Distributed Genetic Algorithm for feature selection in Gaia RVS spectra. Application to ANN parameterization

Diego Fustes, Diego Ordóñez, Carlos Dafonte, Minia Manteiga and Bernardino Arcay

Abstract This work presents an algorithm that was developed to select the most relevant areas of a stellar spectrum to extract its basic atmospheric parameters. We consider synthetic spectra obtained from stellar atmospheres models in the spectral region of the RVS instrument (Radial Velocity Spectrograph) of the European Space Agency's Gaia spatial mission. The algorithm that demarcates the areas of the spectra sensitive to each atmospheric parameter (effective temperature and gravity, metallicity, and abundance of alpha elements) is a genetic algorithm, and the parameterization takes place through the learning of the artificial neural networks. Due to the high computational cost of processing, we present a distributed implementation in both multiprocessor and multicomputer environments.

Diego Fustes
Department of Information and Communications Technologies, University of A Coruña, 15071, A Coruña, Spain, e-mail: dfustes@udc.es

Diego Ordóñez
Department of Information and Communications Technologies, University of A Coruña, 15071, A Coruña, Spain, e-mail: dordonez@udc.es

Carlos Dafonte
Department of Information and Communications Technologies, University of A Coruña, 15071, A Coruña, Spain, e-mail: dafonte@udc.es

Bernardino Arcay
Department of Information and Communications Technologies, University of A Coruña, 15071, A Coruña, Spain, e-mail: cibarcay@udc.es

Minia Manteiga
Department of Navigation and Earth Sciences, University of A Coruña, 15071, A Coruña, Spain, e-mail: manteiga@udc.es

1 Introduction

The objective of the ESA's Gaia Mission is to survey the Milky Way, gathering data that will allow us to study the composition and evolution of our galaxy, as well as advancing extragalactic studies in general. To achieve this Gaia will observe the distribution, kinematics and physical characteristics of stars over a representative fraction of the Galaxy's volume, with the goal of understanding its dynamics and structure. See [1] for further details.

Our research group is a member of Gaia's scientific team (Gaia DPAC, Data Processing and Analysis Consortium), which was created to develop the Gaia data reduction algorithms, including classification and parameterization tasks. In this article we will centre on the determination of stellar atmospheric parameters, like effective temperatures, superficial gravities, overall metallicity and abundances of alpha elements.

Parameterization of main stellar atmospheric properties from a stellar spectrum is a well-known problem in astrophysics. In this work we present the implementation of a genetic algorithm to select the most accurate inputs to train the ANN which subsequently will perform the parameterization of Gaia's RVS spectra.

Some of the most challenging issues are both the high dimensionality of the spectra and the huge number of objects scanned , causing big requirements regarding to computational costs. Therefore we discuss several ways to distribute the computation of ANNs and Genetic Algorithms among a bunch of computers, in order to reach scalable solutions for data-intensive tasks.

2 Gaia RVS synthetic spectra

The synthetic stellar spectra that we are using to perform our tests was compiled by A.Recio-Blanco and P. de Laverny from Niza Observatory, and B.Plez from Montpellier University [6]. The library has a total of 9048 samples, with wavelenghts between 847.58 and 873.59 nm, resolution of 0.0268 nm and a number of 971 points per signal. The dataset was arbitrarily divided into two subsets, in a proportion of 70%-30%. The first subset will be used to train the algorithms, the second for testing.

White noise was added to the synthetic spectra, obtaining datasets with different SNR values: 5, 10, 25, 50, 75, 100, 150, 200 and ∞.

Previous works in this field have demonstrated the suitability of Artificial Neural Networks (ANN) to perform automated parameterization on astronomical archives [2] [5]. Even though that in general terms the obtained results are good, we believe that the challenge of parameterizing Gaia enormous data volume demands the use of highly efficient algorithms.

2.1 Signal processing

A preprocessing stage, previous to the proper process of parameterization was included in order to refine the algorithm performance, to reduce dimensionality and to filter noise. This include the use of wavelet transform [4] and principal component analysis (PCA) [3].

PCA is based in the spectra only and therefore, in principle, it can not be specialized to select those points relevant to predict a specific parameter. To include this functionality, we decided to develop a genetic algorithm which select the relevant features as a function of the parametrization results. In such way, it can specialize in the derivation of each of the parameters. In the following sections, we show how we apply this algorithm to both original spectra (flux vs. wavelength spectra) and Wavelet transformed signal of the spectra.

