第58卷 第5期 天 文 学 报 Vol.58 No.5
2017年9月 ACTA ASTRONOMICA SINICA Sept., 2017

doi: 10.15940/j.cnki.0001-5245.2017.05.005

基于DBSCAN聚类算法的疏散星团NGC 188的3维运动学成员判定*

高新华† 王 超 顾晓清 徐守坤

(常州大学信息科学与工程学院 常州 213164)

摘要 利用疏散星团NGC 188所在天区的1046颗恒星样本的高精度3维(3D)运动学数据(自行和视向速度)测试了DBSCAN (Density-Based Spatial Clustering of Applications with Noise)聚类算法的成员判定效果. 为了避免自行和视向速度的单位不一致带来的影响, 在数据预处理阶段将3个分量的数据统一标准化至[0, 1]区间. 利用第k个最近邻点距离方法分析了1046颗恒星样本在标准化无量纲3D速度空间的分布特征, 再根据第k个最近邻点距离随k值的变化趋势确定了DBSCAN聚类算法的输入参数(Eps, MinPts), 最后利用DBSCAN聚类算法分离出497颗3D运动学成员星. 分析结果表明得到的3D运动学成员星是可靠的.

关键词 疏散星团和星协: 个别: NGC 188, 恒星: 运动学与动力学, 技术: 视向速度, 赫罗图与颜色-星等图

中图分类号: P144; **文献标识码**: A

1 引言

疏散星团是研究恒星演化和银河系结构的重要天体, 原因是疏散星团的年龄、距离、红化、金属丰度等重要参数相对场星而言更容易确定[1−2]. 但疏散星团属于星族I, 绝大多数疏散星团都分布在银盘附近, 场星污染严重, 因而可靠的成员判定是研究疏散星团的前提. 目前被广泛使用的成员判定方法—Vasilevskis-Sanders方法(简称VS方法)是由Vasilevskis等[3]首先提出, 再由Sanders加以改进[4], Zhao等[5]又将VS方法推广至能处理不等精度的自行数据. 从数据挖掘的角度看, VS方法是基于混合高斯模型的聚类方法[6], VS方法假设成员星和场星的自行或视向速度的分布均符合高斯分布, 再通过极大似然法求解出高斯模型中的分布参数, 最后计算出每一颗恒星的成员概率, 通常认为成员概率高的恒星属于星团的可能性也大[5]. VS方法在数学上是严格的且能算出每一颗恒星的成员概率, 但一些学者指出VS方法有不容忽视的局限性[7−8], 当成员星数目远小于场星时可能统计效果不好, 当成员星和场星在速度空间重合时效果也不好. 此外, 利用极大似然法求解分布参数时可能会陷入局部极大值[9]. 我们注意到VS方法适合处

2017-03-08收到原稿, 2017-04-11收到修改稿

*国家自然科学基金项目(11403004)资助

†xhgcczu@163.com

理1维视向速度数据或2维自行数据, 但不适合处理3维(3D)及以上的高维数据, 原因是高维高斯模型需要引入更多的分布参数.

针对VS方法存在的这些问题, 我们尝试利用不依赖模型(无参数)的DBSCAN (Density-Based Spatial Clustering of Applications with Noise)聚类算法[10]在疏散星团NGC 188和NGC 6819的3D速度空间(自行和视向速度)中进行了成员判定, 并且获得了纯净的3D运动学成员[11−12]. 此外, 我们还利用DBSCAN聚类算法对疏散星团NGC 6791的2D颜色-星等数据进行了成员判定, 发现了分裂的主序结构[13−14]. 但我们没有讨论如何确定DBSCAN聚类算法的输入参数(Eps, MinPts), 而仅仅是基于个人经验得到了这两个输入参数, 有可能会导致遗漏部分成员星或混入部分场星. 另一个重要问题是视向速度和自行的单位显然是不同的, 而DBSCAN聚类算法的Eps参数是基于欧氏距离的, 不同单位的数据混用的结果可能是某一分量的数据(例如视向速度)在欧氏距离的计算中占主导地位, 从而会影响聚类效果. 本文的主要目的是尝试解决DBSCAN聚类算法在成员判定时存在的数据单位不同和输入参数问题. 我们的思路是把自行和视向速度均标准化到[0, 1]之间, 再利用第k个最近邻点距离(the k-th Nearest Neighbor Distance, 简称kNND)法估算标准化3D速度空间的(Eps, MinPts)输入参数. 我们选择老年疏散星团NGC 188 ($l = 00^{\rm h}47^{\rm m}28^{\rm s}$, $b = +85°15'18''$)作为测试样本, 原因是WIYN Open Cluster Survey (WOCS)计划专门针对NGC 188进行了观测, 获得了大量高精度的测光、自行和视向速度数据.

