Identification of Peculiar Stars in Open Clusters by using Gaussian Mixture Model and Probabilistic Random Forest Methods

Abstract

In this analysis, we have tried to reproduce the result of a paper by Jadhav et al. (2021) to identify interesting stars from a given star cluster by defining their peculiarity. This was implemented using the Gaussian mixture modelling (GMM) and Probabilistic Random Forest (PRF) algorithms. They first classify all the stars in the field of view based on whether or not they are likely to be a part of the cluster (rather than just the field) and then identify the 'interesting' stars on the colour magnitude diagram which do not follow the trend of the rest of the cluster.

1 Introduction

Star clusters are interesting objects to study because they are collections of stars formed from the same gas cloud. As a result, they share several of the characteristics such as metallicity, age, etc. One approach to their study is to take a very global or statistical approach and study the properties of the cluster as a whole. For example the age, metallicity, total mass of a given cluster etc. These can be predicted using very standard diagnostic tools like colour-magnitude diagrams and isochrone fitting.

This sort of analysis will describe the 'trend' for a given cluster; so a natural second question to ask is why some peculiar stars differ from the trend. More detailed analysis of these outliers is in order. But before we can analyse these outliers, we first need to identify them. This may seem trivial at first, but then there are some difficulties we may face. Firstly, we need to identify, with great confidence, whether a star belongs to a given cluster at all, or is just part of the background. Then, we need to quantitatively put some criteria to identify the outliers.

In this guide we describe a method for identifying such outliers largely based on the method used by Jadhav et al. (2021). We will describe the methods used to do this analysis here.

2 Gaussian Mixture Modelling

Gaussian Mixture Modelling is a technique to determine the probability that any particular star in the field of view of the image of a cluster actually belongs to the cluster being studied. We have used this technique to classify the stars in the given field of view based on their proper motions.

To begin with, we assume that the distribution of proper motions of the stars in the cluster is distributed in a relatively symmetric 2D Gaussian distribution around the 'average' proper motion of the cluster. Apart from that the background or 'field' stars will also have some random proper motions, and we assume that these are distributed as a Gaussian distribution as well. As a result, the distribution for the proper motion in RA and proper motion in declination for a random star will look like the weighted sum of the two gaussians.

So now to determine the full distribution, we are looking for the means and covariance matrices for the two gaussians, and their weights (the sum of the two weights should be 1 always, each weight represents the probability that a random star is part of the respective gaussian distribution, i.e. part of the field or the cluster.) These quantities can be obtained as follows - First assume some random reasonable values for these parameters, and calculate the estimator for each quantity. In particular, we start by estimating the membership probability and the field probability. These are to be estimated as the probability that a star belongs to the cluster (field) Gaussian, given that it does belong to at least one of them.

$$p_c(star) = \frac{w_c N_c}{w_c N_c + w_f N_f}$$
$$p_f(star) = \frac{w_f N_f}{w_c N_c + w_f N_f}$$

Where N_c and N_f are the values of the respective Gaussian distributions for the proper motion value of the star with the guessed parameters from before and w_c, w_f are the guessed weights. Then, based on these probabilities, we can estimate the weights (the sum of the probabilities (over all stars) that the star belongs to the corresponding Gaussian, divided by total number of stars), the mean of each distribution (probability weighted mean of proper motion of each star) and the Covariance as the probability weighted covariance over all stars.

$$w_i = \frac{\sum_{j=1}^{N} p_n^j}{N}$$

$$\mu_i = \frac{\sum_{j=1}^{N} p_n^j \mu^j}{\sum p_n^j}$$
$$\Sigma_i = \frac{\sum_{j=1}^{N} p_n^j (\mu^j - \mu_i)^T (\mu^j - \mu_i)}{\sum p_n^i}$$

Here, i is a stand-in for either c/f (both have same formulae). In μ^j and p^j, the j just stands for the index for addition (the j^{th} star).

Then the authors iterate with these calculated estimators as the new guess values and go on to calculate the new values of each estimator. After sufficiently many iterations, this process should converge to a set of parameters which can be used for the final membership probability calculation for all stars. As a crosscheck to see whether the method has given the correct parameters we can plot a histogram of the distribution of PMRA and PMDec of each star which has a membership probability greater than some threshold, about the value of the cluster mean parameters - these should look roughly Gaussian for most clusters (roughly assuming isotropic shape).

3 Probabilistic Random Forest Method

Random Decision Forests are an ensemble learning method used for classification and regression. This can be visualized as a group of trees and based on the features, it traverses in the tree to reach a final decision at the terminal node. The predicted class of the unlabeled object is the class with the highest probability within the terminal node. While training, using a single tree would cause overfitting the training data and its performance might be poor on unseen data. By training on randomly chosen sets of trees from a given forest (set of trees), this can be avoided.

