CP191

Round 1 Report and Reflection

Md Mahmudunnobe

Minerva Schools at KGI

Abstract

The members of an open stellar cluster are often hard to distinguish from the field star in the same area of sky. This paper summarized the recent methods used for finding membership probability and classified them as astrometric-photometric and parametric-nonparametric approaches based on the input variables and model parameters.Working with this paper showed me the importance of choosing more specific problems, acquiring the necessary knowledge and managing time properly.

Word Count: 70

Introduction

When a group of stars originates from the same interstellar cloud at the same time, they form a stellar cluster. The stars of a specific cluster have almost the same age, same distance, same relative motion compared to us, and the same chemical compositions. But as we look into very far away stellar clusters, it becomes difficult to distinguish the members from the other stars projected in the same direction of the sky, either in the foreground (closer than the cluster) or background (beyond the cluster) of the cluster stars. Many different analytical methods have been developed and still developing to improve the accuracy of finding the membership probability (how likely it is to be a member) for a given star in a large field of stars containing a stellar cluster. And along with other branches of astronomy, this membership detection process also started to use different machine learning approaches since the beginning of the 21st century.

As there is still no single resource available where all different methods are reviewed, classified and compared to each other, I want to work to fulfill this gap of information so that people can easily use the appropriate tools for member identification and focus more on the follow-up analysis of the stellar clusters themselves. As an introductory step, I analyzed the most recent papers, mainly from 2018-2020, regarding cluster membership of open stellar clusters and classified them based on the input variables and the models that have been used.

Word Count: 246

Cluster Membership

The most recent cluster membership approaches can be broadly classified in two groups based on the type of input variables: astrometric and photometric approach. Similarly based on the type of models used, we can also classify them in two groups: parametric and non-parametric approaches. We will discuss them thoroughly in the following sections.

Astrometric Approach

Astrometric variables are the ones which we can detect and quantify from an image of the cluster region. This includes the position of the star in the sky plane (i.e. right ascension, RA and declination, dec), its relative translational motion over time (proper motion), the shift of its position as earth moves around the sun (parallax) that can also be used to calculate the distance etc.

These variables are densely distributed for members of a cluster compared to the random stars in a projected sky area, thus very helpful to use in cluster membership. But one problem is, for distant clusters, the position shifts are too small to detect any proper motion or parallax with most of the instruments. Only recently GAIA (Global Astrometric Interferometer for Astrophysics) mission is providing us these astrometric data in about microarcseconds accuracy, which opens a series of work on open clusters in recent years.

The basic idea for all astrometric approaches is quite simple. Use a model to find the group of stars which have similar RA, dec, proper motion and parallax compared to the other field stars. Among them the most important and most used variable for cluster membership is the proper motion (Balaguer-Núñez, 1999; Yadav et al, 2013). RA and dec is not really a distinguishable variable for the cluster, as all the foreground and background stars along that line

of sight will have similar RA and dec. Parallax is important as all the cluster members should be at the same distance, but it is harder to measure than the proper motion. So if we plot the proper motion along RA and dec for the stars in the field, the cluster members form a clump. The most common method using proper motion for cluster membership is based on Sanders' maximal likelihood model (Sanders, 1971).

Photometric Approach

Photometric variables are the ones which come from the photometric measurements: where we measure the intensity of light from the star in different band filters (i.e. blue, visible, infrared etc.). This includes primarily the flux and the apparent magnitude in different bands and the difference of magnitude between two bands (color index or simply 'color').

If we plot the luminosity (energy emitted per second) and the temperature of a star in its lifetime, it follows a very distinct plot called "HR diagram". Similarly, if we have a group of stars of similar age but varying mass (just like a stellar cluster), we will find them scattered along this HR diagram. Luminosity is related to magnitude and temperature is related with color, so we can express the HR diagram as a color-magnitude diagram. So after observing photometric variables, if we plot a color-magnitude diagram, the members of a cluster should lie along a H-R diagram, where the field stars should have a random position.

The photometric variables are easier to observe than the precise astrometric variables, but they are harder to use to find the membership probability as neither color nor the magnitude is densely distributed for cluster members than the field stars. It is more common to use the photometric variables to verify the members classified by astrometric approach and remove any non-members as they all should lie in the HR diagram (Yen et al 2017, Monteiro & Dias, 2019).

