工程科学学报
Chinese Journal of Engineering

基于自组织聚类的多机协同编批方法

张世辉 金同清 张运杰 周锐 冉华明 周礼亮

Multi-aircraft collaborative batching method based on self-organizing clustering

ZHANG Shihui, JIN Tongqing, ZHANG Yunjie, ZHOU Rui, RAN Huaming, ZHOU Liliang

引用本文:
张世辉, 金同清, 张运杰, 周锐, 冉华明, 周礼亮. 基于自组织聚类的多机协同编批方法[J]. 北科大：工程科学学报, 2024, 46(7): 1269−1278. doi: 10.13374/j.issn2095−9389.2023.10.09.002
ZHANG Shihui, JIN Tongqing, ZHANG Yunjie, ZHOU Rui, RAN Huaming, ZHOU Liliang. Multi−aircraft collaborative batching method based on self−organizing clustering[J]. *Chinese Journal of Engineering*, 2024, 46(7): 1269−1278. doi: 10.13374/j.issn2095−9389.2023.10.09.002

在线阅读 View online: https://doi.org/10.13374/j.issn2095−9389.2023.10.09.002

工程科学学报, 第 46 卷, 第 7 期: 1269−1278, 2024 年 7 月
Chinese Journal of Engineering, Vol. 46, No. 7: 1269−1278, July 2024
https://doi.org/10.13374/j.issn2095-9389.2023.10.09.002; http://cje.ustb.edu.cn

基于自组织聚类的多机协同编批方法

张世辉[1)], 金同清[2)], 张运杰[3)], 周　锐[3)], 冉华明[4)]✉, 周礼亮[4)]

1) 中国航空工业集团公司沈阳飞机设计研究所, 沈阳 110035　2) 中国航天科技创新研究院, 北京 100176　3) 北京航空航天大学自动化科学与电气工程学院, 北京 100191　4) 中国电子科技集团公司航空电子信息系统技术重点实验室, 成都 610036
✉通信作者, E-mail:ranhuaming7245@163.com

摘　要　针对多机协同对抗过程中的编批问题, 设计了一种基于改进自组织迭代聚类的多机协同编批方法. 该方法解决了传统自组织迭代聚类算法中人工参数设置不便利不直观的问题, 能够在给定少数直观超参数条件下, 使多机自主调整聚类过程中所涉及的参数, 最终迭代出合理的编批结果. 首先对高维多机态势信息进行标准化和主成分分析处理, 从而确认新的向量空间; 然后引入密度聚类中的邻域密度判别思想对传统自组织迭代聚类方法的合并和分裂操作进行改进, 优化并减少了传统方法进行分裂和合并操作所涉及的人工参数, 提升了执行编批聚类任务的智能自主性; 最后选取算法评价指标, 使用所提算法以及传统算法对多个人工合成数据以及实际想定场景进行聚类测试并对测试结果进行评价. 人工合成数据仿真表明改进自组织迭代聚类算法在优化聚类过程中的人工参数后仍与原始算法表现出相当的性能, 实际想定场景的编批结果进一步说明了改进自组织迭代聚类算法在具体应用场景中的有效性以及在未来实际场景中的实用性.

关键词　多机协同编批; 高维态势信息; 自组织; 聚类; 超参数
分类号　V221+.3; TB553

Multi-aircraft collaborative batching method based on self-organizing clustering

ZHANG Shihui[1)], *JIN Tongqing*[2)], *ZHANG Yunjie*[3)], *ZHOU Rui*[3)], *RAN Huaming*[4)]✉, *ZHOU Liliang*[4)]

1) Shenyang Aircraft Design and Research Institute of AVIC, Shenyang 110035, China
2) China Academy of Aerospace Science and Innovation, Beijing 100176, China
3) School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China
4) CETC Key Laboratory of Avionic Information System Technology, Chengdu 610036, China
✉Corresponding author, E-mail: ranhuaming7245@163.com

ABSTRACT　This article addresses the bathing problem in multi-machine collaborative operations, proposing a method based on improved self-organizing iterative clustering. This approach circumvents the issues of traditional manual parameter setting in the self-organizing iterative clustering algorithm that is often inconvenient and non-intuitive. The proposed method allows multiple machines to autonomously adjust the parameters involved in the clustering process, given a small number of intuitive hyperparameters. The ultimate goal is to iterate toward reasonable editing results. Initially, this article focuses on selecting feature vectors for the multi-machine collaborative confrontation situation. It applies standardization and principal component analysis to high-dimensional multi-machine situation information to confirm the new vector space. This space mainly encompasses position information in three dimensions and speed information. Subsequently, the paper introduces the concept of neighborhood density discrimination from density clustering. This improves the merging and splitting operations of the traditional self-organizing iterative clustering method. It optimizes and reduces the artificial parameters involved in these operations, enhancing the intelligent autonomy for batch clustering tasks. Before optimization, artificial parameters primarily include the number of expected clusters, minimum number of points within a class, number of iterations,