In the case when there is uncertainty in the data, i.e, uncertainty in the features and corresponding labels, which we obtain quite often in the data , the random forest paradigm might not be able to perform well. Let us introduce the probabilistic approach to this, namely Probabilistic Random Forest (PRF), thereby accounting for the uncertainties. The features for PRF can be represented as probability distributions (PDF) with expectancy values that are equal to the given feature values and corresponding variance. The labels will be probability mass functions (PMF) with each label assigned some probability.

We followed the lines of the original authors to obtain the member probability (mp) from the GMM model and train with (1-mp, mp) as the probability distribution for the labels. Then we use PRF to fit the data. We changed the number of trees and Features (F_10, F_8, F_6) as given in the paper Jadhav et al. (2021). We are able to get the accuracy of 95-98% for the PRFs with more than 140 trees. The plots shown below are for the number of trees = 180.

4 Identification of Peculiar Stars

The authors trained the PRF model based on various parameters like astrometry alone which we call as P_{F6} - because there are 6 parameters, namely, (RA,dec,proper motion in RA, proper motion in dec, parallax and PMR0, i.e. how far off the object's proper motion is, from the average proper motion of the cluster), astrometry as well as photometry (P_{F8} - previous 6 parameters, as well as absolute magnitude and color) and all these along with errors in astrometry (P_{F10} - all parameters in P_{F8} along with how reliable the GAIA astrometry data is, according to the GAIA catalog). Here, note that the average proper motion parameters are simply the mean of the Gaussian obtained from Gaussian mixture modelling in PMR0. Then, from the trained model, we can calculate the membership probability for each star.

From here, the authors discard the stars having low probability in P_{F6} as well as P_{F10} - these are most likely just part of the background, and we should not expect them to follow the trends of the cluster. They classify the stars with good probability in P_{F6} but not in P_{F10} as candidates. These are the stars where more accurate astrometry is needed before we can start drawing conclusions. Finally, they consider the stars, which have high probability in both P_{F6} and P_{F10} as members - these are most likely actually part of the cluster.

Next up - the authors go on to define the 'peculiarity' of a star, this is simply the difference between the probabilities obtained from P_{F6} and P_{F8}. Here, we only need to consider the confirmed member stars from the previous analysis, we are obviously only interested in stars which we know are actually in the cluster. P_{F6} predicts on the basis of astrometry, in other words, position, whether a star is in

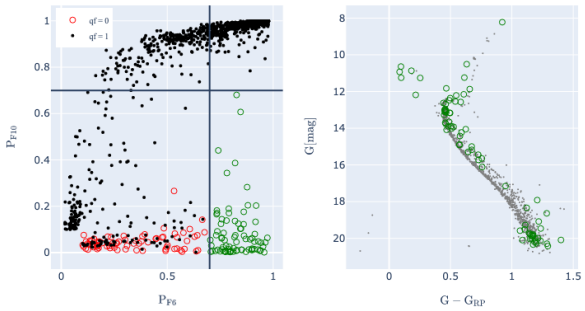

Figure 1
In the left plot, they have compared the P_{F6} and P_{F10} values of stars - allowing them to classify stars as members, candidates or field. The threshold is taken as 0.7. The right plot is a CMD using the members and candidates.

the cluster, whereas P_{F8} also considers photometry, which means P_{F8} also tries to fit the star to the colour-magnitude diagram of the cluster, in a sense. Thus, any star which is actually in the cluster (high P_{F6} score) but not following the colour magnitude diagram trend (comparatively lower P_{F8} score) will have higher values of peculiarity, and can be studied further to study its properties in detail. Again, plotting a colour-magnitude diagram, with the peculiarity represented by a colour code will be helpful to visualise exactly in what way, the star differs from the trend (whether it is a blue straggler or a giant star etc) and can be used to decide what exact studies we want to perform.

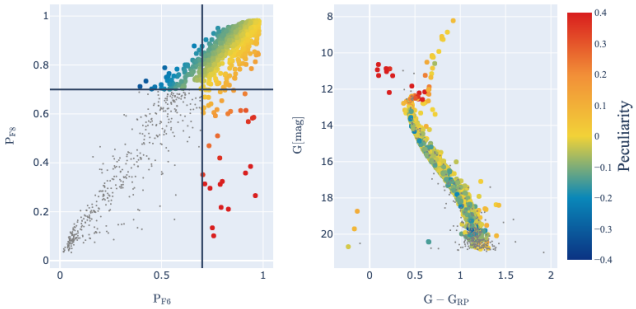

Figure 2
On the left, they have compared the P_{F6} and P_{F8} values of the stars. On the right, they have a CMD based on the Peculiarity of stars

References

Jadhav, V. V., Pennock, C. M., Subramaniam, A., Sagar, R., & Nayak, P. K. 2021, Monthly Notices of the Royal Astronomical Society, 503, 236–253, doi: 10.1093/mnras/stab213