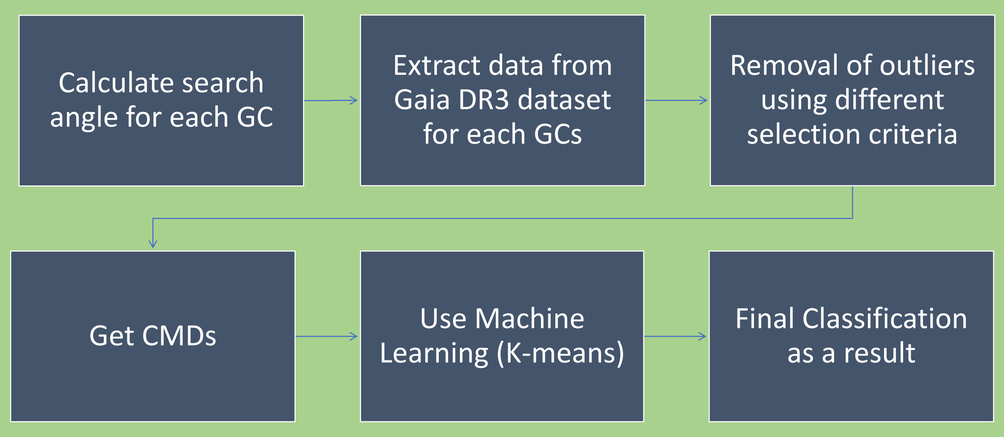
Classify galactic globular clusters based on color-magnitude diagram using an unsupervised machine learning

Bishnu Kumar Sharma & Chow-Choong Ngeow
Graduate Institute of Astronomy, National Central University (Taiwan)

Abstract: We use Gaia DR3 data to construct color-magnitude diagrams (CMDs) for the Galactic globular clusters (GCs). Our aim is to classify the GCs based on CMD morphology, such as main-sequence branch, red giant branch, and horizontal branch, using a pure machine learning (ML) approach, as well as learning the properties of GCs based on ML classification. As a first step, we get CMDs, using the information on position, proper motion, photometry, extracted using the VizieR Queries (astroquery.vizier). We used K-means clustering to find the nearest similar CMD images.

Introduction: We use five parametric astrometric data i.e. position (RA, Dec), proper motion (pmRA, pmDE), parallax and photometry ($G_{BP} - G_{RP}$, G_{mag}) from Gaia DR3(1) and construct CMDs and classify them using Unsupervised ML. We use Convolutional Neural Network(CNN) tensorflow.keras and clustering algorithm (K-means clustering(5)) for image classifications. K-means clustering classifies n images into k number of classes in which each observation belongs to cluster based on nearest centroid.

Flow Chart:



Method:

- We arbitrarily use 2 times half light radius (2) as a search region,
- Simple trigonometry to find the search angle,
- Python module of astroquery.vizier to extract data from Gaia archive,
- Different selection cuts like negative parallax are rejected(3), mean of the positive parallax is chosen, Re-normalize Unit Weight Error(RUWE) < 1.4 (4), epsi <= 2.
- Half mean of Proper motion (PM) is chosen as a distance to select members from the center to the stars (mean pmRA and mean pmDE are chosen as center of cluster).
- We didn't clean CMDs below Main-sequence turnoff as removing those stars may alter our results while performing k-means clustering.
- Scikit learn library, CNN, tensorflow.keras and K-means clustering are used for image classification.

Assumptions: We use CMD images with no scale and labels for our ML model. We assumed that labels and scale may effect our classification results.

Elbow Plot : The elbow plot is the graphical representation of finding possible number of "K" in K-means clustering. It finds the sum of the square of distance between points in a cluster and the cluster centroid.

