基于 DBSCNA 聚类算法的成员星推断模型

陈瑞斌　曹博文　杨子钰

（华北理工大学　河北　唐山　063210）

摘要：毕星团位于金牛座，是距地球最近的疏散星团。其成员星在 300 个以上，有多颗肉眼可见的亮星。依据依巴谷卫星的观测数据，可以相当高的精度测量相关各星的距离和运动情况，以对毕星团进行更加精确的研究。本文通过探究毕星团区别与其他星团的不同特征，对数据集中的数据进行研究和分类，确定了毕星团的成员星，并绘制出了赫罗图。

先根据利用普森公式计算绝对星等，以色指数和绝对星等为轴完成了第一次赫罗图的绘制。考虑到恒星的分类与其在三维空间中的位置有关，由数据集中的赤经、赤纬和视差距计算得到了全部恒星在三维空间中的位置图，查阅资料了解到毕星团为球状星图，推测图中的散点密集处即为毕星团的位置。由此，用 k-dist 确定邻域参数，利用基于密度的 DBSCAN 聚类算法求出聚类中心。利用聚类中心再次对数据集进行二次筛选，建立分类模型，最终确定了毕星团的成员星及其赫罗图。

关键词：赫罗图；DBSCAN 聚类算法；k-dist

Member star inference model based on DBSCNA clustering algorithm

Chen Ruibin Cao Bowen Yang Ziyu

Abstract: The Hyades is located in the constellation Taurus and is the closest open star cluster to Earth. Its member stars are more than 300, and there are many bright stars visible to the naked eye. According to the observation data of the Hipparcos satellite, the distance and motion of the relevant stars can be measured with a high degree of accuracy, so as to conduct a more accurate study of the Hydrangea star cluster. This paper studies and classifies the data in the data set by exploring the different characteristics of the Hyde star cluster from other star clusters, determines the member stars of the Hya star cluster, and draws a H-R diagram.

Firstly, the absolute magnitude is calculated by using the Poussen formula, and the first H-R diagram is drawn with the color index and absolute magnitude as the axes. Considering that the classification of stars is related to their position in three-dimensional space, the position map of all stars in three-dimensional space is calculated from the right ascension, declination and parallax in the data set. According to the data, the Hyaloid star cluster is a spherical star map, It is speculated that the densely scattered points in the figure are the positions of the Hyades. Therefore, k-dist is used to determine the neighborhood parameters, and the density-based DBSCAN clustering algorithm is used to obtain the cluster centers. The data set was screened again by the cluster center, and a classification model was established. Finally, the member stars and their H-R diagrams of the Hyades were determined.

Key words: H-R diagram; DBSCAN clustering algorithm; k-dist;

1. 毕星团成员星的初步判断

赫罗图的 x，y 轴分别对应着恒星的色指数 B-V 与绝对星等 Amag，利用 Matlab 软件便可以粗略绘制出包含所有的原始赫罗图：

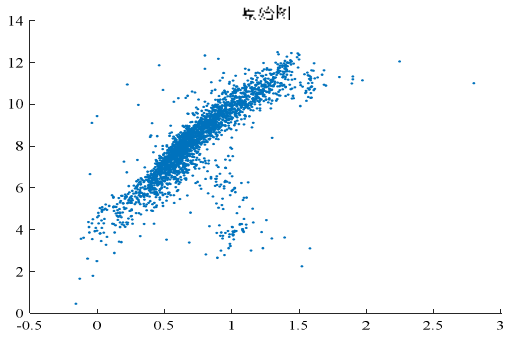

图 1 原始赫罗图

现在我们再选出平均视差在 22 毫角秒左右的亮星，即视差角在[21,23]范围内的亮星。

经过数据筛选处理后的毕星团成员赫罗图如下：

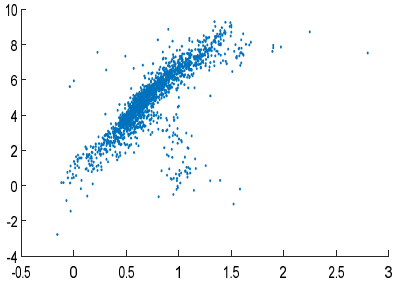

图 2 粗略处理后的赫罗图

在 R 聚类分析过后，我们可以得到恒星的赤经、赤纬和视差在一个三维坐标系中[1]，利用位置和视差数据，可以计算出每颗恒星在三维位置空间中的坐标，坐标中心为太阳，用(a,f,c)分别表示恒星的赤经、赤纬、视差,那么恒星在以 pc 为单位的三维位置空间中的坐标(x, y, z)可以用以下公式计算：

