文章编号：2095-4980(2021)03-0517-06

基于投票法的密度峰聚类算法

黄文康[1]，杨苏杭[2]，范梦婷[2]，原俊青*[2]

(1.浙江经贸职业技术学院 信息技术系，浙江 杭州 310018；2.浙江工业大学 理学院，浙江 杭州 310023)

摘 要：密度峰聚类(DPC)算法采用点的密度与距离属性对数据进行划分。该算法对大多数数据集能获得较好的聚类结果。然而，对于存在交叉、重叠情况的数据集，DPC 算法的最近邻居分配方法将造成较大误差。针对这一缺陷，本文考虑到数据点的大部分邻居属于相同的簇，提出一种多邻居投票的聚类方法。该方法采取多个邻居的投票结果来决定未知点的归属。数值实验表明，基于投票法的密度峰聚类算法在面对点分布存在交叉、重叠情况的数据集时优于 DPC 算法。

关键词：聚类；密度峰；K 最近邻；投票法

中图分类号：TN914.42 **文献标志码：**A **doi：**10.11805/TKYDA2020253

Density peak clustering based on voting method

HUANG Wenkang[1]，YANG Suhang[2]，FAN Mengting[2]，YUAN Junqing*[2]

(1.Department of information technology，Zhejiang Economic & Trade Polytechnic，Hangzhou Zhejiang 310018，China；
2.School of Science，Zhejiang University of Technology，Hangzhou Zhejiang 310023，China)

Abstract：Density Peak Clustering(DPC) divides the data according to the density and distance attributes of points, which can achieve better clustering results for most data sets. However, the nearest neighbor allocation method of DPC will cause large errors for the data sets with overlapping. Aiming at this defect, a multi neighbor voting clustering method is proposed, which uses the voting results of multiple neighbors to determine the ownership of unknown points. Numerical experiments show that the density peak clustering algorithm based on voting method outperforms general DPC when facing overlapping data sets.

Keywords：clustering；density peak；K-Nearest Neighbor；voting method

聚类是一种无监督的机器学习算法。给定一组数据点，使用聚类算法将数据点划分到特定的分类组[1]。同一分类组中的数据点应该具有相似的属性，而不同分类组中的数据点应该具有较大的差异[2-4]。聚类分析是许多领域中常用的统计数据分析技术[5]。在生物信息学中，聚类分析技术为基因和蛋白质信息的分析和提取提供了强有力的手段[6-7]。聚类分析方法具有如下特征：a)简单、直观；b)异常值和特定属性对聚类有较大的影响；c)聚类分析的结果可能有多个解，但最优解的选取需要主观判断和后续分析；d)聚类分析的解依赖于人为选择的聚类变量，增加或删除一些变量可能会实质性地影响到最终结果；e)不需要知道数据的真实类别等信息，聚类分析就能自动将其分成若干类别。

主要聚类分析算法可以分为数据划分法、层次聚类法[8]、密度法[9]、基于网格的方法和基于模型的方法[10-11]等。例如，K-Means 算法[12]和 K-Medoids 算法都属于划分方法。给定一个数据集和需要划分的个数 k，K-Means 算法可以根据某个距离函数把数据重复划分到 k 个簇中，直到收敛为止。K-Medoids 算法使用最接近簇中心的对象来表示每个簇。具有噪声的基于密度的聚类方法(Density-Based Spatial Clustering of Applications with Noise，DBSCAN)[13]属于基于密度的方法，它基于密度阈值来控制簇的生长。该算法将具有足够高密度的区域划分为类，在噪声空间数据库中可以发现任意形状的簇。统计信息网格(STatistical INformation Grid，STING)算

收稿日期：2020-05-29；**修回日期：**2020-08-15

基金项目：国家自然科学基金青年基金资助项目(11601483)

作者简介：黄文康(1988–)，男，硕士，实验师，主要研究方向为机器学习、人工智能。email:389321927@qq.com

*通信作者：原俊青 email:yuanjq@zjut.edu.cn

法属于基于网格的方法，它将空间区域划分为不同分辨力级别的矩形单元，并形成层次结构。高层的每个单元会被划分为多个低一层次的单元，从底层网格开始，逐步计算并存储网格中数据的统计信息。网格建立后，采用类似于 DBSCAN 算法的方法对网格进行聚类。