2 数据和方法

2.1 数据

我们的方法测试需要高精度自行和视向速度数据, Platais等[15]将照相底片和CCD数据相结合, 获得了NGC 188所在天区0.75 deg^{2}范围内视星等最大到$V_{\rm mag}$ =21 mag的7812颗恒星的自行数据, 其中亮于16.5 mag的恒星自行精度好于1.5 mas$\cdot$yr^{-1}, 部分恒星的自行精度甚至达到0.15 mas$\cdot$yr^{-1}. Geller等[16]对距NGC 188中心约30′范围内的1046颗恒星进行了重复光谱观测, 获得了精度好于1 km$\cdot$s^{-1}的视向速度数据, 恒星的星等覆盖范围为12$\leqslant V_{\rm mag} \leqslant$16.5 mag. Geller等[16]的视向速度星表中的1046颗恒星在Platais等[16]的自行星表中均有相对应的自行数据, 这样一共1046颗具有高精度3D运动学数据的恒星样本可用于我们的方法测试. 图1显示了这1046颗恒星样本在3D速度空间中的分布情况, 我们注意到自行和视向速度的单位是不同的, 并且视向速度的覆盖范围大致为−300–100 km$\cdot$s^{-1}, 约为赤经和赤纬方向的自行覆盖范围的4倍. 我们曾用DBSCAN聚类算法对NGC 188进行过3D成员判定[11], 但未考虑自行和视向速度单位不同的重要因素. 通常在聚类分析之前需要对数据进行标准化处理, 目的是将不同单位的属性值统一起来, 避免较大值域的属性值主导聚类结果[17]. 我们拟采用以下公式[17]将3D运动学数据的3个分量均标准化至[0, 1]区间:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}, \tag{1}$$

其中, x是原始数据, $\min(x)$和$\max(x)$分别是原始数据的最小和最大值, x'是标准化后的无量纲值. 标准化后的1046颗恒星样本在无量纲3D速度空间的分布情况如图2所示.

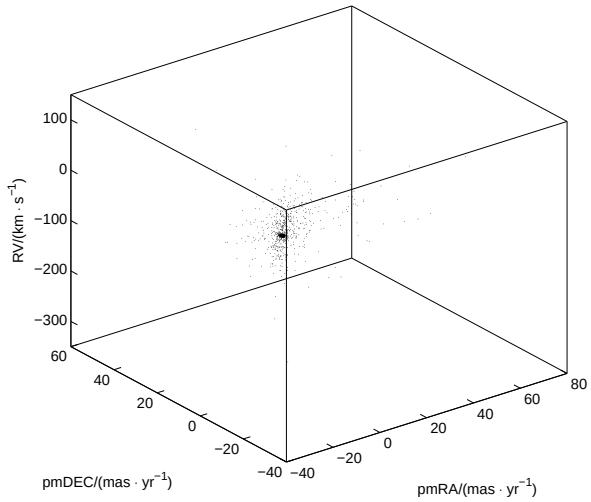

图 1　1046颗恒星样本在3D速度空间(自行和视向速度)的分布. pmRA和pmDEC分别为赤经和赤纬方向自行; RV为视向速度.