收稿日期: 2023−10−09
基金项目: 国家自然科学基金资助项目 (61773031)

upper limit of standard deviation that limits data distribution within a class, and an allowable shortest distance indicator between classes. Post optimization, the artificial parameters are limited to the expected cluster quantity, minimum number of points, and the number of iterations within a single classification. These optimized parameters are relatively intuitive, and the algorithm output does not strongly correlate with the input parameters. Ultimately, the paper selects algorithm evaluation indicators, including Dunn, Davies–Bouldin, silhouette coefficient, and Calinski–Harabasz. It uses these to evaluate the proposed algorithms ISODATA+ and K-MEANS+, along with the original ISODATA algorithm, against multiple artificially synthesized data sets (completely random data, Gaussian-generated data, and sin-type data) and real-world scenarios. The experimental results suggest that while KMEANS+ shows significant advantages owing to multiple manually set hyperparameters, it requires constant debugging when adjusting parameters, which increases the complexity of the task. Compared with the original self-organizing iterative algorithm ISODATA, statistical results show that the improved algorithm has equivalent capabilities to the original algorithm. This demonstrates that the ISODATA+ algorithm maintains good clustering capabilities even after removing some artificial parameters. The batching results from actual scenario tests further illustrate the effectiveness of the improved self-organizing iterative clustering algorithm in specific application scenarios, demonstrating its practicability for future real-world applications.

KEY WORDS multi-aircraft collaborative batching; high-dimensional situation information; self-organization; clustering; hyperparameter

编批问题指代对目标机群进行功能和任务的划分, 是多机协同对抗中的重要一环. 一般而言, 功能和任务相似的飞机在空间坐标上都有一定的联系, 编批也就是按照飞机的位置坐标、速度和航向来进行划分的. 良好的编批算法可以根据敌方当前的空间站位和运动状态, 推算敌方的战术目的并给予针对性的打击.

从形式上看, 编批问题可以数学化为聚类问题, 即给定多个特征向量, 寻找一组中心特征向量, 使得特征向量与中心特征向量的差在某种度量上最小. 聚类问题属于机器学习问题中的无监督学习范畴[1], 其使用的算法大致可以划分为三类. 第一类是基于原型的聚类(Prototype-based clustering)[2−3], 此类算法假设聚类结构能通过一组原型刻画, 在现实聚类任务中极为常用. 通常情况下, 算法先对原型进行初始化, 然后对原型进行迭代更新求解. 采用不同的原型以及不同的求解方式, 将产生不同的算法. 最著名的原型聚类算法有 K-Means 算法, 有学者在朴素 K-Means 算法的基础上做出了改进, 并对其效果加以验证[4]. 第二类为基于密度的聚类(Density-based clustering)[5−6], 此类算法假定聚类结构能通过样本分布的紧密程度确定. 通常情况下, 密度聚类算法从样本密度的角度来考察样本之间的可连接性, 并基于可连接样本不断扩展聚类簇以获得最终的聚类结果. 著名的密度聚类算法有 DBSCAN 算法[7], 有学者在 DBSACN 算法的基础上, 针对原有的算法参数需手动设置导致聚类效果不稳定和精度低的缺点, 提出一种基于 K-dist 图的 DBSCAN 算法参数的自适应确定方法, 称为 X-DBSCAN. 第三种为层次聚类 (Hierarchical clustering)[8−9], 此算法试图在不同层次对数据集进行划分, 从而形成树形的聚类结构. 数据集的划分可以采用“自底向上”或“自顶向下”的策略. 常用的层次聚类有凝聚层次算法等, 文献 [10] 提出了一种名为 CHAMELEON 的层次聚类算法, 分为两步将子图先划分再合并, 最终得到聚类结果. 文献 [11] 详细比对了各类聚类算法的特点, 以及算法的应用短板.

此外, 聚类问题还需考虑的一个重要因素为目标类别的数目[12]. 较为朴素的 K-means 算法需要人为的指定 K, 也就是分类数目[13], 这在分类数目较少或者有分类期望时, 较为容易, 但是针对分类数目较多时很难在初始状态就进行指定. 虽然可以通过迭代自组织数据分析算法(IOSDATA)[14] 进行改进[15], 监控聚类指标评价较低的类, 以及合并数目较少的类, 但是其需要人为设定的参数数量较多且不甚直观, 需要操作人对于数据的整体分布有先验知识, 还有可以改进的空间[16−17].

除了对于聚类算法的研究, 众多学者还对聚类有效性指标(Cluster validity index, CVI)[18−19] 做了大量的研究, 文献 [20−26] 中提及了 Dunn 系列标准、Davies–Bouldin 标准、分割系数 PC 和 CE、Xie–Beni 标准以及 S-Dbw 标准, 从间距, 密度, 噪声抗性, 几何关系和模糊程度等角度衡量.

本文在现有的自组织聚类的处理方案的基础上做出适应性的调整和改进, 提出了针对自组织迭代聚类算法的改进算法来解决编批问题. 首先针对多机协同对抗态势完成特征向量的选取; 然

后对态势信息进行了主成分分析和标准化处理来确认新的向量空间; 最后针对自组织迭代聚类算法人工参数设置不便利不直观的缺陷进行改进, 吸收密度聚类和层次聚类的合并处理机制, 不再依赖人工提供的参数, 在迭代过程中以自组织的方式对所涉及到的参数进行自主调整, 最终完成聚类任务, 并通过仿真验证了改进的算法在多机协同自组织编批问题上的有效性.

1 自组织编批问题描述

自组织编批的主要设计目的是在敌方集群巡航状态下对敌方的攻击意图做出合理判断并辅助我方的迎战任务分配, 其重点是在没有太多人员干预的前提下对敌方编批意图做出合理判断, 最终使得同一编批内无人机之间的差异尽可能小, 不同编批内无人机间的差异尽可能大.

