doi: 10.15940/j.cnki.0001-5245.2025.02.006

3种聚类算法在疏散星团成员星证认中的性能对比*

熊 壮[1] 张 鹏[1,2†] 杨向明[1] 刘高潮[1,2] 刘 迪[1] 李佳朋[3]
田海俊[3‡]

(1 三峡大学理学院 宜昌 443002)
(2 三峡大学天文与空间科学研究中心 宜昌 443002)
(3 杭州电子科技大学理学院 杭州 310018)

摘要 长期以来, 疏散星团成员星的证认问题一直是天文学领域的一个挑战, 由于疏散星团形成及演化的复杂性, 没有统一的方法能准确地确定疏散星团中的成员星. 目标是以3种不同空间分布类型的疏散星团为样本, 选取恒星的位置和运动的五维参数, 通过DBSCAN (Density-Based Spatial Clustering of Applications with Noise)、FOF (Friend of Friend)和STAR GO (Star's Galactic Origin) 3种聚类方法, 对疏散星团进行聚类检测, 量化不同算法的绩效. 研究结果表明, FOF与STAR GO算法对有特殊结构的星团更为适用, 能识别出星团的潮汐或延展结构, 而DBSCAN对星团核心区域成员星的识别更为完整. 旨在星团结构细节与提高成员星识别的完整性之间找到更均衡的算法策略.

关键词 疏散星团和星团: 个别: Trumpler 10, NGC 752, NGC 2232, Tian 2, 方法: 数据分析, 天文数据库: 盖亚卫星数据

中图分类号: P154; **文献标识码**: A

1 前言

疏散星团形成于大型气体云坍缩形成引力束缚的恒星聚集区, 通常被视为恒星演化的实验室. 由于疏散星团的形成特征, 除了形成于天体遗迹(如超新星爆炸的残留气体云等初代大质量恒星爆炸产物)之外的疏散星团, 都具有相似的运动特征、化学成分和年龄, 这些相似点都为恒星的演化提供了重要的参考.

作为宇宙中较为典型的天体群, 近年来, 疏散星团作为研究恒星形成和演化重要性的实证研究对象备受学界瞩目. 如Chen等[1]对疏散星团质量分层的研究, 提出了疏散星团形成的初始条件与弛豫过程结合, 可能导致质量分层. Sun等[2]对疏散星团内双主序的研究, 提出了潮汐作用可能对恒星自转及疏散星团演化方向产生影响. 当然, 疏散星团的研究领域远不止于此. 为深入研究疏散星团的形成和演化, 科研人员为编制疏散星团的星表做了大量工作. 最早可以追溯到公元前2世纪依巴谷(High Precision Parallax Collecting Satellite, Hipparcos)首次利用肉眼编绘的西方星表, 记录了约

2024-03-07收到原稿, 2024-05-16收到修改稿
*国家自然科学基金项目(12373033、12373030), 湖北省自然科学基金项目(2023AFB577)资助
†zhangpeng@ctgu.edu.cn
‡hjtian@lamost.org

一千颗恒星的位置和亮度等信息. 随着近现代科技和先进天文观测技术的发展, 一系列星表逐渐问世, 逐步揭开了恒星的神秘面纱. 例如, 依巴谷卫星[1]在4 yr间发现了约260万颗恒星的位置及相关信息. 此外, 还有著名的NGC (New General Catalogue)星表[2]和IC (Index Catalogue)星表[3]等.

除此之外, 还有更多的研究人员利用不同的数据和方法对不同天区的恒星进行分析, 不断丰富星表数据, 如Dias等[3]、He等[4]和Liu等[5]. Liu等[5]采用基于传统的可变链接长度(通常也称为Friend of Friend, FOF)的方法, 利用Gaia Data Release 2 (Gaia DR2)中五维恒星参数(银经l、银纬b、视差ϖ、经度自行pmra、纬度自行pmdec)对银河系做了普查, 识别出2443个候选星团, 并发现76个新的候选星团, 为银河系疏散星团的普查增添了新的成果. 另外, 还有众多研究者使用不同的方法来搜索天文数据中新的或现有的疏散星团. 如Cantat-Gaudin等[6]通过恒星的光度信息进行K-means聚类识别, Yu等[7]通过运动学的方法使用层次聚类算法对疏散星团进行聚类识别等. 迄今为止, 大多数新星团都是通过DBSCAN (Density-Based Spatial Clustering of Applications with Noise)聚类算法检测到的(如Castro-Ginard等[8–9]、He等[10–11]), 这种方法被证明是一种有效的大数据盲搜方法. 可见, 在针对银河系疏散星团的调查工作方面, 普查能为我们提供大量的候选星团数据, 为了解银河系内星团的分布、特征和数量提供了宏观视角.

众所周知, 星表中的星团往往是批量发现的, 而其中成员星的数量通常相对较少, 这是星表中星团的一个显著特征. 为了更深入地了解和分析星团, 并揭示恒星形成和演化的规律, 普查所获得的星团信息往往是不够详细的, 还需要对星团进行单一的聚类. 例如, 对于恒星相关的研究, 需要更多关于星团成员星的资料. 通过Dias等[3]、He等[4]以及Boffin等[12]对疏散星团NGC 752的研究, 可以发现更多关于成员星的有用信息, 图1是NGC 752的颜色星等图(Color-Magnitude Diagram, CMD), 图中的横坐标表示色指数(代表蓝光波段与红光波段视星等之差, 用$G_{BP}-G_{RP}$表示), 纵坐标表示恒星在G波段的视星等G, 在该图中, 显示了来自于3个星表中关于NGC 752的成员星, 可以发现这些星表中记录的成员星数量有所差异. 特别是在密近双星的识别方面, Boffin等[12]提供了更为完整的数据, 同时也体现星表普查数据与星团单一聚类数据的区别以及对所研究星团选择最佳匹配算法的必要性.

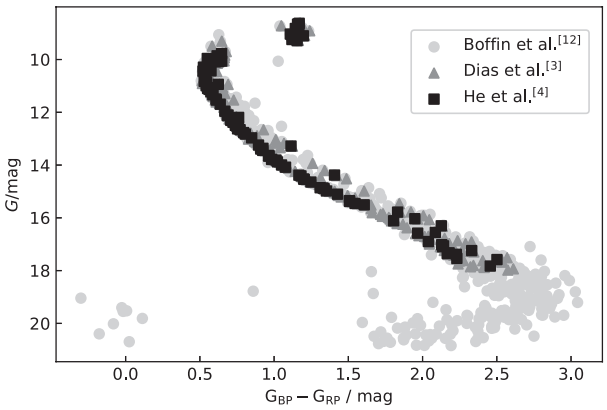

图 1 疏散星团NGC 752的CMD, 圆形表示Boffin等[12]研究NGC 752潮汐结构的成员星, 三角形与正方形分别表示Dias等[3]和He等[4]星表中的NGC 752成员星.

Fig. 1 The CMD of the open cluster NGC 752, where circular symbols represent the member stars of NGC 752 studied by Boffin et al.[12] for its tidal structure, and triangles and squares represent the member stars of NGC 752 from the catalogs of Dias et al.[3] and He et al.[4], respectively.

近年来, 除传统的FOF算法(Geller等[13]、Helmi等[14])外, 还涌现了许多基于FOF的改进算法, 并得到了广泛的研究, 比如Rasera等[15]实现了基于FOF算法的并行计算方法. 这些改进算法显著提升了原始FOF算法的计算效率, 在本文中, 我们将选用常用的DBSCAN算法, 并与在ROCKSTAR (Robust Overdensity Calculation using K-Space Topologically Adaptive Refinement)[16]中实现的改进FOF算法进行对比分析; 除此之外, 还选择了一

[1]https://www.cosmos.esa.int/web/hipparcos
[2]https://in-the-sky.org/data/catalogue.php?cat=NGC
[3]https://in-the-sky.org/data/catalogue.php?cat=IC

种基于自组织映射(Self-Organizing Map, SOM)的STAR GO (Star's Galactic Origin)方法[17], 通过精确评估3种聚类算法在处理不同类型星团数据集时的效能, 可以为找出适用于不同类型星团的最佳匹配算法提供参考, 从而进行更深入的星团研究.

