Neural Computing and Applications (2022) 34:1993–2006
https://doi.org/10.1007/s00521-021-06510-9

ORIGINAL ARTICLE

GUASOM: an adaptive visualization tool for unsupervised clustering in spectrophotometric astronomical surveys

M. A. Álvarez[1] · C. Dafonte[1] · M. Manteiga[2] · D. Garabato[1] · R. Santoveña[1]

Received: 21 January 2021 / Accepted: 27 July 2021 / Published online: 19 September 2021

Abstract

We present an adaptive visualization tool for unsupervised classification of astronomical objects in a Big Data context such as the one found in the increasingly popular large spectrophotometric sky surveys. This tool is based on an artificial intelligence technique, Kohonen's self-organizing maps, and our goal is to facilitate the analysis work of the experts by means of oriented domain visualizations, which is impossible to achieve by using a generic tool. We designed a client-server that handles the data treatment and computational tasks to give responses as quickly as possible, and we used JavaScript Object Notation to pack the data between server and client. We optimized, parallelized, and evenly distributed the necessary calculations in a cluster of machines. By applying our clustering tool to several databases, we demonstrated the main advantages of an unsupervised approach: the classification is not based on pre-established models, thus allowing the "natural classes" present in the sample to be discovered, and it is suited to isolate atypical cases, with the important potential for discovery that this entails. Gaia Utility for the Analysis of self-organizing maps is an analysis tool that has been developed in the context of the Data Processing and Analysis Consortium, which processes and analyzes the observations made by ESA's Gaia satellite (European Space Agency) and prepares the mission archive that is presented to the international community in sequential periodic publications. Our tool is useful not only in the context of the Gaia mission, but also allows segmenting the information present in any other massive spectroscopic or spectrophotometric database.

Keywords Big data astronomy · Self-organizing maps visualization · Computational astrophysics · Artificial neural networks · Astronomical surveys

1 Introduction

It is becoming more and more common for astrophysicists to search for answers regarding the structure and evolution of the Universe by analyzing vast portions of the sky, which contain huge amounts of a variety of astronomical objects. Typically, they want to determine their physical, galactic, or extragalactic nature and extract information about their most relevant properties.

To analyze such data, which are of a complex nature and appear in sets of the order of hundreds of thousands or millions of objects, the use of advanced computing techniques is inevitable. Historically, Astronomy has been at the forefront of the development and implementation of information management services through the Internet, and for more than twenty years, both space missions and most of the large terrestrial observatories have been developing efficient systems of archiving and accessing scientific data. Most of these files are open-access, but their analysis is complex and requires the development of tools based on Statistics and the use of methodologies such as those derived from artificial intelligence (AI), in what has come to be called "Data Mining in Astronomy."

✉ M. A. Álvarez
marco.antonio.agonzalez@udc.es

[1] CIGUS CITIC - Department of Computer Science and Information Technologies, University of A Coruña, Elviña, A Coruña 15071, Spain

[2] CIGUS CITIC - Department of Nautical Sciences and Marine Engineering, University of A Coruña, Paseo de Ronda 51, A Coruña 15011, Spain

Data mining deals with the processes necessary to obtain knowledge in large databases. It is also called KDD ("Knowledge Data Discovery"). The term was coined in the 1990s, when large corporations began to store their data in digital form. Data mining includes methods from disciplines such as Statistics, AI, Databases, and the theory of computational complexity. It focuses mainly on Descriptive Statistics and unsupervised learning, although it also usually includes predictive models obtained through supervised learning. It is evident from their definitions that these fields overlap widely. For example, clustering techniques such as K-means are at the same time Descriptive Statistics, Data Mining, and unsupervised learning techniques, depending on their use. The important issue when solving a problem that requires intelligent automatic processing is to properly choose which techniques to use and how to apply them to the domain.

Unsupervised learning refers to the generation of a model that explains a set of observations without any prior knowledge. The two main applications of this type of learning are unsupervised classification and dimensionality reduction. The vast majority of algorithms that have been developed to solve this problem coincide in the exploitation of the concept of similarity between objects. In this way, the algorithms seek to identify object associations that form partitions of the input dataset. These groups are generated with the aim of maximizing the similarity between the objects belonging to the same group and, at the same time, minimizing the similarity between objects belonging to different groups.

Dimensionality reduction is defined as the process of decreasing the number of variables involved in solving a specific problem, so that it can be addressed more easily, and it can be performed by selecting or extracting features. Usually, characteristics selection is based on supervision, while feature extraction techniques are applied in the field of unsupervised learning. A classic feature extraction technique is principal component analysis, PCA [1, 2], which has been widely used in various fields, including Astrophysics. However, PCA lacks the ability to capture nonlinear relationships between variables in the input space. In recent years, various techniques are commonly used to try to solve this problem, such as PCAs with Kernel, the so-called local linear embedding (LLE), multidimensional scaling (MDS) techniques [3], or t-distributed stochastic neighbor embedding (t-SNE) [4].

Neural networks also contemplate unsupervised learning. The main unsupervised neural networks are ART networks [5], neural gas networks [6] and self-organizing maps or SOMs [7], and other later alternatives [8]. Such networks follow a learning process based on model fitting according to some measurement of similarity. Among all clustering techniques, SOMs provide an additional feature: the nonlinear projection of the set of observations in a space of reduced dimensionality, similar to the LLE, MDS, and t-SNE algorithms. Therefore, SOMs unite the two main branches of unsupervised learning: clustering and dimensionality reduction. The probabilistic counterpart of SOM is the generative topographic maps (GTM) [9], that use a expectation–maximization algorithm (EM). The comparison between both methods is explained in Sect. 9, showing that, in the context of the present study, SOM performs better than GTM.

This work is structured as follows. In Sect. 2, different works related to data mining, unsupervised learning, and SOMs are mentioned. Section 3 explains the SOM output analysis difficulties in Big Data environments and why the existing tools are not suitable in this context. Sections 4 and 5 describe the fundamentals of SOMs and discuss the need to develop a specific interface for visualization and analysis, oriented to astronomical spectrophotometry. Section 6 describes the operations and data preprocessing that have been implemented in our tool, named GUASOM, that are aimed at providing different useful visualizations of the datasets and their classifications using topologically ordered neurons. The different visualization environments will be described using as an example the results obtained on a set of approximately 82000 spectrophotometric observations for a wide variety of astronomical objects from the Gaia mission [10]. Due to the fact that Gaia spectrophotometric observations are under embargo until GDR3 will be released in the second half of 2022, we will show only collective results obtained for the mentioned sample that we have been analyzing for validation purposes. We will use different sets of data, in particular the clustering obtained on the SDSS Legacy survey sample of spectra, to illustrate some of the more specific utilities. In Sect. 7, we will show how the achieved clustering quality can be assessed either using clustering quality indices, or by external validation in the case that well-known representative samples of the data were available. Section 8 briefly discusses GUASOM's performance on a fully photometric survey, the ALHAMBRA survey [11], whereas in Sect. 9, the results obtained from a spectroscopic survey are presented, the SDSS Legacy survey [12, 13]. Finally, Sect. 10 summarizes the characteristics of GUASOM and its implementation as a Web tool and reviews the results obtained in the tested astronomical archives.

2 Related work

In the field of Astronomy and Astrophysics, the application of grouping algorithms is a relatively new concept, despite being a field in which statistics have a large number of applications. The pioneering works in this regard were those presented by [14] and [15], in which unsupervised classification of galaxy and star spectra from the SDSS catalog was carried out. In these works, the authors use the K-means algorithm [16] to classify hundreds of thousands of medium resolution spectra.

LLE has been applied by [17] to the classification of Kepler light curves into morphological types. In the field of spectral classification, LLE has been applied by [18] to the classification of massive protostellar spectra and by [19] to represent SDSS spectra in a three-dimensional space, which can be easily visualized by experts in the domain.

Although SOM networks are common in different fields, they have been of little use in Astronomy [20–23]. Some of our previous works devoted to the analysis of atypical objects [24–26] demonstrated SOMs ability to reveal the properties of a set of spectra when there is little or no prior knowledge.

