CS156: Final Project

Open Stellar Cluster Membership Detection

Minerva Schools at KGI

Introduction:

Open stellar clusters are a group of stars that formed from the same interstellar cloud. All the members of the cluster are concentrated in the same region in the sky, thus have a similar distance from us. They also have similar motion relative to us, both tangentially (along the sky plane) and radially (motion along the line of sight). They have similar chemical properties and all of them are of similar age.

The earliest member detection method includes primarily to find a compact cluster in the 2D sky plane, but it cannot distinguish the members from the foreground and background stars. Later, proper motion and parallax (inversely proportional to distance) are also being used to detect the member. GAIA space mission (Gaia Collaboration, 2020) is the latest mission to measure the position, proper motion, and parallax with very high precision.

Problem Definition

Recently Cantat et al (2020) used GAIA data to detect the members of the open stellar clusters of our galaxy and assign a membership probability to them. They used an unsupervised clustering method called UPMASK (Krone-Martins & Moitinho, 2013), which is primarily based on the k-means clustering algorithm to detect the cluster members and used random sampling to assign the membership probability.

As a follow-up of their paper, I used a supervised clustering model to find any additional members for the open cluster named “pearl cluster” or NGC 3766. As my training set, I used their detected members for 3766 and a random set of non-member field stars taken outside of the cluster radius.

The radius of a specific open cluster is reported quite differently in different literature. The primary reason behind this is the lack of precise data. Earlier observations detected only a few members confidently, thus reported a smaller radius, while later more improved observation found reliable surrounding members, thus a larger radius. Cantat et. al (2018) reported a higher radius for many clusters using the precise GAIA data. So, I took a slightly larger radius than the one Cantat detected as the field of view of my target stars to check if there are any more stars lying in the surrounding who shares the same properties as the identified members.

Solution Specification

Training data

The GAIA (Gaia Collaboration, 2020) and Cantat et al (2020) both datasets are taken from the Vizier Astronomical Database (Ochsenbein, 1996) using open-source Astropy and Astroquery packages of python. Any data with pmra_over_error, pmdec_over_error, or parallax_over_error smaller than 3 are rejected to ensure more precise data. The member stars in the training data are the 1345 stars of NGC 3766 identified by Cantat et al. (2020). The most distant identified member star is around 0.32 degrees away from the center.

To select non-member, I chose a ring region around the center with an inner and outer radius of 0.7 and 0.8 degrees. I took a random sample of an equal number of stars (1345) from this region to balance my training data. The final training dataset has 2690 stars: 1345 members, 1345 non-member (see Appendix Fig 1).

Selected Features

Similar to Cantat et al. (2018; 2020) I used the following three features in my model:

1. pmra: proper motion along the right ascension direction of the sky plane
2. pmdec: proper motion along the declination direction of the sky plane
3. parallax: the shifted angle of the star as the earth moves around the sun, which is used to measure the distance (inversely proportional to distance)

All the members should be clumped together in the 3D space of pmra, pmdec and parallax, while all the non_members should be outside of this clump (See Appendix Fig 2 for their distribution on pmra-pmdec). For an open cluster, of course, they also need to stay in the same position in the sky plane (ra and dec). But I don't need to use them as a feature as I am selecting only the stars within a 0.4-degree radius of the cluster center. I didn't use distance from the cluster center, because as my non-members are chosen from an outer ring area, any ML method will use this distance to get a higher accuracy by just setting a cutoff in this distance parameter.

Choice of Metric

The members of the open clusters are assumed to have similar age and similar chemical structure, thus they are further analyzed to study the star dynamics and other useful properties. Having a large number of member helps to make them more reliable analysis, but if we have non-members which has a very different chemical composition and assumed them to be a member, it misleads the further analysis. Thus it is a common practice to apply strict criteria to avoid any non-members being selected as members. As we don't want any false positives as a cluster member, the metric of our choice is the precision for member class (TP/(TP+FP)).