本文将在第2部分介绍聚类使用的数据, 第3部分介绍了3种聚类算法, 第4部分介绍了聚类算法的聚类相关过程, 最后一部分是3种聚类算法的比较、结论和相关讨论.

2 数据

在数据方面, 疏散星团的无监督搜索结果极大程度上取决于观测数据的质量和数量. 随着观测数据的增多, 观测精度越来越高, 我们可以获得更全面的信息, 进而更容易识别出更多的星团, 甚至可能发现一些新奇或罕见的星团类型. 因此, 观测设备的性能和观测条件对于无监督搜索结果的准确性和可靠性起着至关重要的作用.

Gaia (Global Astrometric Interferometer for Astrophysics)卫星为我们提供了银河系中数量空前的恒星高精度天体测量和测光数据. Gaia DR3 (Gaia Data Release 3)中包含了亮于21星等(G波段)的约18亿个点源, 其中14亿个点源有完整的五维(赤经RA、赤纬DEC、视差ϖ、视向速度$R_{\rm v}$、自行pm)天体测量数据. Gaia EDR3 (Gaia Early Data Release 3)上发表的天体测量和宽带测光的系统误差延续到Gaia DR3上. 每个点源导出参数的不确定性强烈地依赖于点源的亮度. 在本研究中, 为了聚类的完整性, 尽可能多地保留了Gais DR3中的观测源(暗源), 由于暗源的观测误差较大, 大多数暗源对星团的结构等研究的贡献较小.

为增强工作中使用的分析算法的适应性, 本文选择了在近几年内发现的具有明显延展结构和潮汐结构的特殊星团. 分别是一个小延展结构Trumpler 10[18–19]和由疏散星团构成的大延展结构SP1 ("恒星蛇(Snake)"[18–19]星协中的part 1, 其中包含Tian 2和NGC 2232两个疏散星团)以及具有潮汐结构的NGC 752 (详见第4节)的数据作为算法测试样本. Trumpler 10选择"恒星蛇"所在天区$l \in (254°, 280°)$、$b \in (-4°, 4°)$范围内的3万颗源; 对于NGC 752[17], 选择其以日心银道坐标$(l, b) = (136°, -23°)$为中心, 半径r在5°范围内的3万颗源. 而对于SP1, 则选择"恒星蛇"所在天区$l \in (170°, 230°)$、$b \in (-30°, 10°)$范围内视星等(G波段)小于18的22 万颗源(为减小小质量场星对延展结构的识别干扰, 且在聚类识别的测试中也表明小质量场星的干扰很大), 星团的相关参数如表1所示.

本研究中, 采用五维数据(恒星在日心银道直角坐标系的位置(X、Y、Z)及恒星的自行(pmra、pmdec))作为聚类的输入向量, 通过聚类算法输出后, 对结果进行星团的真伪判断, 确定聚类所得到的星团是真的星团, 而非场星的聚集, 再通过等年龄线[4]和自行对结果进行筛选, 最终确定成员星(见4.3节).

3 聚类算法

毫无疑问, 恒星的可靠参数越多、越精确, 聚类效果就会越好, 考虑星团数据的复杂性, 在星团的聚类研究中, 所使用的算法都是基于基础算法的改进算法, 只有高效的算法才能适用于大规模、无监督的疏散星团识别任务. 另外, 在Gaia DR3数据集中, 疏散星团的成员星仅占很小比例, 因此所选用的聚类算法不仅需要能够高效处理大量数据, 还必须具备足够的敏感性, 以识别这些相对稀少的对象.

在本项研究的后续部分, 评估了3种所选聚类算法的性能. 本文为研究中使用的聚类算法设定了几项基准要求. 第一, 算法必须高效, 能够在合理的时间内在配备有较强计算能力的机器上处理整个Gaia数据集. 第二, 算法需要有效地分辨背景恒星, 银河系背景噪声太多, 所以必须能够将这些背景星剔除. 最后, 所选择的算法应该有一个在文献中广泛描述并且可以开源实现的版本(如Yuan等[16]的STAR GO开源代码), 便于验证和复用.

3.1 FOF算法简介

在原始的FOF算法中, 如果两个节点在预先指定的链接长度内, 则将它们附加到同一组中. 通常,

[4]http://stev.oapd.inaf.it/cgi-bin/cmd

这种链接长度是根据平均节点间距离的分数B来选择的, 其常用值为$B \in [0.15, 0.2]$, 在Yuan等[16]的文章中就对基本的FOF算法的B值做了比较, 不同的B值下的FOF对星团的识别能力不同; 在实践中, 要为每个链接长度内的节点确定“朋友”这意味着计算量极大. 本文使用的算法则是基于FOF的ROCKSTAR[16]的算法, ROCKSTAR可将数据分组, 并行处理, 自适应地调节链接长度, 极大地缩小了计算量, 其具体步骤为:

(1)为并行化运行, 根据输入数据信息, 将节点分为几批不同的组;

(2)对于每一组内的节点进行归一化处理, 给出一个自然的相空间度量;

(3)自适应选择一个相空间链接长度, 使70%的群节点以子群的形式链接在一起(当选择90%时, 算法会运行更多的时间, 并且还可能发现伪星团, 当选择50%时会致使没有聚类结果, 故一般选择70%)[16];

(4)对每个子组重复上述过程、重整化, 产生新的链接长度和子级别组;

(5)一旦找到所有满足要求的组(设置组的最小节点数阈值, 这里设置为20), 种子组就会被放置在最低级别的组上, 粒子就会被分层地分配到相空间中最近的种子组上;

(6)一旦节点被分配到组中, 未绑定的节点就会被移除, 最终会形成一个由不同级别的组组成的树状结构.

最后的目标组并不一定是末端的子组, 而是根据经验和输出数据选择合适的批次和子组.

3.2 STAR GO算法简介

STAR GO[16]是一种基于SOM[20]与自适应群识别相结合的无监督学习算法, 其核心是SOM的降维.

3.2.1 SOM

SOM是一种基于竞争学习的无监督机器学习神经网络, 用于数据可视化和将高维空间的数据映射到低维(通常是二维)空间, 同时保持输入数据的拓扑结构. 它通过将输入数据映射到一个由神经元组成的网格中, 实现对数据的聚类和可视化. 每个神经元都有一个权重向量, 表示该神经元在特征空间中的位置.