针对处于巡航阶段的敌方无人机集群, 无人机的北向、东向和地向坐标位置($p_{\mathrm{b}}, p_{\mathrm{e}}, p_{\mathrm{h}}$)和三轴转动坐标欧拉角俯仰角、滚转角、偏航角($\theta, \varphi, \psi$)是现空间意义上的编批划分所必需的特征分量. 同时考虑到速度指向对于判断飞机状态的重要性, 仍然需要添加速度向量相关的三个坐标北向速度、东向速度、地向速度($v_{\mathrm{b}}, v_{\mathrm{e}}, v_{\mathrm{h}}$). 考虑到飞机自身的姿态角对于整体的态势评估的意义不大, 且如果假定飞机在巡航姿态时攻角和侧滑角微小, 那么速度向量一定程度上就能够代表飞机的姿态. 因此最终选定了如下坐标组成单个飞机的态势向量, 第 k 架飞机可以表示为 $\boldsymbol{x}_k = [p_{\mathrm{b}k}, p_{\mathrm{e}k}, p_{\mathrm{h}k}, v_{\mathrm{b}k}, v_{\mathrm{e}k}, v_{\mathrm{h}k}]$, 其中 $p_{\mathrm{b}k}, p_{\mathrm{e}k}, p_{\mathrm{h}k}$ 分别为飞机 k 的北向、东向和地向坐标, $v_{\mathrm{b}k}, v_{\mathrm{e}k}, v_{\mathrm{h}k}$ 分别为飞机 k 的北向、东向和地向速度大小. 下文中不再区分各个物理量的量纲和原始表征符号, 统一使用 x_{ko} 代表第 k 个数据的 o 个特征分量. 输入的态势信息可以表达为态势向量的组合(设敌方共 n 架飞机被我方探测到):

$$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]^{\mathrm{T}} \tag{1}$$

2 自组织编批态势处理

聚类算法大多针对的是量纲一主成分低维度数据, 这类数据的不同维度之间具有相似的数量级, 不同特征之间具有较大的差异性, 且在以欧氏距离为基础的聚类衡量指标下能够正常表征聚类效果. 而针对多机协同对抗场景设计聚类算法时, 目标特征一般会选定飞机的位置向量、速度向量、机头指向等. 根据已知的信息不同, 还可能添加了包括针对对方机型的识别, 各项能力指标的数值化判定, 这些数据有量纲之间的差别, 也可能存在非主成分的特征. 因此在使用聚类算法之前仍然需要进行数据的预处理环节.

自组织编批态势信息的预处理方法包括主成分分析和数据标准化. 主成分分析旨在发现一组正交基向量, 通过在新的基向量上对原始数据进行投影, 使得投影向量的方差最大化. 数据标准化的目的在于消除数据的绝对差异对算法的影响, 通常将每个特征的值范围限制在一定区间内, 并使数据的平均值接近零. 通常, 数据被映射到区间 [−1, 1], 这个过程也被称为“归一化”. 通常情况下, 主成分分析并不保证输出数据的范围, 但其过程涉及各个特征之间的相对数量关系, 因此在进行主成分分析之前需要对数据进行归一化处理. 尽管这样的处理不一定将结果限制在 [−1, 1] 范围内, 但预处理效果明显, 同时也符合主成分分析的要求, 即数据具有零均值.

设某类数据包含 n 个数据, 即该类所有数据为 $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k, \cdots, \boldsymbol{x}_n]^{\mathrm{T}}$, 则该类别数据的标准化处理如下:

$$w_{ko} = \frac{x_{ko} - \bar{x}_o}{\sigma_o} \tag{2}$$

$$\bar{x}_o = \frac{1}{n}\sum_{k=1}^{n} x_{ko} \tag{3}$$

$$\sigma_o = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_{ko} - \bar{x}_o)^2} \tag{4}$$

其中, w_{ko} 为经过标准化处理后的数据, 其表示第 k 个数据在特征 o 上的数值, $\bar{x}_o$ 表示特征 o 的平均值, σ_o 表示特征 o 的标准差.

主成分分析的核心流程如下所述. 首先, 计算输入数据的协方差矩阵如下:

$$\boldsymbol{S} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \tag{5}$$

其中, $\boldsymbol{X}$ 为输入数据, 随后对协方差矩阵 $\boldsymbol{S}$ 进行奇异值分解:

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}} \tag{6}$$

其中, $\boldsymbol{U}$ 是一组新的单位正交基向量, $\boldsymbol{D}$ 是对角矩阵, 其对角线元素为特征值. 特征值越大, 数据在对应的新基向量方向上分布越广泛, 以此作为新的坐标系便可以实现更好的分布效果. 因此选择特征值较大的对应基向量可以更好地捕捉数据的主要变化方向, 从而在新坐标系中实现更好的分类效果. 在这里我们使用经验参数的方式选取 L 对

特征向量和特征值组合作为新的空间基向量, 选取原则如下:

$$\frac{\sum_{o=1}^{L}|\lambda_o|}{\sum_{o=1}^{M}|\lambda_o|} > 0.99 \tag{7}$$

其中, M为特征值的总数量, λ_o为第o个特征值, 即对角矩阵$\boldsymbol{D}$的第o个对角线元素, 上式为保留原有数据的 99% 特征的含义. 经过测试发现数据的每个维度均有价值, 所以本文选定原始数据的所有维度作为待分析维度.

