Final Project

IBM Machine Learning Professional Certificate

Course 4: Unsupervised Machine Learning

Emily Kendall

1 Introduction and Main Objectives

In this work we use unsupervised machine learning techniques to identify Open Clusters (OCs) within observational data gathered by the Gaia Space Telescope (Gaia Collaboration et al., 2016). OCs are a type of star cluster comprising $\mathcal{O}(10) - \mathcal{O}(1000)$ stars that were formed from the same progenitor molecular cloud and have similar age. OCs are particularly useful as tracers of the structure of the Milky Way galaxy, and studies of these objects have produced valuable insights on the history and evolution of the galactic disk.

Various studies have been conducted aiming to group Milky Way stars into individual OCs (Dias et al., 2002; Kharchenko et al., 2013; Cantat-Gaudin et al., 2018; Ratzenböck et al., 2023). While there has been significant overlap between studies, there remain many stars whose cluster association is disputed or unknown. The main objective of this analysis, therefore, is to apply a series of unsupervised clustering techniques to Gaia data in order to independently identify OCs and assign stars to them. This work will enable us to verify previous stellar OC assignments as well as identify stars for which cluster assignments are uncertain, guiding follow-up studies.

2 Dataset Overview

In order to improve the efficiency of this task, we consider only a subset of the data made available in the second Gaia data release (DR2) (Gaia Collaboration et al., 2018). In particular, we utilise the dataset made available in (Cantat-Gaudin et al., 2018) (Hereafter CG18), from which the authors were able to identify 1229 unique OCs. This dataset has been made publicly available through the VizieR catalogue access tool (Ochsenbein et al., 2000).

This dataset contains observations for 401448 individual stars, with each star assigned to one of 1229 unique OCs. In this work, we will perform our own independent clustering analysis and compare our results with these assignments. The columns of the original dataset are described in Table 1.

Of the columns listed in Table 1, we may drop '_RA.icrs' and '_DE.icrs', as these are simply duplicates of 'RA_ICRS' and 'DE_ICRS', respectively. We may also drop columns 'GLON' and 'GLAT', as these are mathematically redundant ('GLON' and 'GLAT' are an alternative way of expressing the same sky position encoded by 'RA_ICRS' and 'DE_ICRS'). Further data cleaning steps are described in the following section.

3 Exploratory Data Analysis

3.1 Data Cleaning

In order to ensure we are using the most reliable, high-quality data, we filter the dataset to exclude any stars which have been observed less than 100 times ('o_Gmag'$<$ 100), as these are likely to have lower quality G-band magnitude estimates and less reliable photometry.

Column	Description
RA_ICRS	Right ascension (ICRS), in degrees
DE_ICRS	Declination (ICRS), in degrees
Source	Gaia DR2 source identifier
GLON	Galactic longitude, in degrees
GLAT	Galactic latitude, in degrees
plx	Parallax, in milliarcseconds (mas)
pmRA	Proper motion in Right ascension ($\mu_\alpha \cos\delta$), in mas/yr
pmDE	Proper motion in Declination, in mas/yr
o_Gmag	Number of G-band observations
Gmag	Gaia G-band magnitude
BP-RP	Gaia colour index (BP magnitude - RP magnitude)
PMemb	Membership probability (0 to 1)
Cluster	Name of the assigned open cluster
SimbadName	Simbad cross-identified name (if available)
_RA.icrs	Duplicate of RA_ICRS
_DE.icrs	Duplicate of DE_ICRS

Table 1: Description of Columns in Gaia Open Cluster Member Catalog J/A+A/618/A93

Following this, we drop the 'o_Gmag' column entirely, as this will have no bearing on cluster assignment.

We then restrict our dataset further to those stars with brighter G-band magnitudes ('Gmag' < 17). In doing so we exclude dimmer sources which tend to have larger uncertainties in proper motion and parallax. This cutoff value also dramatically diminishes the volume of data, making computations faster.

We find that the original dataset contains 2144 null values in the 'BP-RP' column, so we exclude all such rows from the dataset.