Fig. 1　The distribution of the 1046 sample stars in the 3D velocity space (proper motion and radial velocity). pmRA and pmDEC indicate, respectively, proper motions in the right ascension and declination directions; RV indicates radial velocity.

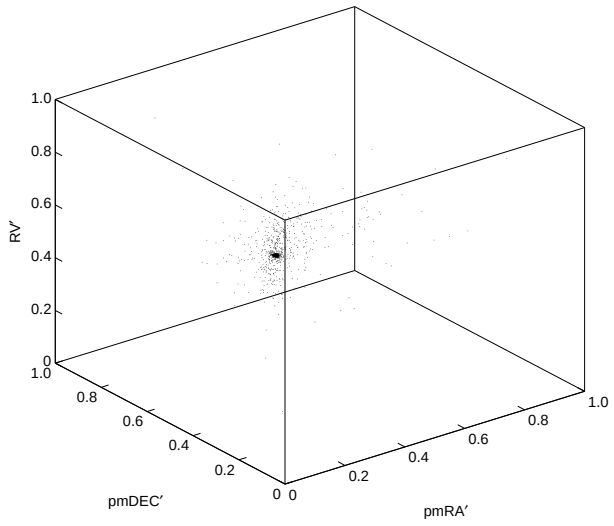

图 2　1046颗恒星样本在标准化3D速度空间的分布. pmRA′和pmDEC′分别为标准化后的赤经和赤纬方向的自行; RV′为标准化后的视向速度.

Fig. 2　The distribution of the 1046 sample stars in the standardized 3D velocity space. pmRA′ and pmDEC′ indicate, respectively, standardized proper motions in the right ascension and declination directions; RV′ indicates standardized radial velocity.

2.2 第k个最近邻点距离分析

(Eps, MinPts)是DBSCAN聚类算法的两个输入参数, 对聚类效果有重要影响. DBSCAN聚类算法的提出者Ester等[10]建议通过统计分析数据集中的点的第k个最近邻距离(kNND)估算DBSCAN的两个输入参数, 并且他们在2D数据集中证实了这种方法的可行性. 为了准确估计这两个输入参数, 我们也利用kNND方法分析成员星和场星在标准化3D速度空间(图2)里的不同分布特征. kNND方法的原理很简单, 计算N颗恒星样本中的第i颗参考恒星与第j颗恒星之间的欧氏距离$\boldsymbol{D}(i,j)$ $(i \neq j)$:

$$\boldsymbol{D}(i,j) = \sqrt{\sum_{n=1}^{3}(x_{in}-x_{jn})^2}, \tag{2}$$

其中, n是数据维数, x_{in}、x_{jn}分别是第i颗参考恒星与第j颗恒星的第n维(n= 1, 2, 3)标准化值, 可获得长度为$N-1$的$\boldsymbol{D}(i,j)$距离序列, 将距离序列升序后很容易获得第i颗参考恒星的kNND值($k = 1, 2, 3, \cdots, N-1$)[18]. 考虑到Ester等[10]并未从理论上分析kNND方法的原理, 并且他们仅在2D数据集上进行了测试, 我们打算从理论上进一步说明3D数据集中kNND法的工作原理. 为了简单起见, 假设星团成员星的总数为N, 在标准化3D速度空间里均匀地分布在有限体积V内, 则任意一颗成员星及其第k个最近邻点所占据的体积约为kV/N ($k \ll N$), 点密度ρ、第k个最近邻点距离D_k以及随k的变化率$\Delta D_k/\Delta k$可用下列公式估算:

$$\rho = \frac{N}{V}, \tag{3}$$

$$D_k \simeq \left(\frac{k}{\rho}\right)^{1/3}, \tag{4}$$

$$\frac{\Delta D_k}{\Delta k} \simeq \frac{1}{3}\rho^{-1/3}k^{-2/3}. \tag{5}$$

(5)式表明成员星的第k个最近邻点距离随k的增加而缓慢增加, 原因是成员星可以看作是3D速度空间中被稀疏场星包围的致密结构. 如果将(5)式用于分析场星, 则不难发现第k个最近邻点距离也随k的增加而增加, 只是增加的幅度要明显大于成员星, 原因是场星的点密度比成员星小得多. 图3表明可以通过逐步增加k值来区分成员星和场星($k = 1, 2, 3, \cdots$), 当k =5时, 成员星的D_k因为增加缓慢还集中在第一个Bin内(这里的Bin是指图3直方图的带宽或窗宽, Bin=0.01), 此时5NND$\leqslant$0.01的恒星样本可以看作是成员星.

2.3 3D运动学成员星

我们发现: 根据kNND的分析结果(图3)可以取(Eps, MinPts) = (0.01, 6)作为DBSCAN聚类算法的输入参数, 取MinPts=6是考虑了恒星自身以及它的5个最近邻恒星之和. 根据DBSCAN聚类算法的原理, 如果任意一颗恒星的0.01倍半径范围内有6颗恒星存在(包含恒星自身), 那么这颗参考恒星就是处于高密度区域的核心点, 它极有可能是星团成员星. 在DBSCAN聚类算法中输入参数(Eps, MinPts) = (0.01, 6)后共获得497个核心点(成员星), 这与Geller等[16]使用VS方法所得到的473颗成员星较吻合.

图4显示了成员星和场星在3D速度空间(自行和视向速度)中的分布情况, 497颗成员星均位于高密度区域. 颜色-星等图(图5)和空间位置分布(图6)也均表明由DBSCAN聚类算法得到的这497颗3D运动学成员星是可靠的, 因为成员星的颜色-星等图清晰地显示了主序、主序拐点、红巨星支以及主序拐点左上方的蓝离散星区, 而成员星的空间分布则明显地表现出向星团中心聚集的趋势. 通过与我们先前工作的比较[11], 我们发现本文中的497颗成员星包含了先前工作中的472颗成员星, 而图7的颜色-星等图表明本工作得到的额外的25颗成员星中绝大多数应该是成员星.

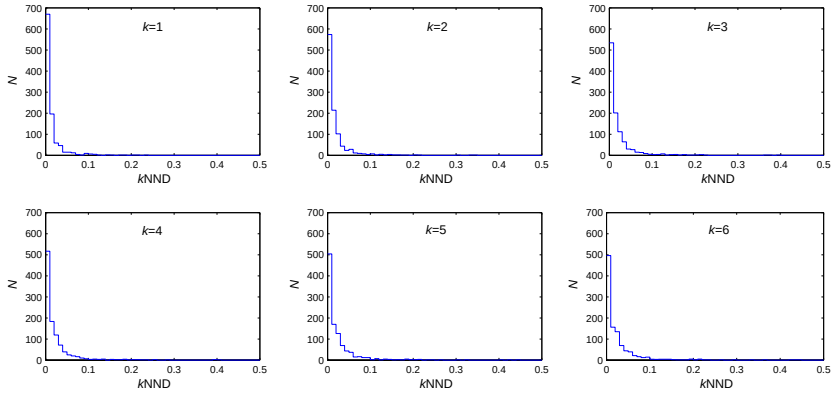

图 3　1046颗恒星样本在标准化3D速度空间的第k个最近邻距离分布(k=1,2,3,4,5,6)

Fig. 3　The kNND distribution of the 1046 sample stars in the standardized 3D velocity space (k=1,2,3,4,5,6)