3 Problem statement

The aim of this work arose from the need to process millions of objects with a clustering technique explained in Sect. 4, in a Big Data scenario in the context of the Gaia mission. In such scenario, we must process all the information in a reasonable time and provide to the community a tool oriented to Astrophysics in order to analyze the information of SOMs related to the physical and statistical properties of the clustered objects.

Different applications were developed for the visualization of SOMs such as Viscovery SOMine GmbH [27], SOMPY [28], SOMToolbox [29], SpiceSOM [30] or another Python implementation of SOM [31]. The comparison presented in Table 1 highlights the fact that most of them just analyze SOMs in a generic way, or they are developed with important restrictions that hinder the use of data from different domains. Furthermore, we are interested in including a series of features that allow an in-depth study of the content of the maps and their neurons (statistical analysis, distance computation, data representation, etc.).

Note that we need to handle all this information in a Big Data scenario, which means that the training process will take several hours or even days, and for that reason, it is important to separate the visualization of the SOMs from the training process, in order to train them in a distributed framework.

We are also interested in developing a versatile tool that adapts to different representations of the data and allows us to visualize the usual parameters in an astronomical survey.

The problem with this information is that at present, there is no visualization tool able to work in a real Big Data scenario or provide specific domain visualizations.

4 Self-organizing maps (SOM)

The main advantage of SOM networks, as mentioned before, is to provide a good quality grouping as well as a non-linear reduction of dimensionality, by projecting the data in a fixed number of clusters (called neurons or units), arranged in a 2D structure, generally a grid with N rows by M columns (Fig. 1). Each neuron has a representative, called a prototype, which is a virtual pattern that best

Table 1 Comparison of relevant features, for the purpose of this work, between existing tools for visualizing SOMs. The bibliographic references are included in brackets

	Tools				
Features	Viscovery [27]	SOMpy [28]	SOMToolbox [29]	SpiceSOM [30]	SOM [31]
Domain oriented	Generic approach	Generic approach	Generic approach	Teaching	Biology
License type	Commercial	Freeware	Commercial	Freeware	Freeware
Big Data Acceleration	No	Python parallelism	No	No	Python parallelism
3D visualizations	No	No	No	Input data	No
Deep cluster analysis	Advanced	Basic	Basic	Basic	Cluster labeling flooding
VO compatibility	No	No	No	No	No
Crossmatch[1]	No	No	No	No	No
Architecture	Desktop	Interactive notebook	Desktop	Desktop	Interactive notebook
Import SOM[2]	False	Python	Matlab	False	Python

[1] The tool allows to perform a crossmatch with external databases.

[2] The tool allows to import a trained SOM for visualization.

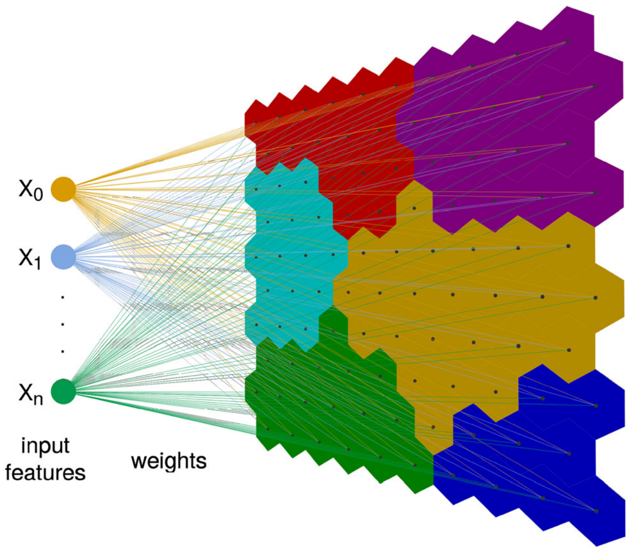


Fig. 1 Self-organizing map architecture

represents the set of input patterns that belong to that group.

The problem to be optimized is finding the best prototypes for all the neurons in the SOM, for which an iterative optimization procedure is followed from a random initialization of the neuron weights. First, for each input observation, the neuron that most closely matches the pattern is activated. This is calculated using the squared Euclidean distance between the pattern and the neuron prototype. Then, the activated neuron and their neighbors are updated according to the activation patterns. The number of neurons in the vicinity of the activated neuron is large in the first few iterations, but it decreases as the iterations continue. In this way, the algorithm begins by ordering the neurons and then goes on to focus on the grouping process, minimizing the residuals (also called quantification errors) between the prototypes and their activation patterns.

An important point in the implementation of SOMs is to define the learning parameters. The maximum number of learning iterations and the neighborhood function can be empirically determined, but the most important parameter to set, the size of the map (the number of neurons), is more difficult to estimate, since it depends entirely on the variance of the data and the size of the dataset. We opted to also determine it experimentally, using the measurement of error in the grouping obtained by the mean quantification error, MQE [32]. MQE measures the average distance between a cluster prototype and the objects that populate it.

There are two main algorithms for the training of SOMs: the online (or sequential) and the batch algorithm. In the online algorithm, the weights are updated with:

$$w_i(t+1) = w_i(t) + \alpha(t)h(t)(x(t) - w_i(t)) \tag{1}$$

where α is the learning rate, h is the neighborhood distance weight, and x is the datapoint at the current iteration t. In the online learning, the winner for an input is found, and the weight vectors are updated immediately. In the batch learning, the updates are deferred to the presentation of the whole dataset. We decided to use the batch learning mode instead of the online mode, taking advance of its greater speed, because it can be parallelized and does not depend on the order in which the patterns are presented to the SOM [33].

Although the MQE index measures the quality of the grouping, we still do not know if the map is correctly ordered according to the input data sets provided. One way to evaluate the order is to visualize the topographic distribution of the data and to compare intercluster and intracluster distances, which leads us to the need to develop a visualization tool customized to display and analyze the information in neurons.

5 SOM Visualization

SOM networks retain information on the distribution of data in their topology, and this fact can help in the analysis stage, by displaying significant information on the map grid. A typical visualization tool for SOMs is the U-Matrix, which shows the distance between prototypes and can serve as a map for data exploration [34]. Large distances between prototypes can be interpreted as gaps between data, and they can help to isolate outlying clusters, while short distances allow to select several neurons for a joined identification of the objects that populate them.

This work therefore presents a tool that is specifically designed for the analysis of massive spectrophotometric surveys and has its origin in Gaia DPAC consortium developments. Gaia datasets are of the order of 10^9 objects, which translate to approximately 1000 Gigabytes, and lead us to develop a tool powerful enough to handle enormous amounts of complex data. Since carrying out calculations in real time is barely possible, we devoted significant efforts to calculate in advance the data oriented to the visualizations. We named this visualization tool “Gaia Utility for the Analysis of Self-Organizing Maps” (GUASOM), despite the fact that we have adapted it to be useful in the analysis of spectrophotometric data in general.

Before going into detail about the system design, we shall introduce some concepts that will be used in what follows:

– Prototype: Representative spectrum or spectral energy distribution of a neuron according to the objects populating it.

- Templates: Models of spectra that represent types of astronomical objects. These models can be obtained by means of real observations or can be provided by synthetic libraries.
- Object-centroid: Observation or object in a neuron that is the most similar one to the neuron prototype.
- Outliers: When the prototype does not represent with sufficient confidence the objects in the neuron because they are very heterogeneous, those objects are named outliers.
- Catalog: Consolidated database of astronomical data that we can use to perform an astrometric crossmatch to retrieve information about the physical nature of as many objects as possible in a neuron. This can allow us to assign a particular label to such a neuron.
- Crossmatch: Search procedure that uses astronomical coordinates and a search radius to retrieve the astronomical information from a catalog for a particular set of objects.
- Template matching: Procedure that compares a set of templates with a given prototype in order to select the most similar one. We defined the Euclidean distance as the similarity metric used to make such a comparison.

5.1 System design

Developing a visualization tool for advanced analysis techniques in a Big Data context is not a simple task. The response time is crucial for this kind of tool, and the resources are limited. Taking this into account, we have to process millions of objects in an acceptable time to the user, and we have to develop an application that is as user-friendly as possible, allowing users to appropriately interpret the results.