Model Selection

As I discussed earlier, the selected features are well-known to be concentrated in a small area for all the cluster members. In 3D space of parallax-pmra-pmdec, the members are very well compacted in a spherically or normally distributed 3D region, while the non-members are randomly distributed outside of this compact region.

Let's compare the common supervised classification methods for this problem. Though our members are likely to be separated from the random non-members, as their separation is not linear but spherical, linear SVM or any other LPM might not be very suitable in this case. A suitable choice of the kernel may be necessary if we want to use SVM.

Also as our members can be assumed as normally distributed in 3D space, another potential candidate could be the simple Naive Bayes model using Gaussian distribution. Another candidate is kNN, but kNN can overfit, and as they used euclidian distance, they would give lower weight to the features with lower values (here parallax).

One of the most suitable models can be the decision tree method, as we have a smaller number of features and they are quite separated in 3D space from any non-members. But using a single tree increases the chance of overfitting and less stable model. A better choice would be to use the random forest, as it can decrease the bias by increasing the variance using a large number of trees. The random forest also allows us to assign the membership probability for a given star using the percentage of trees that classified it as a member.

Testing and Analysis

In order to select the best possible model, I run all of these 5 candidate models and calculate their precision score in the test data (Table 1). SVM results in the poorest precision as expected. As the training and test precision is similar for SVM, there is no overfitting. I used cross-validation to finetune the parameters of the model, whenever necessary. As the random forest gave the highest precision in test data, we process with random forest for this problem.

Model	Precision in Training Set	Precision in Test Set
RBF SVM	0.751	0.762
Naive Bayes	0.879	0.896
kNN	1.00	0.930
Decision Tree	0.937	0.942
Random Forest	0.946	0.968

Table 1: Precision score for all 5 candidate models in the training and test set.

Predictions

As my target set, I took all the stars from the GAIA data that are less than 0.4 degrees from the cluster radius. Then I subtract the identified member stars that are already in my training data. Running this over 21574 target stars, the model predicted another 828 members for NGC 3766 open cluster. The new members closely follow the old members in proper motions and parallax spaces (see Appendix Fig. 3 and 4 for their distribution).

Discussion

The majority of the new members lie on the outer edge of the cluster (Appendix Fig 5), which suggested that the radius of the cluster might be bigger than we thought earlier. But to support the idea more strongly, we need to test this algorithm for a large number of other clusters as well as analyze if the newly found members share the same chemical compositions as the earlier ones. In proper motion (Appendix Fig 6), along with the central clump, we also found some additional members who have a slightly smaller pmra (around -6). But together all (1345+828 =) 2173 stars showed the same structure in the sky plot and proper motion plot as we have for the member as well as in the distribution of the parallax (Appendix Fig 3 - 6).

Code Link:

https://github.com/minerva-schools/cs156-pcw-mahmud-nobe-1/blob/master/assignments/Final_Project_CS156.ipynb

References

Cantat-Gaudin, T., & Anders, F. (2020). Clusters and mirages: Cataloguing stellar aggregates in the Milky Way. Astronomy and Astrophysics. https://doi.org/10.1051/0004-6361/201936691

Cantat-Gaudin, T., Jordi, C., Vallenari, A., Bragaglia, A., Balaguer-Núñez, L., Soubiran, C., Bossini, D., Moitinho, A., Castro-Ginard, A., Krone-Martins, A., Casamiquela, L., Sordo, R., & Carrera, R. (2018). A Gaia DR2 view of the open cluster population in the Milky Way. Astronomy & Astrophysics, 618, A93. https://doi.org/10.1051/0004-6361/201833476