选择输入五维参数(X、Y、Z、pmra、pmdec), 将其映射到二维神经元地图上; 起点是构建$m \times m$像素的二维神经元图, m表示二维神经元地图的边长, 每个神经元在图上的坐标为(I, J), 随机赋予每一个神经元权重向量$\boldsymbol{\omega}$, 其大小与输入向量有相同的维数和范围, 给定一个输入向量$\boldsymbol{v}$, 对于数据中的第i颗星的输入向量$\boldsymbol{v}^i$, 通过寻找$|\boldsymbol{v}^i - \boldsymbol{\omega}|$最小的神经元, 确定与$\boldsymbol{\omega}$最接近的神经元, 这样的神经元被定为最佳匹配单元. 学习过程包括根据所有神经元到位于2D神经图第i颗神经元(I_i, J_i)的最佳匹配单元的距离$d^i_{I,J}$来改进第i个神经元$\boldsymbol{v}^i$的权重向量$\boldsymbol{\omega}$, 其中$d^i_{I,J}$定义为:

$$d^i_{I,J} = \sqrt{(I - I_i)^2 + (J - J_i)^2}\,. \tag{1}$$

第i个神经元的权重向量的变化量$\mathrm{d}\boldsymbol{w}^i_{I,J}$, 由下式给出:

$$\mathrm{d}\boldsymbol{w}^i_{I,J} = \alpha_q(\boldsymbol{v}^i - \boldsymbol{w}^i_{I,J}) \exp\left(-\frac{{d^i_{I,J}}^2}{\sigma_q^2}\right), \tag{2}$$

其中α_q表征学习率, σ_q控制第q次迭代中最佳匹配单元周围神经元的相邻影响. 由(1)式可以看出, 神经元权值的变化对其与最佳匹配单元的距离很敏感. 学习过程使用$\boldsymbol{v}^i$对数据集中的每颗恒星执行, 然后重复学习过程以进行总次数N的迭代. 对于第q次迭代, 对应的α_q和σ_q分别为:

$$\alpha_q = \alpha_0\left(1 - \frac{q}{N}\right), \tag{3}$$

$$\sigma_q = \sigma_0\left(1 - \frac{q}{N}\right), \tag{4}$$

其中$\alpha_0 = 0.3$和$\sigma_0 = \frac{1}{2}\max(m, n)$ (n表示输入数据的维度)的典型基准值, 类似于Geach[21]. 可以发现结果与α_0和σ_0无关, 在该基准值附近变化. 随着迭代次数q的增加, 由于α_q和σ_q的减小, 权值向量的变化量$|\mathrm{d}\boldsymbol{w}^i_{I,J}|$减小, 使得学习过程更加精细化. 当$|\mathrm{d}\boldsymbol{w}^i_{I,J}| \longrightarrow 0$时, 认为学习过程完成, 此时的权重向量即可看作是输入向量在二维空间中的表示.

3.2.2　自适应簇的识别

在将数据应用于SOM之后, 可以利用相邻神经元的权重向量($\boldsymbol{\omega}_{I,J}$)之间的差异来可视化聚类结构, 对位于($I$, J)的神经元的矩阵元素定义为:

$$\boldsymbol{u}_{I,J} = \lg(|\boldsymbol{\omega}_{I\pm1,J} - \boldsymbol{\omega}_{I,J}| + |\boldsymbol{\omega}_{I,J\pm1} - \boldsymbol{\omega}_{I,J}|)\,, \quad (5)$$

其中$\boldsymbol{u}$表示与神经元图对应具有$m \times m$个元素的矩阵, 每一个$\boldsymbol{u}$矩阵的元素对应一个神经元, 图2分别是3个星团对应$u_{I,J}$的可视化及对应的$m \times m$ ($m = 100$)二维神经元图, $\boldsymbol{u}$矩阵的所有元素值如颜色条所示. 每颗恒星都映射到二维神经图的最佳神经元, 同时每个神经元都可以与不止一颗恒星有关, 或者一颗恒星也没有.

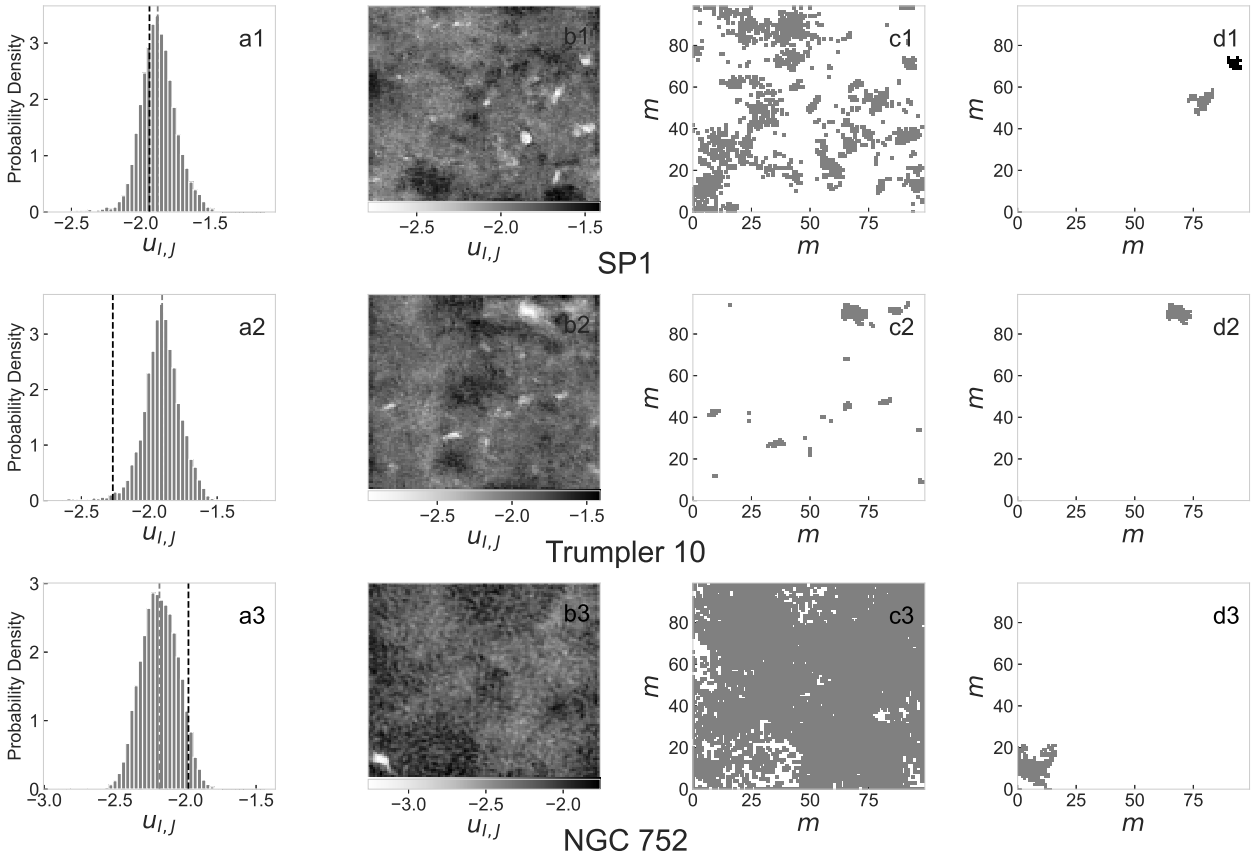

图 2　3个疏散星团在STAR GO中$u_{I,J}$的可视化. 子图(a1)、(a2)、(a3)是$u_{I,J}$的核密度估计直方图, 其中灰色虚线表示所有矩阵的中位数$u_{\rm m}$, 黑色虚线表示所取的阈值$u_{\rm thr}$ ($\boldsymbol{u}$矩阵中所有元素中位数与其3σ的差值); 子图(b1)、(b2)、(b3)是$u_{I,J}$的二维可视化图; 子图(c1)、(c2)、(c3)是当前阈值下的所有神经元图; 子图(d1)、(d2)、(d3)是当前阈值下$u < u_{\rm thr}$的神经元图.

Fig. 2　Visualization of three open star clusters in STAR GO $u_{I,J}$. Subgraphs (a1), (a2), (a3) are the kernel density estimate histogram of $u_{I,J}$, where the gray dashed line represents the median of all matrices $u_{\rm m}$, and the black dashed line represents the threshold taken $u_{\rm thr}$, which is the median of all elements in the $\boldsymbol{u}$ matrix with its 3σ; Subgraphs (b1), (b2), (b3) are two-dimensional visualizations of $u_{I,J}$; Subgraphs (c1), (c2), (c3) are all the neuron graphs under the current threshold, and subgraphs (d1), (d2), (d3) are the neuron graphs under the current threshold of $u < u_{\rm thr}$.