In an only photometric approach, usually we find the random color-magnitude diagram for the field star using a reference image, where the cluster is not present. Then we subtract this from the color magnitude diagram of the (field+cluster) image to find the member stars.

There are two other variables used in cluster membership, which we can discuss under a photometric approach as they are closely related: the radial velocity and the polarization of the light. We can find the radial velocity of stars from the doppler shift obtained by spectroscopy, which is also harder to measure for the distant stars as we need super high resolution, but as all of the cluster member should have similar radial velocity, it classifies the members very accurately, specially when used with other astrometric variables like proper motion (Randich et al, 2017, Fritzewski et al, 2019; Jackson et al, 2020).

The starlight gets polarized as they interact with the asymmetrically aligned dust particles while coming through the interstellar medium. As all cluster members' light travels a similar path to reach us, we can expect them to be polarized similarly compared to the randomly polarized field stars. Though the accuracy is quite less using just only polarization, it was used along with other stronger variables like proper motion (Medhi & Tamura, 2013).

Parametric Approach

Now based on the model parameters, we can also classify the methods in two broad categories: parametric and non-parametric approaches. Parametric models are those models, where we try to map the output variable (membership probability) and the input variables (astrometric or photometric) using some suitable distribution with some specific parameters and try to find the optimal value of the parameters to improve the accuracy of prediction..

Most of the parametric methods used recently for cluster membership are one or other forms of bayesian inference, which relates with the bayes equation:

$$P(\theta|data,\ model) \propto P(data|\ \theta,\ model)\, P(\theta \mid model).$$

where $P(data|\ \theta,\ model)$ is the likelihood, means the probability density using a specific parameter(s) θ for a observed data and $P(\theta \mid model)$ is called the prior or initial educated guess for the parameter distribution and $P(\theta|data,\ model)$ is called the posterior distribution for our parameter.

One common method is the maximum likelihood model, where we try to find the parameter for a distribution which will give the maximum probability density for the observed data. Most maximal likelihood models used astrometric data and/or radial motion as they have a dense distribution for the cluster members. Sampedro and Alfaro (2016) showed how we can use a N-dimensional space with densely distributed N variables to determine the membership probability.

Hidayat et al, (2019) used Markov Chain Monte Carlo (MCMC) method, which approximates the posterior parameter of the the gaussian bivariate distribution of proper motion in RA and dec for both clusters and members from an uninformative initial prior and then constant iteration using bayesian inference, which is then used to get the membership probability. Bossini et al, (2019) also used a similar approach using astrometric data using BASE-9, an analytical tool developed using MCMC and numerical integration techniques for stellar cluster analysis.

Finally some were able to incorporate both the astrometric and photometric approaches together in the bayesian method (Maíz, 2019; Parren et al 2015; Stott, 2018). Stott used the

likelihood for the astrometric variables along with an initial uninformative prior to get a posterior for membership probability. Then he used this distribution as the prior of the 2nd bayes equation along with the likelihood for the photometric variables to determine the final posterior membership probability.

Parren et al (2015) developed a python package called ASteCA for stellar cluster analysis, where they used the likelihood for the joint gaussian distribution of color and magnitude for each pair of the stars in the input data. They estimated the density of field stars from a reference image and estimate the number of field star, n_f multiplying the density with the (cluster+field) image. Then they used the ratio of the number of member star and total star $(1 - n_f/n_{total})$ as the prior and derived the posterior probability of the membership probability for a given star.

Non parametric Approach

Unlike parametric approach we do not assume any specific distribution of our data in a non-parametric approach and we do not try to find the value of the parameters underlying the model. Instead, we are only interested to know the output variable (membership probability) from the input variables and use the color-magnitude diagram to verify the result. The recently used non-parametric methods include the k-means clustering (El aziz et al, 2016), spectral clustering and random forest (Gao 2018).