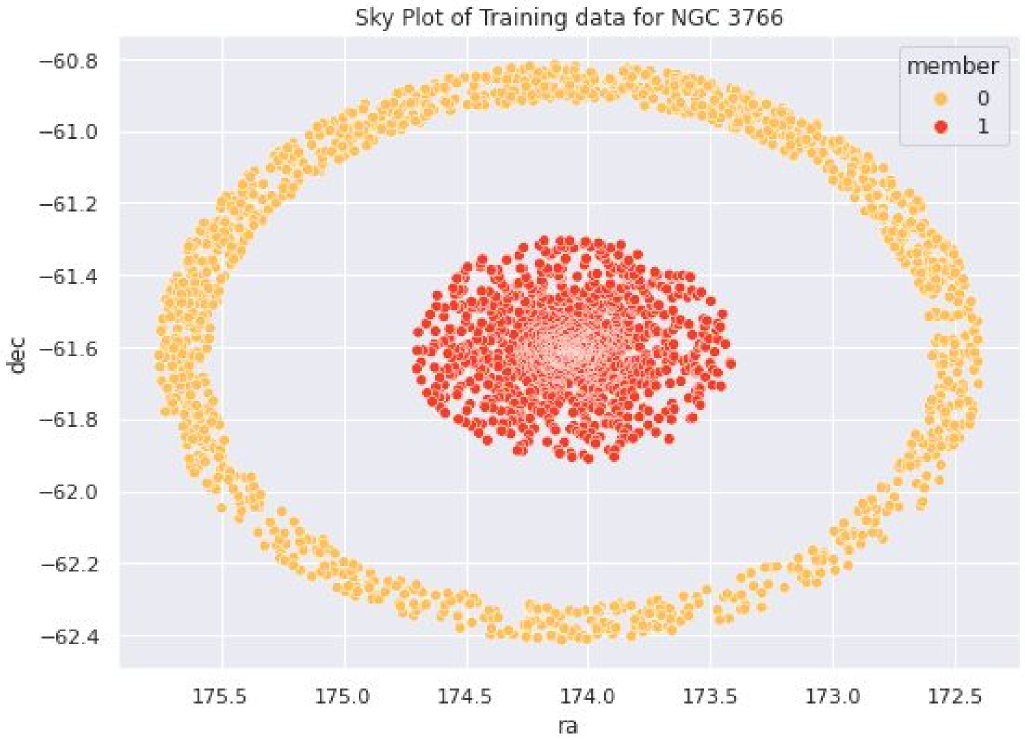
Gaia Collaboration & Brown, A. & Vallenari, A. & Prusti, T. & de Bruijne, J. & Babusiaux, C. & Biermann, M.. (2020). Gaia Early Data Release 3: Summary of the contents and survey properties. https://www.cosmos.esa.int/web/gaia/early-data-release-3

