Highlights on Spanish Astrophysics X, Proceedings of the XIII Scientific Meeting of the Spanish Astronomical Society held on July 16 – 20, 2018, in Salamanca, Spain. B. Montesinos, A. Asensio Ramos, F. Buitrago, R. Schödel, E. Villaver, S. Pérez-Hoyos, I. Ordóñez-Etxeberria (eds.), 2019

Detection of new Open Clusters with *Gaia*.

A. Castro-Ginard[1], C. Jordi[1], X. Luri[1], L. Balaguer-Núñez[1], and T. Cantat-Gaudin[1]

[1] Dept. Física Quàntica i Astrofísica, Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Martí i Franquès 1, E08028 Barcelona, Spain.

Abstract

The publication of the *Gaia* Data Release 2 (*Gaia* DR2) includes precise astrometric data (positions, proper motions and parallaxes) for more than 1.3 bilion sources, mostly stars. This such a vast amount of new data requires the use of machine-learning and data-mining techniques to handle large scale analysis. In particular, the search for open clusters (OCs), groups of stars that were born and move together, located in the disc, is a great example for the application of these techniques.

We explore the performance of a density based clustering algorithm, DBSCAN, to find clusters in the data together with a supervised learning method such as an Artificial Neural Network (ANN) to automatically distinguish between real OCs and statistical clusters.

The development and implementation of this method in a five-dimensional space $(\alpha, \delta, \varpi, \mu_{\alpha^*}, \mu_\delta)$ of the *Tycho-Gaia* Astrometric Solution (TGAS) lead to the proposal of a list of new nearby OCs candidates. This contribution shows the validation of the candidates with *Gaia* DR2 data and a framework designed to be applied to the full *Gaia* DR2 archive.

1 Introduction

The analysis of astronomical catalogues is becoming more complex as the data volume of these catalogues is increasing. For instance, the *Gaia* mission [1] in its first data release (*Gaia* DR1 [2]) contains positions for more than one bilion sources. Even though this large amount of sources, full five-parameter astrometric data is available only for a small subset: the *Tycho-Gaia* Astrometric Solution (TGAS [3, 4]). The TGAS subset represents a perfect scenario to develop and test scientific applications, based on data-mining techniques and machine-learning algorithms, in preparation for larger releases. The use of these techniques is mandatory from the second *Gaia* data release (*Gaia* DR2 [5]) onwards, which contains precise five-parameter astrometric data for more than 1.3 bilion sources, together with three-band photometry.

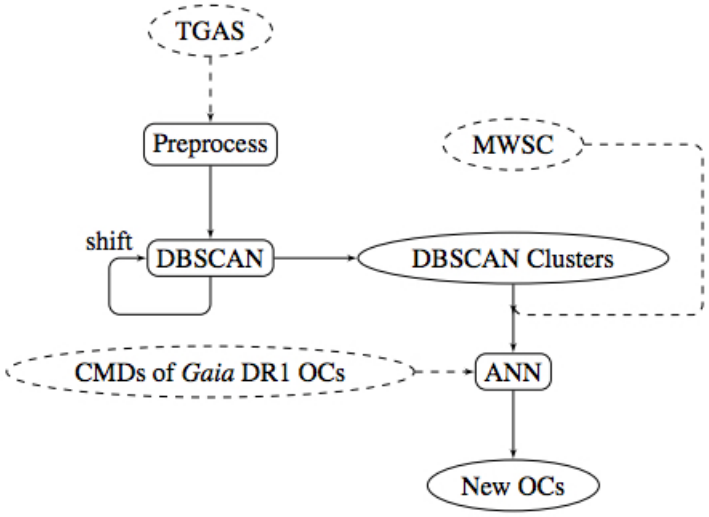

Figure 1: Application of the method represented by a flow chart diagram. Figure taken from Fig. 1 in [6].

We have developed a method [6] to automatically search for overdensities in the five-dimensional astrometric data, *i.e.* positions, parallax and proper motions $(\alpha, \delta, \varpi, \mu_{\alpha^*}, \mu_\delta)$, and decide if they are open clusters (OCs) based on the photometry (G, G_{BP}, G_{RP}). The method is developed and tested on the TGAS subset, with the final goal of its application to the full *Gaia* DR2 archive.

