Performance Comparison of Three Clustering Algorithms for Member Star Identification in Open Clusters: A Postprint

Authors: Xiong Zhuang, Zhang Peng, Yang Xiangming, Liu Gaochao, Liu Di, Li Jiapeng, TIAN Haijun

Date: 2025-04-08T16:02:35+00:00

Abstract

For a long time, the identification of member stars in open clusters has been a challenge in the field of astronomy. Due to the complexity of formation and evolution of open clusters, there is no unified method that can accurately determine member stars within them. The objective is to take three open clusters with different spatial distribution types as samples, select five-dimensional parameters of stellar positions and motions, and perform clustering detection on open clusters through three clustering methods—DBSCAN (Density-Based Spatial Clustering of Applications with Noise), FOF (Friend of Friend), and STAR GO (Star' s Galactic Origin)—to quantify the performance of different algorithms. The research results show that the FOF and STAR GO algorithms are more suitable for clusters with special structures, capable of identifying tidal or extended structures of clusters, while DBSCAN provides more complete identification of member stars in the core regions of clusters. The aim is to find a more balanced algorithmic strategy between cluster structural details and improving the completeness of member star identification.

Full Text

Preamble

Performance Comparison of Three Clustering Algorithms in Open Cluster Member Star Identification

XIONG Zhuang[1], ZHANG Peng[12], YANG Xiang-ming[1], LIU Gao-chao[12], LIU Di[1], LI Jia-peng[3], TIAN Hai-jun[3]

[1] College of Science, China Three Gorges University, Yichang 443002
[2] Center for Astronomy and Space Sciences, China Three Gorges University,

Yichang 443002
[3] College of Science, Hangzhou Dianzi University, Hangzhou 310018

ABSTRACT

For a long time, the identification of open cluster member stars has been a challenge in astronomy. Due to the complexity of open cluster formation and evolution, no unified method exists for accurately identifying member stars within open clusters. This study selects five-dimensional parameters describing the position and motion of stars from three different spatial distribution types of open clusters to evaluate the performance of Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Friend-of-Friend (FOF), and Star' s Galactic Origin (STAR GO) clustering methods in detecting open star clusters. The results show that FOF and STAR GO algorithms are more suitable for clusters with special structures, as they can identify tidal or extended structures, while DBSCAN can more completely identify member stars in the core region of clusters. The aim is to find a more balanced algorithmic strategy between preserving cluster structural details and maintaining the completeness of member star recognition.

Key words: globular clusters: individual: Trumpler 10, NGC 752, NGC 2232, Tian 2; methods: data analysis; astronomical databases: Gaia satellite data

1 Introduction

The identification of member stars in open clusters has long been a fundamental challenge in stellar astronomy. Traditional methods rely on statistical approaches to select stars with similar kinematic and photometric properties from field star contamination. With the advent of large-scale sky surveys, particularly the Gaia mission, the volume and precision of astrometric and photometric data have increased dramatically, enabling more sophisticated clustering analyses.

Previous studies have employed various algorithms for open cluster identification. The Friend-of-Friend (FOF) algorithm, introduced by Geller & Huchra [**?**] and Helmi & de Zeeuw [**?**], has been widely used due to its simplicity and effectiveness in finding overdense regions. More recently, Castro-Ginard et al. [**?**] applied DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to systematically search for open clusters in Gaia data, demonstrating its capability to identify clusters of arbitrary shape. Additionally, Yuan et al. [**?**] developed STAR GO (Star' s Galactic Origin), which utilizes a self-organizing map (SOM) to trace stellar origins, showing promise in identifying extended tidal structures.

However, each algorithm has distinct advantages and limitations. FOF is sensitive to the choice of linking length and may struggle with hierarchical structures. DBSCAN excels at identifying core members but may miss extended tidal features. STAR GO provides detailed structural information but requires careful parameter tuning. This study systematically compares these three algorithms

using well-studied open clusters—Trumpler 10, NGC 752, NGC 2232, Tian 2, and SP1 (“Snake” part 1)—to evaluate their performance across different cluster morphologies.

The Gaia Data Release 3 (Gaia DR3) provides unprecedented astrometric precision, with proper motion uncertainties below 0.1 mas/yr for bright sources (G $<$ 18 mag) and parallax uncertainties of order 0.1 mas. This work utilizes five-dimensional data (positions: RA, Dec; proper motions: pmra, pmdec; and parallax) to characterize cluster membership, focusing on stars with G $<$ 18 mag to ensure reliable astrometric solutions.