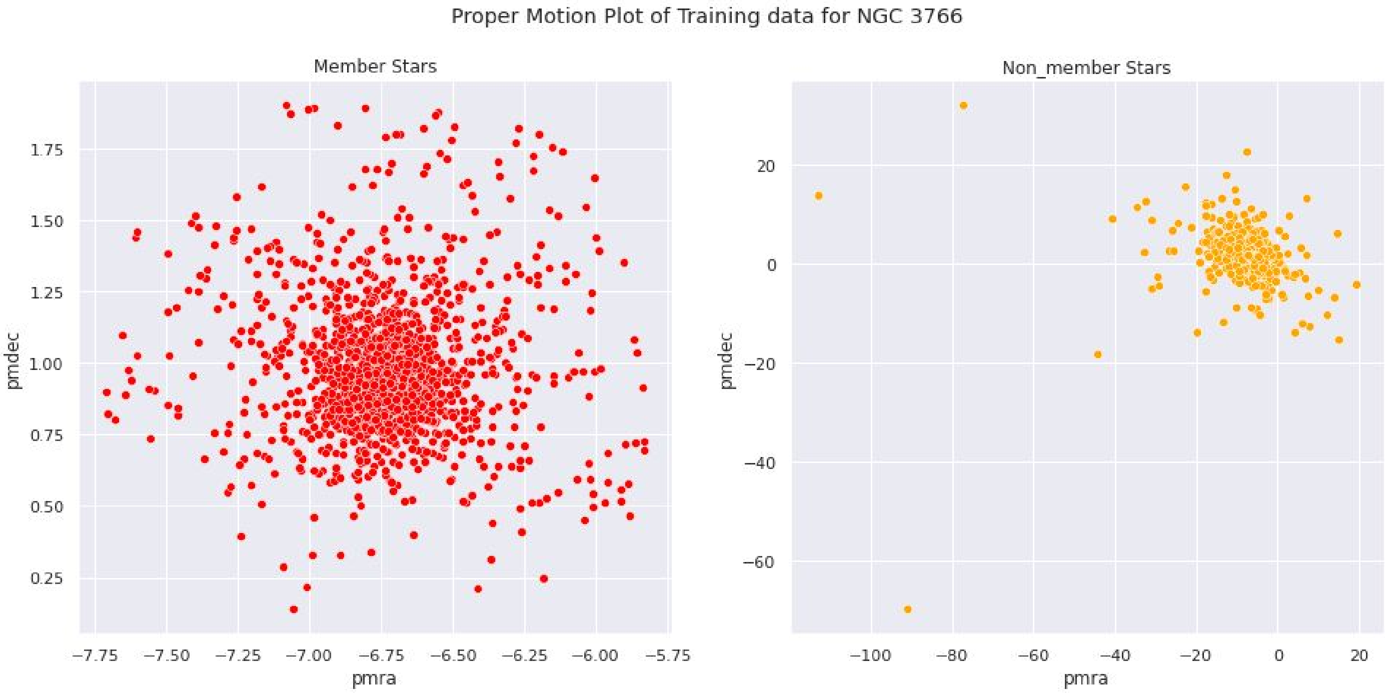
Krone-Martins, A., & Moitinho, A. (2013). UPMASK: unsupervised photometric membership assignment in stellar clusters. Astronomy and Astrophysics. https://doi.org/10.1051/0004-6361/201321143

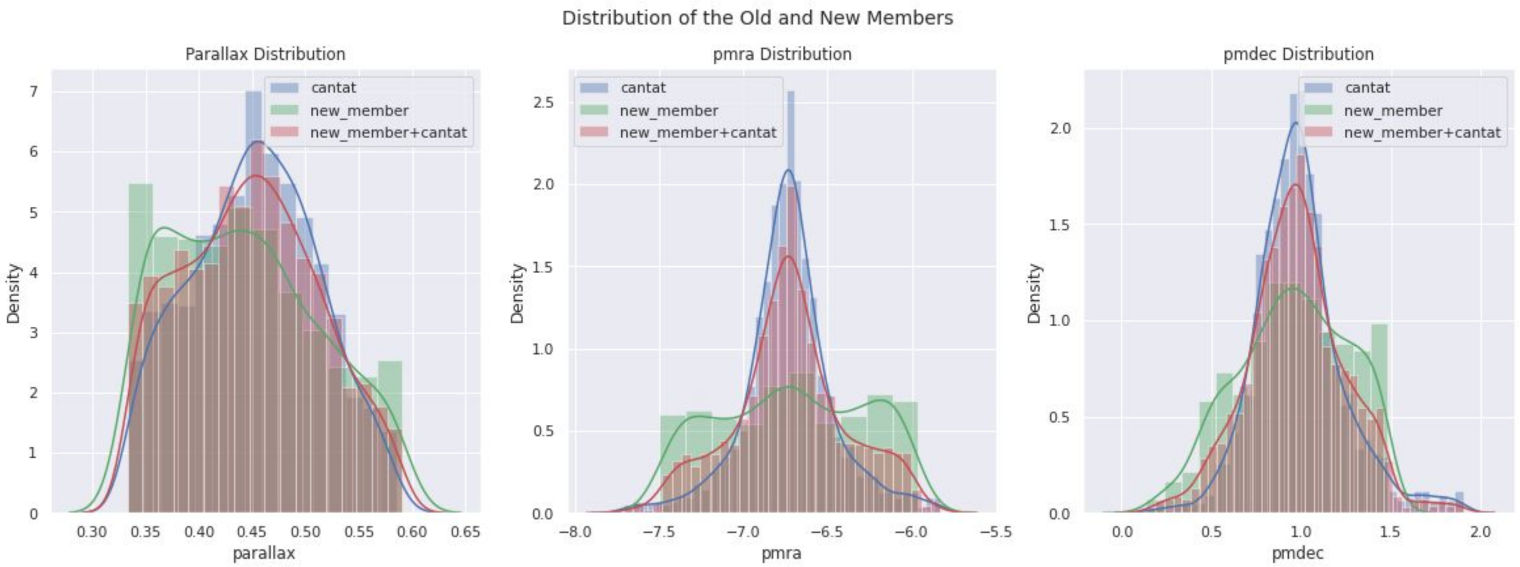
Koehrsen, W. (2018). Hyperparameter Tuning the Random Forest in Python. Retrieved 17 December 2020, from https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