In k-means clustering we randomly select k 'center points' in the n-dimensional space, where n is the number of our input variable. Then we assign each data point to the clusters of the 'center point' from which the euclidean distance is minimum. Once we find k clusters, we find the mean position of each star and set them as new 'center points'. We repeat this until the center

points become fixed and calculate the summation of the variance for each of the clusters. calculate the euclidean distance for each datapoint from these ‘center points’ and classify El aziz et al (2016) used k-means clustering using the astrometric variables to determine the members for NGC 188 and NGC 2266. In k-means clustering, we consider the data in n dimensional spaces, where n is the number of input variables. We calculated the closeness of each star using the euclidean distance between them, and based on the closeness (or distance) we grouped the full dataset into k different clusters. Here, due to the similarity of the astrometric data for the cluster members they will be very close to each other while the field stars will be in a random position. Spectral clustering used the similar idea as in k-means clustering but instead of calculating the euclidean distance in n-dimensional space, it creates graphs while connecting the data points based on the similarity of the input variables and group the datasets into separate closely connected clusters.

Random forest is based on the decision tree approach, where the training dataset is used to build a tree, which later can classify or predict the output of a new data. Random forest, to reduce the variance of a single tree, builds many trees each time using a random subset of the data and random subset of the input variables. Then the final prediction or classification is the average or the majority of all such trees. As a supervised method, random forest needs training data in order to build the model. We can use the existing cluster membership data or use other methods to first create the train data before running it into a random forest to detect more members from the new observations. Gao (2018) used spectral clustering method first to find a small group of member stars for NGC 188 and then used that as a training set in random forest to

find more members and their membership probability (the percentage of trees that classify them as members) for all of them.

Discussion

Using a set of similar variables, astrometric and photometric, a lot of methods have been used so far to determine the members of a cluster, among them only a few recent ones are discussed in the span of this paper. Advance instruments and space missions can improve the accuracy and limits of determine the astrometric variables in recent days as the recent GAIA mission of ESA captures the 3D photometry and astrometry of our galaxy with astonishing details. Using these new data we can try to improve our knowledge about the open clusters in our galaxy in upcoming days.

Also we can try to find some more clever way to use the photometric data by itself or along with astrometry if possible. As the amount of observational data is increasing so much, it is a must to create a model or method that can be applied without a need of visual inspection and should be generalized enough to process the data of different open clusters identically like the ASteCA package. The use of more sophisticated machine learning algorithms including deep learning can be also used in this field as they are being widespread in astrophysics (Vanderplas, 2012), which can possibly do more than just finding membership probabilities if used properly.

Word Count: 2002

HC and LO applications

#audience: I kept in mind that this paper is oriented mainly to capstone professors and all my peers irrespective of their major so that they can read and provide necessary feedback. So the first time I introduced any new terms from astronomy, I explained what it is and how it relates with the context (i.e. proper motion, parallax, color-magnitude diagram etc.). I avoided the jargons and rigorous mathematical derivations, whenever possible and added qualitative explanations instead (i.e. for k-means clustering or random forest etc.)

Word Count: 85

#gapanalysis: As a long term goal, we want to understand if any new methods should be used for membership probability of open clusters. But before that, in this paper, I identified and summarised the recent major methods that are used for cluster membership and provided reasoning for categorizing them using appropriate criteria. Now we also need to analyze the effectiveness of these methods for a better application of gap analysis, but that requires more knowledge and expertise on this topic.

Word Count: 79

Bibliography

Balaguer-Núñez, L., Tian, K., & Zhao, J. (1999). Determination of proper motions and membership of the open clusters NGC 1817 and NGC 1807. Astronomy and Astrophysics Supplement Series, 133(3).

Bossini, D., Vallenari, A., Bragaglia, A., Cantat-Gaudin, T., Sordo, R., Balaguer-Núñez, L., Jordi, C., Moitinho, A., Soubiran, C., Casamiquela, L., Carrera, R., & Heiter, U. (2019). Age determination for 269 Gaia DR2 open clusters. Astronomy & Astrophysics, 623, A108. https://doi.org/10.1051/0004-6361/201834693

El Aziz, M. A., Selim, I. M., & Essam, A. (2016). Open cluster membership probability based on K-means clustering algorithm. Experimental Astronomy, 42(1), 49–59. https://doi.org/10.1007/s10686-016-9499-9

Fritzewski, D., Barnes, S., Strassmeier, K., James, D., Geller, A., & Meibom, S. (2019). Spectroscopic membership for the populous 300 Myr-old open cluster NGC 3532. Astronomy and Astrophysics, 622. https://doi.org/10.1051/0004-6361/201833587

Gao, X. (2018). Memberships, distance and proper-motion of the open cluster NGC 188 based on a machine learning method. Astrophysics and Space Science, 363(11). https://doi.org/10.1007/s10509-018-3453-4

