AUTOMATIC CLASSIFICATION OF STELLAR SPECTRA

ICIAR CARRICAJO, MINIA MANTEIGA OUTEIRO
Departamento de Ciencias de la Navegación y de la Tierra, Universidade da Coruña, E-15011 A Coruña, SPAIN
ALEJANDRA RODRIGUEZ, CARLOS DAFONTE, BERNARDINO ARCAY
Departamento de Tecnología de la Información y de las Comunicaciones, Universidade da Coruña, E-15071 A Coruña, SPAIN

Abstract: We propose and discuss the application of Artificial Intelligence techniques to the classification of stellar spectra. Two types of systems are considered, knowledge-based systems (Expert Systems) and different classes of neural networks. After analysing and comparing the performance of both systems in the classification of stellar spectra, we reach the conclusion that neural networks are more adequate to determine the spectral types and luminosity of stars, whereas knowledge-based systems are more performative in determining global temperatures.

In order to determine the best approach to the classification of each spectrum type, we describe and analyse the performance and results of various neural networks models. Backpropagation networks, self-organising maps and RBF networks in particular were designed and tested, through the implementation of different topologies, to obtain the global classification, spectral type and luminosity of stars. The best networks reached a success rate of approximately 97% for a sample of 100 testing spectra.

The morphological analysis algorithms that were developed in the knowledge-based systems are used to extract and measure spectral features, and to obtain the input patterns of the neural networks. Some networks were trained with this parameterisation, others with flux values of specific spectral zones; it was the first strategy that resulted in a better performance.

Our approach is focused on the integration of several techniques in a unique hybrid system. In particular, signal processing, expert systems, fuzzy logic and artificial neural networks are integrated by means of a relational database, which allows us to structure the collected astronomical data and to contrast the results of the different classification methods.

In addition, we designed several models of artificial neural networks that were trained with synthetic spectra, and included them as an alternative classification method.

The proposed system is capable of deciding the most appropriate classification method for each spectrum, which widely opens the research in the field of automatic classification.

1 Introduction

The MK Spectral classification system (MK system) was proposed by W.W. Morgan and P.C. Keenan with the publication of the first photographic spectral classification atlas, *An Atlas of Stellar Spectra*, [1]. Ever since that publication, the MK system has been revised and refined by Morgan, Keeman and others.

The MK classification system is defined by a set of standard stars and is based on the visual appearance of the spectra.

The classification process is often directly performed by experts, who analyse and classify the spectra by hand. Not only are these manual techniques very time-consuming and involve a great amount of human resources, they also constitute a subjective process, since a given spectrum may be classified differently by different people. These problems could be resolved through the use of computational techniques.

Among the different techniques of Artificial Intelligence, knowledge-based systems and neural networks seem the most appropriate answers to approach the problem of stellar classification. Knowledge-based systems can reproduce the spectral classification reasoning of the experts in the field. Neural networks have already proved their success in classification problems [2]: they are generally capable of learning the intrinsic relations that reside in the patterns with which they were trained.

Some well-known previous works have also applied this Artificial Intelligence technique to the problem of stellar classification ([3], [4]), obtaining different grades of resolution in the classification. Rather than trying to test models that have already demonstrated their suitability, we implement different models of neural networks that allow us to perform a sensibility analysis of this technique in the classification of spectra. We simultaneously try to determine the best learning algorithm and the best network structure for this specific problem.

Having tested both techniques (expert systems and neural networks) we are ready to analyse their respective adaptation to the problem, and to compare their results. Our main objective is the formalisation of a hybrid system that integrates all the developed artificial techniques and is able to choose the most appropriate classification method for each spectrum type. Because it combines several techniques, this type of system is more versatile than a system that uses only one technique, and it presents a greater capability of adaptation to the stellar classification problem.

The different methods, algorithms and techniques that were used for the development of the proposed system are described in the next sections.

2 Morphological Analysis

The spectra are stored in a relational database that was implemented with PostgreSQL running under Linux [5]. As a first step in the pre-processing module, the unclassified spectra are retrieved from the astronomical database and adjusted to flux 100 at wavelength 5450 Å for their comparison with a reference spectral catalogue. We opted for the Silva spectral catalogue [6] because of its completeness and coverage. The 50 spectra of this catalogue are sampled in the range of 3500 to 8900 Å with 5 Å resolution, and scaled to flux 100 at 5450 Å.