3　多机协同自组织编批算法

为实现无人干预的自组织编批, 本文首先以自组织迭代聚类算法(ISODATA)为基础进行聚类. 然后考虑到 ISODATA 算法在追求类间距离和类内数据密集度的平衡时引入了多个参数, 限制了其在实际应用中的广泛使用. 所以本文对其做出适应性改进, 旨在保持较少输入参数的同时, 自动实现更优的聚类结果.

3.1　自组织迭代聚类算法

自组织迭代聚类算法(ISODATA)是在 K-means 算法的基础上做出适应性改进所形成的算法, 同时也集成了层次聚类的思想, 引入分裂和合并两种操作. ISODATA 算法作为一种有中心性质的聚类方法, 具备自动确定聚类数量的能力, 这也是相对于 K-means 的一个显著优势. 其在使用时需要人为设定 5 个参数, 分别是期望聚类数量K_0, 类内最少点数量$N_{\min}$, 迭代次数 iter, 限制类内数据分布程度的标准差上限$\delta_{\max}$, 以及容许的类间最近距离指标$d_{\min}$. 该算法的核心步骤包括两种: 合并和分裂. 此外还有三个常规操作: 中心矫正, 不合规类别删除以及初始化. 总体来看, 自组织迭代聚类算法是在每个迭代轮次内依次执行中心调整和不合规类别删除. 最后分别判断每个类别是否满足分裂条件, 以及寻找可以合并的两个类别. 在详细阐述算法之前, 约定函数$f_{\rm dis}(\boldsymbol{x}_{k'},\boldsymbol{x}_{k''})$代表任意两个数据点$\boldsymbol{x}_{k'}$和$\boldsymbol{x}_{k''}$之间的距离:

$$f_{\rm dis}(\boldsymbol{x}_{k'},\boldsymbol{x}_{k''}) = |\boldsymbol{x}_{k'} - \boldsymbol{x}_{k''}| \tag{8}$$

其中, k'和k''为数据点的编号. 约定第i个类别的聚类中心:

$$\boldsymbol{z}(i) = \frac{1}{m_i}\sum_{k=1}^{m_i}\boldsymbol{d}_k^i \tag{9}$$

式中: $\boldsymbol{d}_k^i$表示第i个类别中的第k数据点, 即$\boldsymbol{d}_k^i \in \boldsymbol{A}(i)$; $\boldsymbol{A}(i)$表示i个类别中的所有数据点的集合, 即$\boldsymbol{A}(i)=\{\boldsymbol{x}_k|$数据点$\boldsymbol{x}_k$在第$i$个类别中$\}$; m_i表示第i个类别中含有的数据点个数, 即$|A(i)| = m_i$. 算法的每个操作的详细表述如下:

(1) 在算法开始时, 以有先验概率的方式随机生成K_0个数据点作为初始聚类中心. 设有n个数据点, 首先计算所有数据点到其他点的距离, 形成邻接矩阵**Dis**, 随机抽取一个数据点$\boldsymbol{x}_k$, 设其余的$n-1$个数据点的被抽取概率为:

$$\rho = \frac{f_{\rm dis}(\boldsymbol{x}_{k'},\boldsymbol{x}_k)}{\sum_{k'=1,k'\neq k}^{n} f_{\rm dis}(\boldsymbol{x}_{k'},\boldsymbol{x}_k)} \tag{10}$$

按照此概率一次性抽取剩余所需的K_0-1个数据点作为数据中心.

(2) 在中心矫正环节, 首先对每个初始类别的数据计算其中心, 并根据新的聚类中心重新分配数据点的归属. 则此时共有K_0个类别, 其中第j个类别的新中心为$\boldsymbol{z}_j$, 第i个类别中的第k个数据点更新后的类别归属为:

$$\arg\min_j(f_{\rm dis}(\boldsymbol{z}_j,\boldsymbol{d}_k^i)) \tag{11}$$

(3) 在不合规类别去除中, 根据事先设定的类中最小数据点数目对每个类别进行检查, 如果类别中的数据点数量不足, 则该类别需要被剔除, 同时将其原有的成员重新分配到其他现有类别中. 即若$|\boldsymbol{A}(i)| \leqslant N_{\min}$, 则第$i$个类别无效, 其中$N_{\min}$为需要的最少数据点数量.

(4) 分裂操作主要检查当前类别是否满足类内标准. 如果满足, 则保留该类别; 如果不满足, 则将其分裂为两个新的类别. 新类别不需要再进行合规验证或进一步分裂校验. 分裂操作的条件有两个, 一是类别内数据点的标准差大于事先设定的期望最大标准差$\delta_{\max}$, 二是预防性地检查类别内数据点数量是否超过最小数据点数量的 2 倍. 对于类别i, 该类有m_i个数据, 首先计算每个特征分量上数据的标准差σ_o:

$$\sigma_o = \sqrt{\frac{1}{m_i-1}\sum_{k=1}^{m_i}\left(w_{ko} - \frac{1}{m_i}\sum_{k=1}^{m_i}w_{ko}\right)^2} \tag{12}$$

从而得到该类别的最大标准差$\delta_i = \max(\sigma_o)$, 若$\delta_i > \delta_{\max}$且$|\boldsymbol{A}(i)| > 2N_{\min}$, 则第$i$个分类需要进行分裂操作. 具体来说, 分裂的主要内容是根据原有类别的信息生成两个新的类中心. 该过程借鉴了 PCA 的思想, 即在原始数据的分布标准差最大方

向上进行分裂. 考虑到 PCA 需要在每个分裂中都求取数据的特征值和特征向量并构建线性映射所需的计算开销较大, 因此这里比较每个特征分量的标准差并寻找某个标准差最大的特征分量作为最终的分裂方向.