Hidayat, Y., Arifyanto, M., Aprilia, & Hakim, M. (2019). An application of the Markov Chain Monte Carlo (MCMC) method to open cluster membership determination. Journal of Physics: Conference Series, 1231(1). https://doi.org/10.1088/1742-6596/1231/1/012029

Jackson, R. J., Jeffries, R. D., Wright, N. J., Randich, S., Sacco, G., Pancino, E., Cantat-Gaudin, T., Gilmore, G., Vallenari, A., Bensby, T., Bayo, A., Costado, M. T., Franciosini, E.,

Gonneau, A., Hourihane, A., Lewis, J., Monaco, L., Morbidelli, L., & Worley, C. (2020). The Gaia-ESO Survey: membership probabilities for stars in 32 open clusters from 3D kinematics. Monthly Notices of the Royal Astronomical Society. https://doi.org/10.1093/mnras/staa1749

Keshav, S. (2007). How to read a paper. ACM SIGCOMM Computer Communication Review, 37(3), 83-84. Retrieved from http://ccr.sigcomm.org/online/files/p83-keshavA.pdf

Maíz Apellániz, J. (2019). Gaia DR2 distances to Collinder 419 and NGC 2264 and new astrometric orbits for HD 193 322 Aa,Ab and 15 Mon Aa,Ab. Astronomy and Astrophysics. https://doi.org/10.1051/0004-6361/201935885

Medhi, B. J., & Tamura, M. (2013). Cluster membership probability: polarimetric approach. Monthly Notices of the Royal Astronomical Society, 430(2), 1334–1343. https://doi.org/10.1093/mnras/sts714

Monteiro, H., & Dias, W. (2019). Distances and ages from isochrone fits of 150 open clusters using Gaia DR2 data. Monthly Notices of the Royal Astronomical Society: Letters, 487(2). https://doi.org/10.1093/mnras/stz1455

Perren, G. I., Vázquez, R. A., & Piatti, A. E. (2015). ASteCA: Automated Stellar Cluster Analysis. Astronomy and Astrophysics, 576. https://doi.org/10.1051/0004-6361/201424946

Randich, S., Tognelli, E., Jackson, R., Jeffries, R. D., Degl'Innocenti, S., Pancino, E., Fiorentin, P. R., Spagna, A., Sacco, G., Bragaglia, A., Magrini, L., Prada Moroni, P. G., Alfaro, E., Franciosini, E., Morbidelli, L., Roccatagliata, V., Bouy, H., Bravi, L., Jiménez-Esteban,

F. M., … Zaggia, S. (2018). The Gaia -ESO Survey: open clusters in Gaia -DR1. Astronomy & Astrophysics, 612, A99. https://doi.org/10.1051/0004-6361/201731738

Sampedro, L., & Alfaro, E. (2016). Stellar open clusters' membership probabilities: An N-dimensional geometrical approach. Monthly Notices of the Royal Astronomical Society, 457(4). https://doi.org/10.1093/mnras/stw243

Sanders, W. L. (1971). An improved method for computing membership probabilities in open clusters. Astronomy and Astrophysics, 14, 226-232.

Stott, J. (2018). Determining open cluster membership: A Bayesian framework for quantitative member classification. Astronomy and Astrophysics, 609. https://doi.org/10.1051/0004-6361/201628568

Vanderplas, J., Connolly, A. J., Ivezic, Z., & Gray, A. (2012). Introduction to astroML: Machine learning for astrophysics. Proceedings - 2012 Conference on Intelligent Data Understanding, CIDU 2012, 47–54. https://doi.org/10.1109/CIDU.2012.6382200

Yadav, R. K. S., Sariya, D. P., & Sagar, R. (2013). Proper motions and membership probabilities of stars in the region of open cluster NGC 3766. Monthly Notices of the Royal Astronomical Society, 430(4), 3350–3358. https://doi.org/10.1093/mnras/stt136

Yen, S., Reffert, S., Schilbach, E., Roser, S., Kharchenko, N., & Piskunov, A. (2018). Reanalysis of nearby open clusters using Gaia DR1/TGAS and HSOY. Astronomy and Astrophysics, 615. https://doi.org/10.1051/0004-6361/201731905