对于输入的每个向量的分量, 本研究计算$\boldsymbol{u}$矩阵2σ (σ为所有矩阵的标准差)的置信区间, 然后进行归一化, 再对归一化输入空间中的数据应用于SOM, 并从结果映射中计算u矩阵的值, 将每个源与其在地图中的最佳神经元关联起来. 接下来的步骤如下:

(1)设定阈值$u_{\rm thr}$, 初步设置$u_{\rm thr}=u_{\rm m}-3\sigma$其中$u_{\rm m}$为$\boldsymbol{u}$矩阵中所有元素的中值(如图2 (a1)、(a2)、(a3)灰色虚线所示);

(2)选取种子组, 选择最多数据点的簇作为种子组;

(3)将其余的簇标记为备选种子组;

(4)增加阈值, 通过逐步增加$u_{\rm thr}$的值, 最大化每个组的大小;

(5)确定阈值, 随着$u_{\rm thr}$的增加, 不可避免地会导致多个簇的合并, 当种子组与另一个组达到临界合并值时, 种子组即为最大化的目标组, 值为最终的$u_{\rm thr}$ (如图2 (a1)、(a2)、(a3)黑色虚线所示).

上面的簇识别算法可以让使用者自适应地找到给定数据集的子结构. 在步骤(4)中增加$u_{\rm thr}$的值是为了最大限度地提高分组恒星的完整性, 不过也同样增加了聚类的精度, 在针对不同类型的星团时, 可通过二维可视化神经元灰度图的亮度来适当调节阈值的大小.

3.3 DBSCAN算法简介

DBSCAN是一种常用的基于密度的聚类算法, 其核心要求是目标数据在原始样本中具有明显的密度区域. 该算法通常被应用于天体的分类和集群分析, 通过基于密度的概念, 能够识别出具有相似密度的天体集群, 并将它们分离出来形成聚类簇. 其步骤为:

(1)选择五维参数作为输入, 确定起始点;

(2)设定邻域半径以搜索邻域范围内的所有点;

(3)如果邻域内点的数量大于预设的阈值, 该点被视为核心点, 如果邻域内点的数量小于阈值, 则将其标记为非核心点, 并归入一个新的簇中;

(4)对于核心点, 遍历其邻域内的所有点, 并将其加入到同一簇中, 对于非核心点, 它们可能被归类为其他簇的边界点, 或者被视为噪声;

(5)重复对所有数据点进行核心点的确定, 最终得到所有分类的簇和噪声项.

在DBSCAN聚类中, 核心在于确定邻域半径和最小恒星数量阈值. 在本工作中, 首先确定最小恒星数量阈值, 然后设置邻域半径. 为确保聚类效果更加精确, 还对簇的数量进行设置. 以期望在对源进行聚类的过程中只出现一个目标簇, 但事实上其他簇可能对目标簇造成干扰. 通常根据经验设定最小恒星数量阈值ε, 来初步判断簇的数量, 如果有多个目标簇, 就要进行后续的判断.

4 算法分析

通过Gaia的数据, 本节使用了以上3种算法来对几种不同结构(在银经、银纬空间中, Trumpler 10[15–16]呈紧密的团状, 但带有延展的“小尾巴”, SP1[18]是由两个疏散星团组成的延展结构, 其中NGC 2232带有延展的“长尾巴”, 在日心银道直角坐标系中, NGC 752[13]具有清晰潮汐尾的结构)的疏散星团进行聚类, 以获取3种算法对于不同结构的聚类效果, 通过对成员星的确定, 来对每种聚类结果进行分析及对比, 其结果将在最后的表中展示.

4.1 算法实现

4.1.1 DBSCAN

在对星团的聚类过程中, 经验性的设置星团的最低恒星数目, DBSCAN会输出不同满足条件的簇, 所以对阈值ε的设置中要保持当簇只有一个, 且簇中恒星数量最大时, 阈值与簇如表1所示, 其中ε_1是最低阈值, 即有且只有一个簇的情况, ε_2是最高阈值, 即只有一个簇的情况且簇中恒星的数量达到最大的情况.

4.1.2 FOF

ROCKSTAR[15]会自动调整成员之间的链接长度, 并根据链接长度将其分成不同层次, 通常根据经验选择位于中间批次的组别. 随着子组的增加, 每层成员之间的关联关系会逐渐减弱. 对于Trumpler 10、SP1、NGC 752, 分别选择第8、6、4个组的数据作为聚类输出.

表 1　两个星团及延展结构的相关参数
Table 1　Parameters related to the two clusters and the extended structure

Cluster \ Parameter	(l,b)[1]	parallax/mas[2]	Age/Myr	$\mu_{\rm thr}$	ε_1[3]	ε_2[4]
Trumpler 10	$(262.86^\circ, 0.58^\circ)$	2.33	45	−2.26	17.51	24.12
SP1	$(212.47^\circ, -5.02^\circ)$	3.16	35	−1.99	40.82	42.26
NGC 752	$(136.95^\circ, -23.28^\circ)$	2.26	1.75×10^3	−2.05	16.02	16.67

[1] The center coordinates of the cluster in the heliocentric galactic coordinate system;
[2] Mean parallax;
[3] The DBSCAN algorithm identifies the lowest threshold for a cluster;
[4] The DBSCAN algorithm identifies the highest threshold for a cluster (only one cluster and the largest number of stars).

4.1.3　STAR GO

两个星团及延展结构的阈值归纳在表1中. 在确定阈值的过程中, 首先选择两倍σ的置信区间选择阈值$u_{\rm thr}$, 此时会输出u矩阵的二维可视化图像及所选区域的神经元图, 在增加阈值的过程中, 观察u矩阵二维可视化选中部分进行调节, 如图2从左至右每一列的顺序, 当图2 (d1)、(d2)、(d3)中的面积最大, 且不与其他区域相连时, 簇的恒星数量达到最大.

4.2　聚类星团的真伪判断

在星团的CMD中, 通过主星序的宽窄在一定程度上可以确定聚类星团是否可靠. 本文用参数值r_n[5]来描述星团主星序的宽窄程度, 一般来说, 一个真实星团的主星序带倾向于狭窄, 而伪星团的主星序则表现为宽, r_n的值越小, 主星序带越窄, r_n的定义为:

$$r_n = \left|\frac{v_1}{v_2}\right|, \tag{6}$$

其中v_1与v_2是协方差矩阵$\boldsymbol{M}$的两个特征值, 且$|v_1| < |v_2|$, 协方差矩阵$\boldsymbol{M}$的定义为:

$$\boldsymbol{M} = \begin{pmatrix} \overline{x_i x_i} & \overline{x_i y_i} \\ \overline{x_i y_i} & \overline{y_i y_i} \end{pmatrix}, \tag{7}$$

其中x_i表示第i颗星的色指数与所有恒星色指数平均值之差, y_i表示第i颗星的视星等与星团平均视星等的平均值之差, 其值为$x_i = (\mathrm{G_{BP}} - \mathrm{G_{RP}})_i - \overline{\mathrm{G_{BP}} - \mathrm{G_{RP}}}$, $y_i = G_i - \overline{G}$. 由于光度不确定度扩大了暗端质谱, 为消除污染, 这里选择G小于18的星进行r_n值的计算, r_n值如表2所示, r_n值均较小(一般r_n的阈值为0.1), 这表明本研究的聚类结果是可靠的.