The resulting dataset contains no null values other than in the 'SimbadName' column, which we will not be using for clustering. Therefore, we do not need to impute any missing values. The 'SimbadName' column, while incomplete, may be useful in comparing the results obtained here with previous studies of open clusters, so we retain this column in our original dataset, but remove it from the working data fed into our clustering models.

3.2 Feature Engineering and Feature Selection

3.2.1 Tangential Velocity

It is sometimes informative to include the linear velocity of astrophysical objects in addition to angular parameters during analysis. To this end, we will combine several existing parameters to create an estimate of the tangential velocity. The tangential velocity represents the star's speed across the plane of the sky. It is computed according to the following formula:

$$\mathrm{Vtan} = 4.74 \times \frac{\sqrt{\mathrm{pmRA}^2 + \mathrm{pmDE}^2}}{\mathrm{plx}} \tag{1}$$

We compute the tangential velocity for each star in our cleaned dataset. We then store these values in a new column labelled 'Vtan'. Stars with very small (or negative) parallax values tend to have large uncertainties in distance, and their tangential velocities according to the above formula become extreme. We therefore exclude stars with parallax < 0.01 from our analysis. Furthermore, tangential velocity values $>$ 1000km/s are likely due to bad parallax measurements or data artifacts, so we also exclude these.

We will investigate whether the inclusion of tangential velocity improves the performance of clustering algorithms in the following sections. We note, however, that because 'Vtan' is computed from 'pmRA', 'pmDE', and 'plx', inclusion of this feature may instead create unnecessary redundancy in the data, and may also overweight the effects of one or more of its component features due to their intrinsic correlations. We discuss this further in Section 4.

3.2.2 Positional Encoding

As we will be using a variety of clustering models which rely heavily on distance metrics, it is important to represent stellar positions in a suitable format. Right ascension ('RA_ICRS') and declination ('DE_ICRD') are spherical coordinates which can lead to distorted distance measures and discontinuities (right ascension wraps around at 0/360°). A better approach is to combine these two features into three-dimensional Cartesian coordinates (x,y,z) on the unit sphere. This ensures that angular distances are preserved, while enabling clustering algorithms to operate effectively using metrics such as Euclidean distance. In particular, we create three new columns in our dataset called 'x', 'y', and 'z', and remove 'RA_ICRS' and 'DE_ICRD'. The new columns are computed as follows:

$$\begin{aligned} \mathrm{x} &= \cos(\mathrm{DE}) \cdot \cos(\mathrm{RA}) \\ \mathrm{y} &= \cos(\mathrm{DE}) \cdot \sin(\mathrm{RA}) \\ \mathrm{z} &= \sin(\mathrm{DE}) \end{aligned} \tag{2}$$

where it is noted that columns 'RA_ICRS' and 'DE_ICRD' must first be converted from degrees into radians before performing the coordinate transformation above.

3.2.3 Skewed Distributions

It is often the case that the performance of clustering algorithms such as K-Means improves when highly skewed features are transformed via a logarithmic or Box-Cox transformation. Such transformations can help to prevent extreme values from dominating distance-based clustering.

In our case, both parallax ('plx') and tangential velocity ('Vtan') show strongly-tailed distributions with many small values and few very large outliers, suggesting that they may benefit from logarithmic or Box-Cox transformation. Therefore, we create two transformed features, 'boxcox_plx' and 'log_Vtan'. The results of these transformations are illustrated in Figure 1.

However, it is important to recognise that such transformations may also negatively impact clustering performance, and we must exercise caution here. In particular, when the tail of a skewed feature represents a physically meaningful phenomenon, transforming it may hide that phenomenon from the clustering algorithm. In our case, it is possible that transforming 'plx' and 'Vtan' will lead to less clean separation between more distant clusters, as they will bunch together in the data. We therefore retain the original skewed features in addition to their transforms, and compare clustering performance in each case in Section 4.