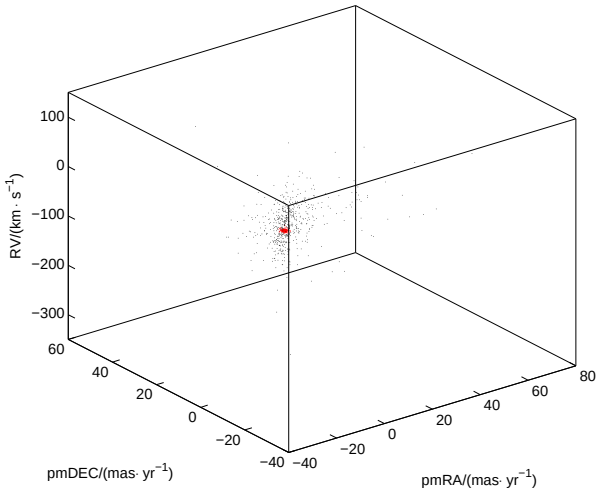

图 4　497颗3D运动学成员星(红点)和场星(黑点)在3D速度空间(自行和视向速度)的分布情况

Fig. 4　The distribution of the 497 3D kinematical members (red dots) and field stars (black dots) in the 3D velocity space (proper motion and radial velocity)

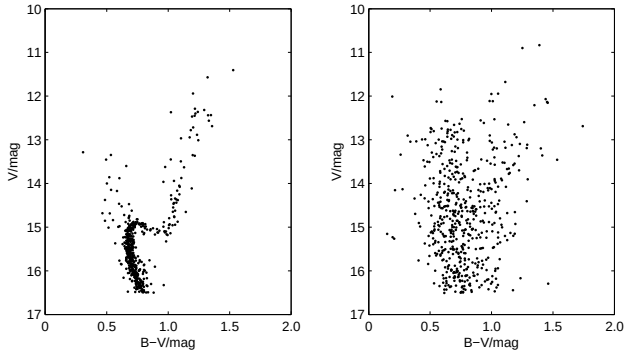

图 5　497颗3D运动学成员星(左)和场星(右)的颜色-星等图

Fig. 5　The color-magnitude diagrams of the 497 3D kinematical members (left) and field stars (right)

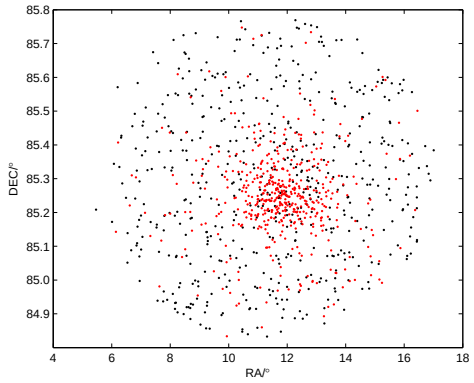

图 6　497颗3D运动学成员星(红点)和场星(黑点)的空间分布情况

Fig. 6　The spatial distribution of the 497 3D kinematical members (red dots) and field stars (black dots)

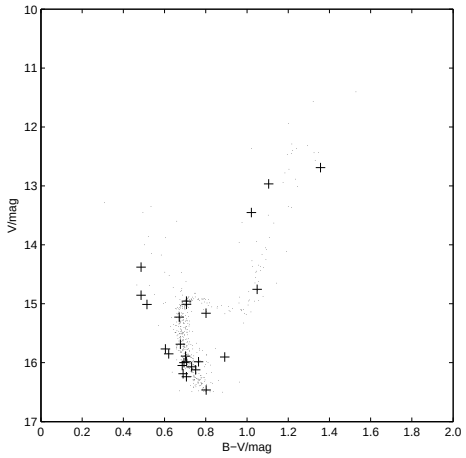

图 7　我们先前工作中得到472颗3D运动学成员星(点)和本工作中增加的25颗成员星(“+”)的颜色-星等图

Fig. 7　The color-magnitude diagram of the 472 3D kinematical members in our previous work (dots) and the extra 25 members obtained in this work (crosses)

3　讨论与结论

本文是我们利用DBSCAN聚类算法进行疏散星团成员判定的后续工作, 主要目的是解决3D运动学数据单位不同和输入参数问题. 我们将不同单位的自行和视向速度统一标准化至[0, 1]区间, 有效地避免了单位不同可能对聚类造成的影响. 在确定DBSCAN聚类算法的输入参数(Eps, MinPts)时, 我们借鉴了Ester等[10]处理2D数据的思路, 并从理论上简单分析了3D速度空间成员星和场星的第k个最近邻距离随k变化的规律, 即成员星的kNND随k的增长速度远小于场星, 我们发现这一规律有助于确定合适的输入参数.