In order to address the first issue, we developed a software that processes in advance all the data that will be visualized in the application. This is the most important feature of the visualization tool, and it is explained in Sect. 6. Regarding the second problem, we decided to develop the tool with Web technologies, due to their flexibility and easy access to the information. Such a development is explained in Sect. 6.1.

The factors that we take into account when we design this software are the following:

- We defined a set of data models for the neurons as well as for the map and for some specialized visualizations. These models contain the relevant information for the different graphics. Each map will have some properties, a set of $N \times M$ neurons, as well as some precomputed information for each visualization.
- Processing all the required calculations will involve data from millions of objects. For that reason, we decided to optimize, parallelize, and even distribute these calculations in a cluster of machines.
- The application must load all the information as quickly as possible. Hence, we store all the models in binary files, which can be loaded faster. The map information is stored separately from the information of some specialized visualizations, because in that way, we have the data in different structures, and it is easy to generate and retrieve them.
- The design of the tool is based on a client-server system [35] with a Web client and a REST server [36]. In that way, the user can access the application using a Web browser, and the server can be located in a machine or in a cluster with enough capabilities to handle all the data with a short response time.
- We decided that JSON (JavaScript Object Notation) is a good option to pack the data between server and client, because it is an easy and light format that is appropriate to use for the types that we usually manage: numerical and categorical values.

6 Data preprocessing

Any user of a tool aimed to analyze millions of objects will not expect to obtain the raw data directly, but some preprocessed information or even statistical data, and some graphics that help to perform such an analysis. In a Big Data environment, those calculations are demanding in computational power and in processing time. For that reason, we decided to implement a preprocessing stage and to store the most relevant results. This considerably reduces the time required to access such an information, which is especially important if the application is online and free to access.

Concurrent access is expected, requesting for different visualizations involving a wide variety of data, and the potential users of an interactive tool will not use the application if they must wait for a long time to obtain the results; that is why it must run fluid and fast. This way, the main objective is to prepare the data in a simple, sorted, and efficient data structures that contain only the information that the users request from the application at any time, allowing a faster processing in the Web browser.

To do so, we analyze all different visualizations to determine the operations and data required to represent them. Then, we calculate such operations in advance and organize data directly in array structures that the visualization tool needs to render. Once the data are prepared, we store them in binary files in order to load them faster. See Fig. 2.

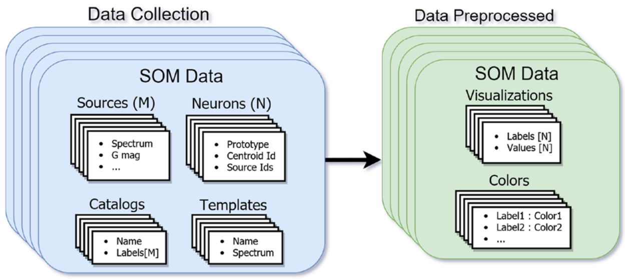


Fig. 2 Data preprocessing scheme representing the transformation of the data performed in the preprocessing stage

Some of such preprocessing tasks that have to be performed are the following:

- Intercluster distances: Several map visualizations rely on the concept of distance between neurons. It is important to pre-calculate the average distance between a neuron and its neighbors.
- Statistical information about the objects populating a neuron: In the case of Gaia spectrophotometric data, we derive basic statistics (mean, standard deviation, bias) of both astrometric measurements (parallaxes, proper motions, galactic coordinates, etc.) as well as photometric ones (magnitudes in the different bands). For other surveys, a different set of observables can be considered.
- Neuron representatives and some illustrative examples: Every neuron prototype and centroid need to be computed from sets containing even millions of objects. It is not operative to display the full content of a neuron, we prefer to pre-select a small number of objects to be displayed, including those that best and worst fit the prototype.
- Templates: The application manages different sets of templates whose suitability to fit the prototype is calculated for each of the neurons. For each set of templates we need to calculate the distance between the prototype of a neuron to all the available templates that are preprocessed separately and stored in independent files, in order to identify which is the most similar one that will be represented in the visualization.
- Crossmatch results: The application must store the information about the physical nature of the objects that was retrieved after performing a crossmatch with external catalogs. Such an information is stored in one file per catalog, and to be able to integrate it with the visualization tool we have to preprocess it.

All the preprocessing tasks result in several files for each one of the maps; these files are stored in a hierarchy of directories on the server side that can be accessed locally, improving the performance of the system.

6.1 Visualization tool

We decided to develop a visualization tool using Web technologies because they make things easier for users. They do not need to care about the platform of the application, the format of the data, or how to download and deploy both the application and the data for the analysis. All the information will be available through the Internet, and the server is responsible for all the computational tasks involved. Any user can connect to the visualization tool using a Web browser.

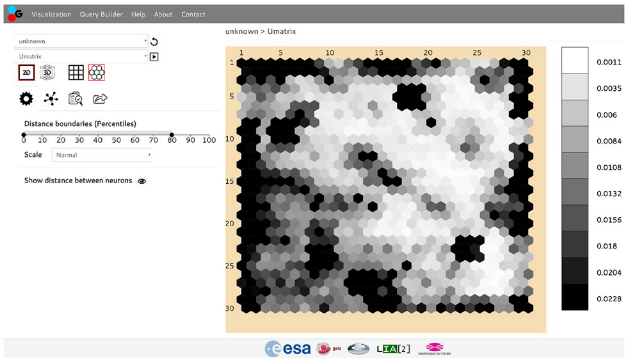
Figure 3 shows the interface of the application, where the left part is devoted to controlling the different features of the tool and the rest of the area is used to represent the map of neurons of a SOM.

Several map visualizations are available for the user for a smooth analysis of the data. Our application provides the classical representations for SOMs:

- Umatrix: As mentioned before, this representation shows the distance between the neuron prototypes, where less distance means more similarity. This is useful to identify groups of neurons populated by objects with similar SEDs. In our application, the user can control the boundaries of the distance between neurons through a slider, with the objective of exploring the inner structure of the map. Figure 3 shows this representation, and the slider in the control section.
- Hits: It shows the number of objects for each neuron, allowing to identify dense regions in the map. An example of this visualization is shown in Fig. 15.

The specific domain visualizations that we have implemented are:

- Parameter distribution: This visualization shows the distribution of a particular parameter of the domain in the map, displaying the average values calculated in each neuron.



unknown

Umatrix

unknown > Umatrix

Distance boundaries (Percentiles)
0 10 20 30 40 50 60 70 80 90 100

Scale Normal

Show distance between neurons

Fig. 3 User interface of GUASOM displaying the Umatrix representation of a validation spectrophotometric sample of Gaia data

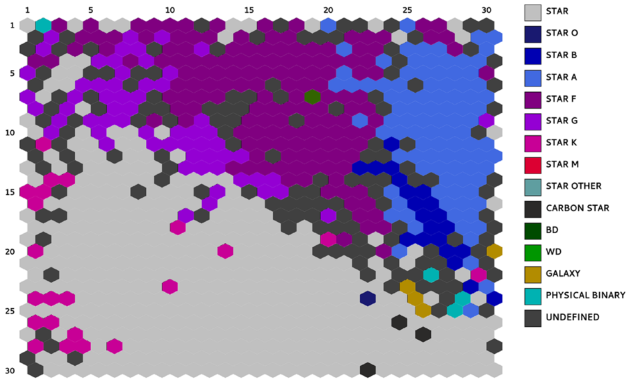


Fig. 4 Map with neurons displaying labels obtained by crossmatch of a set of Gaia spectrophotometric observations with Simbad database. The same dataset is used in Figs. 5, 6, and 7

- Catalog labels (Fig. 4): This graphic shows the representative label of each neuron according to a specific catalog chosen by the user. The labels of the objects were obtained through the crossmatch procedure mentioned before. The user can control the qualified majority limit that the label has to reach to be representative through a slider.
- Template labels (Fig. 5): It is similar to the catalog labels visualization, but in this case, we use the representative label of each cluster based on a template. We select the template that best fits with the prototype using the Euclidean distance. One slider allows the user to control the distance between the prototype and the template in order to decide the adjustment threshold between them that allows assigning the label with sufficient confidence.
- Category distribution (Fig. 6): In this representation, the distribution of a unique type of object is shown. The user can select the category to be displayed between a set of labels, according to the templates and catalogs available for the map. With this graphic, the user can easily observe the regions of the map containing objects of the chosen type.
- Color distribution (Fig. 7): It shows the distribution of the color of the objects in the map, derived as the difference in magnitudes between two photometric bands. For Gaia, the bands are those corresponding to the two photometers, BP and RP, and the color is calculated as $G_{BP} - G_{RP}$, but for other astronomical archives other bands could be visualized accordingly.
- Novelty: This visualization shows the distance between a selected template and the prototype. Less distance means less novelty because the template associated with this neuron is quite similar to the prototype, so it refers to a well-known object type. The user can select the set of templates to render.