Rodriguez 和 Laio 提出了一种新型的聚类算法[14]，即基于快速搜索和发现密度峰值进行聚类(DPC)。DPC 算法可以自动选出样本的聚类中心。此外，该算法能够摆脱一般聚类算法对数据的严格要求，实现对任意分布的数据的高效聚类。DPC 算法选择聚类中心的原则是：具有较高的局部密度，与其他高密度点之间的距离较大。该算法只需要针对数据的 2 个量进行计算。对比以往各种聚类算法，DPC 算法可以自动识别聚类个数，并获取较优的聚类结果。

但 DPC 算法也有一些不足。DPC 算法基于以下假设：具有较高的局部密度并且与高密度点有较大距离的点是聚类中心。聚类中心找到后，剩余的每个点被分配到它的有更高密度的最近邻所属聚类。这种样本分配策略使得 DPC 算法能自动确定类簇数和聚类中心，但也容易造成连带错误效应，一旦一个样本分配错误，便会导致一系列的样本类簇分类错误。尤其是在一些数据集中，点的分布会存在交叉、重叠的情况，此时点之间的距离都很小且差别不大，如果仍然根据单个最近邻居点分配，就显得太过片面，造成较大的误差，也会使连带错误效应更加明显。为解决 DPC 算法的这一缺陷，本文在 DPC 算法的基础上提出了一种新的聚类方法，称为基于投票法的密度峰聚类(Density Peak Clustering based on Voting Method，DPC-VM)。DPC 算法用单个最近邻决定目标点的分类，而 DPC-VM 算法采取多个邻居点进行投票的方式来决定点的归属。对于每个点，都要选取若干邻居点。这些邻居点必须满足 3 个条件才能参与投票：密度比该点大；距离最近；到该点的距离小于截断距离。对于满足条件的邻居点，每次随机取出一部分点进行投票得到一个标签。重复该选取和投票过程若干次，得到总的标签集。最后对标签集再做一次投票，根据得到的标签决定该点最终的所属类。

1　密度峰聚类算法

DPC 算法[15-17]定义聚类中心的特征是：具有较高的局部密度，与更高密度点的距离较大。鉴于聚类中心的这一重要特性，该算法在此部分借助于决策图实现。要做出决策图，对于每一个数据点，需要计算 2 个量：点的局部密度和点到局部密度更高的点的距离。

对于点的局部密度，有截断核函数和高斯核函数这 2 种计算方法。假设存在一个数据集 $X=\{x_1,\cdots,x_i,\cdots x_n\}$，对每个数据点 x_i，它的局部密度ρ_i可以表示为：

$$\rho_i=\sum_j \chi\left(d_{ij}-d_{\mathrm{c}}\right) \tag{1}$$

式中：如果 $x<0$，$\chi(x)=1$，否则，$\chi(x)=0$；d_{ij}是数据点 x_i 和 x_j 之间的欧几里得距离[18]；d_{c}是截断距离，它可以被定义为：

$$d_{\mathrm{c}}=d_{\lceil N\times q\rceil}\in D=\{d_1,d_2,\cdots,d_N\} \tag{2}$$

式中：D 是任意两点之间的距离集，并且集合 D 中的距离按升序排列；N 是 D 中包含数据量；q 是手动设置的百分比；$\lceil\cdot\rceil$表示向上取整。

式(1)使用截断核函数来计算数据点的局部密度。当使用高斯核函数，点的局部密度表示为：

$$\rho_i=\sum_j \exp\left(-\frac{d_{ij}^2}{d_{\mathrm{c}}^2}\right) \tag{3}$$

对于数据点 x_i，δ_i是从 x_i 到任何密度比 x_i 大的点之间的距离的最小值，它被定义为：

$$\delta_i=\min_{j:\rho_j>\rho_i}\left(d_{ij}\right) \tag{4}$$

对密度最大的点 x_k，δ_k 表示为：

$$\delta_k=\max_j\left(d_{kj}\right) \tag{5}$$

在计算出每个点的密度和距离后，它们可以被绘制成横坐标为 ρ 和纵坐标为 δ 的决策图。从决策图上可以观察到聚类中心和其他点的分布特征。图 1 所示是数据集 *R*15 的原始点分布图，图 2 和图 3 分别是数据集 *R*15 用 DPC 算法聚类后的决策图和点分布图。从决策图中可以发现，具有高 δ 值和高 ρ 值的点是聚类中心，具有高 δ 值和低 ρ 值的点可以视为由单个点构成的异常簇，即异常点。聚类中心找到后，剩余的每个点被分配到更高密度的最近邻所属聚类。聚类分配只需一步，不需要迭代。