We do not transform any other features in our dataset, despite some not being normally distributed. In particular, we do not expect sky positions ('x', 'y', 'z') to be normally distributed, and transforming these columns individually would destroy their combined geometric meaning. Furthermore, because the Gaia G-band magnitude ('Gmag') is already logarithmic in flux by definition, an additional logarithmic transform would distort its physical meaning, so we do not perform any transformation here. Finally, we find that 'pmRA', 'pmDE', and 'BP-RP' are already approximately normally distributed, so we do not transform these features.

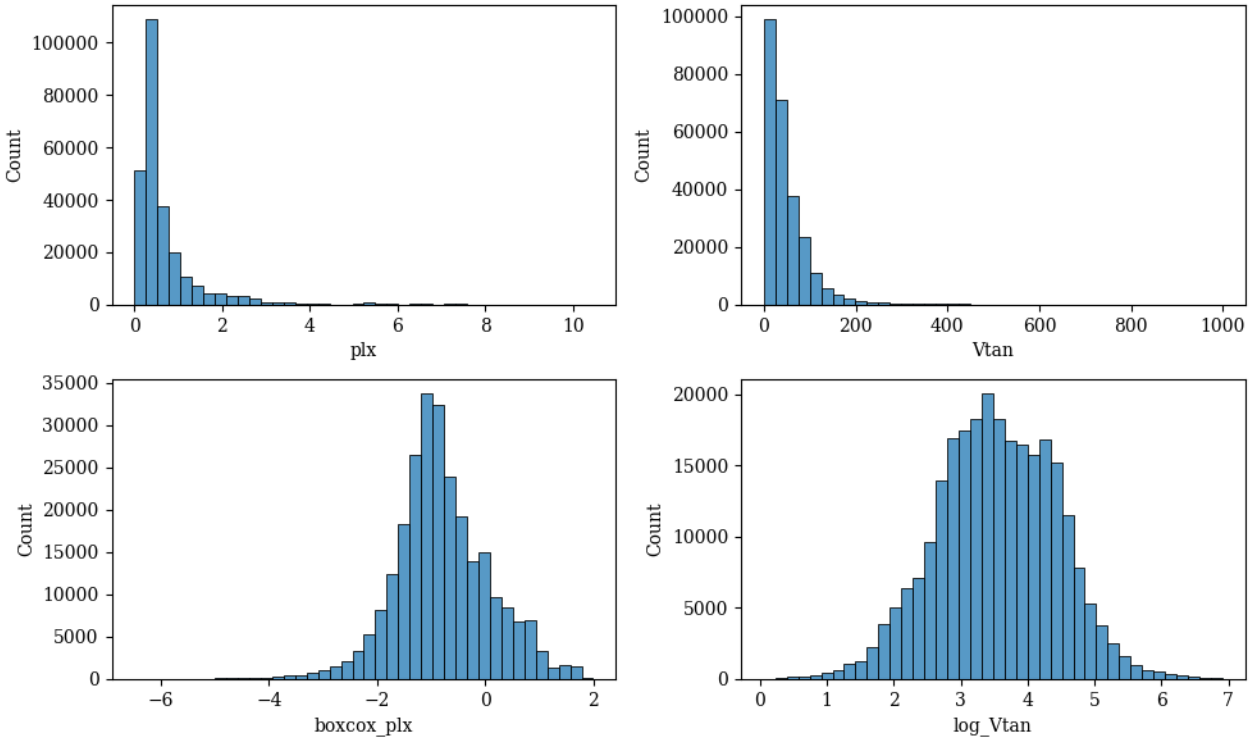


Figure 1: Distributions of parallax and tangential velocity before (top) and after (bottom) transformations to correct skewness.

3.2.4 Feature Exclusions

We note that inclusion of features 'Gmag' and 'BP-RP' are likely to worsen clustering performance. Indeed, while G-band magnitude is, overall, an informative predictor of OC distance, this information is already largely encoded in the parallax, and inclusion of 'Gmag' therefore introduces redundancy. Furthermore, while OCs typically exhibit a coherent colour-magnitude sequence, stars in different parts of the sequence can be far apart in ('Gmag', 'BP–RP') space (Hu & Zhou, 2022). Therefore, inclusion of these features may lead to clustering algorithms splitting along cluster isochrones rather than keeping the entire cluster together. We investigate this possibility in Section 4.

