulati, R. K., 2000, Proceedings of ADASS IX (in press).
Gulati, R. K., Gupta, R., Gothoskar, P. & Khobragade, S. 1994, ApJ, 426, 340
Jones, L. A. 1996, Ph.D. Thesis, North Carolina University, Chapel Hill
Jones L. & Worthey G. 1995, ApJ, 446, 31
Odewahn, S. C. & Nielsen, M. L. 1994, Vistas in Astronomy, 38, pp. 281-285.
Preparata, F. P. & Shamos, M. I. 1985, Computational Geometry: An Introduction (Springer-Verlag)
Rose J. A. 1994, AJ, 107, 206
Sodré, L. & Cuevas, H. 1994, Vistas in Astronomy, 38, pp 287-291
Storrie–Lombardi, M. C., Irwin, M. J., von Hippel, T., Storrie–Lombardi, L. J. 1992, MNRAS, 259, 8p
Szalay, A. S. 1999, The Sloan Digital Sku Surveys, In: Computing in Science and Engineering, 54
Worthey, G., Faber, S. M., Gonzalez, J. J. & Burstein, D. 1994, ApJS, 94, 687

Olac Fuentes and Ravi K. Gulati, Instituto Nacional de Astrofísica, Optica y Electrónica, Luis Enrique Erro # 1, Santa María Tonantzintla, Puebla, 72840 México, (fuentes, gulati@inaoep.mx).

The Seventh Texas-México Conference on Astrophysics: Flows, Blows and Glows (Austin, Texas, 6-8 April 2000)
Editors: William H. Lee and Silvia Torres-Peimbert