2 Method

Figure 1 shows a diagram of the methodology used to identify possible new OCs. Using TGAS as our initial database, we apply an unsupervised learning algorithm such as DBSCAN [7] to detect groups of stars showing an overdensity in the five-parameter space. Then, these overdensities are classified into statistical clusters or physical OCs using an Artificial Neural Network (ANN [8]), which identifies an isochrone on a Color Magnitude Diagram (CMD). In this case, because the TGAS subset is purely astrometrical data, the CMD is built using the photometric data from the *Two Micron All Sky Survey* catalogue (2MASS [9]).

2.1 Preprocessing

Before the application of the method, we select a region of the sky where we expect to find most of the clusters. According to existing OCs catalogues such as the DAML [10] and MWSC [11], most of the clusters are found at $|b| < 20$deg (96% and 94% respectively). In addition, we reject stars with high or negative parallaxes (selecting only stars with 0mas$\leq \varpi \leq$ 7mas) or with high proper motions ($|\mu_{\alpha^*}|, |\mu_\delta| > 30$mas·yr^{-1}); this facilitates the determination of the DBSCAN parameters with no loss of generality since these conditions would make an OC

easily detectable.

The region of study is further divided into smaller rectangles of size L deg. This second division is done to reduce the volume of data in each region in order to reduce computational time; and to define a more representative density of field stars, so the algorithm can search for a significant overdensity in that region. As a last step, because the algorithm computes the distance between pairs of stars in the five-dimensional parameter space, and decides if they are clustered or not based on these distances, we standarise the star parameters (to have mean zero and variance one) so their weights in the process are the same.

2.2 DBSCAN

DBSCAN is a density-based clustering algorithm that identifies overdensities in the parameter space as clusters. The definition of what DBSCAN considers a cluster depends on two parameters: ϵ and *minPts*. The parameter *minPts* refers to the minimum number of members of a cluster, while ϵ refers to the radius of the hypersphere (in the parameter space) centred in each star where this *minPts* members have to be located (see Fig. 2 of [6]).

The determination of the *minPts* parameter, together with L, is left to be optimised using simulations (see Sect. 3 of [6] for details). The values for these parameters found to be optimal in this case are: $L = \{12, 13, 14, 15, 16\}$ and *minPts*$= \{5, 6, 7, 8, 9\}$.

For the determination of ϵ, we take advantage of the fact that the distance from a star to its k_{th} nearest neighbour from stars belonging to the cluster is smaller than from stars belonging to the field. Figure 2 shows an example of the determination of ϵ in a region around a known cluster, the red line (ϵ) separates the stars belonging to the cluster (green) from the stars belonging to the field (orange).

2.3 Identification of OCs

DBSCAN finds clusters in an statistical sense, they can be real OCs or just statistical clusters. To distinguish between these two posibilities, we use an ANN with a multilayer perceptron arquitecture with one hidden layer. The ANN is able to recognise the isochrone pattern in the CMD, and therefore identify real OCs among all the candidates. This is achieved by training the ANN with examples of CMDs of real OCs. Since we work with TGAS data, the OCs examples are those from the *Gaia* DR1 [12]. The test CMDs are classified with a precision of a 97.05% to the right class, OC or statistical cluster.

Because the ANN is trained with clusters from [12], we expect to find clusters with the same characteristics. The OCs used to train the ANN are nearby clusters with ages from 40 to 850 Myr and no significant differential extinction.

3 Results

The whole method is applied to the TGAS data, and after the removal of coincident clusters with [11], we end with a list of 31 probable OC candidates (see Table 1 in [6], cluster candidates