3 Methods

3.1 FOF Algorithm Implementation

The Friend-of-Friend algorithm groups stars into clusters based on a linking length criterion. Two stars are considered “friends”if their distance in parameter space falls below a specified linking length, and clusters form through transitive linking of friend pairs. The choice of linking length critically affects the results: too small a value fragments the cluster, while too large a value merges unrelated field stars.

In this study, we implement FOF using two approaches: the ROCKSTAR algorithm [**?**, **?**], originally designed for halo finding in cosmological simulations, and the Star Cluster Hunting Pipeline (SHIP) [**?**]. ROCKSTAR employs an adaptive linking length that varies with local density, making it suitable for identifying both compact cores and extended structures. SHIP uses a fixed linking length optimized for open cluster detection in Gaia data.

For each cluster, we calculate the linking length based on the expected velocity dispersion and spatial extent. The linking length in proper motion space is typically set to 2–3 times the median uncertainty, while the spatial linking length scales with the cluster’ s tidal radius. This adaptive approach allows FOF to trace extended tidal tails while maintaining compact core identification.

4 Algorithm Details and Results

4.1.1 DBSCAN Parameters

DBSCAN identifies clusters as dense regions separated by sparse regions. The algorithm requires two key parameters: , the maximum distance between neighboring points, and MinPts, the minimum number of points required to form a dense region. In this work, we set based on the median astrometric uncertainty of the sample, typically 0.5–2 mas in parallax space and 0.1–0.5 mas/yr in proper motion space. MinPts is set to 5–10, depending on the expected cluster richness.

DBSCAN classifies points as core points, border points, or noise. Core points have at least MinPts within distance ; border points are reachable from core

points but do not themselves satisfy the MinPts criterion; noise points are neither core nor border points. This classification allows DBSCAN to identify clusters of arbitrary shape and handle field star contamination effectively.

4.1.2 FOF Parameters

The FOF algorithm' s primary parameter is the linking length, denoted as l_{link}. In multidimensional parameter space, we define a distance metric that combines spatial and kinematic separations:

$$d_{ij} = \sqrt{\left(\frac{\Delta\theta_{ij}}{\sigma_\theta}\right)^2 + \left(\frac{\Delta\mu_{ij}}{\sigma_\mu}\right)^2 + \left(\frac{\Delta\varpi_{ij}}{\sigma_\varpi}\right)^2}$$

where $\Delta\theta_{ij}$, $\Delta\mu_{ij}$, and $\Delta\varpi_{ij}$ are the angular, proper motion, and parallax differences between stars i and j, respectively, and σ_θ, σ_μ, σ_ϖ are the characteristic scales for each dimension. Stars are linked if $d_{ij} < l_{\text{link}}$.

For the clusters studied here, we adopt $l_{\text{link}} = 0.2$ for compact clusters like NGC 2232 and $l_{\text{link}} = 0.5$ for extended structures like Trumpler 10 and SP1. These values are determined empirically by maximizing the recovery of known members while minimizing field star contamination.

4.3 Age Selection and Isochrone Filtering

To refine cluster membership, we apply age-based photometric filtering using isochrones. For each cluster, we select stars within a specific age range based on their position in the color-magnitude diagram (CMD). Trumpler 10 and SP1, being young associations, use an age range of 5–120 Myr. NGC 752, an older open cluster, is best fitted with an age of 1.75 Gyr, and we select stars within $±$1 mag of the isochrone.

The CMD filtering is performed after the initial clustering to remove field stars that happen to have similar kinematics but differ in age or evolutionary stage. We use Gaia' s GBP and GRP bands to construct the CMD, with the following transformation:

$$x_i = (G_{BP} - G_{RP})_i - \langle G_{BP} - G_{RP} \rangle_{\text{cluster}}$$
$$y_i = G_i - \langle G \rangle_{\text{cluster}}$$

where the angle brackets denote the cluster's median color and magnitude. Stars deviating by more than 0.1 mag from the isochrone are rejected.