3.3 Summary of Cleaned Dataset

After performing our data cleaning and feature engineering steps, the resulting dataset contains 257288 rows. While we do not use the 'PMemb', 'Cluster' or 'SimbadName' columns from the original data for clustering, we retain these in a separate dataset for further analysis. In particular, we use the 'Cluster' column to ensure that our reduced dataset still provides good coverage across the clusters identified in CG18. We illustrate the distribution of our cleaned data set among those clusters in Figure 2. In particular, we find that we retain data for all 1229 of the previously identified clusters, each with > 5 members. This coverage will enable us to effectively compare our own clustering results with the previous cluster assignments.

The features we use for our clustering are described in Table 2. Note that we will experiment with various combinations of these features, and will discuss which feature combinations give

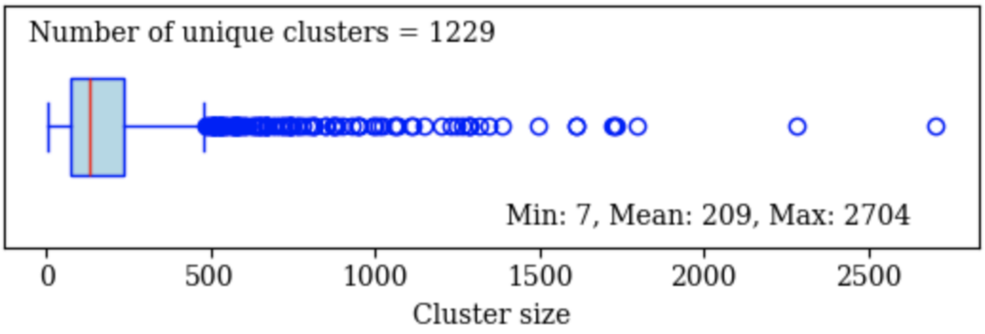
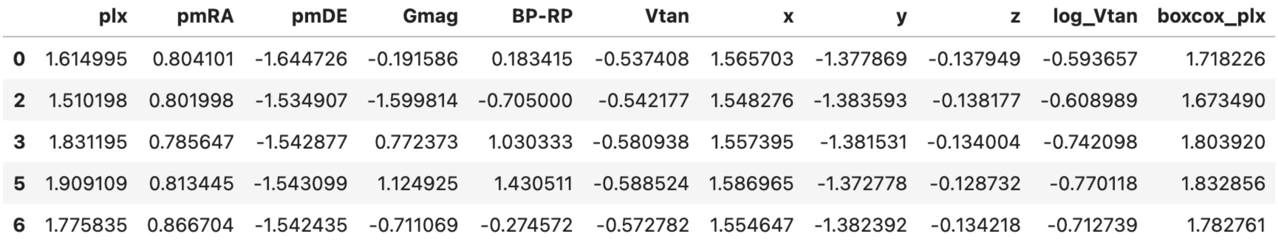


Figure 2: Distribution of our subset of data across original clusters

the best results in section 4. All features are float types, so we do not need to perform any one-hot encoding of categorical features. However, these features exist across a broad range of scales, so feature scaling is an important preprocessing step. All features listed are scaled using Scikit-Learn StandardScaler. This ensures that all features are given equal weight in distance-based clustering algorithms. We illustrate the first few rows of the dataset after all preprocessing steps have been completed in Figure 3. Note that we have retained the index from the original dataset so that we can compare our results to the clusters found in CG18.