As illustrated in the examples of Gaia spectrophotometric test data shown in Figs. 4, 6, 7, and especially 5, these visualizations provide a remarkable improvement in

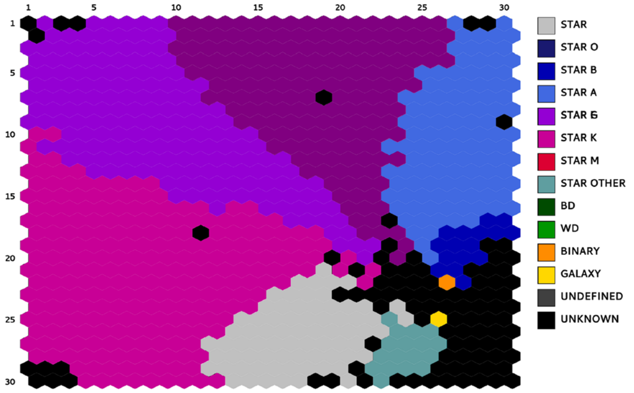


Fig. 5 Map displaying neurons colored according to the template labels that best fit every neuron for the same test sample of spectrophotometric Gaia data as in the previous figure

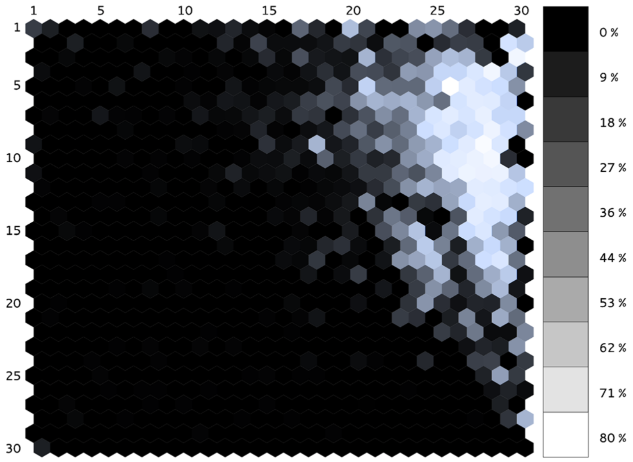


Fig. 6 Example of category “Star A” (objects with stellar spectral type A) distribution

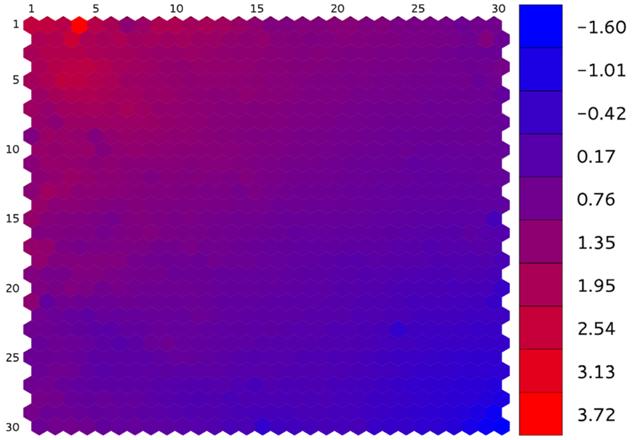


Fig. 7 Example of $G_{BP} - G_{RP}$ color distribution in a test sample of Gaia spectrophotometric data

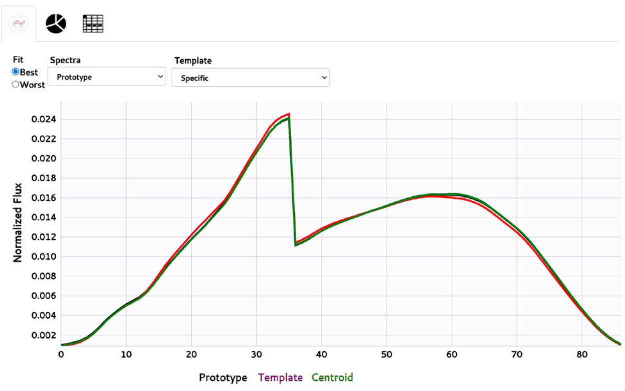
detection of groups of neurons with similar properties as well as those areas with atypical objects or artifacts. The strength of the tool lies in its ability to explore the neurons and the objects assigned to them by means of some specific visualizations:

- Spectra (Fig. 8): It shows the matched template, the object-centroid, and the prototype of a particular neuron. The user can also visualize the spectra of those objects in a neuron that best and worst fit the prototype.
- Population (Fig. 9): It shows the frequency of the different types of objects in the neuron according to the available templates or catalogs.
- Statistical summary (Fig. 10): The summary shows a table with the statistical information available for a neuron.

Some extra functionalities allow a complete interaction with other tools and databases. The first one is the cross-match utility, where the user can select any source in a neuron and perform a crossmatch with a selection of external catalogs to retrieve the available information about the source. At this moment, we provide access to three of the most used database catalogs: Simbad [37], SkyServer [38], and Aladin [39, 40].

The second utility is the integration of the Simple Application Messaging Protocol (SAMP) [41], which is the most common protocol of the Virtual Observatory in Astrophysics [42], to communicate the visualization with other astrophysical applications. The user can select several objects assigned to one or various neurons and send them to another tool using this protocol. Both the celestial coordinates or the spectra can be shared.

SAMP has a hub-based architecture, where the hub is a service used to route all messages between clients, but by default, it allows only local connections, which means that only applications on the same machine can communicate. We improved this protocol by adding new features to the hub in order to allow communications over the Internet, exchanging data between applications on different machines and locations, providing authentication (even centralized by means of an LDAP directory server) and an SSL-ciphered layer.



Fit ● Best ○ Worst

Spectra: Prototype

Template: Specific

Fig. 8 Example of representative spectra of a neuron. The map contains a test sample of spectrophotometric Gaia data

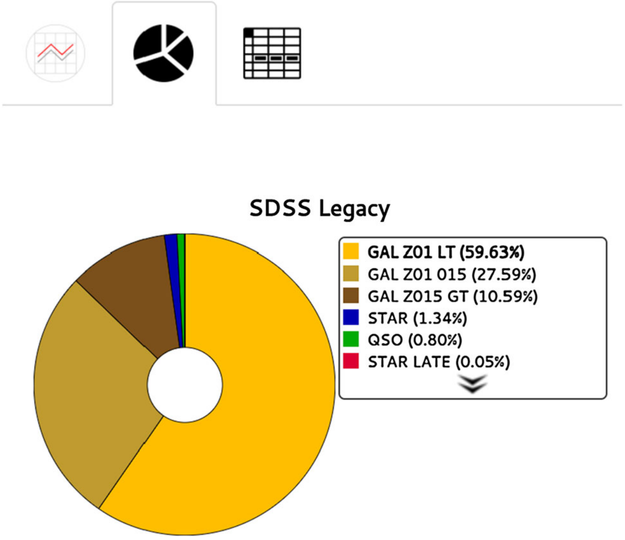
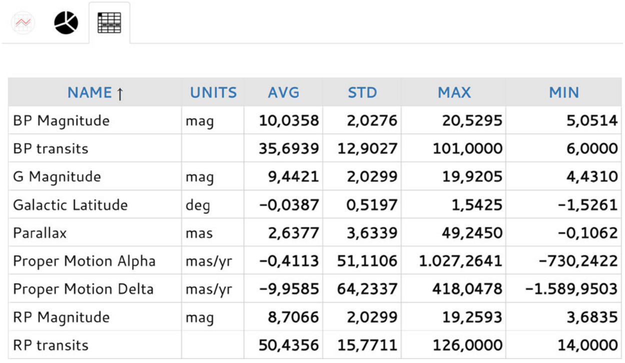


