Analysis and Knowledge Discovery by Means of Self-Organising Maps for Gaia Data Releases

M. A. Álvarez[1], C. Dafonte[1], D. Garabato[1], and M. Manteiga[2]

[1] Universidade da Coruña (UDC), Dept. de Tecnologías de la Información y las Comunicaciones, Elviña, 15071 A Coruña, España
{marco.antonio.agonzalez,dafonte,daniel.garabato}@udc.es,

[2] Universidade da Coruña (UDC), Dept. de Ciencias de la Navegación y de la Tierra, Paseo de Ronda 51, 15011 A Coruña, España
manteiga@udc.es

Abstract. A billion stars: this is the approximate amount of visible objects estimated to be observed by the Gaia satellite, representing roughly 1% of the objects in the Galaxy. It constitutes the biggest amount of data gathered to date: by the end of the mission, the data archive will exceed 1 Petabyte. Now, in order to process this data, the Gaia mission conceived the Data Processing and Analysis Consortium, which will apply data mining techniques such as Self-Organizing Maps. This paper shows an useful technique for source clustering, focusing on the development of an advanced visualization tool based on this technique.

Keywords: Gaia mission, European Space Agency, Data mining, Artificial Intelligence, Self-Organizing Maps visualizations

1 Introduction

Gaia is one of the key missions of the European Space Agency (ESA), which will conduct a census of the Milky Way with unprecedented accuracy. It was successfully launched on December 19th, 2013. During the scientific operation Gaia will observe all visible objects with magnitudes $6 < G < 20$, an estimated one billion objects, representing approximately 1% of the objects in the Galaxy (the largest amount achieved to date). Thus, it can develop a detailed 3D map that will allow us to answer several open questions about the composition, formation, and evolution of the Milky Way. Gaia will observe not only stars, but all types of astronomical objects with apparent brightness within the limits of the satellite, either galactic (asteroids, comets), or extragalactic (other galaxies, supernovas, quasars).

In order to analyze the data from the Gaia mission, the European Space Agency organized the Data Processing and Analysis Consortium (DPAC) which is composed of hundreds of scientists and engineers. DPAC is divided into nine

Coordination Units (CUs). The present work is dedicated to algorithm development in CU8, which is responsible for source classification and astrophysical parameters (AP) estimation, as well as in CU9, with data mining and visualization developments, both responsible for providing the platforms for computer processing and data visualization.

Several methods will be applied to the observed objects, in order to obtain a wide variety of relevant information for analysis and data classification. Since the Gaia mission is the first astronomical mission that will unbiasedly observe the entire sky down to magnitude 20, a large number of outliers are expected to be found. This paper explains how these types of objects are to be classified, by means of an unsupervised learning technique that is known as Self-Organizing Maps.

In order to provide a user-friendly environment for this technique that is readily available to the scientists community, we developed a useful interface for analysis and data visualization that is supposed to become operational in the summer of 2016, when the first Gaia Data Release will become public.

2 Self-Organizing Maps

Self-organizing maps (SOM) come from the branch of competitive neural networks. They were proposed by Kohonen in 1988 [8]. Since then, they remain the quintessential unsupervised ANN. In fact, to date there have been published over 5,000 articles related to SOMs.

These maps are obtained by projecting a multidimensional continuous input space into a discrete two-dimensional output space. That is, the input dataset is projected onto a set of neurons, which are arranged topologically in a lattice. Generally, a 2D lattice is used for simplicity of calculation and subsequent visualization, but a lattice in 3 or more dimensions can be specified.

Neurons are the main components, which are trained by means of competitive learning; in that sense, each time that an input feeds the network all neurons compete between them to represent this element, but there is only one winner, the best representative neuron, which is calculated through a distance measure. In our case, euclidean distance proved to be the best in result terms, also in computational time.

This technique has the advantage of projecting the dataset into a two (or three)-dimensional grid, where each neuron is related to its neighbourhood, so that the grid preserves the input topology. This topological preservation facilitates the data exploration and highlights properties of the data at hand [7]. Thus, a SOM calculates a reduced set of prototypes and a topological relationship between them. Additionally, it can be quickly trained and it has been demonstrated to be very robust in the presence of noise. SOMs have not yet been widely applied to the field of astronomy, but an introduction to their possible applications can be found in [6].