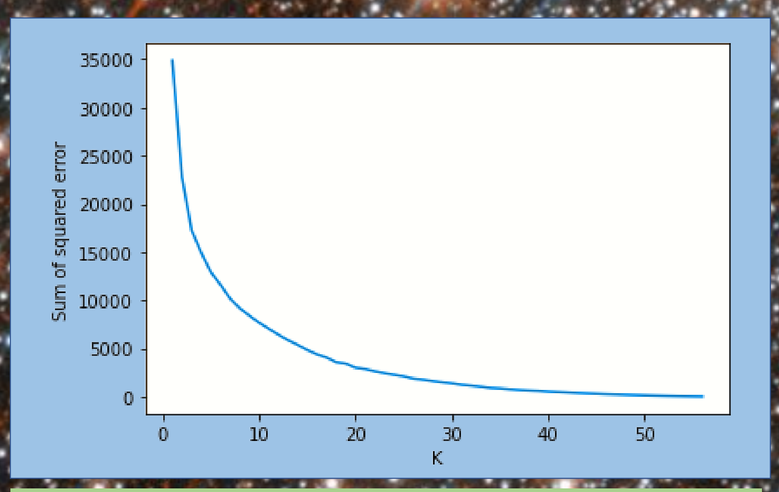


Fig1: An Elbow plot.

Some of the similar CMDs we present here with proper scaling after getting results from ML model.

