DECOCC: Deep Embedded Clustering for Open Cluster Characterization with Gaia DR2/3 Data

Carlos D. Álvaro (Meteológica), César A. Guzmán (IAC, UNAD), J. Álvaro (S&P-EPS, SEA)

Leiden eas 2021 Virtual

A novel method to characterize Open Clusters (OC). Our method employs a model built on ***Artificial Neural Networks*** (ANNs). More specifically, we adapted a state of the art model, the ***Deep Embedded Clustering*** (DEC - Xie, Girshick & Farhadi, 2016) model for our purpose. The developed method aims to *improve classical state of the arts techniques*. We *improved not only in terms of computational efficiency* (with lower computational requirements), but in *usability* (reducing the number of hyperparameters to get a good characterization of the analyzed clusters). For our experiments, we used the ***Gaia DR2 database*** as the data source, and compared our model with the clustering technique K-Means. *Our method achieves good results*, becoming even better (in some of the cases) than current techniques.

1 DATA ACQUISITION

We used the OpenCLUST catalog to select stellar regions of the sky and then downloaded all the data from Gaia DR2 database regarding those regions.

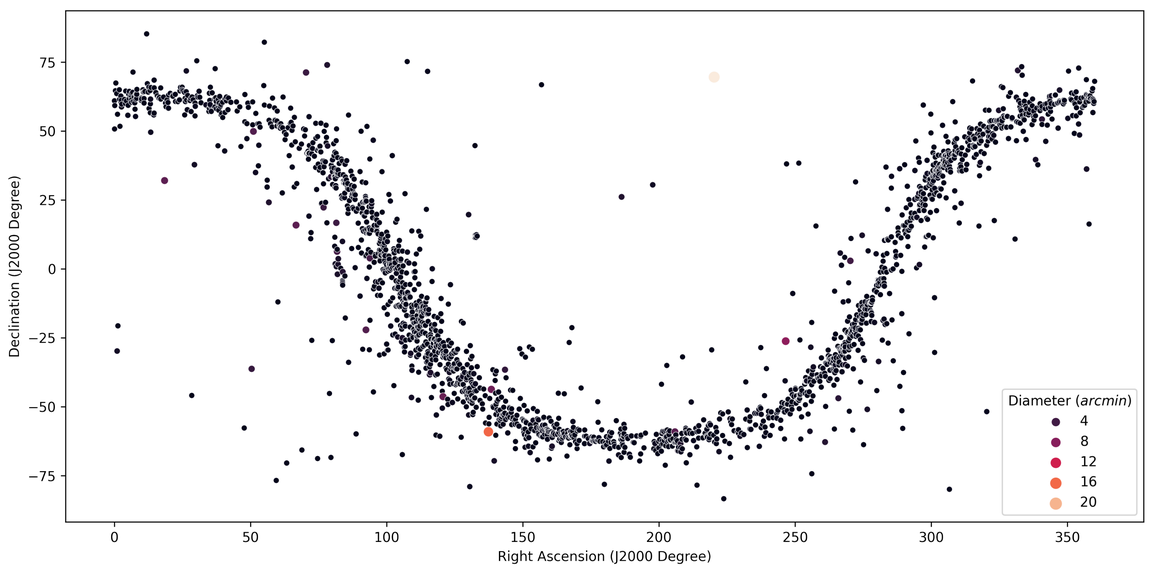

Figure 1: OpenCLUST catalogue distribution

2 FEATURE SELECTION

For computational efficiency purposes, we operated as few features as possible while maintaining a good characterization of the studied clusters. Also, instead of using the raw variables downloaded from Gaia, we corrected some of them by applying simple transformations.

Feature	Description	Computation
parallax (*mas*)	Parallax (ω)	RAW feature
pmra_corr (*year*$^{-1}$)	Proper motion in right ascension divided by parallax	$\frac{pmra}{parallax}$
pmdec_corr (*year*$^{-1}$)	Proper motion in declination divided by parallax	$\frac{pmdec}{parallax}$
pmmod (*mas·yr*$^{-1}$)	Proper motion modulus	$\sqrt{pmra\_corr^2 + pmdec\_corr^2}$

Table 1: Features provided to DECOCC

We used Vmag and BP-RP features outside of our model to build the HR diagrams and identify the open cluster on each case.

3 K-MEANS DRAWBACKS

K-Means is a simple clustering model that only requires a single parameter: the number of clusters to make. However, while by increasing the number of groups we are able to better isolate the OC we are looking for, more clusters arises. This fact complicates our task to distinguishing the cluster of interes. Therefore, we need a new method that creates fewer groups but capable of isolate the desired open cluster.