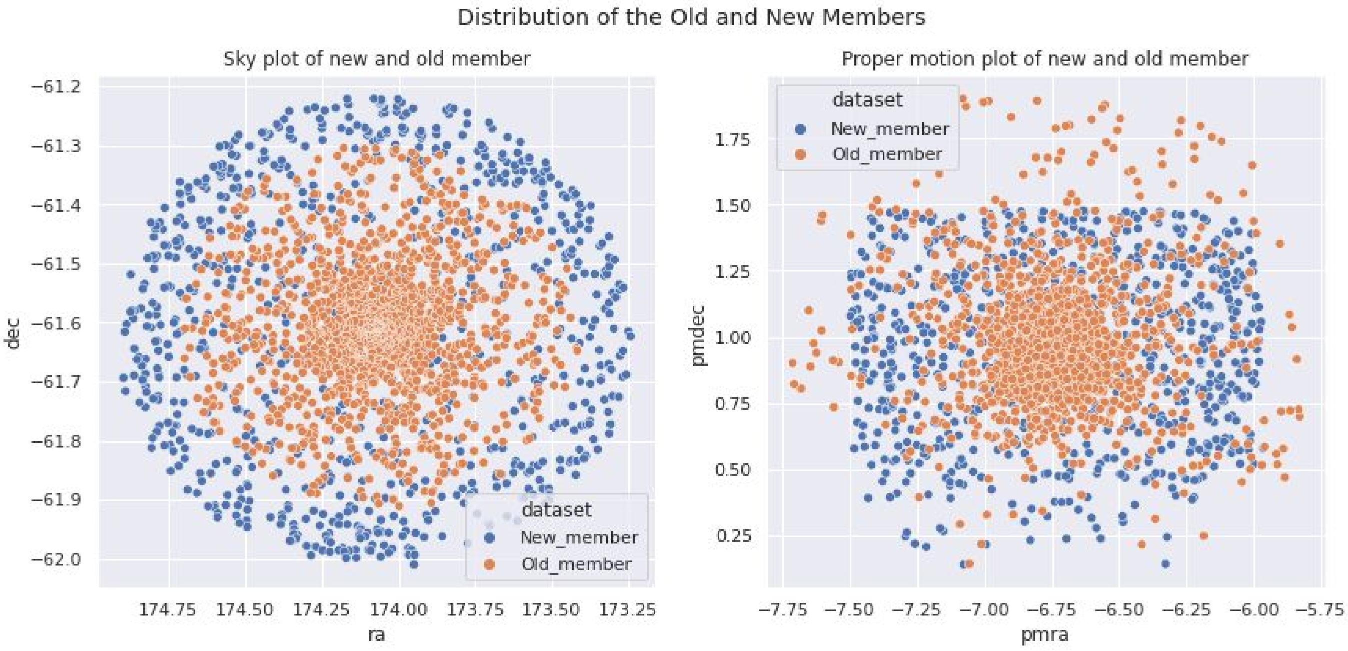
Ochsenbein, F. (1996). The VizieR database of astronomical catalogues. CDS, Centre de Données astronomiques de Strasbourg. https://doi.org/10.26093/CDS/VIZIER