Fig. 9 Distribution of objects populating a neuron among classes: galaxy (GAL), star, and quasar (QSO). For GAL class, different subclasses are considered based on the redshift range



NAME ↑	UNITS	AVG	STD	MAX	MIN
BP Magnitude	mag	10,0358	2,0276	20,5295	5,0514
BP transits		35,6939	12,9027	101,0000	6,0000
G Magnitude	mag	9,4421	2,0299	19,9205	4,4310
Galactic Latitude	deg	-0,0387	0,5197	1,5425	-1,5261
Parallax	mas	2,6377	3,6339	49,2450	-0,1062
Proper Motion Alpha	mas/yr	-0,4113	51,1106	1.027,2641	-730,2422
Proper Motion Delta	mas/yr	-9,9585	64,2337	418,0478	-1.589,9503
RP Magnitude	mag	8,7066	2,0299	19,2593	3,6835
RP transits		50,4356	15,7711	126,0000	14,0000

Fig. 10 Example of a statistical summary of the relevant parameters of the objects populating a neuron

Carrying out all these changes required developing a new version of the hub that was named SAMP+. In order to maintain the backward compatibility with the original version, which is widely used, we allow users to connect hubs in a nested fashion using a bridge mode. In that way, we can place a central hub where our application is connected and other applications and hubs can connect as well to exchange data.

Table 2 shows the main differences between the original SAMP hub and the new one.

Figure 11 shows the architecture of the new and the old version of the connected hub.

Table 2 Comparison between SAMP and SAMP+ improved implementation

	SAMP	SAMP+
Features		
Area	Local machine	Internet
Authentication	None	User/Password
Security	None	SSL
Hub nesting	Disabled	Enabled

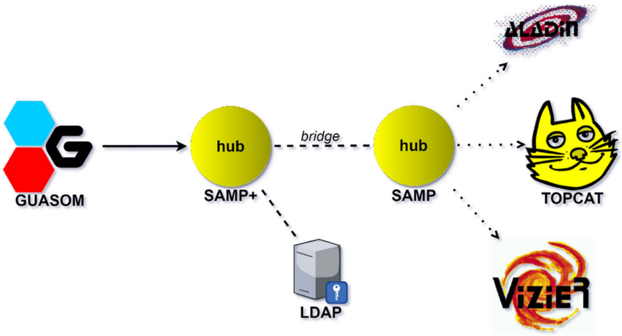


Fig. 11 Architecture of SAMP+, our improved version of SAMP

7 Clustering quality assessment

We evaluate the quality of the achieved clustering by means of external and internal criteria, as explained in the following sections.

7.1 External quality assessment

Verifying the performance of the processing techniques and, in this case, the quality of the obtained clustering is a fundamental task to ensure product quality. Although SOM networks address unsupervised grouping, it is always recommended to validate the performance in the domain through a set of labeled reference samples, using any external evaluation metrics that can be found in the bibliography [43]. Generally, those indices are defined for binary classification problems, in which a distinction is made between two possible classes, one positive and one negative, while for multiclass classification tasks, one of the most extended methods to analyze the resulting predictions is the confusion matrix. A confusion matrix is a useful tool for evaluating the success of a classification algorithm. It is a table in which each column represents an object class, for which we compute the percentage of objects falling into clusters where the predominant class corresponds to each object represented in the rows. The last row shows the number of objects per class in the input dataset. This type of external validation is useful to evaluate the performance of the clustering algorithm on a well-known set of data, for instance a validation dataset.

7.2 Internal quality assessment

When it is not possible to resort to any a priori information about the physical nature of the objects that populate the neurons, because such nature is unknown, clustering quality assessment based on classification success rates cannot be performed. In this case, internal parameters can be used to measure the quality of the grouping and to validate the SOM algorithm performance. In a Big Data environment where large datasets are analyzed, as in the case of Gaia mission, the computation of distances among all the observations in the input dataset is not feasible due to its high computational cost, both in time and memory. For this reason, we have chosen to carry out a descriptive approach to analyze the quality of the grouping, based on the intra-cluster distances.

We have selected three parameters that allow us to describe the distance distribution of all objects in a particular neuron: the width of the distribution according to the value of the FWHM (Full Width at Half Maximum), the skewness that measures the asymmetry of the distances distribution, and the kurtosis that measures the level of concentration of distances. A high quality clustering will result from neurons with low values for the parameter FWHM and with large positive values for both the skewness and for the kurtosis. In order to quantify any neuron clustering quality, we must establish a relative ranking of the values of those indices in any particular map. Thus, for example, the quality of the FWHM parameter is established in relation to the average value obtained for the 10% of neurons with lower FWHM. To facilitate the interpretation of these indices, we have defined a new categorical index L based on the values obtained for the three parameters. We empirically established seven categories to rank L, according to the values of each three parameters in 6 quartiles (95, 90, 75, 50, 32, and 10). Thus, if all the indices are in the 95th quantile, L will take value 0; if all are in the 90th quantile, then L will correspond to category 1, and so on up to category 6, which will correspond to those neurons whose quality indicators are outside the lowest quantile considered, 10. A visualization of L is included in GUASOM, and an example is shown in Fig. 12.

8 SOM clustering of ALHAMBRA survey photometry

Advanced Large Homogeneous Area Medium Band Redshift Astronomical (ALHAMBRA) [11] is a photometric survey developed in the Calar Alto Observatory (Spain)

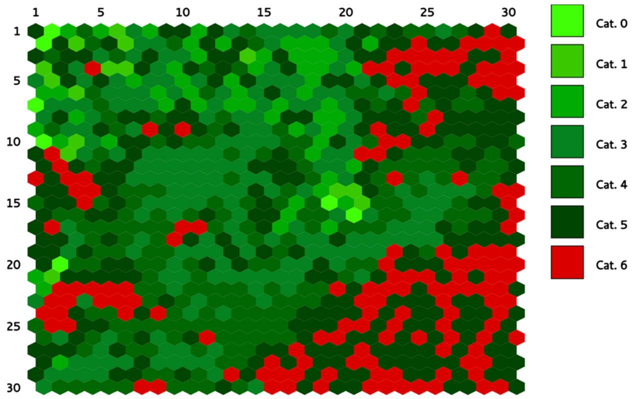


Fig. 12 GUASOM clustering quality categories obtained for SDSS legacy survey

between 2005 and 2012. It was conceived to study the cosmic evolution by measuring photometric redshifts of galaxies until a visible limiting magnitude close to 23.5 dex in several selected regions of the sky. ALHAMBRA archives contain approximately 400,000 sources, mainly galaxies and stars. They used a set of 20 filters covering the wavelength interval 3500 to 9700 A, plus three additional filters in the near infrared and a synthetic optical filter built from the previous ones. The archives contain the processed information of the sources which, among others, consist of the calibrated photometry, redshift determinations, and a statistical flag which uses information about the point spread function (PSF) of the sources to provide a probabilistic classification between stars and galaxies.

The input that we used to feed the SOMs consisted in the magnitudes observed in the 24 filters [44] described above, while the probabilities provided by the catalog (to be a star, galaxy, or unknown based on the PSF of the images) served as external information for labeling the clusters. The clustering obtained by GUASOM was analyzed in [45]. To summarize, GUASOM was able to reduce the dimension of the classification problem to a small number of neurons, whose prototypes represent the full population. Even more, when we represent in a color–color diagram (Fig. 13) the neuron prototypes, we see that GUASOM is capable of efficiently segmenting both types, without the need to resort to information on the PSF of the sources.