Our analysis considers 10 bands, 9 lines and the relevant relationships between them as classification parameters. Given the fact that a spectrum is a signal that relates wavelengths to energy fluxes, we included signal processing techniques to search and measure the spectral features [7]. The implemented algorithms are mainly based on continuum estimation and energy calculation.

To calculate the intensity of each line accurately, we estimate the local spectral continuum. The signal is smoothened with a low pass filter that excludes the peaks in an interval around the sample where the line was detected. This filter is implemented by a five-point moving average method that selects the five more stable fluxes. That is

$$C_j = \left(\frac{\sum_{j-n}^{j+n} E_i * X_i}{N} \right) \ , \tag{1}$$

where C_j is the estimation of the continuum for the sample j, E_i is the flux in the sample i, N is the number of values used in the moving average method to calculate the local spectral continuum, and X is a binary vector that indicates the representative fluxes of the spectral continuum in the zone. $X_i = 1$ if E_i is a flux value representative of the local spectral continuum, and $X_i = 0$ if E_i is a peak. The intensity is positive for the absorption lines and is negative for the emission lines.

As for the molecular bands, we only have to measure their energy to decide whether they are significant. In this case, the upper threshold line for each band is calculated by means of linear interpolation between the fluxes in the limits of the interval defined for each band. Then, the area between this line and the axis of abscissas is calculated with a discrete integral, and the area that surrounds each band is calculated by integrating the flux signal between the extremes of the band. Finally, the flux of the band is obtained by subtracting both calculated energies. That is

$$B_{lr} = \int_l^r L(\lambda_i) - \int_l^r E(\lambda_i) \ , \tag{2}$$

where B_{lr} is the value of the band between the samples l and r, L is the projection line, E is the flux function, λ the wavelength, l the left limit of the band and r the

right limit. The obtained value becomes more negative as the band becomes deeper and wider, so positive or negative fluxes close to zero are not considered bands.

The morphological analysis module was developed in C++ [8]. The graphical options include the representation of the absorption/emission lines and the molecular bands that were detected during the analysis. This facility is essential when debugging the system with the help of the human experts.

3 Automatic Classification

3.1 Expert System

In this computational approach, the classification module simulates the manual process of stellar classification. Since the human reasoning in this field includes uncertainty and imprecision, we designed an expert system that combines traditional production rules with credibility factors [9] and fuzzy logic [10]. The manual classification criteria are the result of the experience of experts in classifying spectra and are combined in a forward reasoning. Our study considers approximately 200 classification criteria.

As a previous step to the design of the expert system, we carried out a sensibility analysis of the classification parameters in order to define the different fuzzy sets, variables and membership functions. We analysed the parameters of the spectra from the reference catalogue by means of the previously described algorithms, and determined the different spectral types that each parameter discriminates. As a final result of this analysis, we defined as many fuzzy variables as levels of classification (global, type and subtype) for each luminosity class, as well as the fuzzy sets and membership functions determined by the values of the spectral features in the spectra from the guiding catalogue.

The developed expert system stores the information that is necessary to initiate the reasoning process in the *base of facts*. This descriptive knowledge of the spectra is represented by means of frames [11], that is, objects and properties structured by levels. We opted for this model because it is the simplest and most adequate to transfer the analysis data to the classification module, and because it allows the equivalence between analysis data and knowledge. The knowledge of the base of facts includes general information, e.g. the name of the stars, and the results of the morphological analysis, i.e., the value of the classification parameters.

The real parameters of spectral classification and the limit values of each type and subtype were included in the expert system in the shape of fuzzy rules. The *base of rules* is the part of the system where the human classification criteria are reproduced. We adopted production rules of the IF-THEN type to implement this module because they easily reproduce the reasoning followed by the experts in the

field. The conditions of these rules refer to the values of the parameters that are stored in the current base of facts (working memory). The conclusions allude to the three levels of spectral classification. In this way, this module actively communicates with the base of facts.

We have used the Shortliffe and Buchanan methodology [9] to carry out an evolution that includes fuzzy sets and membership functions, contextualised for each spectral type and allowing superposition between them. In addition, we obtain the spectral classification of stars with a probability value that indicates the confidence grade. Sometimes this module can conclude an alternative classification of the spectra, in the case of obtaining a first classification with a significantly small truth value.

Since it could be interesting for the user to follow the reasoning process, the system includes an explanation module in the rules base which reveals how the system reached a final conclusion.

The classification module was developed in OPS/R2 [12].