表 2　各星团及聚类算法对应的r_n值($G < 18$)
Table 2 The r_n value of each cluster and clustering algorithm ($G < 18$)

Algorithm \ Cluster	Trumpler 10	SP1	NGC 752
DBSCAN	0.01871	/[1]	0.03248
FOF	0.00437	/	0.01214
STAR GO	0.00845	/	0.01801

[1] Since SP1 consists of two open clusters, the r_n value was not calculated

4.3　星团成员星的确定

通过聚类算法对疏散星团聚类之后, 针对3个星团, 应用3种不同的筛选条件. 在对Trumpler 10的筛选中, 选择年龄在5∼120 Myr[18] (图3是3种算法在对3个星团聚类过程中通过等年龄线筛选前后的CMD, 图3的第1、2行中等年龄线的金属丰度等相关参数与星团的最佳拟合年龄一致)的数据, 再把pmra和pmdec限制在2σ内, 得到算法

对星团的聚类结果; 在对NGC 752的初步筛选中, NGC 752的最佳拟合年龄为1.75 Gyr左右, 因为NGC 752属于年老的疏散星团, 不适合用等年龄线进行筛选, 由于在主星序上方0.75 mag位置是不可分辨双星, 故而这里设置在最佳拟合等年龄线的1 mag范围(如图3 (a3)、(b3)、(c3)所示)进行筛选, 接着设置筛选pmra和pmdec在1σ范围内的源. 而对于延展结构SP1, 仅用5∼120 Myr[18]的等年龄线(与星团的最佳年龄拟合线的金属丰度Z同为0.02)进行筛选, 如图3 (a2)、(b2)、(c2)所示. 表3则是3个疏散星团的初步聚类结果、筛选过程及星团恒星的数量, 图4则显示了星团聚类最终结果的视向速度图, 由此也可以判断出图中所属恒星很大概率为成员星, 不过由于含有金属丰度恒星很少且误差较大, 故而未进行金属丰度的认定.

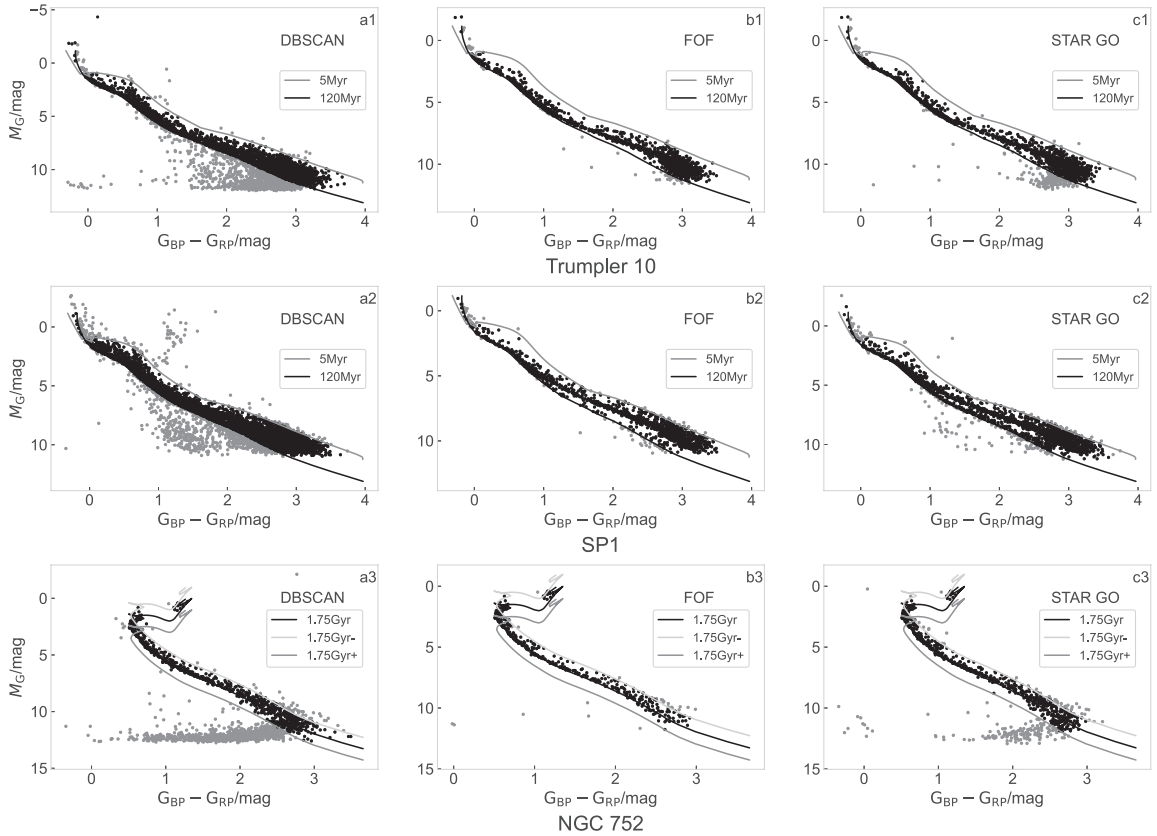

图 3 星团的初步筛选图. Trumpler 10和SP1均选择5∼120 Myr范围内的星, NGC 752则选择最佳拟合等年龄线1 mag范围内的星. 其中灰色点表示算法输出的聚类结果, 黑点表示通过等年龄线筛选之后的部分. 第3行中的1.75 Gyr+与1.75 Gyr−, 分别表示在1.75 Gyr年龄线上绝对星等变化1 mag之后的曲线.

Fig. 3 Preliminary screening of star clusters. Both Trumpler 10 and SP1 select stars in the range of 5∼120 Myr, while NGC 752 selects stars in the range of 1 mag that best fit the isoage line. The gray dots represent the clustering results output by the algorithm, and the black dots represent the parts filtered by the equal age line. The 1.75 Gyr+ and 1.75 Gyr− in the third row represent the curve after a 1 mag change in absolute magnitude on the 1.75 Gyr age line, respectively.

4.4 同类算法对比

Liu等[5]运用了另一种基于FOF的Star cluster Hunting Pipeline (SHIP)聚类方法, 并识别出2443个星团, 图5是本文对Trumpler 10和NGC 752基于FOF算法和DBSCAN算法的聚类结果与同类算法的聚类结果对比图, 图6是本文对NGC 2232和Tian

2基于FOF算法和DBSCAN算法的聚类结果与同类算法的聚类结果对比图. 图5 (a)、(b)展示了本文所采用基于FOF的ROCKSTAR算法与Liu等[5]基于FOF的SHIP算法进行的星团聚类对比图. 可以明显观察到, 本研究基于FOF的ROCKSTAR算法在星团成员星完备性方面优于Liu等人的算法, 星团恒星数量如表4所示, 为便于比较, 均选择M_G小于18的恒星. 通过表4的数据和图5 (a)、(b), 可以看出ROCKSTAR在对Trumpler 10的聚类方面优于Liu等[5]; 而对NGC 752这一潮汐结构的聚类则各有优劣, 特别是主星序上方的密近双星部分, 两者互有缺失. 而SP1属于延展结构, SP1的NGC 2232和Tian 2都是单独存在于星表中, 为方便比较, 对于NGC 2232和Tian 2本文分别选择在星团半径为1.205°和2.397°内的源, 其对比结果与Trumpler 10和NGC 752的对比结果类似, CMD如图6所示.