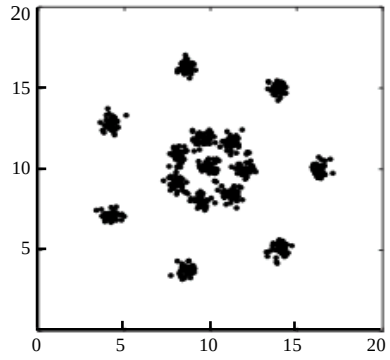

Fig.1 Original point distribution of data set *R*15
图 1 数据集 *R*15 的原始点分布图

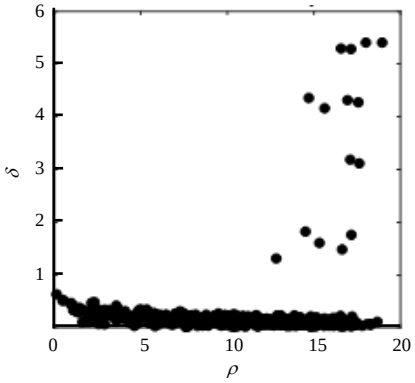

Fig.2 Decision diagram of data set *R*15
图 2 数据集 *R*15 的决策图

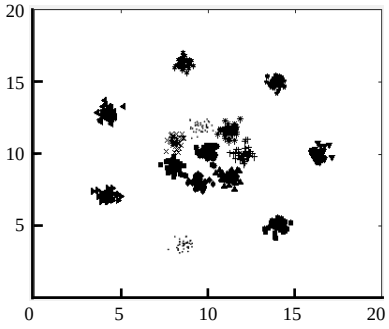

Fig.3 Point distribution of data set *R*15 after clustering
图 3 数据集 *R*15 聚类后的点分布图

2 基于投票法的密度峰聚类

在 DPC 算法中，各个点与更高密度的最近邻点属于同一类，即根据单个邻居点来决定目标点的分类。当遇到数据集中的点分布存在交叉、重叠的情况时，用 DPC 算法的最近邻居分配法显然不合理，这就需要根据更多邻居点来决定目标点的归属。

本文提出的算法基于 K 最近邻(K-Nearest Neighbor，KNN)的思想。KNN 算法的核心思想是，在特征空间上，如果一个样本的 k 个最近邻样本中的大多数属于某一个类别，则该样本也属于这个类别。由于 KNN 方法主要根据周围有限的邻近样本，而不是根据判别类域的方法来确定所属类别，因此适用于类域交叉或重叠较多的待分样本集。KNN 算法的基本做法是，给定测试实例，基于某种距离度量找出训练集中与其最靠近的 k 个实例点。然后基于这 k 个最近邻的信息来进行预测。通常在分类任务中可使用“投票法”，即选择这 k 个实例中出现最多的标记类别作为预测结果。

根据 KNN 算法的思想，本文在 DPC 算法的基础上提出了一种新的聚类算法 DPC-VM[19]。即对点分类时，参照其周围多个邻居的信息，而非单一的邻居信息。

假设存在一个数据集 $X=\{x_1,\cdots,x_i,\cdots x_n\}$，对每个点 x_i 的邻居点进行筛选，看是否满足投票要求的 3 个条件：第一，邻居点的密度比该点大；第二，离该点的距离最近；第三，到该点的距离小于截断距离(截断距离采用第二部分中的截断距离 d_c 的 n 倍)。满足这些条件的点，才是目标点周围且对点的分类影响较大的点。

对于点 x_i 周围满足条件的邻居点 x_j，规定从中最多选取 k_m 个邻居点。然后，从 k_m 个邻居点中随机取出 $k_i(1\leqslant k_i\leqslant k_m)$ 个点进行投票，得到一个标签，标签表示邻居点所属的类别。重复上一个步骤 t 次，每次进行投票的点和个数都不同，可以得到 t 个标签并作为一个总的标签集。最后，对总的标签集再做一次投票，得到的标签作为该点的最终归类。关于 DPC-VM 算法实现的伪代码如下所示。

Require：数据集 $X=\{x_1,\cdots,x_i,\cdots x_n\}$，截断距离 d_c，邻居点的个数 k_m 和重复投票次数 t

For $i=1:n$ do
  For $j=1:n$ do
    计算第 i 个点和第 j 个点的欧几里得距离
  End for
End for
For $i=1:n$ do
  计算第 i 个点的密度
End for
For $i=1:n$ do
  If 第 i 个点不是密度最大点 then
    计算第 i 个点到其具有更高密度的最近邻居点的距离 δ_i
  End if
End for
对于密度最大的点，δ_k 为距离最大值
作出决策图并选择具有高 δ 值和高 ρ 值的点为聚类中心