3.2 Artificial neural networks

The application of neural networks to the problem of stellar classification requires a complete and consistent set of spectra that constitute the basis on which those networks are designed and tested. Our research team selected 285 spectra from public catalogues [6] [13], and a number of spectra from various telescopes; this selection covers all the known types and luminosities and guarantees a continuous transition of the spectral features between each spectral type and its adjacent types. In order to obtain the best possible generalisation of the networks, the training set was built with approximately 50% of the spectra of each spectral type and luminosity, leaving the remaining 50% for the validation and testing of the networks. The input patterns include the measurement of 25 spectral features that can be divided into three categories:

- Absorption and emission lines: including hydrogen, helium and metallic lines (Ca, K).
- Molecular bands: hydrogen and carbon absorption bands.
- Rates between lines: CH-K rates, He-H rates.

These spectral features are extracted and measured by means of the signal processing algorithms described in Sect. 2. Once the spectral analyser obtains the input values, they are normalised and presented to the neural network. We have standardised the inputs of the networks by applying a specific sigmoidal function to each parameter:

$$1/\left(1+e^{-(a*x+b)}\right) \text{ with } a>0\ . \tag{3}$$

The input patters of the proposed models consist of the complete set or a subset of the 25 normalised spectral parameters, although for some networks we have considered full spectral zones in the training process.

The neural networks that were used in the experimentation are based on both supervised and non-supervised learning models, in particular backpropagation, Kohonen and Radial Basis Functions (RBF) networks. The networks were trained with these three models and by applying different topologies, including global and hierarchic approaches; we have also implemented several enhanced learning algorithms. The topologies, the learning functions and the results obtained by these networks are described below.

Backpropagation Networks

Backpropagation is a supervised learning algorithm that belongs to the general feed-forward model. This model is based on two learning stages: forward propagation and backward propagation.

Training a feed-forward neural network with supervised learning consists of presenting a set of input patterns that are propagated forward by the net until the activation reaches the output layer. This is phase is called the forward propagation phase. When the activation reaches the output layer, the output is compared with the teaching input (provided in the input patterns). The error, or difference between the output and the teaching input of a target output unit, is then used together with the output of the source unit to compute the necessary changes of the link between both units. Since the errors are propagated backwards, this phase is called backward propagation [14].

We have made use of three different backpropagation learning algorithms:

- Standard backpropagation: this very common learning algorithm updates the weights after each training pattern.
- Enhanced backpropagation: this algorithm uses a momentum term that introduces the old weight change as a parameter for the computation of the new weight change.
- Batch Backpropagation: in standard backpropagation, an update step is performed after each single pattern, whereas in batch backpropagation all the weight changes are summed over a full presentation of all the training patterns (one epoch). Only then, an update with the accumulated weight changes is performed.

We have tested the three backpropagation learning algorithms for the spectral types, spectral subtypes and luminosity classes. As for the topology, the different implemented networks are shown in Table 1. These topologies were tested for the three backpropagation learning algorithms.

Network	Input patterns	Hidden layer
Type/subtype	25 spectral parameters	10
Type/subtype	25 spectral parameters	5x5
Type/subtype	25 spectral parameters	10x10
Type/subtype	16 spectral parameters	10
Type/subtype	16 spectral parameters	10x5x3
Type/subtype	400 flux values	100x50x10x3
Luminosity	25 spectral parameters	10x10
Luminosity	16 spectral parameters	10x5x2

Table 1: Results of various Topologies for Backpropagation Networks.

The aforementioned networks were implemented with the Stuttgart neural network Simulator (SNNS v.4.1).

Kohonen Networks

Kohonen's Self-Organising Map (SOM) algorithm is based on non-supervised learning. SOMs constitute a unique class of neural networks, because they construct topology-preserving mappings of the training data there where the location of a unit carries semantic information [15].

Self-Organising maps consist of two unit layers: a one-dimensional input layer and a two-dimensional competitive layer, organised as a 2D grid of units. Each unit in the competitive layer holds a weight vector that, after training, resembles a different input pattern.

The learning algorithm for the SOM networks meets two important goals: the clustering of the input data, and the spatial ordering of the map, so that similar input patterns tend to produce a response in units that are close to each other in the grid. In the learning process, the input pattern vectors are presented to all the competitive units in parallel, and the best matching unit is chosen as a winner.

We have tested Kohonen networks for the spectral types/subtypes and luminosity classes, using bidimensional maps from 2x2 to 24x24 units.