Due to the vast amount of data to be handled, the processing time is an important aspect to keep in mind. This is why we developed the algorithm using distributed computing through Hadoop and Spark frameworks. Moreover, our algorithm has already been integrated into the SAGA's pipeline[10] at CNES in France with the goal of being available for scientific work. In addition, we developed a version of the algorithm which includes symbolic data[3].

3 Classification Tool

As it was commented before, the amount of data to be processed is enormous, in fact most of the tasks related with Gaia will be processed on a cluster; as a result, we developed the algorithm of the SOM in such a way that it can be executed on a distributed system. Previous works in our laboratory show that we developed and successfully tested a Hadoop[11] cluster version of the algorithm [4][5][9] for CU8, but now in CU9 it has been decided to migrate to Spark[12]. We therefore developed a new version of the algorithm to be deployed on this latter architecture.

By means of SOMs we would be able to analyze and classify the outliers provided by the satellite, but the main objective of our work is to provide a useful tool, which allows the scientific community to perform data analysis and classifications. This tool must allow users to request for complex analysis, in terms of computational time and the amount of data involved, in an easy way and through a friendly interface. Taking this into account we need at least two different modules, the user interaction module and the computation module, in a Client-Server architecture.

The users interact with the first module, the client, which is responsible for data visualizations and for getting information through communication messages with the second module, the server, that implements all the logic for data processing.

The server module has to be able to communicate with the cluster in order to request for the trainings and to retrieve the related files. We decided to implement a Rest (Representational State Transfer) service using Spring; for communications between server and clients we decide to use JSON (JavaScript Object Notation).

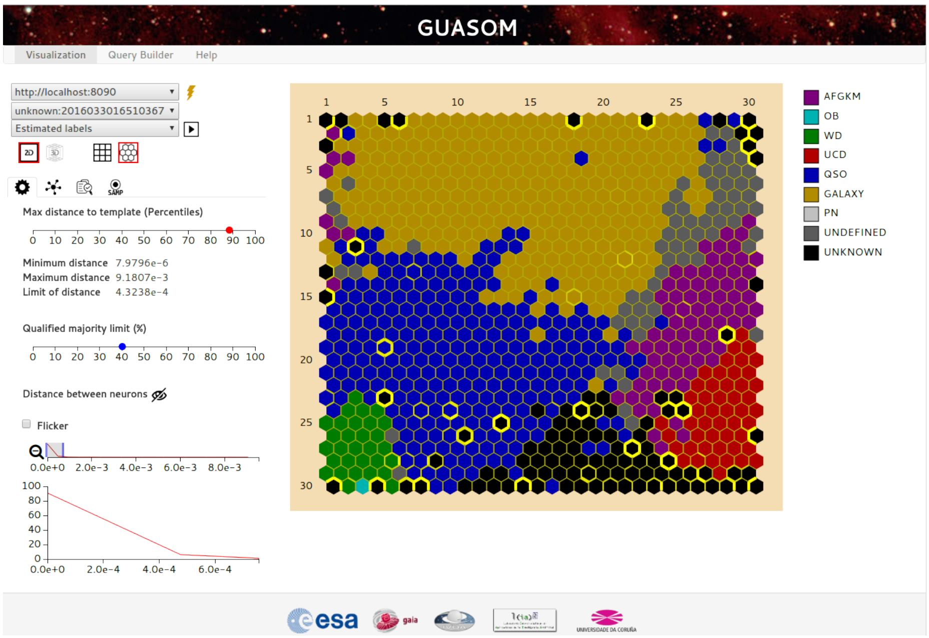
The client application has to be powerful, useful, complete, compatible with other Gaia tools, and easy to access and use. We decided to develop and implement a Web application, working with Apache Tapestry because of its potential and, to provide useful features for plots and statistics, JavaScript, 3js and SVG libraries were really helpful. We also integrated SAMP[2] (Simple Application Messaging Protocol) to allow data exchange with other Gaia tools, as described below.

4 Features

Once the application is finished, several visualization tools are available to unveil the data's physical nature and distribution. Users can visualize different representations of the entire SOM as well as some graphics for each neuron individually, studying as such the data from different levels.

As we can see in Fig. 1, the interface has two defined areas: the left panel, where some controls and options are available, and the rest of the space, which is used for the SOM and neuron visualizations.