(5) 最后为合并操作, 根据最小中心距离来衡量是否可以合并两类别. 即若 $f_{\mathrm{dis}}(z_i, z_j) < d_{\min}$, 则可以合并 i, j 两类, 其中, $d_{\min}$ 为类间最近容许距离. 合并后的中心为:

$$z_{\mathrm{new}} = \frac{|\boldsymbol{A}(i)|z_i + |\boldsymbol{A}(j)|z_j}{|\boldsymbol{A}(i)| + |\boldsymbol{A}(j)|} \tag{13}$$

最终, 算法根据迭代次数和聚类有效性指标来确定何时停止迭代, 给出聚类结果. 自组织迭代聚类算法不再需要人为指定聚类数量, 但同时引入了用于衡量类内数据聚集程度的标准差上限 $\delta_{\max}$ 和类间最近容许距离 $d_{\min}$. 这在高维空间中难以直观确定且难以估计, 此外这些指标的确定与数据的先验分布紧密相关, 而聚类算法难以预知数据的真实分布状态.

3.2 改进自组织迭代聚类算法

针对自组织迭代聚类算法的人工参数高维、物理意义模糊且需要外界指定的缺陷, 引入密度聚类思想, 在此基础上提出一种改进的自组织迭代聚类方法, 可以自主调整聚类过程涉及到的相关参数, 减少了需要进行人工手动调节的参数数目, 能够更加智能自动执行编批聚类任务.

密度聚类根据密度来确定类别, 而不仅仅是数据的距离, 对聚类大小和形状具有鲁棒性. 基本思想是通过密度判据逐渐的将未纳入归类的点纳入已认定的归类点. 密度判据的核心思路为: 通过周围的临近数据点的分布状态判断某个点与临近点是否属于一类. 本文中, 改进自组织迭代聚类算法基于密度聚类的特点针对性优化了分裂和合并操作所涉及的外界参数. 在判断是否需要进行分裂和合并操作时, 引入基于邻域密度的判别法来决定是否合并. 在全局统计的基础上, 对于单个特征的方差最大的簇进行分裂操作.

针对自组织迭代聚类算法的合并操作, 可能会遇到强行分簇的情况, 如图 1 所示, 其中不同的颜色代表了分簇后的结果, 即每种颜色代表一种分簇类别. 这是因为预设的聚类中心过多, 在经过多轮次调整后, 原本应当聚类为一簇的数据点被强行分为多簇. 为解决此类情况, 借鉴密度聚类的评价方式, 并结合边缘点的情况, 可以考虑首先计算每个聚类中每个点的密度, 随后评估两个类之间的数据点最小距离, 并将两个密度相关值和距离相关值做比较, 以此为标准判定两个类别是否可以合并.

图 1 典型合并情况示意图

Fig.1 Illustration of a typical merger scenario

具体而言, 针对已有分类 i, j, 合并判据表示为 $r_{ij} < \mathrm{avr}_i$ or $r_{ij} < \mathrm{avr}_j$, 则合并 i, j 两类. r_{ij} 表达第 i, j 两类之间的最小点间距:

$$r_{ij} = \min(f_{\mathrm{dis}}(\boldsymbol{d}_k^i, \boldsymbol{d}_{k'}^j)) \tag{14}$$

其中, $\boldsymbol{d}_k^i \in \boldsymbol{A}(i)$, $\boldsymbol{d}_{k'}^j \in \boldsymbol{A}(j)$. avr_i, avr_j 分别为第 i, j 两类内部核心点的距离平均值, 以剔除噪声点的影响, 第 i 类数据内部核心点的距离平均值计算如下:

$$\mathrm{avr}_i = \mathrm{average}(f_{\mathrm{dis}}(\boldsymbol{d}_k^i, \boldsymbol{d}_{k'}^j)),\ \ \mathrm{if}(f_{\mathrm{dis}}(\boldsymbol{d}_k^i, \boldsymbol{d}_{k'}^j) < \mathrm{avr}_i') \tag{15}$$

$$\mathrm{avr}_i' = \mathrm{average}(f_{\mathrm{dis}}(\boldsymbol{d}_k^i, \boldsymbol{d}_{k'}^j)) \tag{16}$$

式中, avr′ 为第 i 类内所有数据点的点间距平均值. 在二维数据时的示意图如图 2 所示, 图中不同的颜色代表了不同的分簇类别, 同时展示出 i, j 两类类内部的点间距平均值.