$$d = \quad (1)$$

$$x = d\cos a \quad (2)$$

$$y = d\sin a \quad (3)$$

$$z = d \quad (4)$$

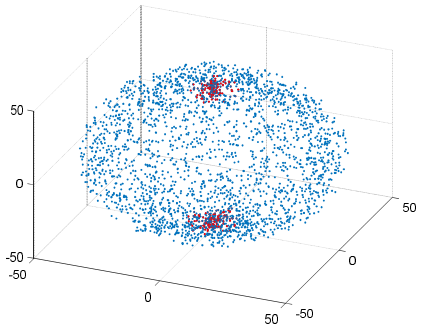

图 3 恒星在三维位置空间的分布对比图

图中的标红位置的散点分布十分集中，且为一个球形，我们基本可以确认该位置即为毕星团所在的位置。

在宇宙中，星团与星团之间的距离远大于星团内恒星的距离。故而可以在剩下的 2679 个恒星数据中依据距离太阳系的距离以及天球赤道坐标系的赤经和赤纬数据中聚类确认毕星团的成员星。而后通过视星等及距离计算绝对星等，星等与光度的对数成正比，照度与距离成平方反比关系。

根据普森公式计算绝对星等　：

$$m_2 = m_1 + 5 - 5 \quad (5)$$

其中m_1为视星等，R 为恒星距离 1000/Plx。以恒星的色指数为 x 轴，绝对星等为 y 轴，再次利用 Matlab 软件便可以进一步绘制出剔除更多点后的赫罗图：

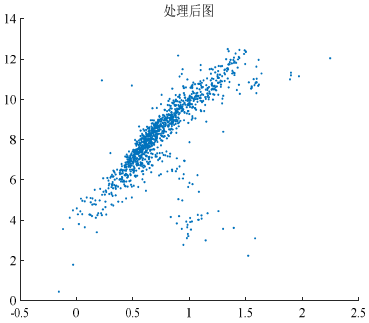

图 4 进一步处理后的赫罗图

2. 利用 DBSCAN 聚类算法的最终判断

在根据恒星的赤经（RA）、赤纬（DE）和视差角（Plx）的公式计算出的 2719 颗恒星在三维空间中的位置后，我们可以看到在图中有散点十分密集的区域，也就是恒星团。通过查阅资料[2]，毕星团几乎是球形的，它的成员星的分布十分的密集，我们有理由推断上述的三维散点图中十分密集的区域就是毕星团的位置，为了证明我们的推断，采用聚类效果较好的 DBSCAN 算法对赤经和赤纬这两个数据进行聚类分析。

在聚类之前，我们需要确定邻域参数(□,MinPts)，由于样本数据量较大，邻域参数的确定较为复杂，我们引入一个新的算法来解决这个问题——k 近邻算法，画出 k-dist 来确定邻域参数[3]。

邻域参数（Eps，MinPts）是 DBSCAN 聚类算法的两个输出参数，对聚类效果有重要影响。DBSCAN 聚类算法的提出者建议统计分析数据集中的点的第 k 个最近邻距离（kNND）估算[4]DBSCAN 的两个输出参数。为了准确估计这两个输入参数，我们利用 kNDD 方法分析成员星和场星在标准化 3D 速度空间里的不同分布特征。

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$E\ (i,\ j)\ (i \neq j)$：

$$E(i,\ j) = \sqrt{\sum_{n=1}^{3}\ (a_{in} - a_j} \quad (6)$$

其中，n 为数据维数，a_{in}，a_{jn}分别为第 i 颗参考恒星和第 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 E（i，j）距离序列，将距离序列升序后就可以得到第 i 颗参考星的 kNND 值（k=1，2，3……，N－1）。