Feature	**Description**
`plx, boxcox_plx`	Separates stars by distance
`pmRA`	Captures star motion in RA direction
`pmDE`	Captures star motion in Dec direction
`Gmag`	Distinguishes between bright and faint stars
`BP-RP`	Indicates stellar type / temperature
`Vtan, log_Vtan`	Characterises linear velocity
`x, y, z`	Representation of sky position suitable for Euclidean distance metrics

Table 2: Final Feature Set for Clustering of Gaia Data



	plx	pmRA	pmDE	Gmag	BP-RP	Vtan	x	y	z	log_Vtan	boxcox_plx
0	1.614995	0.804101	-1.644726	-0.191586	0.183415	-0.537408	1.565703	-1.377869	-0.137949	-0.593657	1.718226
2	1.510198	0.801998	-1.534907	-1.599814	-0.705000	-0.542177	1.548276	-1.383593	-0.138177	-0.608989	1.673490
3	1.831195	0.785647	-1.542877	0.772373	1.030333	-0.580938	1.557395	-1.381531	-0.134004	-0.742098	1.803920
5	1.909109	0.813445	-1.543099	1.124925	1.430511	-0.588524	1.586965	-1.372778	-0.128732	-0.770118	1.832856
6	1.775835	0.866704	-1.542435	-0.711069	-0.274572	-0.572782	1.554647	-1.382392	-0.134218	-0.712739	1.782761

Figure 3: First few rows of the preprocessed scaled data which will be used for clustering. We will experiment with several different feature combinations from this dataset.

4 Models and Results

In this section we use a variety of unsupervised clustering methods to separate the Gaia data into individual OCs. As a guide to measure performance, we compare our results to the cluster assignments found in CG18 via the following procedure:

1. Use the specified clustering method to produce a dictionary of clusters in which keys are cluster IDs and values are lists of indices of stars assigned to each cluster.
2. For each star, find its associated cluster in CG18 using its unique index. Thus build a separate dictionary of clusters for comparison.
3. Find the best mapping between our own clusters and those of CG18 by computing the Jaccard distance between each pair of clusters, and finding the 1:1 mapping which minimises total Jaccard distance using scipy.optimize.linear_sum_assignment.
4. Once the optimal cluster mapping is found, compute the proportion of stars whose cluster assignment according to CG18 is the same as our own.

While we use the above procedure as a broad performance indicator it is important that the previous assignments are not interpreted as undisputed 'ground truth' labels. Indeed, while many of the OCs identified in CG18 are known astrophysical objects, catalogues of individual stars belonging to these objects are far from complete. Furthermore, the authors identify 60 new objects whose status have not been verified. We describe the results of three different clustering methods below.

4.1 K-Means Clustering

The authors of CG18 utilise a K-means based method to identify OCs. In particular, they identify 1229 unique clusters. They use only 'pmRA', 'pmDE' and 'plx' to perform the initial clustering, and then check whether the distribution of stars in each identified cluster is more concentrated than a random distribution to give a binary yes/no as to whether this is a true OC.

We adopt a slightly different approach, using a K-means classifier alone without a separate binary separation phase. We use 1229 as our target cluster number for our own K-means classifier. We experiment using different feature combinations to obtain the best classifier. The results are described in Table 3.

Feature combination	**Clustering similarity**
`[pmRA, pmDE, log_Vtan, x, y, z, boxcox_plx, Gmag, BP-RP]`	23%
`[pmRA, pmDE, Vtan, x, y, z, plx, Gmag, BP-RP]`	23%
`[pmRA, pmDE, Vtan, x, y, z, plx]`	54%
`[pmRA, pmDE, plx]`	38%
`[pmRA, pmDE, boxcox_plx]`	33%
`[pmRA, pmDE, plx, x, y, z]`	**62%**
`[pmRA, pmDE, boxcox_plx, x, y, z]`	51%

Table 3: Similarity of various different K-means clustering outcomes with the clusters identified in CG18. Different feature combinations are chosen in each run.

Table 3 offers a number of valuable insights. As anticipated, we see that transforming parallax using Box-Cox does not improve clustering performance, and instead leads to decreased performance. This is likely due to the decreased separability of distant clusters in the transformed data. We also see that inclusion of the tangential velocity feature tends to decrease performance, likely due to the intrinsic correlations between tangential velocity and parallax. Finally, we see that inclusion of G-band magnitude and BP-RP dramatically reduces clustering performance. This is because OCs tend to display coherent colour-magnitude sequences, however individual stars within a cluster may lie far from one another in this 2-dimensional subspace. Overall, therefore, clustering performance is best for the following feature combination: 'pmRA', 'pmDE', 'plx', 'x', 'y', 'z'.

