Kalpa Publications in Computing

Volume 14, 2023, Pages 73–76

Proceedings of V XoveTIC Conference. XoveTIC 2022

Dust extinction removal from BP/RP (Blue Photometer/Red Photometer) spectra gathered from the Gaia Space Telescope using machine learning algorithms*

Lara Pallas-Quintela[1], Melissa Ness[23], Minia Manteiga[1], and Carlos Dafonte[1]

[1] Research Group LIA2, Department of Computing, CITIC Research Center, Universidade da Coruña, 15071 A Coruña, Spain
lara.pquintela@udc.es, minia.manteiga@udc.es, carlos.dafonte@udc.es
[2] Department of Astronomy, University of Columbia, New York, NY, 10027, USA
[3] Center for Computational Astrophysics, Flatiron Insitute New York, NY, 10010, USA
melissa.ness@columbia.edu

Abstract

The current project aims to tackle with Gaia's BR/RP spectra distortion caused by interstellar dust, called reddening or extinction, which makes data not to be correctly classified. For such, it is proposed a machine learning algorithm that is able to learn how to correct such effect making use of denoising autoencoders. In addition, it was also developed a method to estimate the extinction degree, since for almost any spectra that is going to be corrected, such value was not computed. The previous tasks are going to be resourceful at our research group, since we take part at the Gaia project and deal with outlier spectra. In this way, we will be able to do a finer data-preprocessing prior to their classification.

Keywords: reddening - denoising autoencoders - BP/RP spectra - Gaia - machine learning

1 Introduction

The Gaia Space Mission is the cornerstone from the European Space Agency (ESA) and aims to build a 3D map of the Milky Way of about a billion of stellar and non-stellar objects. One of the Gaia working packages our research is involved in is the Outlier Analysis (OA) Working Package ([7]), where the main goal is to estimate the spectral type of those BP/RP spectra which were classified as spurious data by other Gaia working packages. When the 3^{rd} data release was published, known as Gaia DR3, there were published around 219 million BP/RP spectra, of which 56 million OA had to process since they were considered to be outliers ([2]).

*We wish to acknowledge the support received from the Centro de Investigación de Galicia "CITIC", funded by Xunta de Galicia and the European Union (European Regional Development Fund-Galicia 2014-2020 Program), by grant ED431G 2019/01.

We also acknowledge Xunta de Galicia for financing this research under the PhD grant program of the Xunta de Galicia (grant code ED481A 2021/296).

To infer the spectral type of such sources, it was already developed an unsupervised neural-network algorithm based on Self-Organizing Maps (SOM) ([4]). However, since these data is considered to be noisy, it is crucial to tune them at the preprocessing stage for refining the current classification published for the first time at DR3.

One of the well known problems that arise with BP/RP spectra is extinction or reddening, which shifts the waves so that it tends to look redder than it actually is. There are currently developed some extinction laws that, for a certain reddening degree, they state a way to remove the reddening effect for such waves. Nonetheless, of those 56 million elements OA process, the extinction value is only estimated for 4M sources, since, as previously stated, most of the spectra are noisy and it is not possible to accurately infer the reddening value following classic procedures. As a consequence, it is not easy for OA to get rid of such distortion and make a finer classification. That is why in this project, we tested several denoising autoencoder configurations that learn how reddening behaves on spectra to get rid of it.

2 Materials and Methods

So as to mitigate the extinction distortion, it was created a model using denoising autoencoders ([1]), which learn how noise behaves on data and removes it as accurately as possible. At a further stage, and following extinction laws, the extinction degree was also estimated as this value was published only for 4 million out of 56 million sources OA deals with.

In order to build the AI model, there were used 2,713,404 spectra with different reddening degrees. To build the dataset, spectra with a residual value of reddening of at most 0.003 were selected, and afterwards, they were manually redden following Fitzpatrick extinction law [6] (as shown at Figure 1), using spectra with a reddening degree between 0 and 2, with gaps of 0.1. In this way, the algorithm will have not only to learn a wide range of spectra but also, spectra without any degree of reddening. Before starting training the model, to better understand the behaviour of the dataset, it was made a comparison of the residual value between the original and the redden spectra to see which wavelengths were more affected by reddening and observing that the biggest differences appear between 400 and 500nm.

Finally, to improve the performance of the autoencoder, the input dataset was smoothed with a Gaussian filter of $sigma = 1$.

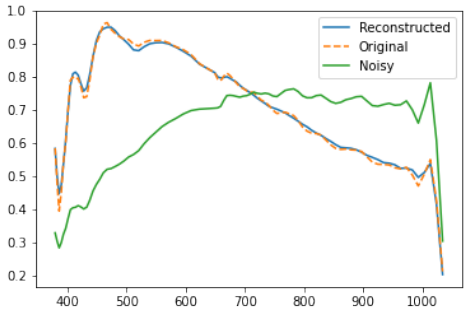

Figure 1: Comparison of the original spectrum against the redden and the reconstructed version

3 Results, conclusions and future work

For testing the performance of the denoising autoencoders, the parameters shown at Table 1 were calculated. Apart from the capability of the algorithm to reconstruct spectra, it is quite important that the training process is fast, since it will be necessary to run it at Apsis ([3]), the Gaia's data processing system. So, even though the shown configurations give similar results, the last one is able to learn the denoising process not only faster, but also with less errors. That is why, that is the one chosen to build the model.