9 SOM clustering of SDSS legacy survey spectroscopy

SDSS Legacy survey [13] spectra contain a magnitude-limited sample of galaxies, a near-volume-limited sample of galaxies called Luminous Red Galaxies (LRG), a magnitude-limited sample of quasars, and a sample of stars. For each one of these classes, objects are selected from an imaging database for subsequent spectroscopic observation according to a target selection algorithm. In the Legacy main database, there are almost a million spectra for the two classes of galaxies and more than 100,000 spectra of quasars. The spectra have a wavelength coverage of 3800-9200 Å, with spectral resolution in the range 1800-2200 Å. An example is shown in Fig. 14.

SDSS employs a spectroscopic pipeline to automatically classify a given spectrum during data processing[1]. All objects are classified (SDSS “specClass” property) as either a quasar, high-redshift quasar, galaxy, star, late-type star, sky or unknown (this class represents those objects without associated label). Additionally to the spectroscopic classification, the pipeline provides other properties such as photometry in bands u, g, r, i, and z, astrometry, and the value of the spectroscopic redshifts (z) that can be useful to analyze GUASOM segmentation results.

GUASOM clustering results on SDSS Legacy main survey are shown in Figs. 14, 15, 16, and Tables 3 and 4. Figure 9 shows the category distribution within a particular neuron for the SDSS Legacy labels specified above. This visualization allows to assess GUASOM clustering according to the SDSS classification for that specific neuron. Figure 15 displays the hits distribution along the map, which allows to identify regions of interest according to their density. Note the region in the lower right part of the map composed of sparsely populated neurons containing sky observations and distributed over a rather large area in the map.

Figure 16 displays the category distribution of objects among the main astronomical classes, labeled according to SDSS spectroscopic classification. The “galaxy” object class contains the largest set of objects, and we combined SDSS information about the object classification with the values listed for the spectroscopic redshifts, establishing three redshift categories, namely $z < 0.1$, $0.1 \leq z \leq 0.15$, and $z > 0.15$. Figure 16 demonstrates that the tool distributes the different types and sub-types of objects with an ordered topology. It is remarkable that the sky spectra are well separated from other object types, precisely in the low-density region pointed out before, the lower right corner of the map.

It should be noted that in some places, neurons in gray color appear near the border of two well-defined categories. These regions, labeled as “undefined,” are composed by neurons where the qualified majority necessary to assign a class label has not reached a pre-established threshold (in this case 80%).

A quantitative measurement of the quality of the clustering obtained with GUASOM can be obtained through the confusion matrix, presented in Table 3, as well as from

[1] See https://classic.sdss.org/legacy/ for more details

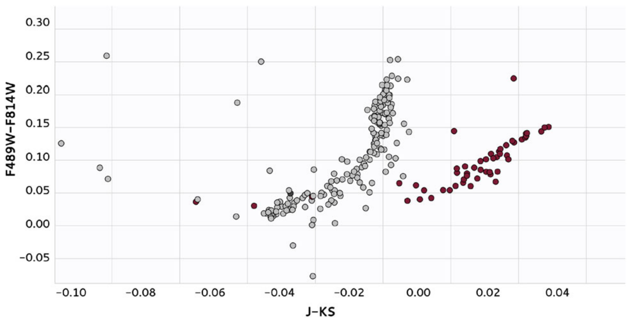


Fig. 13 Example of color–color visualization for ALHAMBRA survey. The circles in gray represent stars, and the ones in magenta represent galaxies

the performance metrics figures in Table 4. Without going into detail about the scientific validation of the results presented therein, the values displayed in the diagonal cells are always greater than 87% except for the "unknown" class, but for this class, GUASOM is able to identify a suitable class for approximately 66% of them, proving that it is strongly effective in classifying SDSS Legacy observations in astronomical types as well as in separating useful astronomical observations from sky spectra or other observational artifacts.

By way of comparison, Table 4 also shows the classification results of GUASOM main alternative, GTM technique, on the same dataset. As we mentioned in the introduction, GTM technique is the probabilistic counterpart to SOM, and it can represent an interesting option to perform unsupervised classification or even to cross-validate GUASOM results. The version of GTM that we have used runs sequentially in Python, while our GUASOM tool

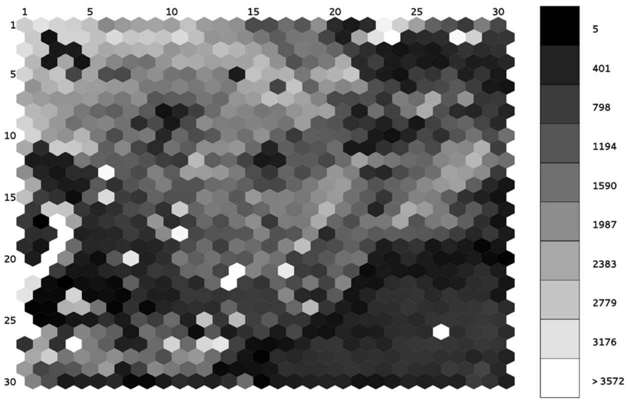


Fig. 15 Hits distribution. SDSS Legacy main survey sample

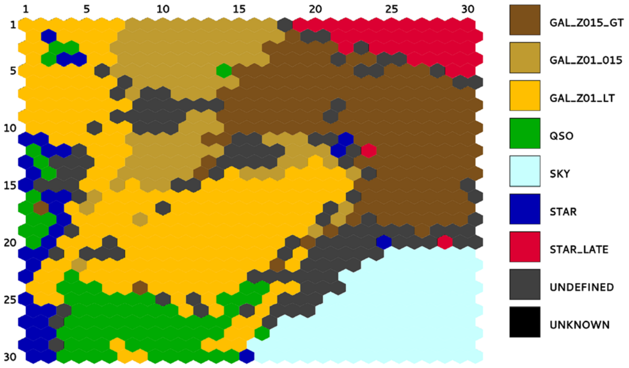
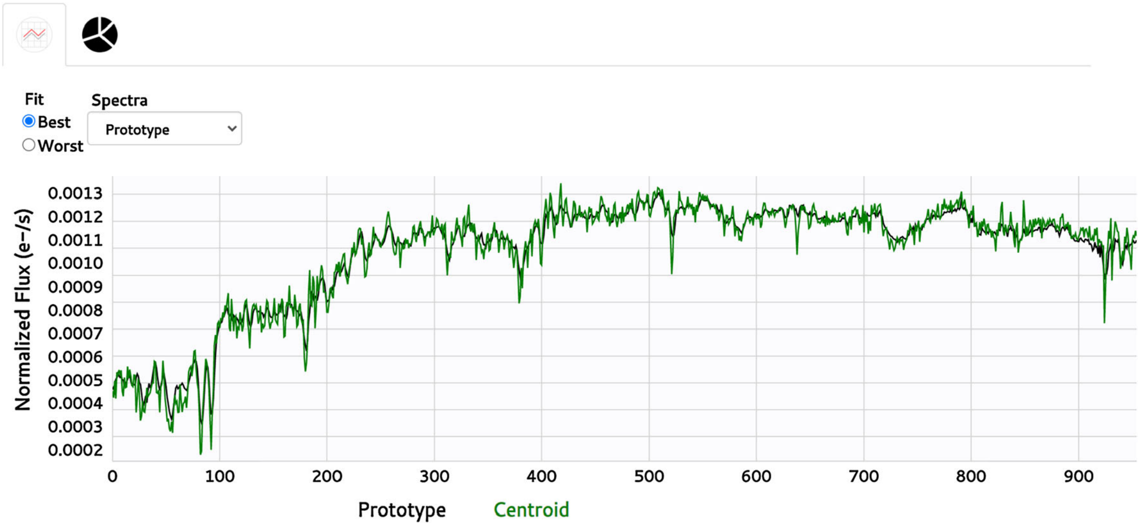


Fig. 16 GUASOM map of the SDSS Legacy main survey showing the category distribution of objects among the main astronomical classes from the SDSS spectral pipeline. For the "galaxy" class, there is a larger set of objects, and the classification labels have been combined with three redshift categories, showing that the tool distributes the different types and subtypes of objects with an ordered topology