For $i=1:n$ do
　　计算第 i 个点的符合条件的邻居点集合
End for
For $i=1:n$ do
　　If 第 i 个点有符合条件的邻居点 then
　　For $j=1:t$ do
　　随机选择任意数目的符合条件的邻居点进行投票，并记录出现最频繁的标签类别
　　End for
　　点被分配到出现最频繁的标签类别
　　Else
　　点被分配到具有更高密度的最近邻居点所属聚类
　　End if
End for
Ensure：聚类结果

3　实验

在这个部分，DPC-VM 算法将在真实数据集 Seed,Haberman,Vertebral,Ecoli,Iris 和 Wine 上进行实验，并将 DPC-VM 算法在聚类上的准确性表现和 DPC 算法进行比较。本文中，准确性使用 Acc 指标[20]来衡量：

$$Acc=\frac{\sum_{i=1}^{k}a_i}{n}, \tag{6}$$

式中：a_i 是第 i 个簇中正确分类的点数；k 是簇的个数；n 表示数据集中的点数。

对于 Acc，较高的值意味着更好的聚类质量。当它们的值为 1 时，表示聚类结果完全正确。

图 4~图 9 分别给出了 DPC-VM 算法在这 6 个常用数据集上的决策图。同时也查询了其他几种聚类算法(增强的快速密度峰聚类算法(Enhanced Fast Density-Peak-based Clustering，E-FDPC)[21]、K-Means 算法[22]和模糊 C-均值算法(Fuzzy C-means algorithm，FCM)[23-24]在这些真实数据集上的准确性表现。DPC-VM 算法、DPC 算法和这 3 种聚类算法的准确性比较如表 1 所示。