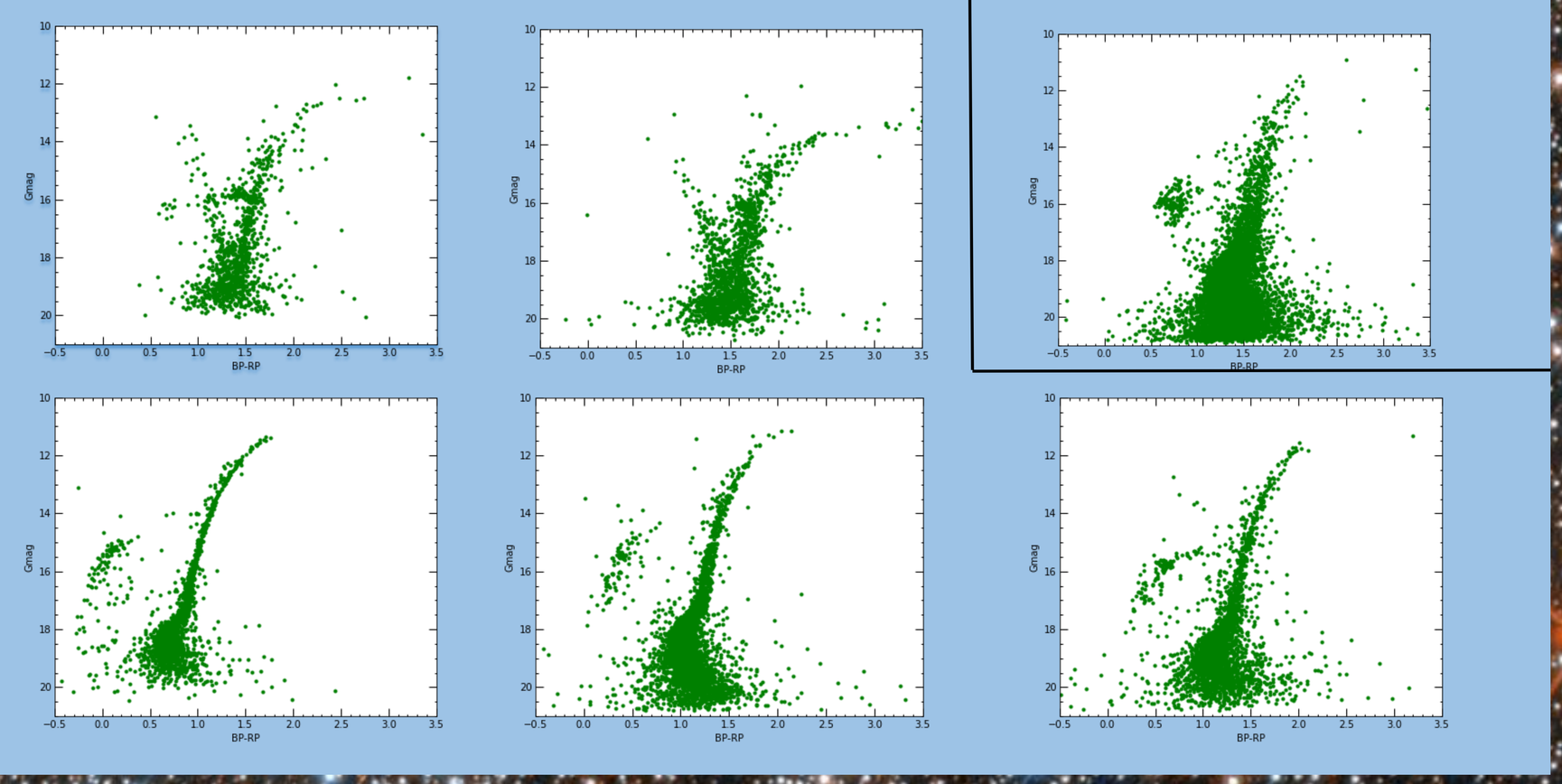


Fig3: Two CMDs from top left corner (NGC 6712, NGC 5927) are in a group, while lower 3 CMDs(NGC 6205, NGC 6254, NGC 4833) are in a group and a CMD at top right corner (NGC 4833) is CMD group with only one member.

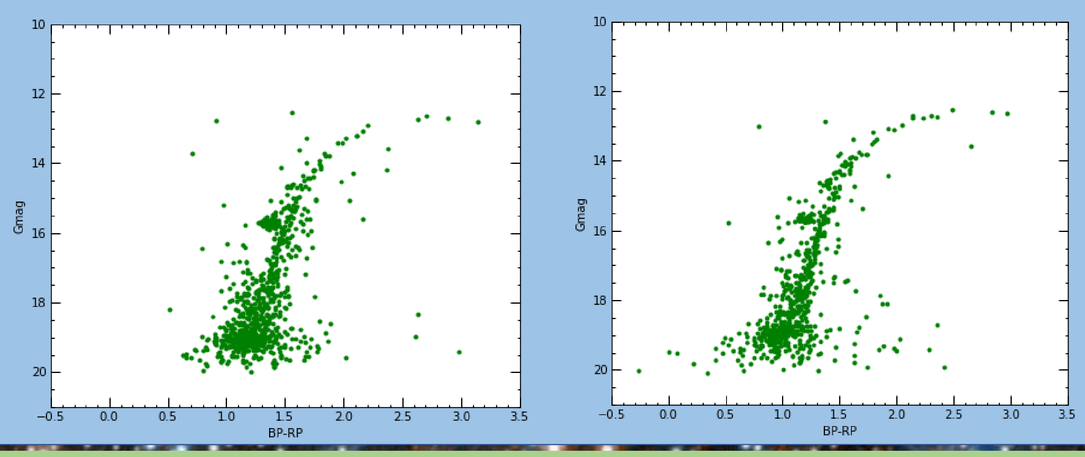


Fig2: NGC 6637 and NGC 6624 are in same group

Results and discussions: We used elbow method to get K-number of clusters. Elbow plots seems to flatten at around 28. Accordingly, we get 28 different clusters out 57 CMDs.

Conclusion: We used K-means clustering for image similarity model. We got 28 groups from our image datasets. A group may contains either single or multiple CMDs .

Future work: We are planning to study their properties like mass, age, metallicity etc. We want to know similarities and the differences between GCs of the same group. Also, we will study what similar CMD morphology reveals about the clusters in a group.

References:

1.Gaia Collaboration: A. Vallenari, A.G.A. Brown et. al. 2022, A&A
2.https://people.smp.uq.edu.au/HolgerBaumgardt/globular/parameter.html
3.X. luri, A.G. A Brown, L.M. Sarro et. al. 2018, Astronomy and Astrophysics
4.Gaia Collobaration, T. Antoja, P.J. McMillan, G. Kordopatis et. al. 2021, A&A
5. Hartigan, J.A. & Wong, M.A(1979) A k-means clustering algorithm. JSTOR: Applied Statistics, 28, 100—108.