CP195: Prospectus

A Semi-supervised Machine Learning Approach for Open Star Cluster Member Determination

Md Mahmudunnobe

Minerva University

Background

Star Cluster: Stars are formed when a large number of gas and dust particles accumulates in a smaller region. When a group of stars are born from the same dust clouds at the same time and stay close to each other due to mutual gravity, it is called a star cluster (Fig 1). All the stars of a cluster have similar age, similar distance from us, similar position in the sky and similar motion across the sky (Platais, 2009). As they comes from the same cloud, they also have same chemical composition, but it is harder to measure.

If we can understand the dynamics of star cluster, it will help us to study star formation mechanism and how the stars evolve. But first we need to confidently find the members of a cluster. The major problem is to differentiate the field stars (non member stars) that are projected in the same direction (Kharchenko, 2005).

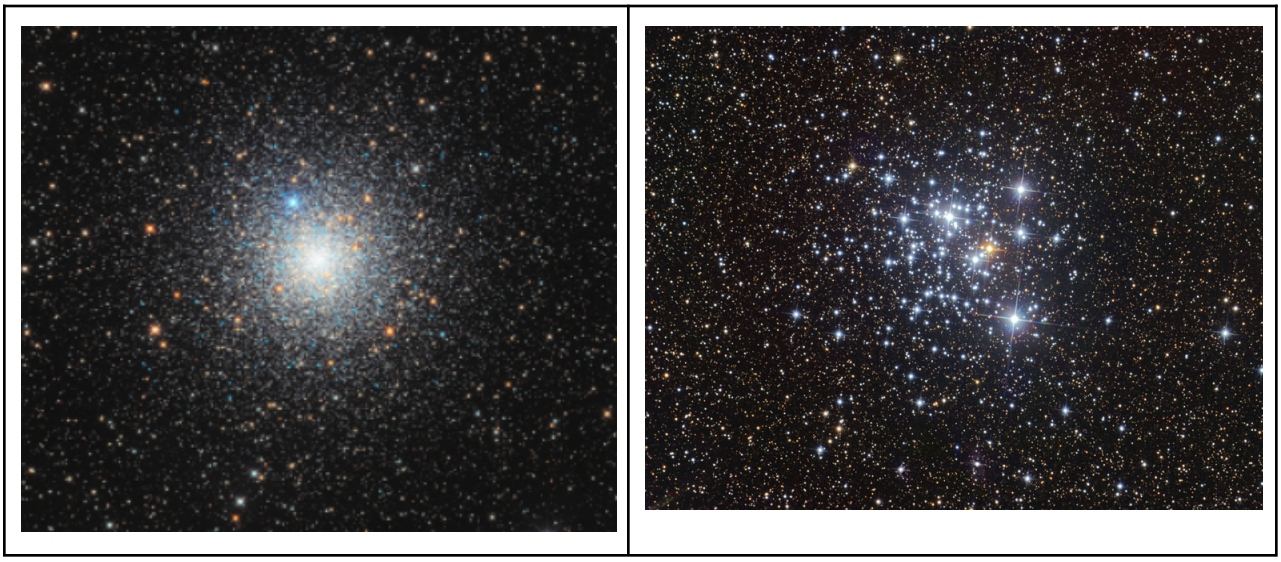

Figure 1: Star Cluster *NGC 6752* (2020) and *Jewel Box Cluster* (2010) showing that the stars in a clusters are densely situated in a close region, where they are held together by mutual gravity. (Taken from Astronomy Picture of the Day, APOD)

Machine Learning Models

Machine Learning (ML) models are algorithm trained using a old dataset that can find pattern or make predictions from a new unseen dataset. Depending on the availability of a labelled training data, it is divided into two models.

When we do not know any labels for our training data, we need to use unsupervised ML model, that can only find the underlying pattern. For example, if we have 6 apples and 4 oranges, but do not know the labels, then the unsupervised ML models can only find two group of objects without predicting their names. They are easier to use but it is harder to evaluate the model performance as we do not know the true labels (James et al., 2017).

But if we have labels for all our training data, then we can use a supervised models. For the same example, a supervised model will predict if an object is an apple or an orange. This type of model is only possible to use when we have labels, but it allows us to get more confident results as well as to correctly find the model performance (James et al., 2017).

For star cluster membership identification, most common method is the unsupervised model, as we usually only have the observed data (position, velocity, etc.) for the stars but we do not know if they are member or not.