There is a useful option for representing the distance of each neuron to its neighbours. This option draws the lines between neurons with different width accordingly to the distance between them. More distance means more width because these neurons are less similar, giving the impression of a wall between them (1).



http://localhost:8090

unknown:2016033016510367

Estimated labels

Max distance to template (Percentiles)

Minimum distance 7.9796e−6

Maximum distance 9.1807e−3

Limit of distance 4.3238e−4

Qualified majority limit (%)

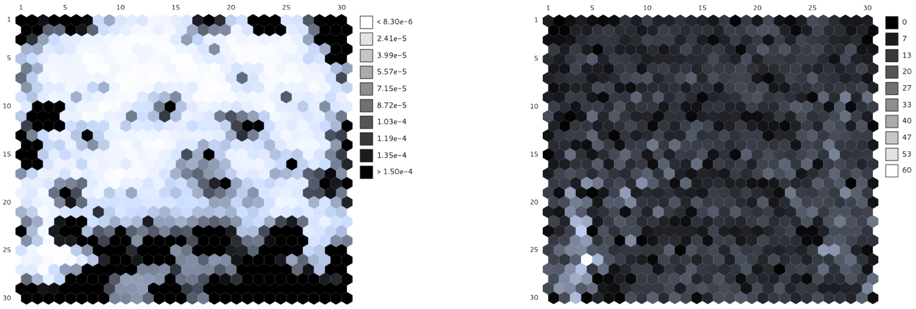
Distance between neurons

□ Flicker

Fig. 1: Capture of the application with Gaia Labels visualization and the feature of distance between neurons activated

Classical SOM visualizations

- *U-Matrix* (2a), which displays the distances among clusters. A dark color corresponds to a large distance and thus a gap between the clusters in the input space, whereas a light color between the clusters means that they are close to each other in the input space. Light areas can be thought of as dense regions in the input space, while dark areas correspond to more sparse ones.
- *Hits* (2b), which shows the number of observations falling in each neuron, represented by a color in a gray scale, so that a dark color indicates that the neuron contains few observations, while a light color indicates a high number of observations. This display is helpful to visualize the data density in each region of the SOM network and, hence, of the input space.



(a) U-Matrix

(b) Hits

Fig. 2: Classic Views

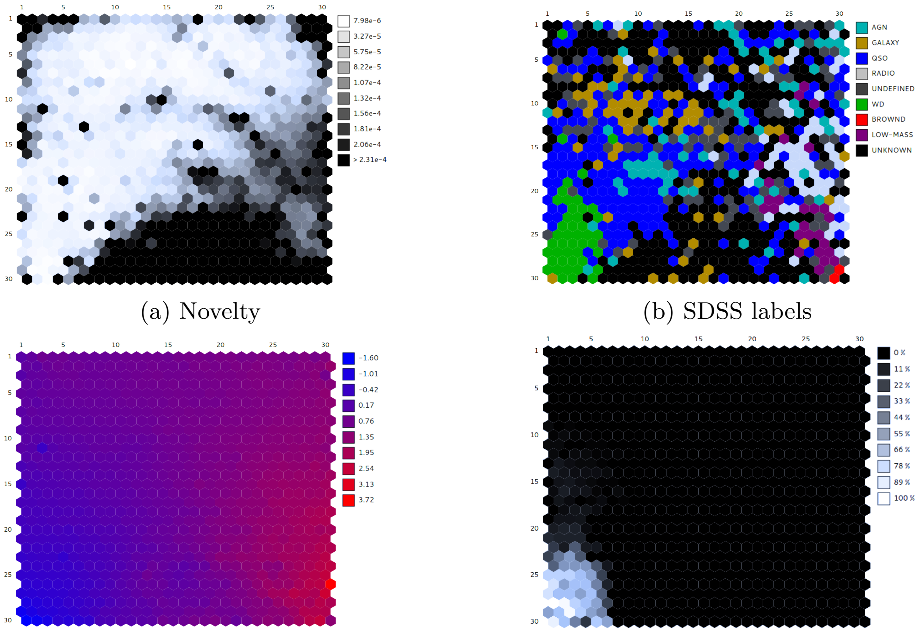
Specialized SOM visualizations

These are more specialized visualizations, oriented towards exploring different features related with Astrophysics and its particular parameters. At this point it is convenient to mention that each neuron or cluster has a representative, called prototype, which is a virtual pattern that better represents or resembles the set of input patterns belonging to such a cluster. The neurons are labelled by comparing their prototype to different templates, which were obtained from different libraries such as SDSS Outliers Library spectra.