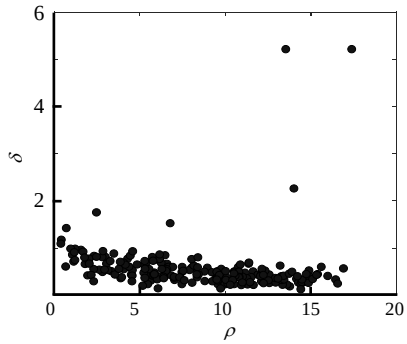

Fig.4 Decision diagram of data set Seed
4 数据集 Seed 的决策图

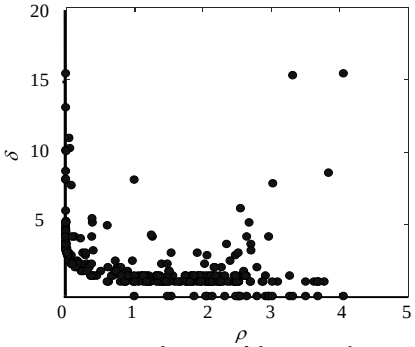

Fig.5 Decision diagram of data set Haberman
图 5 数据集 Haberman 的决策图

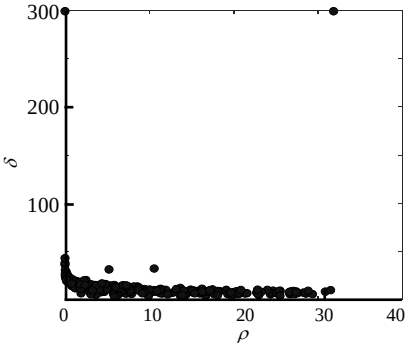

Fig.6 Decision diagram of data set Vertebral
图 6 数据集 Vertebral 的决策图

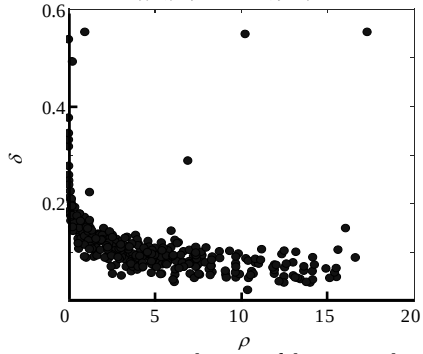

Fig.7 Decision diagram of data set Ecoli
图 7 数据集 Ecoli 的决策图

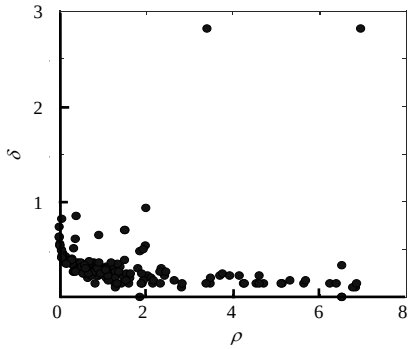

Fig.8 Decision diagram of data set Iris
图 8 数据集 Iris 的决策图

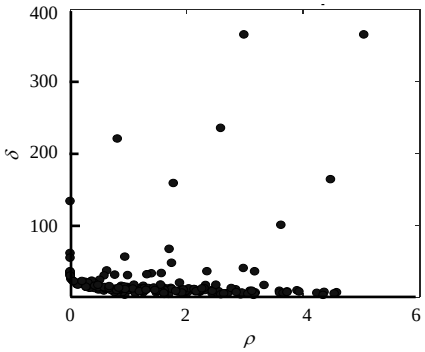

Fig.9 Decision diagram of data set Wine
图 9 数据集 Wine 的决策图

在表 1 中，通过调整总判定投票次数 t，可以使得在用 DPC-VM 算法测试数据集 Seed, Haberman,Vertebral,Ecoli,Iris 和 Wine 时的准确性趋于稳定。表 1 中所用的 DPC-VM 算法在不同数据集上测试出的准确性不一定是表现最好的值，但皆为相对稳定且有明显提高的值(出现次数最多的值)。

表 1 数据集的实验结果
Table1 Experimental results of data sets

algorithm	Seed	Haberman	Vertebral	Ecoli	Iris	Wine
DPC-VM	**90.00**	**55.56**	**75.81**	**79.17**	84.67	56.18
DPC	88.57	53.59	56.13	78.87	83.33	56.18
E-FDPC	40.48	17.32	–	–	19.33	14.61
K-means	89.14	51.30	–	–	82.80	**70.22**
FCM	89.52	50.98	–	–	**89.33**	68.54

观察表中的数据，可以发现在不同数据集上，DPC-VM 算法相比于其他算法的准确性提高程度不同。以 DPC 算法为主要比较对象，则 DPC-VM 算法在 Vertebral 数据集上的 Acc 提高较为明显。这是因为 Vertebral 数据集中类别与类别之间交叉的部分较多，使用 DPC-VM 算法很好地改善了这部分点的分类问题。而对于数据集 Wine，使用 DPC-VM 算法几乎没有改善精确度。这是因为数据集 Wine 是高维稀疏数据集。由于点很稀疏，导致符合条件的邻居点几乎是空集，相当于没有进行重新分配，而是按照原来的算法分配到最近邻所属类。

总的来说，DPC-VM 算法在 Acc 上的表现基本优于其他算法。即使对于数据集 Wine，使用 DPC-VM 算法测试时的 Acc 至少不低于 DPC 算法。Seed,Haberman,Vertebral,Ecoli 和 Iris 都是类边缘的点存在交叉、重叠情况的数据集，这说明 DPC-VM 算法在点分布存在交叉、重叠的数据集上表现较优。

4 结论

本文提出了一种新型密度峰聚类算法 DPC-VM，该算法运用了 KNN 算法的思想，采取多个邻居点进行投票的方式来决定未知点的归属。这些邻居点要参与投票必须满足 3 个条件：第一，密度比该点大；第二，距离最近；第三，到该点的距离小于截断距离。对于满足条件的邻居点，每次随机取出一部分点进行投票得到一个标签，并重复随机选取和投票的操作。最后对得到的总标签集再做一次投票来决定点最终的所属类。实验证明了 DPC-VM 算法在面对点分布存在交叉、重叠情况的数据集时的有效性。

未来工作方向主要是以下几个方面。首先，探索一些应用于高维数据集时具有高精确度的算法；其次，探索减少对参数 d_c 的依赖的方法；第三，采用网格方法降低 DPC 算法面对大数据集时的计算复杂度。

参考文献：

[1] 朱道广,李弼程. 一种基于随机化视觉词汇和聚类集成目标分类[J]. 太赫兹科学与电子信息学报, 2014,12(2):276-283. (ZHU Daoguang,LI Bicheng. An object categorization approach based on randomized visual vocabulary and clustering aggregation[J]. Journal of Terahertz Science and Electronic Information Technology, 2014,12(2):276-283.)

[2] JORDAN M I,MITCHELL T M. Machine learning:trends,perspectives and prospects[J]. Science, 2015,349(6245):255-260.

[3] XU R. Survey of clustering algorithms[J]. IEEE Transactions on Neural Networks, 2005,16(3):645-678.