为了从理论上进一步说明 3D 数据集中 kNND 法的工作原理，假设成员星的总数为 N，在标准化 3D 速度空间里均匀分布在有限体积 V 内，那么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（$k \ll N$），第 k 个邻点的距离E_k以及随 k 的变化率$\Delta E_k/\Delta k$、点密度ρ可以用下列公式估算：

$$\rho \quad (7)$$

$$E_k \quad (8)$$

$$\frac{\Delta E_k}{\Delta k} \approx \frac{1}{3}\rho^{-1/3}k^{-2/3} \quad (9)$$

(9)式表示成员星第 k 个最近邻点距离随 k 的增加而缓慢增加，因为成员星可以看作是 3D 速度空间中被系数场星包围的紧密结构。通过对上式分析不难得出第 k 个最近邻点距离也是随 k 的增加而增加的，只不过增加的速度要明显大于成员星，原因是场星的密度比成员星要小很多。

k-dist 图中的横坐标表示数据对象（一个恒星）与它的第 k 个最近的对象间的距离；纵坐标为对应于某一 k-dist 距离值的数据对象的个数。

当 k 变化时，注意横坐标，横坐标随着 k 的增长先震荡，后稳定。在 k=3 和 k=4 时稳定了下来，于是我们取 k=3（或 4）作为 Minpts 参数的值。

邻域参数 Minpts 的值已经确定，接下来确定□的值，直接应用 DBSCAN 算法对 k=3 进行聚类，□的值根据聚类的效果确立为 0.24。

邻域参数的值确立后，聚类的最终结果如下图：

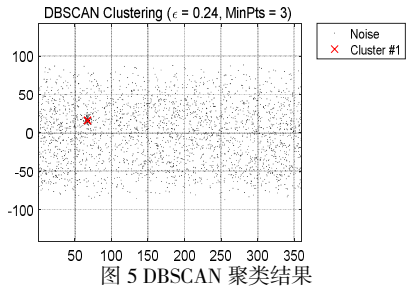

图 5 DBSCAN 聚类结果

聚类中心已经确定，根据聚类中心的赤经和赤纬重新建立筛选的条件，再次利用绝对星等和色指数建立赫罗图，最终确立的毕星团的成员星有 312 个，如下图所示：

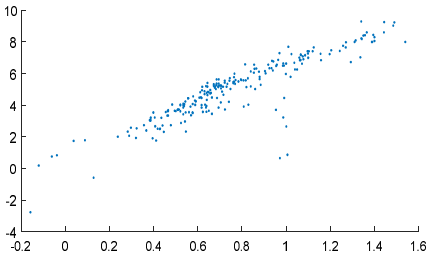

图 6 最终赫罗图

同样的，在这 312 个确立的毕星团的赫-罗图后，我们再利用它们的赤经（RA）、赤纬（DE）和视差角（Plx）的公式重新建立起恒星的三维空间位置，如下图所示：

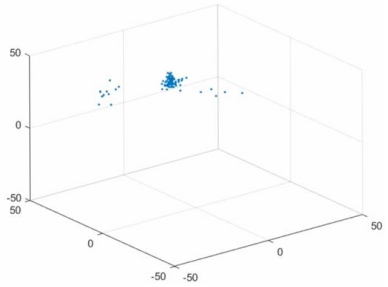

图 7 三维位置图

该图与建立的全部数据的三维空间位置图相比，可以直观的看到，我们确立的毕星团的成员星确实是这一密集的星团，这也再次佐证了赫-罗图的正确性。

参考文献：

[1]刘永利. 用赫罗图阐释恒星演化[J]中国科技纵横,2009(11):481-482. DOI:10.3969/j.issn.1671-2064.2009.11.356.

[2]温学诗. 昴星团和毕星团[J]. 太空探索,2006(1):58-61.

[3]徐守坤,王超,庄丽华,高新华.DBSCAN 聚类算法在 Gaia-DR2 中检测疏散星团的研究[J].天文学报,2018,59(05):17-27.

[4]高新华,王超,顾晓清,徐守坤.基于 DBSCAN 聚类算法的疏散星团 NGC 188 的 3 维运动学成员判定[J].天文学报,2017,58(05):67-74.