Fit
● Best
○ Worst

Spectra
Prototype

Fig. 14 Example of representative spectra for a galaxy in SDSS Legacy survey

Table 3 Confusion matrix obtained by comparing GUASOM clustering results with those published by SDSS spectroscopic classification software in the SDSS Legacy main survey

		Class							
		Star	Star Late	Galaxy Z01_LT	Galaxy Z01_015	Galaxy Z015_GT	Quasar	Sky	Unknown
Prediction	Star	**88520**	1154	1222	415	757	201	58	1517
		88.07%	*2.48%*	*0.35%*	*0.20%*	*0.30%*	*1.94%*	*0.10%*	*10.34%*
	Star Late	2107	**42981**	77	40	688	254	33	565
		2.10%	*92.55%*	*0.02%*	*0.02%*	*0.27%*	*0.24%*	*0.06%*	*3.85%*
	Galaxy Z01_LT	3200	140	**333834**	15621	7049	6280	34	1348
		3.18%	*0.30%*	*95.87%*	*7.34%*	*2.77%*	*6.06%*	*0.06%*	*9.18%*
	Galaxy Z01_015	561	62	7273	**188898**	6785	1201	0	150
		0.56%	*0.13%*	*2.09%*	*88.78%*	*2.66%*	*1.16%*	*0.00%*	*1.02%*
	Galaxy Z015_GT	3133	1554	1489	7101	**235622**	2953	19	2006
		3.12%	*3.35%*	*0.43%*	*3.34%*	*92.49%*	*2.85%*	*0.03%*	*13.67%*
	Quasar	2456	243	4118	630	3197	**90417**	72	1827
		2.44%	*0.52%*	*1.18%*	*0.30%*	*1.25%*	*87.21%*	*0.13%*	*12.45%*
	Sky	33	131	23	12	94	109	**56492**	2295
		0.03%	*0.28%*	*0.01%*	*0.01%*	*0.04%*	*0.11%*	*98.52%*	*15.64%*
	Unknown	497	173	187	62	552	446	628	**4966**
		0.49%	*0.37%*	*0.05%*	*0.03%*	*0.22%*	*0.43%*	*1.10%*	*33.83%*
	Total	100507	46441	348223	212779	254746	103674	57340	14678
		8.83%	*4.08%*	*30.59%*	*18.69%*	*22.38%*	*9.11%*	*5.04%*	*1.29%*

Table 4 Performance metrics obtained by GUASOM on the classification of SDSS Legacy main survey. Global metrics performance figures are shown both for GUASOM as well as GTM in order to compare the clustering quality obtained by both methods

	Precision	Recall	F1-Score
Class metrics			
Star	0.93	0.88	0.90
Star Late	0.92	0.93	0.92
Galaxy Z01_LT	0.91	0.96	0.93
Galaxy Z01_015	0.92	0.89	0.90
Galaxy Z015_GT	0.93	0.92	0.93
Quasar	0.88	0.87	0.88
Sky	0.95	0.99	0.97
Unknown	0.66	0.34	0.45
Global metrics			
SOM	0.91	0.92	0.91
GTM	0.70	0.70	0.69

has been optimized and runs in a distributed manner, so it is difficult to compare the execution times. When we compare the values of the global metrics that illustrate the success rate of each of the techniques, we see that in the present application, SOM clustering quality is considerably better than GTM.

10 Conclusions

GUASOM, an adaptive visualization tool for unsupervised classification in large astronomical spectrophotometric surveys, has been described and its usefulness and performance demonstrated by conducting unsupervised classification of several large sets of astronomical data, both spectrophotometric as well as spectroscopic. Our goal was to facilitate the expert analysis of the clustering results by means of specialized visualizations, a task impossible to achieve by means of generic tools. We designed a client-server that handles the data treatment and computational tasks to give responses as quickly as possible and used JSON to pack the data between the server and the client. We optimized, parallelized, and evenly distributed the necessary calculations in a cluster of machines. By applying our clustering tool to several databases, we demonstrated the main advantages of an unsupervised approach: the classification is not based on pre-established models, thus allowing the "natural classes" present in the sample to be discovered, and it is suited to isolate atypical cases, with the important potential for discovery that this entails. The client-server architecture separates the computational part from the visualization part, making things much easier for the users, and it also allows us to manage and secure the data.

As shown, this visualization tool enhances the potential of the SOMs by allowing an in-depth study of the information, giving the experts a tool to perform more complex analysis. Additionally, we believe that the improvements made in the SAMP protocol can be of great interest because they allow us to secure the connection, to authenticate the users that exchange the information, and to connect applications running on different machines. Furthermore, it is possible to maintain backward compatibility while expanding connection possibilities.

At this moment, the Gaia Data Release 1 [46], Data Release 2 [47], and Early Data Release 3 [48] have been published, but Gaia spectrophotometry will not be available until Gaia Data Release 3 (DR3), which is scheduled for 2022. A specific tool version, called "GUASOM flavor DR3," will be published to the community in order to facilitate the analysis of one of the modules included in the Gaia software pipeline, the Outlier Analysis module [49], which is responsible for analyzing the classification outliers. GUASOM allows the exchange of data with other tools by means of the Virtual Observatory standard [42].

Acknowledgements This work made use of the infrastructures acquired with grants provided by the State Research Agency (AEI) of the Spanish Government and the European Regional Development Fund (FEDER), RTI2018-095076-B-C22. We acknowledge support from CIGUS-CITIC, funded by Xunta de Galicia and the European Union (FEDER Galicia 2014-2020 Program) through grant ED431G 2019/01 and research consolidation grant ED431B 2021/36. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC), https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High Performance Computing at the University of Utah. The SDSS website is www.sdss.org. SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration. We also want to acknowledge Alhambra survey funded by the Spanish Goverment under Grant AYA2006-14056.

Funding Open Access funding provided thanks to the Universidade da Coruña/CISUG agreement with Springer Nature.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Pearson K (1901) LIII On lines and planes of closest fit to systems of points in space. Lond Edinb Dublin Philos Mag J Sci 2(11):559–572. https://doi.org/10.1080/14786440109462720
2. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24:498–520
3. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500):2323–2326. https://doi.org/10.1126/science.290.5500.2323
4. Hinton GE, Roweis S (2003) Stochastic Neighbor Embedding. In: S. Becker, S. Thrun, K. Obermayer (eds.) Advances in Neural Information Processing Systems, vol. 15. MIT Press . https://proceedings.neurips.cc/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf
5. Carpenter GA, Grossberg S (2010) Adaptive resonance theory, pp. 22–35. Springer US, Boston, MA . https://doi.org/10.1007/978-0-387-30164-8_11
6. Martinetz T, Schulten K (1991) A "Neural-Gas" network learns topologies. Artif Neural Netw 1:397–402
7. Kohonen T, Schroeder M, Huang T (2001) Self-Organizing Maps, 3rd edn. Springer Series in Information Sciences
8. Chang J, Wang L, Meng G, Xiang S, Pan C (2017) Deep Adaptive Image Clustering. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5880–5888. https://doi.org/10.1109/ICCV.2017.626
9. Bishop CM, Svensén M, Williams CKI (1996) GTM: A Principled Alternative to the Self-Organizing Map. In: Proceedings of the 9th International Conference on Neural Information Processing Systems, NIPS'96, p. 354–360. MIT Press, Cambridge, MA, USA
10. Collaboration G, Prusti T, de Bruijne JHJ, Brown AGA et al (2016) The Gaia mission. Astron Astrophys 595:A1. https://doi.org/10.1051/0004-6361/201629272
11. Molino A, Benítez N, Moles M et al (2014) The ALHAMBRA survey: Bayesian photometric redshifts with 23 bands for 3 deg2. Mon Notices R Astron Soc 441(4):2891–2922. https://doi.org/10.1093/mnras/stu387
12. Stoughton C, Lupton RH, Bernardi M et al (2002) Sloan digital sky survey: early data release. Astron J 123(1):485–548. https://doi.org/10.1086/324741
13. Abazajian KN, Adelman-McCarthy JK, Agüeros MA et al (2009) The seventh data release of the sloan digital sky survey. Astrophys J Suppl Series 182(2):543–558. https://doi.org/10.1088/0067-0049/182/2/543
14. Sánchez Almeida J, Aguerri JAL, Muñoz-Tuñón C, de Vicente A (2010) Automatic unsupervised classification of all sloan digital sky survey data release 7 galaxy spectra. Astrophys J 714(1):487–504. https://doi.org/10.1088/0004-637X/714/1/487
15. Sánchez Almeida J, Allende Prieto C (2013) Automated unsupervised classification of the sloan digital sky survey stellar

spectra using k-means clustering. Astrophys J 763(1):50. https://doi.org/10.1088/0004-637X/763/1/50

16. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, pp. 281–297. University of California Press, Berkeley, Calif. https://projecteuclid.org/euclid.bsmsp/1200512992
17. Matijevič G, Prša A, Orosz JA et al (2012) Kepler Eclipsing Binary Stars. III. Classification of Kepler eclipsing binary light curves with locally linear embedding. Astron J 143(5):123. https://doi.org/10.1088/0004-6256/143/5/123
18. Ward JL, Lumsden SL (2016) Locally linear embedding: dimension reduction of massive protostellar spectra. Mon Not R Astron Soc 461(2):2250–2256. https://doi.org/10.1093/mnras/stw1510
19. Vanderplas J, Connolly A (2009) Reducing the dimensionality of data: locally linear embedding of Sloan galaxy spectra. Astron J 138(5):1365–1379
20. Naim A, Ratnatunga U, Griffiths E (2009) Galaxy morphology without classification: self-organizing maps. Astrophys J Suppl Series 111:357. https://doi.org/10.1086/313022
21. Geach JE (2012) Unsupervised self-organized mapping: a versatile empirical tool for object selection, classification and redshift estimation in large surveys. MNRAS 419:2633–2645. https://doi.org/10.1111/j.1365-2966.2011.19913.x
22. Way M, Klose C (2012) Can self-organizing maps accurately predict photometric redshifts? Publications of the Astronomical Society of the Pacific 124. https://doi.org/10.1086/664796
23. Carrasco Kind M, Brunner RJ (2014) SOMz: photometric redshift PDFs with self-organizing maps and random atlas. Mon Notices R Astron Soc 438(4):3409–3421. https://doi.org/10.1093/mnras/stt2456
24. Ordóñez-Blanco D, Arcay B, Dafonte C et al (2010) Object classification and outliers analysis in the forthcoming Gaia mission. Lect Notes Essays Astrophys 4:97–102
25. Fustes D, Dafonte C, Arcay B et al (2013) SOM ensemble for unsupervised outlier analysis. Application to outlier identification in the Gaia astronomical survey. Expert Syst Appl 40(5):1530–1541. https://doi.org/10.1016/j.eswa.2012.08.069
26. Fustes D, Manteiga M, Dafonte C et al (2013) An approach to the analysis of SDSS spectroscopic outliers based on self-organizing maps: Designing the outlier analysis software package for the next Gaia survey. Astron Astrophys. https://doi.org/10.1051/0004-6361/201321445
27. Alvarez-Betancourt Y, Garcia-Silvente M (2014) An overview of iris recognition: a bibliometric analysis of the period 2000–2012. Scientometrics. https://doi.org/10.1007/s11192-014-1336-1
28. Moosavi V, Packmann S (2015) SOMPY: a self organizing map library in Python. https://github.com/sevamoo/SOMPY
29. Oy L, Vesanto J, Himberg J, Alhoniemi E, Parhankangas J (2000) Som toolbox for matlab. Helsinki University of Technology. https://citeseer.ist.psu.edu/vesanto00som.html
30. Thang C (2007) Spice Neural Network https://spiceneuro.wordpress.com/english/
31. Bouvier G, Desdouits N, Ferber M, Blondel A, Nilges M (2014) An automatic tool to analyze and cluster macromolecular conformations based on self-organizing maps. Bioinformatics 31(9):1490–1492. https://doi.org/10.1093/bioinformatics/btu849
32. Pölzlbauer G (2004) Survey and comparison of quality measures for self-organizing maps. Proceedings of the Fifth Workshop on Data Analysis (WDA04) pp. 67–82
33. Fort JC, Letrémy P, Cottrell M (2002) Advantages and drawbacks of the Batch Kohonen algorithm. In: ESANN, pp. 223–230
34. Kaski S (1997) Data exploration using self-organizing maps. Acta Polytech Scand 82:57
35. Israel J, Mitchell J, Xerox Corporation, Palo Alto Research Center, Sturgis H (1978) Separating Data from Function in a Distributed File System. CSL.: Xerox Palo Alto Research Center. Xerox Palo Alto Research Center. https://books.google.es/books?id=gYs3HAAACAAJ
36. Fielding RT, REST, (2000) Architectural Styles and the Design of Network-based Software Architectures. University of California, Irvine http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm
37. Wenger M, Ochsenbein F, Egret D et al (2000) The SIMBAD astronomical database: the CDS reference database for astronomical objects. Astron Astrophys Suppl Series 143(1):9–22
38. Szalay AS, Gray J, Thakar AR et al (2002). The SDSS Skyserver: Public Access to the Sloan Digital Sky Server Data. In: Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, pp. 570–581. ACM, New York, NY, USA. https://doi.org/10.1145/564691.564758
39. Bonnarel F, Fernique P, Bienaymé O et al (2000) The ALADIN interactive sky atlas. A reference tool for identification of astronomical sources. Astron Astrophys 143:33–40. https://doi.org/10.1051/aas:2000331
40. Boch T, Fernique P (2014) Aladin Lite: Embed your Sky in the Browser. In: N. Manset, P. Forshay (eds.) Astronomical Data Analysis Software and Systems XXIII, Astronomical Society of the Pacific Conference Series, vol. 485, p. 277
41. Taylor M, Boch T, Taylor J (2015) SAMP, the simple application messaging protocol: letting applications talk to each other. Astron Comput 11(PB):81–90. https://doi.org/10.1016/j.ascom.2014.12.007
42. Quinn PJ, Barnes DG, Csabai I et al (2004) The International Virtual Observatory Alliance: recent technical developments and the road ahead. In: Quinn PJ, Bridger A (eds) Optimizing Scientific Return for Astronomy through Information Technologies, vol 5493. International Society for Optics and Photonics, SPIE, pp 137–145. https://doi.org/10.1117/12.551247
43. Tharwat A (2018) Classification assessment methods. Appl Comput Inf 17(1):168–192. https://doi.org/10.1016/j.aci.2018.08.003
44. Aparicio Villegas T, Alfaro EJ, Cabrera-Caño J et al (2010) The ALHAMBRA photometric system. Astronl J 139(3):1242–1253. https://doi.org/10.1088/0004-6256/139/3/1242
45. Garabato D, Manteiga M, Dafonte C, Álvarez MA (2017) Guasom Analysis Of The Alhambra Survey. In: Early Data Release and Scientific Exploitation of the J-PLUS Survey, p. 16 . https://doi.org/10.5281/zenodo.1041764
46. Collaboration G, Brown AGA, Vallenari A, Prusti T et al (2016) Gaia data release 1. Summary of the astrometric, photometric, and survey properties. Astron Astrophys. https://doi.org/10.1051/0004-6361/201629512
47. Collaboration G, Brown AGA, Vallenari A, Prusti T et al (2018) Gaia data release 2. Summary of the contents and survey properties. Astron Astrophys. https://doi.org/10.1051/0004-6361/201833051
48. Collaboration G, Brown AGA, Vallenari A, Prusti T et al (2020) Gaia early data release 3: Summary of the contents and survey properties. Astron Astrophys. https://doi.org/10.1051/0004-6361/202039657
49. Bailer-Jones CAL, Andrae R, Arcay B et al (2013) The Gaia astrophysical parameters inference system (Apsis) - pre-launch description. Astron Astrophys 559:A74. https://doi.org/10.1051/0004-6361/201322344

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.