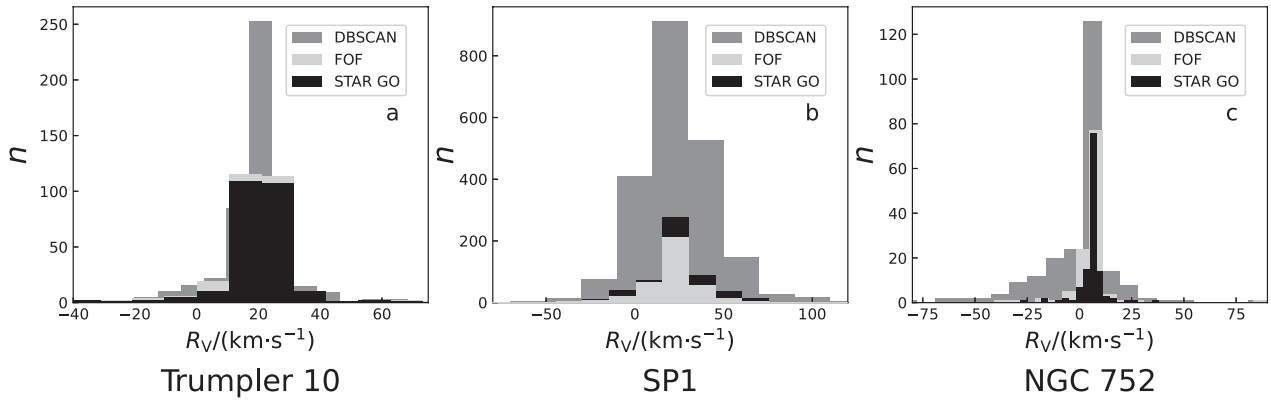

图 4　星团成员星的视向速度图

Fig. 4　Radial velocity map of cluster member stars

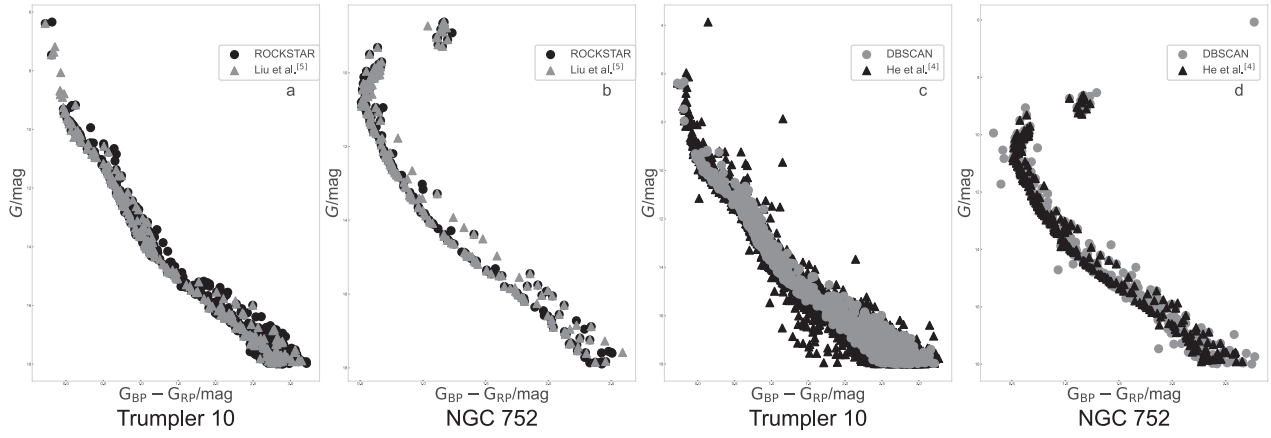

图 5　同类算法聚类对比图(Trumpler 10 & NGC 752). 图(a)、(b)是本文基于FOF算法的ROCKSTAR算法与Liu等[5]基于FOF的SHIP算法聚类结果的CMD. 图(c)、(d)是本文的DBSCAN算法与He等[4]改进的DBSCAN算法聚类结果的CMD.

Fig. 5　Cluster comparison graph of similar algorithms (Trumpler 10 & NGC 752). Panels (a) and (b) present the CMDs of clustering results obtained by the ROCKSTAR algorithm based on the FOF method in this study and the SHIP algorithm based on the FOF method by Liu et al.[5], respectively. Panels (c) and (d) present the CMDs of clustering results from the DBSCAN algorithm in this study and the improved DBSCAN algorithm of He et al.[4], respectively.

表 3 3个疏散星团筛选步骤的恒星数量

Table 3 The number of stars in the cut step for the three open clusters

Algorithm \ parameter	Cluster	Step 1[1]	Step 2[2]	Step 3[3]
Trumpler10	DBSCAN	4024	2632	1797
	FOF	1213	1151	952
	STAR GO	1408	1178	978
SP1	DBSCAN	6632	5080	/[4]
	FOF	1335	1254	/
	STAR GO	1873	1668	/
NGC 752	DBSCAN	1582	722	445
	FOF	393	368	190
	STAR GO	808	600	285

[1] Preliminary clustering results;

[2] Equal age wire cutting; The filter for Trumpler 10 is 5 ∼ 120 Myr and the filter for NGC 752 is the best fitting age for the cluster ± 1 mag;

[3] Proper Motion cutting the pmra and pmdec screening range of Trumpler 10 is (μ, $\pm 2\sigma$), the pmra and pmdec screening range of NGC 752 is (μ, $\pm 1\sigma$);

[4] In SP1, proper motion screening was not done because there were two open clusters.

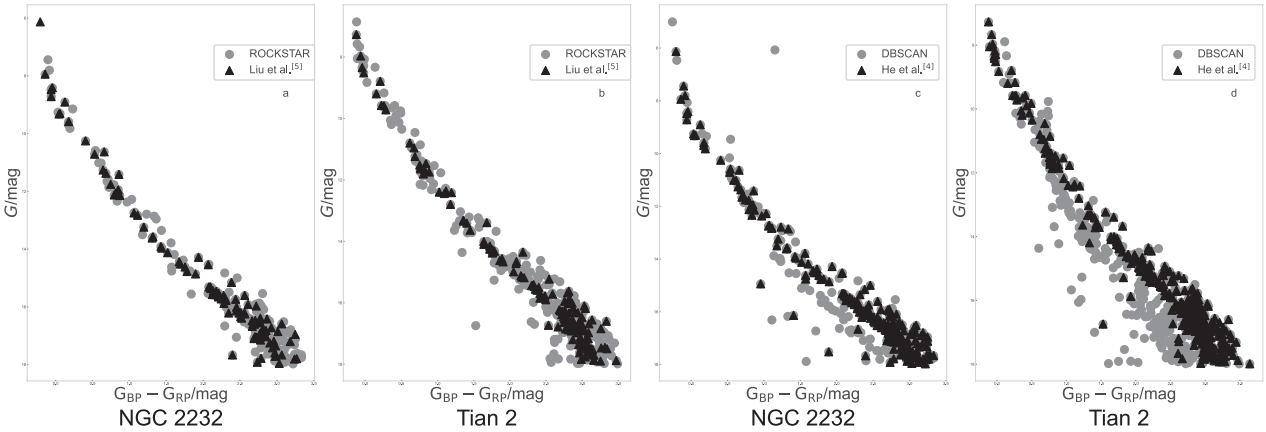

图 6 同类算法聚类对比图(SP1). 图(a)、(b)是本文基于FOF算法的ROCKSTAR算法与Liu等[5]基于FOF的SHIP算法聚类结果的CMD. 图(c)、(d)是本文的DBSCAN与He等[4]改进的DBSCAN 算法聚类结果的CMD. 最后一个子图主星序下方之所以会出现灰色部分, 是因为此时的聚类结果没经过筛选, 在筛选之后下方突出的灰色主星序消失.

Fig. 6 Clustering comparison plot for similar algorithms (SP1). Panels (a) and (b) present the CMDs of clustering results obtained by the ROCKSTAR algorithm based on the FOF method in this study and the SHIP algorithm based on the FOF method by Liu et al.[5], respectively. Panels (c) and (d) present the CMDs of clustering results from the DBSCAN algorithm in this study and the improved DBSCAN algorithm of He et al.[4], respectively. The gray portion below the main sequence in the last subplot appears because the clustering results have not been filtered. After filtering, the prominent gray main sequence at the bottom disappears.