While 'Gmag' and 'BP-RP' should not be used directly during the clustering process, they are nevertheless valuable as a separate means to assess cluster validity. In particular, we may use these features to produce colour-magnitude diagrams (CMDs) for the clusters we identify, and compare these with known properties of OCs. In Figure 4, we illustrate the CMDs for two sample clusters[1], for which our results are in close agreement with those of CG18. Both clusters demonstrate the characteristic isochrone of an OC. This is an indication that the clustering process is robust.

In Figure 5, we present CMDs for two moderately populated clusters. While there is broad agreement between our results and those of CG18, our cluster assignments appear to recover a larger portion of the characteristic open-cluster isochrone, particularly at brighter magnitudes. This may indicate that the CG18 memberships are incomplete, although further analysis is required to confirm this.

Finally, in Figure 6, we illustrate CMDs for two smaller OCs which exhibit significant mismatch between our results and CG18. Given that we map our clusters to those of CG18 by minimising the global Jaccard distance between pairs of clusters, it is possible that some pairings, particularly for smaller clusters, are incorrect. Further investigation into the differences between small clusters is left for future work.

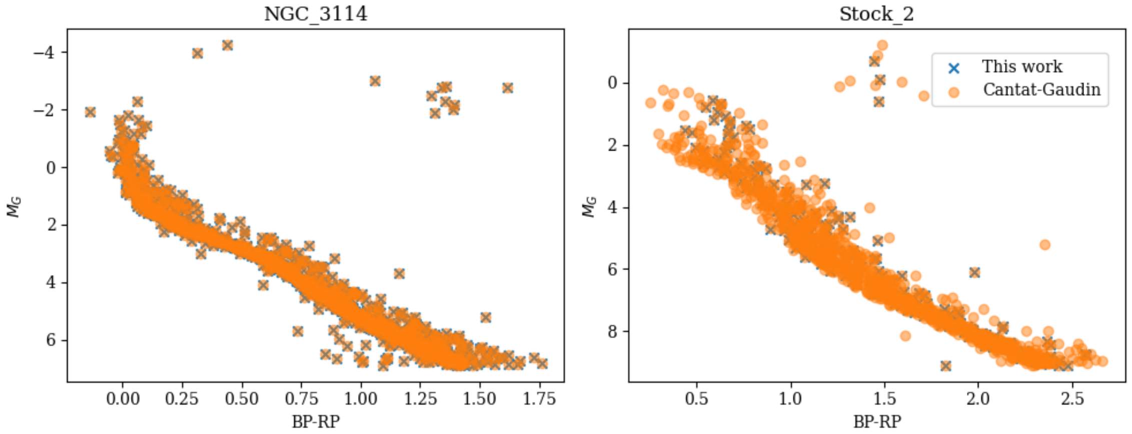


Figure 4: CMDs for two large OCs. Our results are in close agreement with those of CG18, and both CMDs demonstrate the characteristic isochrone of an OC.

4.2 DBSCAN Clustering

In addition to K-means clustering, we also apply a DBSCAN clustering pipeline to the data. We use the optimal feature set identified in Table 3 to train our model, and choose model hyperparameters in order to produce a total number of clusters similar in magnitude to the fiducial value of 1229. Using $\epsilon = 0.085$ and a minimum sample value of 3 for core points, we identify 1144 individual clusters in the data, with 9632 points (3.7%) being classified as outliers. We find an overall cluster assignment similarity of 51% as compared to CG18. While it is possible that hyperparameters could be further tuned to obtain closer agreement, we prefer the K-means approach due to the direct comparability with the clustering pipeline utilised in CG18.

[1] Note that we do not use Gaia A_G values to correct for interstellar extinction in producing these plots.