- *Novelty* (3a). This SOM plot shows the novelty of the neurons. A dark color corresponds to the more novel neurons, calculated by means of the distance between template and prototype. Less distance means less novelty because the observations of this neuron are quite similar to the prototype, and the prototype is well known.
  item *Simbad labels* (3b). This visualization shows the representative class for each neuron. In this case, SOM clusters receive a color in function of the most frequent SIMBAD [1] identification. In the figure, black clusters do not have objects identified in SIMBAD, a grey color is assigned to clusters with a similar frequency among two or more classes, or with no classes greater

or equals to the qualified majority limit. This value can be specified by the user.

- *Color distribution* (3c), which allows to evaluate the network organization according to the stars' colors, which are directly linked to the temperatures. Color distribution $G_{BP} - G_{RP}$ is obtained by subtracting the magnitudes corresponding to the integrated flux respectively of RP and BP spectra.
- *Gaia labels* (1), which represents the distribution of astronomical object classes obtained with Gaia photometric simulations. The color assigned to each cluster was set in function of the predominant class of the objects belonging to it.
- *Category distribution* (3d). This visualization shows the distribution of a unique category, selected by the user, on the map, allowing to determine which neurons have objects of this category.



(a) Novelty

(b) SDSS labels

(c) Color distribution

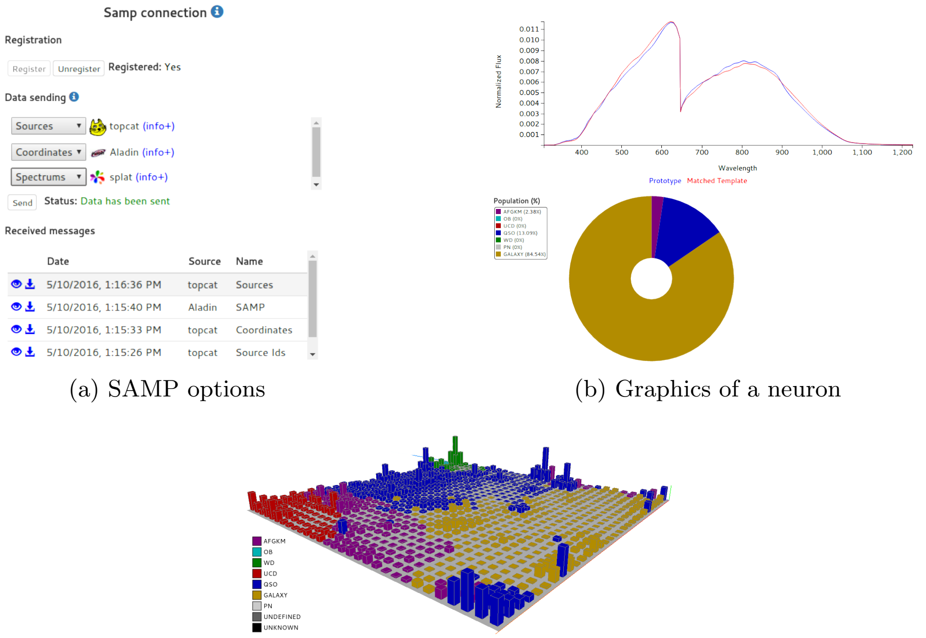
(d) White Dwarf distribution

Fig. 3: Specialized views

These visualizations provide a remarkable improvement in detection of isolated areas, simplifying the identification of weird objects and similar ones.

An extra option that we have developed is the possibility to represent some of these graphs in 3D, allowing the user to visualize some combinations of the previous visualizations. For instance, a user could visualize a combination of the Umatrix and the Gaia Labels representations (4c).

For each neuron, a user could visualize the graphic of the prototype with the matched template (the template that is more similar to the prototype), in order to analyse their similarity, the distribution of the different type of objects which belongs to this neuron (4b) and, in addition, some extra information which is shown in the interface.