表 4 与星表聚类算法结果比较表($G < 18$)

Table 4 Comparison table with the results of the star list clustering algorithm ($G < 18$)

Algorithm \ Cluster	Trumpler 10	NGC 752	NGC 2232	Tian 2
Liu et al.[5]	271	190	103	395
ROCHSTAR	635	183	220	147
He et al.[4]	3310	285	276	548
DBSCAN	1143	411	219	358

图5 (c)、(d)是He等[4]基于DBSCAN的改进算法与本文的DBSCAN算法的聚类对比图. 在He等[4]的聚类中, 去掉了$M_{\rm G}$大于18的微弱源, 通过恒星的位置和自行做为聚类的原始参数, 本文也运用了恒星的位置和自行作为聚类的输入向量. 如图5 (c)、(d)和图6 (c)、(d)所示, 除在Trumpler 10的聚类中He等[4]的聚类数量多于本研究的DBSCAN数量之外, 对于NGC 752和NGC 2232的聚类数量均相当, 在图6 (d)中, DBSCAN出现两条主星序, 这是由于这里对比使用的DBSCAN数据是未经筛选直接聚类输出的聚类结果数据, 在通过年龄和自行的筛选之后, 下面灰色部分消失(为场星).

4.5 本文算法之间的交叉对比

3种算法对3个疏散星团的聚类最终结果交叉图详见图7, 该图展示了对于同一星团, 不同算法的聚类结果交叉情况("D F S"是3种算法的交叉部分, 而其他3种颜色是对应算法与另外两种算法交集之外的补集). 通过视差图观察, 可以推断在3种算法中, 聚类的最终结果所展示的恒星很大概率是星团的成员星, 表明本研究中的算法对星团的聚类是成功的, 亦表明3种算法最终所得到恒星很大概率是星团的成员星. 表5是3种算法对3个疏散星团的聚类量化表, 与图7对应, 详细展示了各算法所识别的成员星数量, 可以发现DBSCAN算法聚类的成员星数量最多, STAR GO的成员星数量次之, FOF所识别的成员星最少.

表 5 3种算法对3个疏散星团的聚类量化表

Table 5 The quantitative clustering table of three clustering algorithms for three open star clusters

Algorithm \ Cluster	Trumpler 10				SP1				NGC 752			
Join Type	T[1]	X[2]	C[3]	Time[4]	T	X	C	Time	T	X	C	Time
DBSCAN	1797	692	1105	0–5	5080	1050	4030	0–5	554	134	420	0–5
FOF	952	692	260	5–10	1254	1050	204	5–10	162	134	28	5–10
STAR GO	978	692	286	316	1668	1050	618	749	281	134	147	126

[1] The algorithm corresponds to the total number of stars in the cluster;

[2] The number of stars at the intersection of the three algorithms;

[3] The algorithm corresponds to the number of complementary stars in the cross section of the cluster;

[4] The running time here refers to the running time (in unit of minutes) of the entire process after adjusting the threshold.

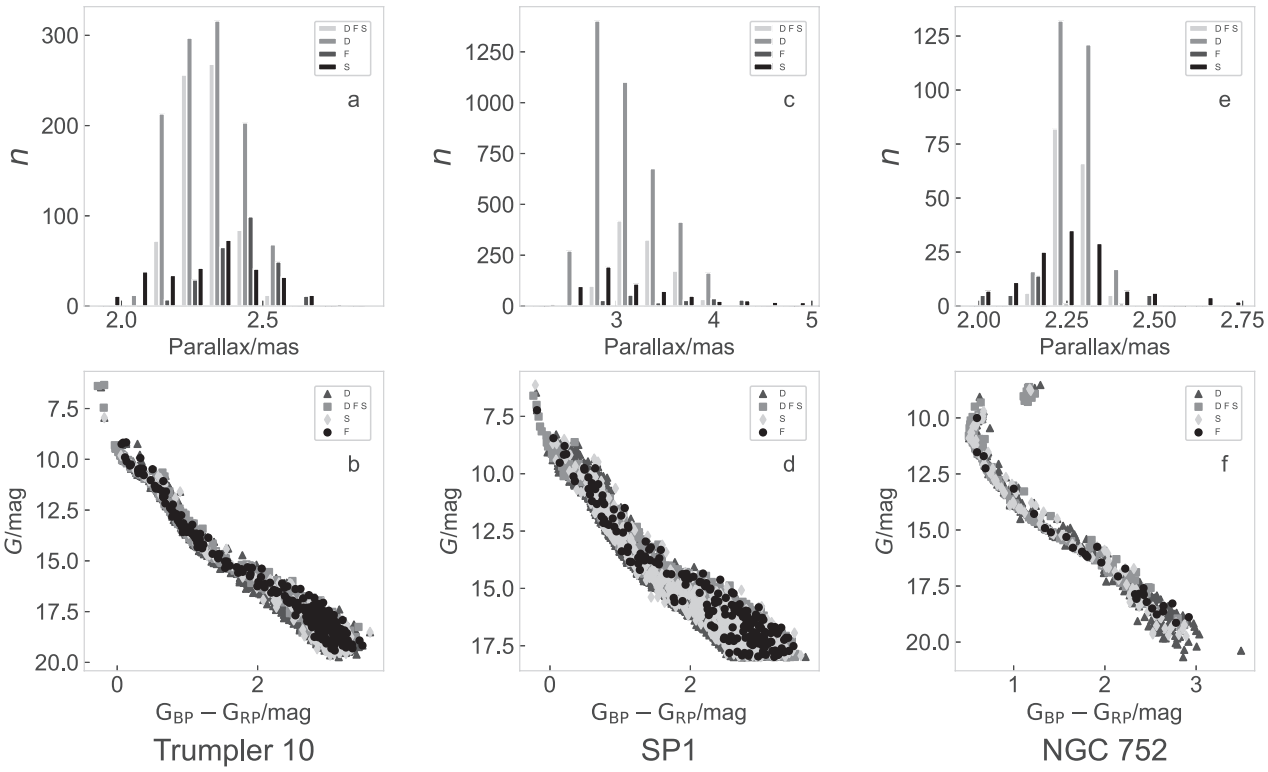

图 7 不同算法的3个星团聚类交叉结果. 这3列分别是3种算法对Trumpler 10、SP1、NGC 752聚类结果作交叉比对的视差直方图及CMD.

Fig. 7 The cross-results of the clustering algorithms in the three star clusters. The three columns present the parallax histograms and CMD for the intersection of clustering results from Trumpler 10, SP1, and NGC 752 obtained using the three algorithms, respectively.

图8是3种算法与3个结构的聚类结果空间分布图, 其中图(a)是3种算法对Trumpler 10的聚类结果在赤经赤纬中的展示, 图(b)、(c)分别是3种算法对NGC 752的聚类结果在赤经赤纬和直角坐标空间中的展示, 图(d)、(e)、(f)分别是3种算法对SP1的聚类结果在银经银纬中的展示. 在图8 (a)、(b)、(c)、(d)中, 均呈现出DBSCAN的空间团状密度聚类特征, 从图8 (d)可以看出对于SP1中的NGC 2232的“长尾巴”部分有较为模糊的尾巴部分, 而对于同样有延展的Trumpler 10 “小尾巴”部分则没有在图8 (a)中体现出来, 同样对于NGC 752的潮汐尾部分, DBSCAN亦没识别出来; 通过图8 (a)、(b)、(c)、(e)、(f)均可发现FOF与STAR GO识别出了星团的特殊结构.

5 结论与讨论

基于Gaia DR3数据的恒星参数, 本研究利用DBSCAN、FOF、STAR GO这3种算法完成了对3种特殊结构的聚类, 在研究中, Trumpler 10与SP1都是年轻星团, 含有许多小质量恒星, 综合其他物理原因, 导致分别呈现较高聚集度和延展结构的特征; 而NGC 752的年龄较大, 小质量恒星可能通过质量分离或被抛出星团而减少, 导致潮汐尾非常明显, 在图8 (d)、(e)、(f)这3幅图中, 这种表现尤为明显.