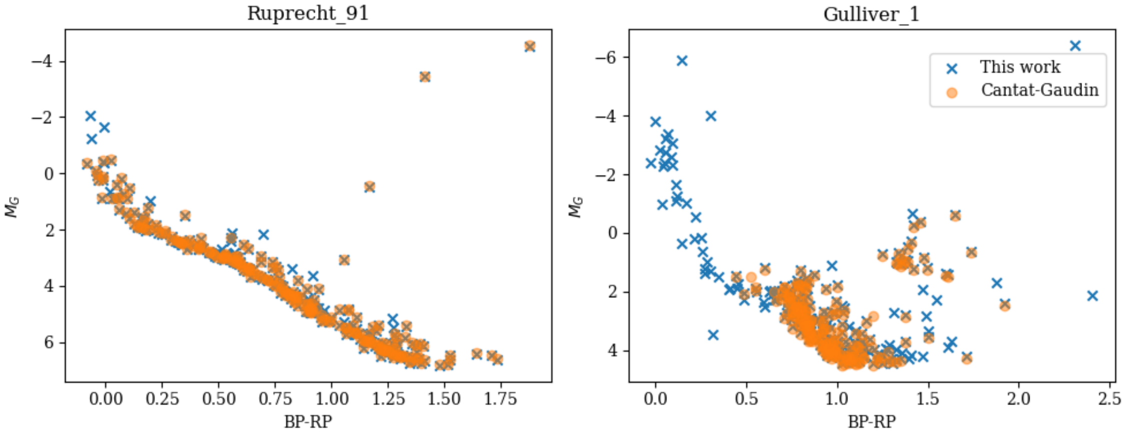


Figure 5: CMDs for two medium sized OCs. Our cluster assignments include stars spanning a broader section of the isochrone than those of CG18.

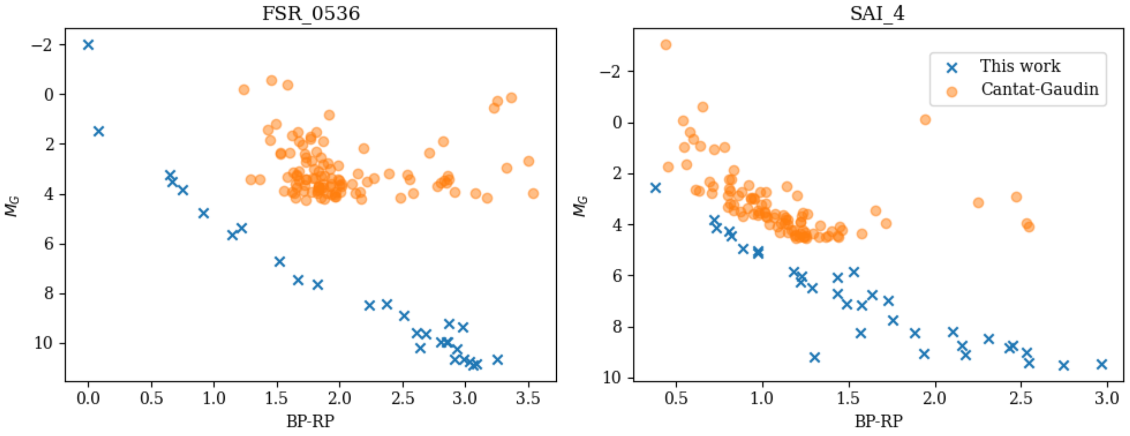


Figure 6: CMDs for two smaller OCs. Each OC has been matched with a cluster in our dataset using linear sum assignment of the Jaccard similarity matrix. The disparity between these cluster assignments may indicate an incorrect cluster mapping.

4.3 Mean Shift Clustering

Finally, we apply a mean shift clustering model to the data. Again, we use the optimal feature set identified in Table 3 to train our model. We estimate the bandwidth for the mean shift model using a quantile value of 0.00075, yielding a bandwidth of 0.184. The mean shift algorithm initialised with this bandwidth finds 1217 separate clusters, and we find an overall cluster assignment similarity of 63% when compared to CG18.

Thus, the performance of the Mean Shift algorithm appears comparable to the K-means algorithm when assessed based on similarity to CG18. However, the results of our K-means clustering pipeline and our mean shift clustering pipeline are only 49% similar. This indicates that while there is a stable core of easily clustered points, there is also a significant number of borderline points which different clustering algorithms handle differently.