Hypothesis Development

For any research project, we always need to come up with a well-formed hypothesis, that we can use to design the research. A good hypothesis has to be very specific, plausible, testable and falsifiable. As a rule of thumb, we can divide the development of a hypothesis into four components.

Evidence: The hypothesis process starts by looking into prior knowledge and/or observations. We can study the past literatures to find the current progress and/or gaps that can lead to a potential hypothesis. Similarly, often we can analyze and visualize observed data to support our hypothesis.

Cause and effect: This is the main statement of the hypothesis. We need to be very specific about what are our predictors and outcomes and they should be measurable. For example, '*the climate change hampers the biodiversity'* is not very specific and it is not clear how we will measure the variables. A improved version will be, '*the increase in* CO_2 *decreases the bird species richness (the number of different species present) in California forests*'.

Mechanism: Next important piece is to explain the plausible mechanism of the cause and effect. Usually this part consists of some assumptions and/or arguments. The assumptions and premises has to be supported by the prior data and any logical arguments need to be valid.

Iterative Process: All hypotheses should work as a stepping stone to progress the knowledge in the field. First we need to define how we will test our hypothesis and specify what results will support and what will refute our hypothesis. For both cases, we need to lay out what can we learn from the study and the possible future steps.

Capstone Project Hypothesis

Evidence: First, I reviewed the literature to find the current progress and gaps. Most of the past studies used only unsupervised model to find members of the star clusters, but there was no common quantitative metric to measure model performance. Only a few paper used an additional supervised model and they found more members (Gao, 2018; Castro-Ginard et al.,

2018). In another hand, machine learning studies shows that supervised models are better to validate and to get more confident members (Jain et al., 1999, Reddy et al., 2018).

Cause and effect: Based on the evidence, I argued that using a well-justified and optimized semi-supervised model (combination of an unsupervised and a supervised model) *(cause)* would increase the model performance and number of retrieved member compared to an unsupervised model *(effect)*. I later defined what metrics I would use for model performance.

Mechanism: An unsupervised model works best when we have more members and less field stars. Thus to increase the confidence, we need to apply it for a smaller region, resulting in smaller number of member. In a semi-supervised model, we can use this smaller but reliable member group as our training data and apply an additional layer of a supervised model. As we now have training data, we can explore a larger region to get more members with same confident level.

Iterative Process: Lastly, I define how can I interpret the results, i.e. the model performance and number of members for the semi-supervised and the traditional unsupervised models. For example, if both of them increases for semi-supervised model, the hypothesis is true. Then as a next step we can study if the specific algorithms used in semi-supervised model need to be changed depending on the age of the cluster. On the other hand, if the model performance is lower, the hypothesis is false and false. Then we could focus on developing a more improved common quantitative metric for unsupervised models.

Reference

Castro-Ginard, A., Jordi, C., Luri, X., Julbe, F., Morvan, M., Balaguer-Nuñez, L., &Cantat-Gaudin, T. (2018). A new method for unveiling open clusters in Gaia New nearby open clusters confirmed by DR2. Astronomy and Astrophysics, 618. https://doi.org/10.1051/0004-6361/201833390

Gao, X. (2018a). A Machine-learning-based Investigation of the Open Cluster M67. The Astrophysical Journal, 869(1). https://doi.org/10.3847/1538-4357/aae8dd

Jain, A. K., Murty, M., & Flynn, P. (1999). Data clustering: A Review ACM Computing Surveys, vol. 31. Google Scholar, 264-318. http://users.eecs.northwestern.edu/~yingliu/datamining_papers/survey.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An introduction to statistical learning (Chapter 2). New York [etc.]: Springer. ISBN-10: 1461471370

Jewel Box Cluster (2010) Astronomy Picture of the Day. Retrieved from https://apod.nasa.gov/apod/ap100817.html

Kharchenko, N. V., Piskunov, A. E., Röser, S., Schilbach, E., & Scholz, R. D. (2005). Astrophysical parameters of Galactic open clusters. Astronomy & Astrophysics, 438(3), 1163-1173. https://doi.org/10.1051/0004-6361:20042523

NGC 6752 (2020) Astronomy Picture of the Day. Retrieved from https://apod.nasa.gov/apod/ap200123.html

Platais, I. (2009). Star clusters. In Astrometry for astrophysics: Methods, Models, and Applications (pp. 360–367). Cambridge University Press. https://doi.org/10.1017/CBO9781139023443.026