由于DBSCAN是基于密度的聚类算法, 使其结果呈现出空间团状分布特征, 在增加成员星识别的同时, 损失了对特殊结构识别的性能.

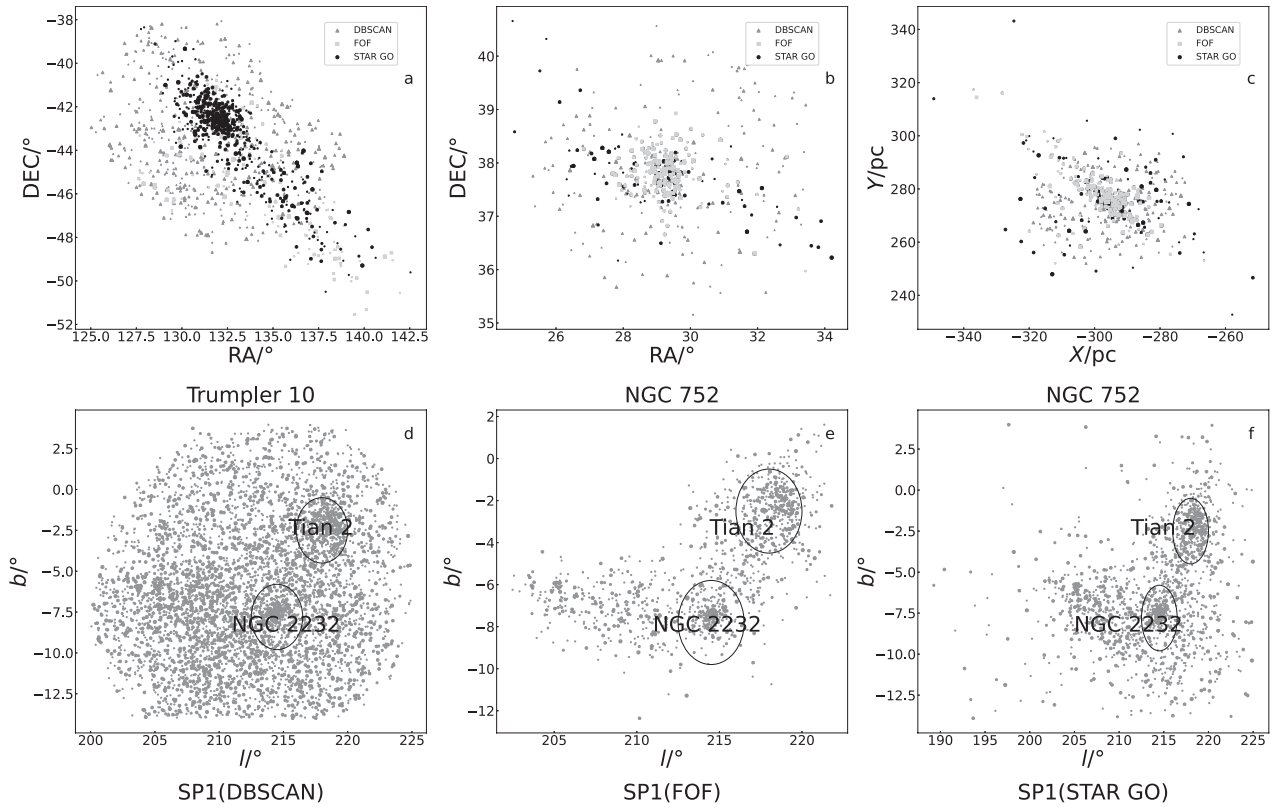

图 8　3种算法与3个结构的聚类结果空间分布图. 图(a)显示了3种算法对Trumpler 10的聚类结果, 其中, FOF和STAR GO算法聚类结果的延展部分更突出. 图(b)和图(c)是3种算法对NGC 752的聚类结果, 疏散星团NGC 752的潮汐结构在图(c)比在图(b)中更明显. 图(d)、(e)、(f)是SP1对应3种算法的聚类结果, 图中圆圈是延展结构的星团部分, 星团的延展部分在图(e)和图(f)中更为明显. 这里将恒星的绝对星等分为4个等级(<3, 3∼6, 6∼8, $\leqslant 8$), 散点越大视星等越低.

Fig. 8　Spatial distribution of clustering results of three algorithms and three structures. Panel (a) shows the clustering results of Trumpler 10 by three algorithms, among which the extended part of the clustering results of FOF and STAR GO algorithms is more prominent. Panels (b) and (c) are the clustering results of the three algorithms for NGC 752. The tidal structure of open cluster NGC 752 is more obvious in panel (c) than in panel (b). Panels (d), (e), and (f) are the clustering results of SP1 corresponding to the three algorithms. The circle in the figure is the cluster part of the extended structure, and the extended part of the cluster is more obvious in panel (e) and panel (f). Here the absolute magnitude of the star is divided into four grades (<3, 3∼6, 6∼8, $\leqslant 8$), the greater the scatter, the lower the apparent magnitude.

在基于FOF的ROCKSTAR算法的聚类中会发现有些层输出簇的个数与被标记数目不同, 这就表明FOF已自动筛选掉一部分, 排除了连接空间之外的源, 并将每组数据自动按不同连接长度进行识别, 使输出极为简单, 这是FOF的一大优点. 同时, 虽然FOF对星团成员星的数量识别结果很少, 但相比于另外两种算法, FOF对星团的特殊结构识别却毫不逊色.

相比于DBSCAN和FOF, STAR GO中涉及到五维向量的SOM, 代码运行所需要的时间较长[5], 且运行时间随着源的增加而增加(如表5); FOF和STAR GO算法在潮汐结构和延展结构识别方面各有优劣. FOF的聚类结果对潮汐结构和延展结构的聚类更清晰, 尤其是对Trumpler 10及SP1中NGC 2232附近的“尾巴部分”, 而STAR GO聚类的成员星数量相对较多.

[5]所用电脑的CPU为12 th Gen Inter (R) Core (TM) i5-12450H, 核心数: 8, 线程数: 12, 基准频率: 2.00 GHz, 3级缓存: 12 MB, 内存: 16 GB DDR5 4800 MHz, 硬盘: 477 GB SSD, 显卡: NVDIA GeForce RTX 3050 Laptop GPU.

综上所述, 图8展示的疏散星团空间结构中, FOF显示出对星团形态聚类的精度较高, 图8 (e)、(f)显示STAR GO星团形态聚类的灵敏度稍弱于FOF, 而DBSCAN仅适用于空间分布较集中的疏散星团聚类. 再考虑3种算法的时间效率, FOF时间效率最高, 而STAR GO时间效率最差. 总之, 在对星团进行未知识别时, 基于FOF的ROCKSTAR算法在星团的成员星和星团的特殊类型聚类方面都较为稳定, 为更大限度地增加成员星则可使用DBSCAN算法, 而STAR GO方法则介于两者之间.

致谢 本研究利用了欧洲航天局盖亚卫星(https://www.cosmos.esa.int/gaia)的数据.

参 考 文 献

[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: 871
[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: 45
[9] Castro-Ginard A, Jordi C, Luri X, et al. A&A, 2022, 661: 118
[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: 109
[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

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

XIONG Zhuang[1] ZHANG Peng[1,2] YANG Xiang-ming[1] LIU Gao-chao[1,2] 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 the field of astronomy. Due to the complexity of the formation and evolution of open cluster, there is no unified method to accurately identify the member stars in open cluster. The objective is to select five dimensional parameters of the position and motion of stars from three different spatial distribution types of open star clusters. This study evaluates 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 the FOF and STAR GO algorithms are more suitable for clusters with special structure, and can identify the tidal or extended structure of the cluster, while DBSCAN can identify the member stars in the core region of the cluster more completely. The aim is to find a more balanced algorithm strategy between the details of the cluster structure and the integrity of the member star recognition.

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