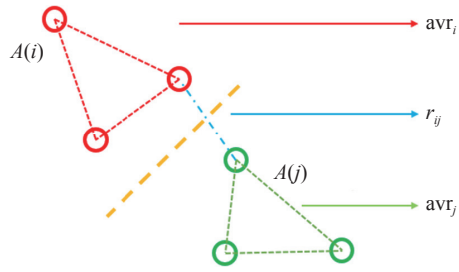


图 2 边缘合并方法示意图

Fig.2 Illustration of the edge-merging method

针对自组织迭代聚类算法的分裂操作, 由于点间距离的大小并不能体现类间的关系, 所以首先引入全局数据的分布假设: 假设聚类最优解存在且当聚类为最优时, 每一个特征的类内数据标准差分布不违背高斯分布的 3σ 原则. 这里是在考虑用当前的全局的聚类有效性评估当前的每个聚类的有效性. 其直观的几何意义为, 若某一类的类内数据紧密程度与其他的类内的数据紧密程度相

差较大, 则该类需要进行拆分处理, 如图 3 左半部分所示, 将类内数据紧密程度与其他的类内的数据紧密程度相差较大的类进行拆分处理, 图中不同的颜色代表了不同的拆分类别.

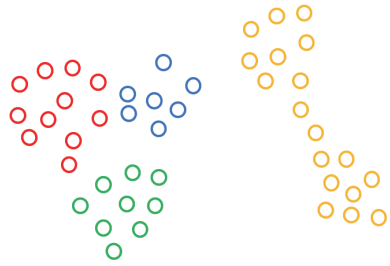


图 3　拆分情况示意图

Fig.3　Illustration of the split situation

若当前将n个数据划分成了K_{current}个类别, 则对类进行拆分处理的判据可以表达为如下形式:

$$\begin{cases} \overline{x}_o = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} x_{ko}, \\ \sigma_{io} = \sqrt{\dfrac{1}{m_i - 1}\displaystyle\sum_{k=1}^{m_i}\left(x_{ko}^{i} - \overline{x}_o\right)^2}, \\ \overline{\sigma}_o = \dfrac{1}{K_{\text{current}}}\displaystyle\sum_{i=1}^{K_{\text{current}}} \sigma_{io}, \\ \text{std}(\sigma_o) = \sqrt{\dfrac{1}{K_{\text{current}} - 1}\displaystyle\sum_{i=1}^{K_{\text{current}}} (\sigma_{io} - \overline{\sigma}_o)^2} \end{cases} \tag{17}$$

其中, m_i表示第i类包含的数据个数, x_{ko}^{i}表示第i类中第k个数据的第o个特征分量, σ_{io}表示类i在第o个特征分量的标准差, $\overline{\sigma}_o$表示σ_{io}的平均值, $\text{std}(\sigma_o)$表示σ_{io}的标准差, 由此可以得到拆分条件为, 若$|\sigma_{io} - \overline{\sigma}_o| > 3\text{std}(\sigma_o)$, 则第$i$类需要拆分.

至此, 通过分裂判据和合并判据, 完整的给出了改进版的自组织迭代分配算法的计算过程. 改进的算法仅仅需要输入期望聚类数量K_0, 单个分类内部的最少点数量$N_{\min}$, 迭代次数 iter 三个参数, 且三个参数本身含义较为直观, 算法的输出不会与输入的参数强相关. 改进自组织迭代聚类算法的运行流程如图 4 所示.

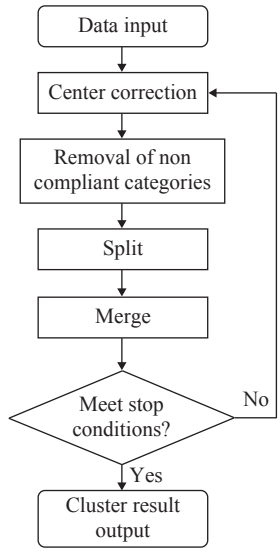


图 4　改进自组织迭代聚类流程图

Fig.4　Improved flow chart of self-organized iterative clustering

4　仿真验证与分析

针对本文所提出的改进自组织迭代聚类算法, 首先选取合适的聚类有效性指标用来评价算法性能, 然后构造测试数据对聚类算法进行了评价和测试, 最后使用实际场景中生成的数据进行编批, 以测试聚类算法有效性.

4.1　聚类有效性指标

下面选取几种常用的聚类有效性指标计算方法[11−16], 以N指代输入聚类数据总数, 以C代表聚类的类别数目, c_i代表第i类.

(1) Dunn 指标.

$$D(c) = \min_{i=1,2,\cdots,C}\left\{\min_{j=i+1,\cdots,C}\frac{g_{\text{Dunn}}\left(c_i, c_j\right)}{\max\limits_{t=1,2,\cdots,C}\text{diam}\left(c_t\right)}\right\} \tag{18}$$

式中, $g_{\text{Dunn}}(c_i, c_j)$为$c_i$和$c_j$的边缘距离, $\text{diam}(c_t)$为类c_t直径. 可以看出, Dunn 指标越大, 聚类效果越好.

(2) Davies–Bouldin(DB) 指标.

$$\text{DB} = \frac{1}{C}\sum_{i=0}^{C}\max_{j=1,2,\cdots,C;i\neq j}\frac{s\left(c_i\right) + s\left(c_j\right)}{g_{\text{DB}}\left(c_i, c_j\right)} \tag{19}$$

可以看出 DB 指标可以综合评价第i类和第j类的类内紧密度$s(c_i)$、$s(c_j)$以及类间距离$g_{\text{DB}}(c_i, c_j)$. 对于 DB 指标而言, 数值越小代表聚类效果越好.

(3) 轮廓系数(Silhouette coefficient, SC).