RBF Networks

Networks based on Radial Basis Functions (RBF) combine non-supervised learning for hidden units and supervised learning in the output layer. The hidden neurons

apply a radial function (generally Gaussian) to the distance that separates the input vector and the weight vector that each one stores, called centroid [14].

We tested the RBF algorithm for the spectral types, spectral subtypes and luminosity classes. As for the topology, the different networks that were implemented are shown in Table 2.

Network	Input patterns	Hidden layer
Type/subtype	25 spectral parameters	16
Type/subtype	25 spectral parameters	8
Type/subtype	25 spectral parameters	4
Type/subtype	16 spectral parameters	8
Type/subtype	16 spectral parameters	4
Type/subtype	400 flux values	124
Luminosity	25 spectral parameters	8
Luminosity	16 spectral parameters	8

Table 2: Results of various Topologies for RBF Networks.

3.3 Synthetic Spectra

We used version 2.56 of the SPECTRUM software, written by Richard O. Gray, to generate a set of synthetic spectra. Spectrum is a stellar spectral synthesis program that computes the emerging flux from a stellar atmosphere under the assumption of Local Thermodynamic Equilibrium (LTE). It considers most of the atomic and molecular transitions in the optical spectral region 3500 Å to 6800 Å, suitable for computing synthetic spectra with temperatures between approximately 4500K and 20000K. The details on the physics included in SPECTRUM can be found in [16].

We also selected the atmospheric models set that was calculated by Robert Kurucz [17]. Each model is characterised by four parameters: effective temperature, T_{eff}, metallicity $[M/H]$, microturbulence velocity V_{micro} and surface gravity, log g. These parameters must be specified to generate each of the synthetic spectra. We generated a total of 170 solar metallicity spectra with effective temperatures ranging from 4000K to 15000K, and the surface gravity, log g, with values of 0.5, 1.0, 1.5, 2.0, 2.5, 3.0. This set of synthetic spectra covers the spectral range K-B with luminosity classes I, II and III (giants and super giants).The synthetic spectra were generated with a sufficiently small wavelength step, 0.02 Å, and a microturbulence velocity of 2.0 Km s^{-1} over the 3500-6800 Å range.

The neural networks that were used for the experimentation are based on supervised learning models. We tested various topologies and enhanced learning algorithms on backpropagation networks, as can be seen in Table 3.

The input patterns of the nets are the 659 flux values (from 3510 Å to 6800 Å); the output is a continuous function of the effective temperature. The networks were trained with the whole set of 170 synthetic spectra and tested with the spectra from the Silva catalogue [6].

Network	Input patterns	Hidden layers
Type/subtype	659 flux values	10x5x3
Type/subtype	659 flux values	5x5
Type/subtype	659 flux values	100x50x10x3

Table 3: Results of various Topologies for Synthetic Spectra.

4 Hybrid System

After analysing the performance of the proposed techniques, we integrate them into a unique system, implemented in C++, that guarantees a reliable, consistent and adapted classification of stars.

The final analysis and classification system can be divided into three main logical modules, as shown in Figure 1.

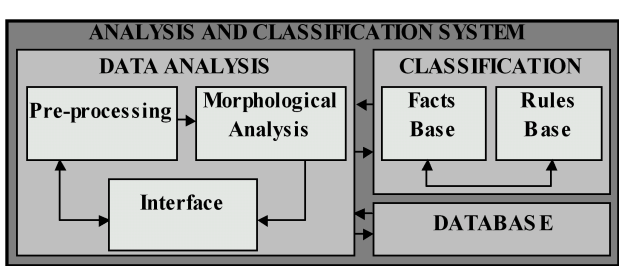

Figure 1: General System Scheme.

The relational database was developed to securely store and organise astronomical data. The data analysis module makes an exhaustive morphological analysis (calculation of maxima, minima, energy, etc.) of the spectra, treating them as a temporal series in order to obtain numerical parameters. The classification module is based on the development of expert systems and artificial neural networks that obtain the temperature and luminosity of stars through the parameterisation that resulted from the morphological analysis.

5 Results

According to the classification made by the human experts involved in this project, the automatic hybrid system is able to classify stars with an error rate below 20%.

Table 4 shows a comparison between the developed automatic classification systems and two human experts. We have analysed the performance of each technique for each classification level: global temperature of the star (early, intermediate, late), spectral type (BAFGKM), and luminosity level (I,III,V). In the neural network approach, ambiguous classifications were considered errors (outputs in [0.45, 0.55]). As for the expert systems, classifications with a low probability (less than 75%) were excluded.