Samp connection

Registration

Register | Unregister | Registered: Yes

Data sending

Sources ▾ topcat (info+)
Coordinates ▾ Aladin (info+)
Spectrums ▾ splat (info+)

Send | Status: Data has been sent

Received messages

	Date	Source	Name
👁 ⬇	5/10/2016, 1:16:36 PM	topcat	Sources
👁 ⬇	5/10/2016, 1:15:40 PM	Aladin	SAMP
👁 ⬇	5/10/2016, 1:15:33 PM	topcat	Coordinates
👁 ⬇	5/10/2016, 1:15:26 PM	topcat	Source Ids

(a) SAMP options

(b) Graphics of a neuron

(c) Combination of Umatrix and Gaia Labels

Fig. 4: Special actions

Another useful feature is the cross match section, where a user could select any of the observations belonging to a selected neuron and search for this observation on different external data bases, such as Simbad or SkyServer, based on its astronomical coordinates.

Finally, in order to make our application compatible with other tools such as Aladin, Topcat, Vaex, etc., we integrate the SAMP protocol, allowing users to select a neuron or a set of neurons and send their information to others. At the same time, our application could receive messages coming from other tools, giving the user the possibility to download or highlight this information on a specific SOM visualization (4a).

5 Conclusion

This application was developed for the future data releases of the Gaia mission. We have already presented it in several meetings of Gaia, causing great expectancy, and we believe that the features of this application are really helpful for the desired purpose.

Due to the amount of data that Gaia will provide, data mining techniques are widely important for the treatment of these data. We are aware of the processing time required for the analysis, and have been optimizing our algorithm and our application so as to minimize the processing and the response time. In future works, we plan to include GPU processing in order to reduce the times even more.

We developed this tool in order to make this Artificial Intelligence technique, Self-Organizing Maps, more accessible to the astrophysical and astronomical communities, providing a useful interface for the analysis of Gaia data.

References

1. SIMBAD Astronomical Database, http://simbad.u-strasbg.fr/simbad/
2. Simple Application Messaging Protocol, http://www.ivoa.net/documents/SAMP/
3. del Coso, C., Fustes, D., Dafonte, C., Nóvoa, F.J., Rodríguez-Pedreira, J.M., Arcay, B.: Mixing numerical and categorical data in a self-organizing map by means of frequency neurons. Applied Soft Computing 36, 246 – 254 (2015), http://www.sciencedirect.com/science/article/pii/S1568494615004512
4. Fustes, D., Dafonte, C., Arcay, B., Manteiga, M., Smith, K., Vallenari, A., Luri, X.: SOM ensemble for unsupervised outlier analysis. Application to outlier identification in the Gaia astronomical survey. Expert Syst. Appl. 40(5), 1530–1541 (Apr 2013), http://dx.doi.org/10.1016/j.eswa.2012.08.069
5. Fustes, D., Manteiga, M., Dafonte, C., Arcay, B., Ulla, A., Smith, K., Borrachero, R., Sordo, R.: An approach to the analysis of SDSS spectroscopic outliers based on self-organizing maps. Astronomy & Astrophysics 559, A7 (2013), http://dx.doi.org/10.1051/0004-6361/201321445
6. Geach, J.E.: Unsupervised self-organized mapping: a versatile empirical tool for object selection, classification and redshift estimation in large surveys. MNRAS 419, 2633–2645 (Jan 2012)
7. Kaski, S.: Data Exploration Using Self-Organizing Maps. Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series (82) (Mar 1997)
8. Kohonen, T.: Neurocomputing: foundations of research. chap. Self-organized formation of topologically correct feature maps, pp. 509–521. MIT Press, Cambridge, MA, USA (1988), http://dl.acm.org/citation.cfm?id=65669.104428
9. Ordóñez, D., Dafonte, C., Varela, B.A., Manteiga, M.: HSC: A multi-resolution clustering strategy in Self-Organizing Maps applied to astronomical observations. Appl. Soft Comput. 12(1), 204–215 (2012), http://dx.doi.org/10.1016/j.asoc.2011.08.052
10. Veronique Valette, Kader Amsif: CNES Gaia Data Processing Centre, a complex operation plan. The 12th International Conference on Space Operations (June 2012), http://www.spaceops2012.org/proceedings/documents/id1291264-Paper-001.pdf
11. White, T.: Hadoop: The Definitive Guide. O'Reilly Media, Inc., 1st edn. (2009)
12. Wills, J., Owen, S., Laserson, U., Ryza, S.: Advanced Analytics with Spark: Patterns for Learning from Data at Scale. O'Reilly Media, Inc., 1st edn. (2015)