从第i类数据随机抽取一定数量的数据样本y_i, 计算该数据样本到类c_i内其他数据样本的距离平均值a_i, 称为簇内不相似度; 计算y_i到不同类c_j的所有样本的平均值u_{ij}, 取其最小值得到簇间不相似度u_i, 则有轮廓系数定义如下:

$$\text{SC}(i) = \frac{u_i - a_i}{\max(a_i, u_i)} \tag{20}$$

可以看出, $\text{SC}(i)$接近 1 则第i个类别聚类合理, 接近−1 则第i个类别聚类不合理, 若在 0 附近, 则认为类别i属于两类的过渡点. 可以通过样本的$\text{SC}(i)$取平均来衡量第i个类别聚类的合理程度.

(4) Calinski–Harabasz (CH) 指数.

该指数通过比较簇内样本到簇中心点的距离平方和和分簇中心到数据中心的距离平方和, 来表征数据整体离散程度和簇内离散程度的关系.

$$\mathrm{CH}(k)=\frac{\mathrm{tr}(B_k)(N-C)}{\mathrm{tr}(W_k)(C-1)} \tag{21}$$

式中: tr($\boldsymbol{X}$)为矩阵的迹, $\boldsymbol{B}_k$为类间协方差矩阵, $\boldsymbol{W}_k$为类内协方差矩阵. 可以看出 CH 指标越大代表簇内自身越紧密, 类间越发分散.

4.2 聚类有效性分析实验

聚类算法有效性分析实验针对 K-means+、ISODATA 以及本文提出的 ISODATA+算法, 分别针对完全随机数据、高斯生成数据和 sin 型数据 3 类人工数据进行实验, 每类实验进行 100 次重复测试, 同时分别记录三种算法的 Dunn、DB、SC、CH 指标在不同实验中的最优频数和最差频数.

(1) 实验 1. 每个测试采用 50 个二维随机分布数据点, 每个维度均匀分布在 0 ~ 1 区间内. 某典型分布图样和实验结果如图 5 所示, 随机数据聚类指标最优频数统计和最差频数统计如表 1 和表 2 所示.

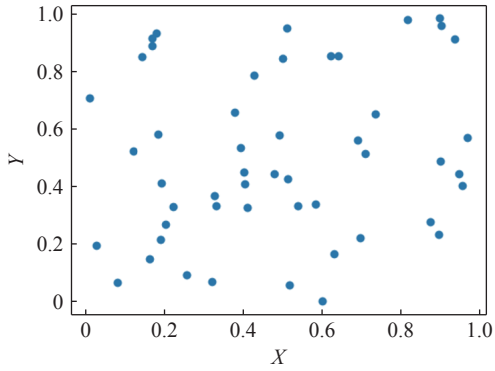


图 5 随机聚类实验数据分布图

Fig.5 Data distribution in a random clustering experiment

表 1 随机数据聚类指标最优频数统计

Table 1 Optimal frequency statistics for indicators of random data clustering

Optimal frequency of indicators	Dunn	DB	SC	CH
ISODATA+	16	33	36	13
ISODATA	6	17	14	9
KMEANS+	78	50	50	78

(2) 实验 2. 随机生成N_{super}个二维高斯分布的中心点参数, 每个分布生成N/N_{super}个二维数据点, 共 50 个数据点, 分布在 0 ~ 1 区间内. 某典型分布图样和实验结果如图 6 所示, 随机数据聚类指标最优频数统计和最差频数统计如表 3 和表 4 所示.

表 2 随机数据聚类指标最差频数统计

Table 2 Worst frequency statistics for indicators of random data clustering

Worst frequency of indicators	Dunn	DB	SC	CH
ISODATA+	47	32	31	39
ISODATA	44	35	37	45
KMEANS+	9	33	32	16

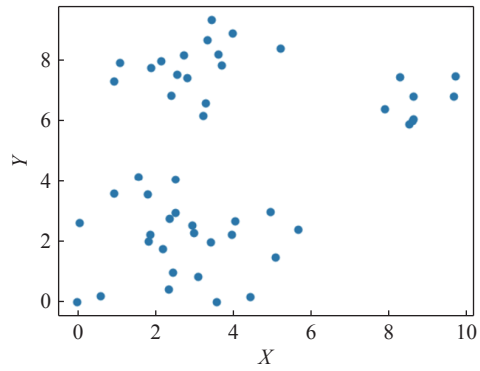


图 6 高斯分布聚类实验数据分布图

Fig.6 Gaussian distribution clustering experimental data distribution

表 3 高斯分布数据聚类指标最优频数统计

Table 3 Optimal frequency statistics of clustering index for Gaussian distributed data

Optimal frequency of indicators	Dunn	DB	SC	CH
ISODATA+	14	10	0	7
ISODATA	25	15	4	14
KMEANS+	61	75	96	79

表 4 高斯分布数据聚类指标最差频数统计

Table 4 Worst frequency statistics of clustering index for Gaussian distributed data

Worst frequency of indicators	Dunn	DB	SC	CH
ISODATA+	45	57	62	55
ISODATA	36	32	36	39
KMEANS+	19	11	2	6

(3) 实验 3. 每组测试采用 50 个二维数据点构成 sin 函数曲线并增加高斯噪声, 数据分布在 0 ~ 1 区间内. 某典型分布图样和实验结果如图 7 所示, 随机数据聚类指标最优频数统计和最差频数统计如表 5 和表 6 所示.