Layers	Execution time (minutes)	Euclidean Distance	Mean Squared Error	Standard deviation
[87, 29, 87]	145	0.004	0.0008	0.157
[87, 43, 29, 43, 87]	100	0.002	0.0007	0.155
[87, 43, 29, 16, 29, 43, 87]	89	0.002	0.0006	0.152

Table 1: Performance results for different denoising autoencoders layouts

Moreover, as illustrated in Figure 1, the autoencoder can not only get rid of reddening, but also smooth spectra. This is beneficial for the training performed at OA, since having spectra with less features, will make such process to perform a faster and more accurate clustering with the SOM algorithm.

The reddening value was computed following Fitzpatrick's extinction law. Nonetheless, it was also tested a supervised data-driven method called The Cannon ([8]) which is used for inferring astrophysical parameters from spectra. Even though it worked pretty well with temperature and gravity, it was not able to infer reddening from our dataset, as show at Figure 2

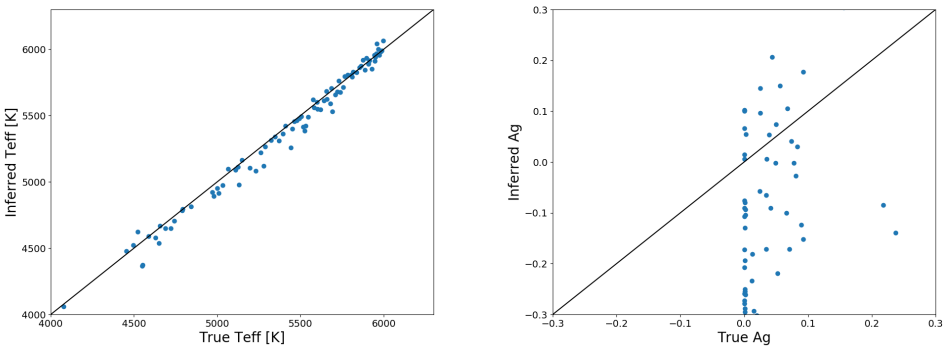

Figure 2: The Cannon results, estimated vs real temperature and extinction

In the future, the current work will also be tested with a disentangling procedure based on ([5]), which is a competitive AI algorithm trying to get rid the effect chemical and physical parameters influence the shape of spectra. It could not only learn to extract reddening, but also its value.

References

[1] Ebtesam Almazrouei, Gabriele Gianini, Corrado Mio, Nawaf Almoosa, and Ernesto Damiani. Using AutoEncoders for Radio Signal Denoising. In *Proceedings of the 15th ACM International Symposium on QoS and Security for Wireless and Mobile Networks*, Q2SWinet'19, pages 11–17, New York, NY, USA, November 2019. Association for Computing Machinery.

[2] F. De Angeli, M. Weiler, P. Montegriffo, D. W. Evans, M. Riello, R. Andrae, J. M. Carrasco, G. Busso, P. W. Burgess, C. Cacciari, M. Davidson, D. L. Harrison, S. T. Hodgkin, C. Jordi, P. J. Osborne, E. Pancino, G. Altavilla, and M. A. Barstow. Gaia data release 3: Processing and validation of BP/RP low-resolution spectral data. Publisher: EDP Sciences.

[3] C. a. L. Bailer-Jones, R. Andrae, B. Arcay, T. Astraatmadja, I. Bellas-Velidis, A. Berihuete, A. Bijaoui, C. Carrión, C. Dafonte, Y. Damerdji, A. Dapergolas, P. de Laverny, L. Delchambre, P. Drazinos, R. Drimmel, Y. Frémat, D. Fustes, M. García-Torres, C. Guédé, U. Heiter, A.-M. Janotto, A. Karampelas, D.-W. Kim, J. Knude, I. Kolka, E. Kontizas, M. Kontizas, A. J. Korn, A. C. Lanzafame, Y. Lebreton, H. Lindstrøm, C. Liu, E. Livanou, A. Lobel, M. Manteiga, C. Martayan, Ch Ordenovic, B. Pichon, A. Recio-Blanco, B. Rocca-Volmerange, L. M. Sarro, K. Smith, R. Sordo, C. Soubiran, J. Surdej, F. Thévenin, P. Tsalmantza, A. Vallenari, and J. Zorec. The Gaia astrophysical parameters inference system (Apsis) - Pre-launch description. *Astronomy & Astrophysics*, 559:A74, November 2013. Publisher: EDP Sciences.

[4] Carlos Dafonte, Daniel Garabato, Marco A. Álvarez, and Minia Manteiga. Distributed fast self-organized maps for massive spectrophotometric data analysis †. 18(5):1419. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

[5] Damien de Mijolla, Melissa Kay Ness, Serena Viti, and Adam Joseph Wheeler. Disentangled Representation Learning for Astronomical Chemical Tagging. *The Astrophysical Journal*, 913:12, May 2021. ADS Bibcode: 2021ApJ...913...12D.

[6] Edward L. Fitzpatrick. Correcting for the Effects of Interstellar Extinction. *Publications of the Astronomical Society of the Pacific*, 111(755):63–75, January 1999.

[7] Daniel Garabato, Minia Manteiga, Lara Pallas-Quintela, Marco A. Álvarez, and Carlos Dafonte. Outlier analysis online documentation.

[8] M. Ness, David W. Hogg, H. W. Rix, Anna. Y. Q. Ho, and G. Zasowski. The Cannon: A data-driven approach to Stellar Label Determination. *The Astrophysical Journal*, 808:16, July 2015. ADS Bibcode: 2015ApJ...808...16N.