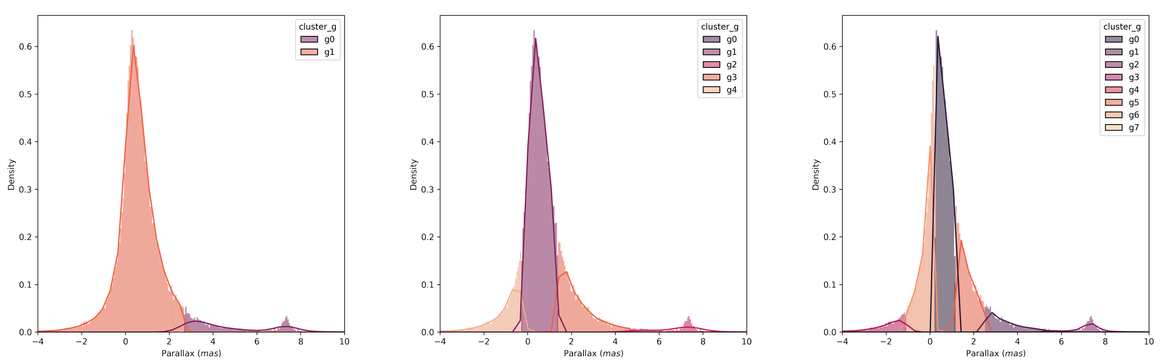

Figure 2: K-Means with Melotte 22 located at ~7.5 mas

4 DECOCC

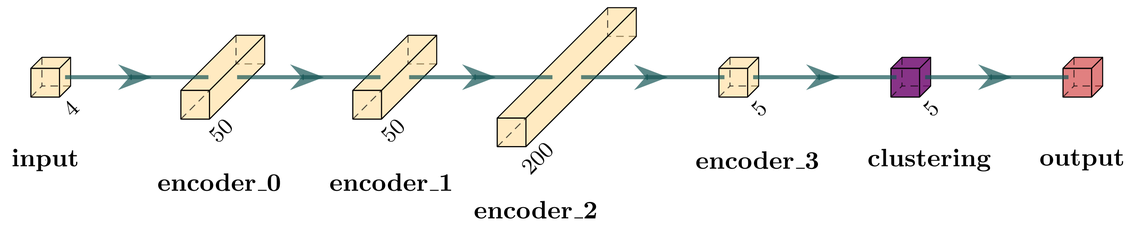

Figure 3: DECOCC neural network default parameterization

DECOCC ***is an adaptation of DEC model*** to the Open Cluster Characterization problem. DECOCC's Neural Network layers are simpler than those in DEC model due to the smaller number of features in our problem. DEC is ***composed by an autoencoder*** (encoder/decoder layers) ***and a clustering layer*** that begins with an initial clustering provieded by K-Means and then clusters stars from a given region.