Technique	Global Temp.	Spectral Type	Luminosity
Human Expert A	99%	92%	81%
Human Expert B	95%	85%	70%
Expert Systems	96.5%	88%	65%
Expert Systems with fuzzy logic	98.6%	90.3%	78.2%
Backpropagation Networks	97%	95.4%	81%
Kohonen Networks	80%	65%	60%
RBF Networks	95%	93%	79%

Table 4: Performance of the tested Classification Techniques

In the synthetic approach, the networks were tested with spectra from the Silva catalogue [6]. The preliminary results show that the net is able to correctly classify 80% of the input spectra (with a maximum deviation of 300 K in the worst cases).

6 Conclusions

This paper has presented an approach to the automation of the spectral analysis and classification process, by means of a hybrid system that provides users with a comfortable tool for spectra processing.

By integrating signal processing, knowledge-based techniques, fuzzy logic and artificial neural networks, we obtained a very satisfactory emulation of the current classification process.

Finally, all the artificial techniques were integrated into a hybrid system that determines the most appropriate classification method for each spectrum. This implies that our hybrid approach becomes a more versatile and more flexible automatic technique for the classification of stellar spectra.

The final system classifies more than 80% of the tested stars, confirming the conclusion that neural networks are more performative in determining the spectral types

and luminosity of stars, whereas knowledge-based systems obtain a higher performance in determining the global temperature.

As an additional research, we have generated synthetic spectra and used them to train backpropagation networks that determine the temperature of stars on the basis of full spectral regions. The obtained results encourage us to continue in this direction: it offers the advantage that physical properties, such as effective temperatures, gravity or metal contents, could eventually be extracted from the classified spectra.

At present, we are refining the expert systems towards new aspects of the spectral analysis; we are working on the design of new neural networks, based on synthetic spectra, that refine the current classification system; and we are completing the development of our stellar database, STARMIND (http://starmind.tic.udc.es), to make it accessible through the Internet. Our aim is to enable users worldwide to store and classify their spectra, and directly contribute to improve the adaptability and accuracy of our automatic analysis and classification system.

Acknowledgements: The authors acknowledge support from grants AYA2000-1691 and AYA2003-09499, financed by the Spanish Ministerio de Ciencia y Tecnología.

References

[1] Morgan, W. W., Keenan, P.C., Kellman, E. 1943, An Atlas of Stellar Spectra with an outline of Spectral Classification. University of Chicago Press, Chicago
[2] Haykin, S. 1994, Neural Networks: A Comprehensive Foundation, Macmillan College Publishing, New York
[3] Weaber, W.B., Torres-Dodgen, A.V. 1995, ApJ, 446, 300
[4] Singh, H.P., Gulati, R.K., Gupta, R. 1998, MNRAS, 295, 312
[5] Momjian, B. 2000, PostgreSQL: Introduction and Concepts, Addison-Wesley Pub Co.
[6] Silva, D.R., Cornell, M.E. 1998, PASP, 110, 863
[7] Kalouptsidis, N. 1997, Signal Processing Systems: Theory and Design, Wiley & Sons
[8] Hollingworth, J., Butterfield, D., Swart, B., Allsop, J. 2000, C++ Builder 5.0. Developer's Guide, SAMS
[9] Buchanan, B., Shortliffe, E. 1984, Ruled-Based Expert Systems, Adisson-Wesley
[10] Zadeh, L. 1965, Fuzzy Sets. Information and Control, 8, 338
[11] Sowa, J. F. 1999, Knowledge Representation: Logical and Computational, Brooks Cole Publishing Co
[12] Forgy, C. L. 1986, The OPS'83 User's Manual System Version 2.2., Production Systems Technologies Inc.
[13] Pickles, A.J. 1992, ApJ, 81, 865
[14] Hilera, J.R., Martínez, V. 1995, Redes Neuronales Artificiales: Fundamentos, modelos y aplicaciones; RA-MA Eds, Madrid

[15] Kohonen, T. 1987, Self-Organization and Associative Memory, Springer-Verlag, New York
[16] Spectrum: A Stellar Spectra Synthesis Program. [Online]. Available: http://www1.appstate.edu/dept/ physics/spectrum (2003)
[17] Kurucz, R. L. 1979, ApJ, 40, 1