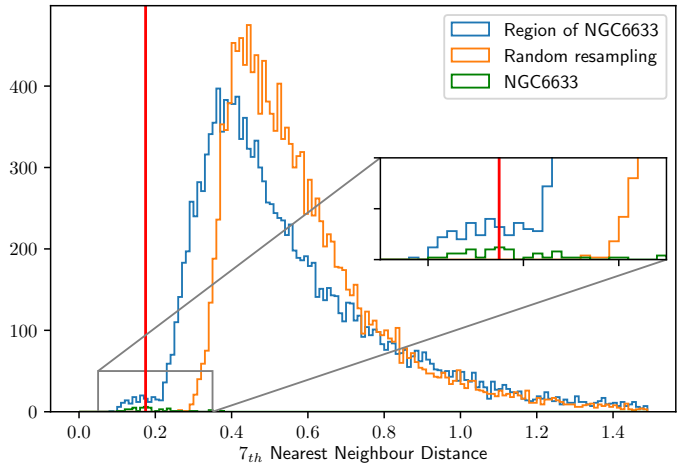

Figure 2: Histogram of 7_{th}-NN distances of a region around NGC6633 (in blue), stars belonging to NGC6633 (in green) and random realization of field stars in that region (in orange). Figure taken from Fig. 3 in [6].

are sorted by number of detections through the explored parameters). Each of the OC candidates is then analysed using *Gaia* DR2 data, which provides more precise astrometric data, photometry more precise than that of 2MASS catalogue and the availability of those parameters down to magnitude $G \sim 21$.

In order to confirm or discard each OC candidate, the DBSCAN algorithm is executed on a *Gaia* DR2 region around the expected centre of the candidate. With this last step, we are able to confirm 23 OC candidates with members down to magnitude $G \leq 17$. These 23 confirmed OCs represent a 70% of the proposed candidates; 100% of the initial OCs candidates found with $N_{\text{found}} \geq 5$ among the explored parameters are confirmed, while for $N_{\text{found}} < 5$ we are able to confirm 59% of the initial candidates. Mean values for position, parallax and proper motion, as well as for radial velocity when available, can be found in Table 2 of [6]. General comments and comments on individual proposed new OCs can be also found in [6].

4 Conclusions

We describe an automated data-mining method for the detection of OCs. The method is based on the use of machine learning techniques such as a clustering algorithm, DBSCAN, to detect overdensities in astrometric data; and a classification algorithm, an ANN, to distinguish between statistical clusters and real OCs.

The application of the method to TGAS data allows the proposal of 31 new OCs candidates, of which 23 (around 70%) are validated using *Gaia* DR2 data.

Looking forward to the application of the method to the all-sky *Gaia* DR2, we have to optimize the parameters L and $minPts$ to account for the larger stellar densities. As well, the better characterization of known OCs [13] with DR2, provides a wider training set for the ANN step.

Acknowledgments

This work was supported by the MINECO (Spanish Ministry of Economy) through grant ESP2016-80079-C2-1-R (MINECO/FEDER, UE) and MDM-2014-0369 of ICCUB (Unidad de Excelencia 'María de Maeztu'). This work has made use of the data from the European Space Agency (ESA) mission Gaia, processed by the Gaia Data Processing and Analysis Consortium (DPAC). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement.

References

[1] Gaia Collaboration (Prusti, T. et al) 2016, A&A 595, A1

[2] Gaia Collaboration (Brown, A.G.A., et al.) 2016, A&A, 595, A2

[3] Lindegren, L., Lammers, U., Bastian, U., et al. 2016, A&A, 595, A4

[4] Michalik, D., Lindegren, L. & Hobbs, D. 2015, A&A, 574, A115

[5] Gaia Collaboration (Brown, A.G.A., et al.) 2018, A&A, 616, A1

[6] Castro-Ginard, A., Jordi, C., Luri, X., et al. 2018, A&A, 618, A59

[7] Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. 1996, in Proc. of the Second International Conf. on Knowledge Discovery and Data Mining, KDD96 (AAAI Press), 226

[8] Bishop, C.M. 1995, Neural Networks for Pattern Recognition (New York, NY, USA: Oxford University Press, Inc.)

[9] Skrutskie, M.F., Cutri, R.M., Stiening, R., et al. 2006, The Astronomical Journal, Volume 131, Issue 2, pp. 1163-1183

[10] Dias, W.S., Alessi, B.S., Moitinho, A., & Lépine, J.R.D. 2002, A&A, 389, 871

[11] Kharchenko, N.V., Piskunov, A.E., Schilbach, E., Röser, S., & Scholz, R.-D. 2013, A&A, 558, A53

[12] Gaia Collaboration (can Leeuwen, F. et al.) 2017, A&A, 601, A19

[13] Cantat-Gaudin, T., Jordi, C., Vallenari, A., et al 2018, ArXiv e-prints [arXiv:1805.08726]