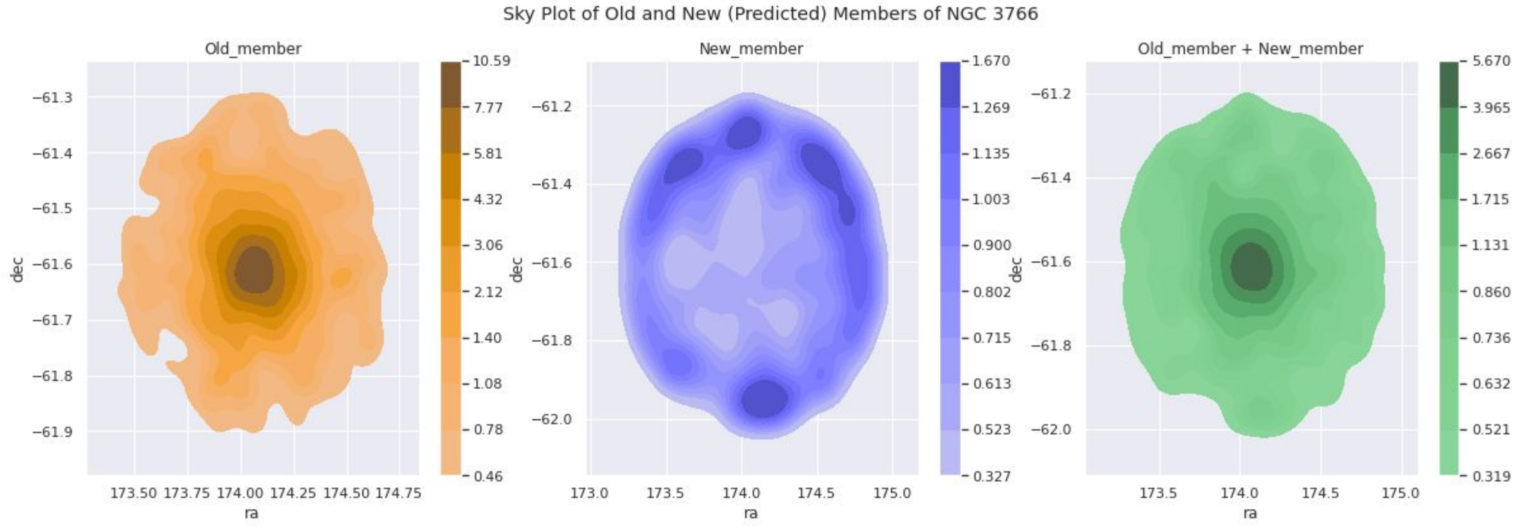
Appendix


Sky Plot of Training data for NGC 3766
member
● 0
● 1
dec
−60.8
−61.0
−61.2
−61.4
−61.6
−61.8
−62.0
−62.2
−62.4
175.5
175.0
174.5
174.0
173.5
173.0
172.5
ra