3 Genetic Algorithm for feature selection

Evolutionary computation is based on processes which can be observed in nature, like the reproduction of the species and the survival of the strongest individuals. Genetic algorithms are iterative processes where the best individuals of a population are selected to reproduce and pass to the next generation. In our case we configured the genetic algorithm as follows:

1. Initialize population: A population of predefined size is generated. Each individual is represented by a chromosome which is composed by a binary alphabet. Each 1 value represents that the pixel at this position in the spectra is exactly selected. We assigned a probability of 30% for the pixel selection because we look for reducing the dimensionality in such factor.
2. Evaluate fitness: The evaluation of the fitness of an individual begins with the application of the Chromosome's mask to both the train set and the test set. After that, a Feed Forward Neural Network is created and trained during a specified number of epochs. Finally we perform a parametrization test and we calculate the fitness as the inverse of the mean error obtained through the tests.
3. Parent selection: 50% of the population is selected with the classic method of the roulette in order to reproduce among them.
4. Crossover: In pairs, the parents are crossed to generate one son and one daughter. The crossover is performed by means of mixing the parent's alphabets.
5. Mutation: The children generated in the previous stage can suffer mutations in some of their mask's pixels with a probability of 5%.
6. Selection: First we select the 10% of the best individuals (among the current population and the generated children) to pass to the next generation. Individuals needed to complete the population size are selected again with the roulette method. The algorithm pass to the next generation, beginning at step 2.

3.1 Distributed computation of fitness function

Most of the computation charge of the described algorithm is due to the cost of performing an ANN training with the RVS spectra, when calculating the fitness function. In order to improving the efficiency of computation, we have dedicated lot of efforts in distribute the computation charge among several CPUs, of which we have extracted several levels of distribution:

- ANN distribution: Neural Networks are massive parallel systems, in the sense of that the calculations of one net layer can be performed concurrently. Our experiments indicate that the cost of software threading and synchronization is big in comparison with the benefits of distribution. Probably a hardware solution will reach better improvements.
- ANN Learning distribution: The learning can be distributed as long as it is configured in batch mode. In our case, the online mode has demonstrated better behaviour, so this type of distribution has been rejected.
- Fitness distribution: Since the genetic algorithm population is composed by a bunch of individuals, the fitness evaluation can be distributed with few restrictions. In this case we opted for distributing the individuals and both the training set and the test set to each CPU. We have implemented this model through OpenMP and MPI in the case of C++ and Apache Hadoop in the case of Java.

4 Results

Table 1 Results of applying the genetic algorithm with SNR200 in several domains and with several configurations.

Domain	**Parameter**	**Population size**	**ANN iterations**	**Mean error**	**Std. deviation**	**Pixels**
Wavelength	Teff	28	100	111,26	173,32	289
Wavelength	logg	28	100	0,172	0,238	284
Wavelength	[Fe/H]	28	100	**0,109**	0,177	285
Wavelength	[α/Fe]	28	100	**0,059**	0,08	285
Wavelet	Teff	28	100	114,08	191,15	283
Wavelet	logg	28	100	0,167	0,243	305
Wavelet	[Fe/H]	28	100	0,126	0,201	298
Wavelet	[α/Fe]	28	100	0,066	0,087	278
Wavelet	Teff	64	100	**106,29**	176,8	285
Wavelet	logg	64	100	0,162	0,229	291
Wavelet	[Fe/H]	64	100	0,116	0,189	301
Wavelet	[α/Fe]	64	100	0,06	0,083	297
Wavelet	Teff	28	1000	108,89	182,47	283
Wavelet	logg	28	1000	**0,158**	0,229	287
Wavelet	[Fe/H]	28	1000	0,113	0,187	305
Wavelet	[α/Fe]	28	1000	0,062	0,081	314

The results of feature selection in both Wavelength and Wavelet domains for the estimation of stellar parameters are presented in table 1. We show the results of four experiments performed when training a test set at SNR200, where we configure the algorithm with different population sizes and ANN training steps. Note that we show the mean error and the standard deviation of an ANN trained during 5000 epochs with the spectra reduced by the ROI selection. The best results (lower errors) are highlighted.

References

1. Gaia information web page
   http://www.rssd.esa.int/index.php?project=GAIA
2. Kaempf, T., Willemsen, Bailer-Jones, C. and de Boer, K. (2005). Parameterisation of rvs spectra with artificial networks first steps. *10th RVS workshop. Cambridge.*
3. Harinder, Gulati, and Gupta (1998). Stellar spectra classification using principal component analysis and artificial neural networks. *MNRAS*, 295
4. Mallat, S. (1989). A theory of multiresolution signal decomposition: The wavelet representation. *Proc. IEEE Trans on Pattern Anal. and Math. intel.*, 11(7):674-693.
5. Ordonez, D., Dafonte, C., Arcay, B. and Manteiga, M. (2008). Parameter extraction from RVS stellar spectra by means of artificial neural networks and spectral density analysis. *Lecture Notes in Artificial Intelligence*, 5271:212-219.
6. Recio-Blanco, A., Bijaoui A. and de Laverny, P. (2006), *MNRAS* 370, 141.