5 Key Findings and Insights

Overall, we find that both the K-means and mean shift approaches yield clusters with reasonable ($>60\%$) similarity to the OCs identified in CG18. The performance of DBSCAN is somewhat lower in this regard, but is still in agreement with ($> 50\%$) of the CG18 classifications. While the results of CG18 are a useful benchmark for model performance, they are not themselves 'ground truth' labels. Indeed, it is important to note that CG18 includes a membership probability column ('PMemb'), which indicates the confidence associated with each cluster assignment. The mean value of 'PMemb' across the entire CG18 dataset is 0.65. Therefore, the differences between our clustering results and those of CG18 are unsurprising.

It is reasonable to expect that our cluster assignments are more likely to agree with those of CG18 when the latter have higher membership probabilities. Indeed, we find that for stars whose assignments are the same between our K-means model and CG18, the average value of 'PMemb' is 0.69. By contrast, for stars whose cluster assignments differ, the average value of 'PMemb' is significantly lower, at 0.45.

Discrepancies between our cluster assignments and those of CG18, particularly in the case of smaller clusters, serve as a useful guide for follow-up studies. Meanwhile, the broad agreement between our results and CG18 in more densely populated OCs serves as a confirmation of membership for these OCs, and allows for greater confidence in the estimation of their global parameters from measured properties of member stars.

6 Next Steps

There are many avenues for extension and improvement of the current work. First, we note that while we have employed three different clustering techniques here, there are other techniques such as hierarchical agglomerative clustering and Gaussian mixture models which we have yet to explore. It may be informative to apply a wide variety of clustering techniques to identify subsets of data which are sensitive to the underlying algorithm, and focus follow-up observations here. For stars whose cluster assignment is less clear, it may also be useful to use an ensemble voting approach to determine the most likely assignment.

Another avenue for extension of the current work is to explore how the clustering assignments may change when we use more recent observational data. In this work we have used Gaia's second data release (DR2). We have done this to ensure consistency with CG18. However, the Gaia Collaboration has since made more data available through a third data release (DR3). This more recent data is likely to provide improved estimates of features such as G-band magnitude and parallax due to improved calibrations and longer time baselines, and may help to resolve some of the inconsistencies between our results and those of CG18.

We would also like to study how clustering results change when the K-means algorithm is initialised with a different number of target clusters. We have used a fiducial cluster number of 1229 in this work in order to provide a direct comparison with CG18. However, it is possible that some of the putative OCs identified in CG18 are spurious, and follow-up studies are required to verify their existence and properties. Studies of these candidate OCs may involve applying clustering algorithms to the relevant subset of the current data to discern whether these clusters are indeed cleanly separable.

Finally, we would like to compare our findings to the results of other studies, such as (Dias et al., 2002; Kharchenko et al., 2013). Many of the objects identified in these studies are putative OC's requiring further investigation, and our results may help to confirm or challenge these discoveries.

References

Cantat-Gaudin, T., Jordi, C., Vallenari, A., et al. 2018, , 618, A93, doi: 10.1051/0004-6361/201833476

Dias, W. S., Alessi, B. S., Moitinho, A., & Lépine, J. R. D. 2002, , 389, 871, doi: 10.1051/0004-6361:20020668

Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, , 595, A1, doi: 10.1051/0004-6361/201629272

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, , 616, A1, doi: 10.1051/0004-6361/201833051

Hu, K., & Zhou, Q. 2022, in Journal of Physics Conference Series, Vol. 2386, Journal of Physics Conference Series (IOP), 012069, doi: 10.1088/1742-6596/2386/1/012069

Kharchenko, N. V., Piskunov, A. E., Schilbach, E., Röser, S., & Scholz, R. D. 2013, , 558, A53, doi: 10.1051/0004-6361/201322302

Ochsenbein, F., Bauer, P., & Marcout, J. 2000, , 143, 23, doi: 10.1051/aas:2000169

Ratzenböck, S., Großschedl, J. E., Möller, T., et al. 2023, , 677, A59, doi: 10.1051/0004-6361/202243690