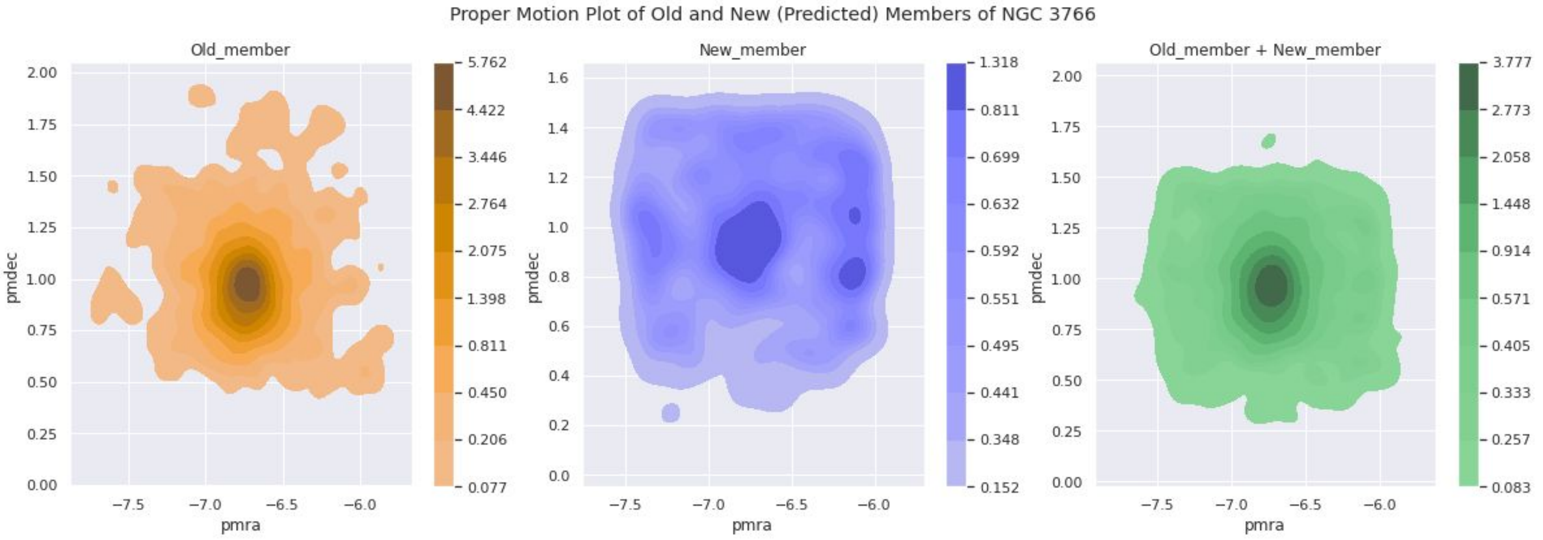

Proper Motion Plot of Training data for NGC 3766
Member Stars
pmdec
1.75
1.50
1.25
1.00
0.75
0.50
0.25
−7.75
−7.50
−7.25
−7.00
−6.75
−6.50
−6.25
−6.00
−5.75
pmra
Non_member Stars
pmdec
20
0
−20
−40
−60
−100
−80
−60
−40
−20
0
20
pmra


Distribution of the Old and New Members
Parallax Distribution
pmra Distribution
pmdec Distribution
cantat
new_member
new_member+cantat
Density
parallax
0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65
0 1 2 3 4 5 6 7
pmra
−8.0 −7.5 −7.0 −6.5 −6.0 −5.5
0.0 0.5 1.0 1.5 2.0 2.5
pmdec
0.0 0.5 1.0 1.5 2.0
0.0 0.5 1.0 1.5 2.0


Distribution of the Old and New Members
Sky plot of new and old member
Proper motion plot of new and old member
dataset
New_member
Old_member
dec
ra
−61.2 −61.3 −61.4 −61.5 −61.6 −61.7 −61.8 −61.9 −62.0
174.75 174.50 174.25 174.00 173.75 173.50 173.25
pmdec
pmra
1.75 1.50 1.25 1.00 0.75 0.50 0.25
−7.75 −7.50 −7.25 −7.00 −6.75 −6.50 −6.25 −6.00 −5.75



Sky Plot of Old and New (Predicted) Members of NGC 3766



Proper Motion Plot of Old and New (Predicted) Members of NGC 3766