5 RESULTS & CONCLUSIONS

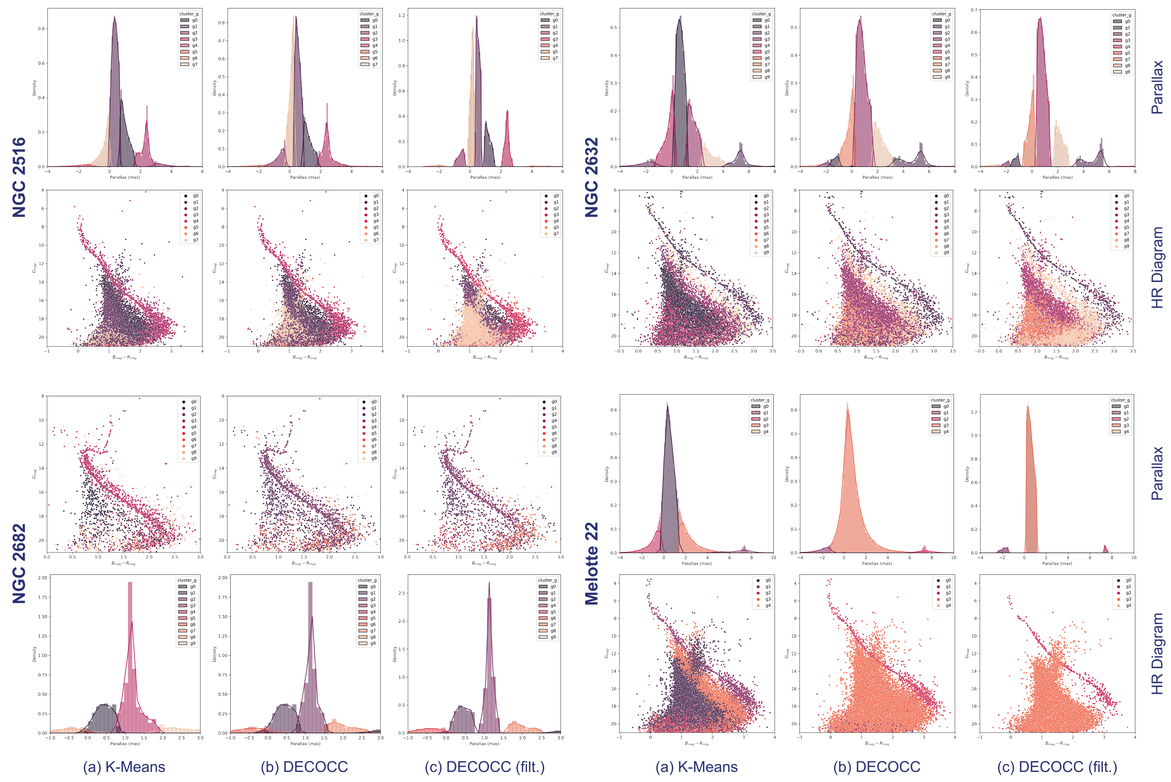

Figure 4: Comparison among K-Means, DECOCC and DECOCC filtered results for NGC 2516, NGC 2632, NGC 2682 and Melotte 22 characterization

	Method	μ_α (*mas·yr*$^{-1}$)	μ_δ (*mas·yr*$^{-1}$)	ω (*mas*)	# stars
NGC 2516	Simbad	-4.6579 ± 0.0075	11.1517 ± 0.0075	2.4118 ± 0.0006	1727
	K-Means	-4.344 ± 0.14	9.507 ± 0.19	2.268 ± 0.01	1542
	DECOCC	-4.426 ± 0.17	9.952 ± 0.20	2.436 ± 0.01	1532
	DECOCC (filt.)	-4.502 ± 0.14	10.114 ± 0.17	2.392 ± 0.004	1072
NGC 2632	Simbad	-36.047 ± 0.110	-12.917 ± 0.066	5.371 ± 0.003	–
	K-Means	-26.352 ± 0.82	-15.828 ± 0.76	5.394 ± 0.03	629
	DECOCC	-20.012 ± 0.69	-14.742 ± 0.58	4.686 ± 0.03	894
	DECOCC (filt.)	-21.571 ± 0.74	-14.234 ± 0.61	4.719 ± 0.03	714
NGC 2682	Simbad	-10.9737 ± 0.0064	-2.9396 ± 0.0063	1.1325 ± 0.0011	1194
	K-Means	-8.616 ± 0.15	-3.710 ± 0.16	1.196 ± 0.01	1374
	DECOCC	-8.926 ± 0.15	-3.550 ± 0.15	1.144 ± 0.005	1238
	DECOCC (filt.)	-9.619 ± 0.13	-3.317 ± 0.13	1.140 ± 0.003	990
Melotte 22	Simbad	19.997 ± 0.127	-45.548 ± 0.101	7.364 ± 0.005	1326
	K-Means	20.25 ± 0.95	-38.01 ± 1.08	7.23 ± 0.06	1378
	DECOCC	23.67 ± 1.29	-46.23 ± 1.50	8.04 ± 0.09	878
	DECOCC (filt.)	19.50 ± 0.41	-44.23 ± 0.39	7.42 ± 0.005	438

	Hyperparameter	Value
NGC 2516	Number of clusters	8
	Clustering Layer	[50, 50, 60]
	Kernel Initializer Seed	2
	Quantil Threshold	0.15
NGC 2632	Number of clusters	10
	Clustering Layer	[50, 50, 40]
	Kernel Initializer Seed	10
	Quantil Threshold	0.1
NGC 2682	Number of clusters	10
	Clustering Layer	[50, 50, 40]
	Kernel Initializer Seed	10
	Quantil Threshold	0.1
Melotte 22	Number of clusters	5
	Clustering Layer	[50, 50, 200]
	Kernel Initializer Seed	11
	Quantil Threshold	0.1

We have ***achieved a model*** to characterize open clusters, comprising an Artificial Neural Network, which is ***neither parameterized nor supervised***. It does ***not require complex hardware***. The model ***works well in identifying and characterizing open clusters***. However, some improvements must be made to improve its accuracy and precision.

cdalvaro