[4] KANFMAN L,ROUSSEEUW P J. Finding groups in data:an introduction to cluster analysis[M]. New York:Wiley, 1990.

[5] 陈强. 基于聚类技术的多阈值图像分割技术[J]. 太赫兹科学与电子信息学报. 2018,16(4):715-718. (CHEN Qiang. Multi-threshold image segmentation based on clustering method[J]. Journal of Terahertz Science and Electronic Information Technology, 2018,16(4):715-718.)

[6] SI Yaqing,LIU Peng,LI Pinghua,et al. Model-based clustering for RNA-seq data[J]. Bioinformatics, 2014,30(2):197-205.

[7] 贾瑞玉,宋飞豹,汤深伟. 双精英遗传策略的基因聚类算法[J]. 小型微型计算机系统, 2020,41(7):1375-1380. (JIA Ruiyu,SONG Feibao,TANG Shenwei. Gene clustering algorithm based on double elite genetic strategy[J]. Microcomputer System, 2020,41(7):1375-1380.)

[8] MURTAGH F,CONTRERAS P. Algorithms for hierarchical clustering: an overview[J]. Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery, 2012,2(1):86-97.

[9] BANFIED J D,RAFTERY A E. Model-based Gaussian and non-Gaussian clustering[J]. Biometrics, 1993(49):803-821.

[10] MCPARLAND D,GORMLEY I C. Model based clustering for mixed data:clustMD[J]. Advances in Data Analysis and Classification, 2016,10(2):1-15

[11] ROKACH L. A survey of Clustering Algorithms[J]. Data Mining and Knowledge Discovery Handbook, 2009,16(3):269-298.

[12] MACQUEEN J. Some methods for classification and analysis of multivariate observations[J]. Proc. of Berkeley Symposium

on Mathematical Statistics & Probability, 1967:281–297.

[13] ESTER M,KRIEGEL H P,SANDER J,et al. A density-based algorithm for discovering clusters in large spatial databases with noise[J]. Proceeding of International Conference on Knowledge Discovery & Data Mining, 1996(34):226–231.

[14] RODRIGUEZ A,ALESSANDRO L. Machine learning:clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191):1492–1496.

[15] AMJAD S S,ABDULRAHMAN L,PARHAM M,et al. Dynamic graph-based label propagation for density peaks clustering[J]. Expert Systems with Applications, 2019(115):314–328.

[16] MIN Xiangqiang,HUANG Yi,SHENG Yehua. Automatic determination of clustering centers for“Clustering by Fast Search and Find of Density Peaks”[J]. Mathematical Problems in Engineering, 2020(34):1–11. DOI:10.1155/2020/4724150.

[17] XU L,ZHAO J,YAO Z,et al. Density peak clustering based on cumulative nearest neighbors degree and micro cluster merging[J]. Journal of Signal Processing Systems, 2019,91(10):1219–1236.

[18] WANG X F,XU Y. Fast clustering using adaptive density peak detection[J]. Statistical Methods in Medical Research, 2015, 26(6):6099482015.

[19] HUANG Z. Clustering large data sets with mixed numeric and categorical values[C]// Proceedings of the 1st Pacific–Asia Conference on Knowledge Discovery and Data Mining(PAKDD). Singapore:[s.n.], 1997:21–34.

[20] 谷留全,柴瑞林,王平心. 基于投票理论的三支聚类分析[J]. 计算机科学与应用, 2019,9(12):2349–2356. (GU Liuquan, CHAI Ruilin,WANG Pingxin. Three branch cluster analysis based on voting theory[J]. Computer Science and Application, 2019,9(12):2349–2356.)

[21] JIA S,TANG G,ZHU J,et al. A novel ranking-based clustering approach for hyperspectral band selection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2016,54(1):88–102.

[22] SPATH H. Cluster dissection and analysis,theory[M]. Chichester,England:Ellis Horwood, 1985.

[23] PAL N R,PAL K,KLLERE J M,et al. A possibilistic fuzzy c-means clustering algorithm[J]. IEEE Transactions on Fuzzy Systems, 2005,13(4):517–530.

[24] 高云龙,王志豪,潘金艳,等. 基于自适应松弛的鲁棒模糊 C 均值聚类算法[J]. 电子与信息学报, 2020,42(7):1774–1781. (GAO Yunlong,WANG Zhihao,PAN Jinyan,et al. Robust fuzzy C-means clustering algorithm based on adaptive relaxation [J]. Acta Electronica Sinica, 2020,42(7):1774–1781.)