DBSCAN是一种对噪声不敏感的基于密度的聚类算法, 不需要事先对恒星分布进行模型假设, 所以能在数据空间识别任意形状的高密度区域. DBSCAN聚类算法关注的重点是数据空间的点与点之间的距离关系而非VS方法的模型假设和参数求解, 因而DBSCAN聚类算法无需复杂的数学计算, 更容易处理高维数据. 值得一提的是DBSCAN聚类算法不依赖模型的特点使得它不仅可用于疏散星团的成员判定, 还可用于在更大的数据空间搜寻未知的高密度结构. 最近, Bhattacharya等[19]利用DBSCAN聚类算法分析了疏散星团Czernik 20和NGC 1857的空间形态特征, 发现一个先前未知的超密结构, 这说明DBSCAN聚类算法具备发现未知结构的能力. 但DBSCAN聚类算法无法算出每一颗恒星的成员概率, 并且要求数据有较高的精度.

参 考 文 献

[1] Friel E D. ARA&A, 1995, 33: 381
[2] Chen L, Hou J L, Wang J J. AJ, 2003, 125: 1397
[3] Vasilevskis S, Klemola A, Preston G. AJ, 1958, 63: 387
[4] Sanders W L. A&A, 1971, 14: 226
[5] Zhao J L, He Y P. A&A, 1990, 237: 54
[6] 周志华. 机器学习. 北京: 清华大学出版社, 2016: 206
[7] 赵君亮. 天文学进展, 1987, 4: 41
[8] Cabrera-Cano J, Alfaro E J. A&A, 1990, 235: 94
[9] 王家骥. 中国科学院上海天文台年刊, 1997, 18: 45
[10] Ester M, Kriegel H P, Sander J, et al. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996, 96: 226
[11] Gao X H. RAA, 2014, 14: 159
[12] Gao X H, Xu S K, Chen L. RAA, 2015, 15: 2193
[13] 高新华, 陈力, 侯振杰. 天文学报, 2013, 54: 439
[14] Gao X H, Chen L, Hou Z J. ChA&A, 2014, 38: 257
[15] Platais I, Kozhurina-Platais V, Mathieu R D, et al. AJ, 2003, 126: 2922
[16] Geller A M, Mathieu R D, Harris H C, et al. AJ, 2008, 135: 2264
[17] Tan P, Steinbach M, Kumar V. 数据挖掘导论(完整版). 范明, 等, 译. 北京: 人民邮电出版社, 2011: 39-40
[18] Gao X H. RAA, 2016, 16: 184
[19] Bhattacharya S, Mahulkar V, Pandaokar S. A&C, 2017, 18: 1

Determining 3D Kinematical Members of the Open Cluster NGC 188 with the DBSCAN Clustering Algorithm

GAO Xin-hua WANG Chao GU Xiao-qing XU Shou-kun
(School of Information Science and Engineering, Changzhou University, Changzhou 213164)

Abstract In order to obtain clean cluster members in the three-dimensional (3D) velocity space (radial velocity and proper motion), we construct a standardized dimensionless 3D velocity space. We use the k-th nearest neighbor distance (kNND) method to estimate the input parameters of the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm based on the assumption that the kNNDs of cluster members slowly increase with increasing the k value. Finally, we use the DBSCAN clustering algorithm to obtain 497 candidate members, which are very likely the cluster members.

Key words open clusters and associations: individual: NGC 188, stars: kinematics and dynamics, techniques: radial velocities, Hertzsprung-Russell (HR) and C-M diagrams