4.4 Comparison of Clustering Results

[Figure 3: see original paper] illustrates the preliminary screening process for three representative clusters. The gray points show the raw clustering output, while black points indicate stars retained after isochrone filtering. For NGC 752, the 1.75 Gyr isochrone clearly separates cluster members from field stars, with the $±$1 mag tolerance capturing the main sequence and turn-off region.

The spatial distribution of clustering results reveals significant differences between algorithms. [Figure 8: see original paper] shows the projected sky positions of members identified by DBSCAN, FOF, and STAR GO for Trumpler 10, NGC 752, and SP1. For Trumpler 10 (panel a), FOF and STAR GO detect extended structures beyond the compact core, while DBSCAN concentrates on the dense central region. For NGC 752 (panels b and c), the tidal tails are more prominent in STAR GO and FOF results, particularly when using the ROCKSTAR implementation. For SP1 (panels d–f), the extended "snake-like" structure is best recovered by STAR GO and FOF, with DBSCAN missing the low-density extensions.

4.5 Performance Metrics and Discussion

To quantify algorithm performance, we compute the recovery rate r_n for each cluster and method, defined as the fraction of high-probability members (G < 18 mag) retained after all filtering steps. presents r_n values for Trumpler 10, SP1, and NGC 752. DBSCAN achieves the highest r_n for compact cores due to its density-based nature, while FOF and STAR GO show lower r_n but capture more extended members.

The computational cost varies significantly: DBSCAN scales as $\mathcal{O}(n \log n)$ with spatial indexing, while FOF' s naive implementation scales as $\mathcal{O}(n^2)$ but can be reduced to $\mathcal{O}(n \log n)$ using kd-trees. STAR GO, involving SOM training, is the most computationally intensive, requiring $\mathcal{O}(n \cdot k)$ operations where k is the number of map nodes.

In summary, no single algorithm excels in all aspects. DBSCAN is optimal for identifying compact, well-defined clusters where completeness of the core is paramount. FOF, particularly with adaptive linking lengths, provides a good balance between core recovery and extended structure detection. STAR GO offers the most detailed structural information but at higher computational cost and with more complex parameter tuning. For comprehensive open cluster studies, we recommend a hybrid approach: use DBSCAN for initial member selection and FOF or STAR GO for detailed structural analysis of the outskirts.

References

[1] Chen L, de Grijs R, Zhao J L. AJ, 2007, 134: 1368
[2] Sun W J, Li C Y, Deng L C, et al. ApJ, 2019, 883: 182
[3] Dias W S, Alessi B S, Moitinho A, et al. A&A, 2002, 389:

[4] He Z H, Wang K, Luo Y P, et al. ApJS, 2022, 262: 7
[5] Liu L, Pang X Y. ApJS, 2019, 245: 32
[6] Cantat-Gaudin T, Jordi C, Vallenari A, et al. A&A, 2018, 618: 93
[7] Yu H, Shao Z Y, Diaferio A, et al. ApJ, 2020, 899: 144
[8] Castro-Ginard A, Jordi C, Luri X, et al. A&A, 2020, 635:
[9] Castro-Ginard A, Jordi C, Luri X, et al. A&A, 2022, 661:
[10] He Z H, Li C Y, Zhong J, et al. ApJS, 2022, 260: 8
[11] He Z H, Xu Y, Hao C J, et al. RAA, 2021, 21: 93
[12] Boffin H M J, Jerabkova T, Beccari G, et al. MNRAS, 2022, 514: 3579
[13] Geller M J, Huchra J P. ApJ, 1982, 257: 423
[14] Helmi A, de Zeeuw P T. MNRAS, 2000, 319: 657
[15] Rasera Y, Alimi J, Courtin J, et al. Invisible Universe: Proceedings of the Conference. New York: American Institute of Physics, 2010: 1134
[16] Behroozi P S, Wechsler R H, Wu H Y. ApJ, 2013, 762:
[17] Yuan Z, Chang J, Banerjee P, et al. ApJ, 2018, 863: 26
[18] Tian H J. ApJ, 2020, 904: 196
[19] Wang F, Tian H J, Qiu D, et al. MNRAS, 2022, 513: 503
[20] Kohonen T. Self-organizing Maps. 3rd ed. Berlin: Springer, 2010: 30
[21] Geach J E. MNRAS, 2012, 419: 2633

Note: Figure translations are in progress. See original paper for figures.