从上述的三组实验的指标最优频数统计结果来看, KMEANS+因为拥有人工设置的多个超参数的优势以及和相同时间内更多的迭代轮次会表现出较大优势, 正是由于其参数数目过多, 导致在调节参数时需要进行不断的调试, 增加了工作的复杂性.

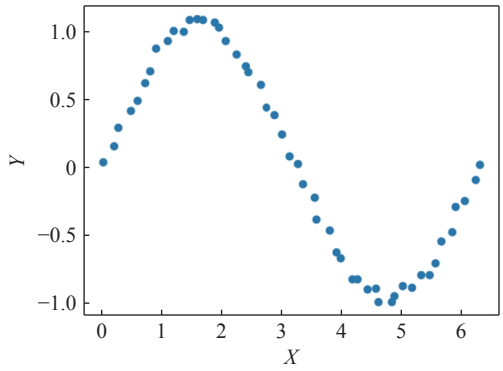


图 7　特殊数据分布

Fig.7　Illustration of special data distribution

表 5　类 sin 分布数据聚类指标最优频数统计

Table 5　Optimal frequency statistics for clustering index of sin-like distributed data

Optimal frequency of indicators	Dunn	DB	SC	CH
ISODATA+	52	2	0	44
ISODATA	45	1	0	56
KMEANS+	3	97	100	0

表 6　类 sin 分布数据聚类指标最差频数统计

Table 6　Worst frequency statistics for clustering indicators of sin-like distributed data

Worst frequency of indicators	Dunn	DB	SC	CH
ISODATA+	14	33	79	0
ISODATA	9	67	21	0
KMEANS+	77	0	0	100

相比于原始的自组织迭代算法 ISODATA, 最优频数统计结果表明改进后的算法在很大程度上能够做到与原始的算法能力相当, 甚至在某些情况下优于原始算法, 证明了去除某些人工参数之后的 ISODATA+算法仍然具备良好的聚类能力.

4.3　编批有效性分析实验

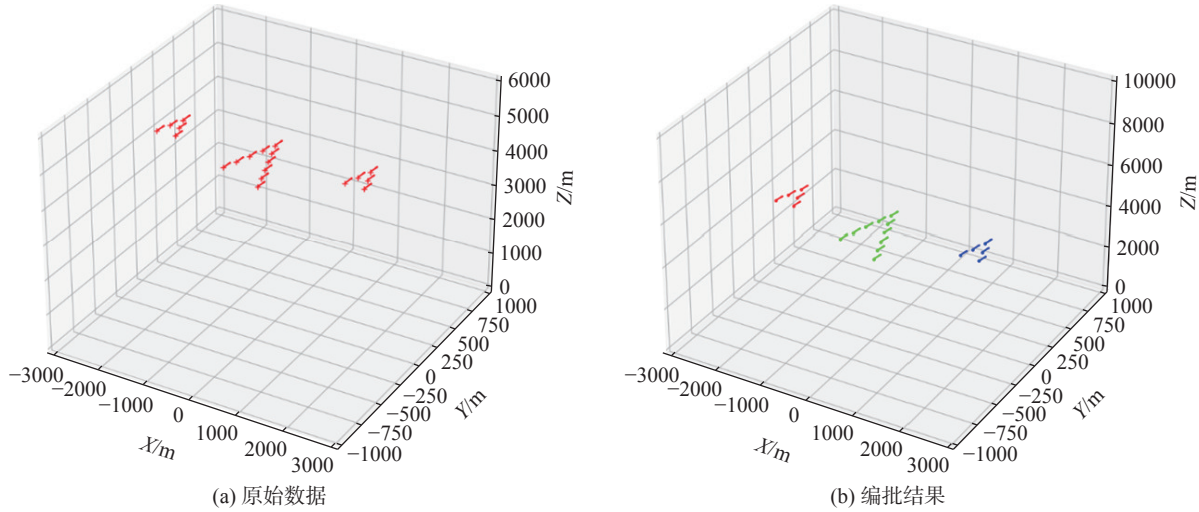
本节将改进的自组织迭代聚类算法应用到编批问题上, 使用实际的飞行数据验证编批算法的有效性, 下面给出典型场景以编队巡航的无人机群编批效果.

场景想定如下, 敌军飞行编队实施针对我方多个目标的大规模侦查任务, 包含共计 20 架无预先标注信息的飞行器, 为了同时打击我方多个目标, 敌机不断的变换编队划分和编队队形试图迷惑我方智能指控系统. 第一阶段 20 架敌机以 3 个巡航编队在绝对高度 5000 m, 以速度 400 m·s^{-1} 向我方重要目标靠近. 在探测到我方拦截编队后, 敌机划分为 3 个不同高度, 6 个不同编队, 以不同的战术编队形态向我方逼近. 截取场景中的常规巡航、迎敌两个稳定瞬态和编队状态变换时的不稳定瞬态作为原始数据, 得到的编批结果如下所示, 编批结果进一步说明了本文所提出算法在具体应用场景中的有效性以及在未来实际场景中的实用性.

在常规巡航状态下, 敌机在 5000 m 高度以人字形分三个编队巡航飞行, 如图 8(a) 所示, 根据本文算法将敌机编为 3 批, 如图 8(b) 所示.

在迎敌状态下, 敌机群迅速拉高、前出, 占据有利空战态势, 同时预留了伴飞的后备编队, 如图 9(a) 所示, 根据本文算法将敌机编为 3 批, 如图 9(b) 所示.

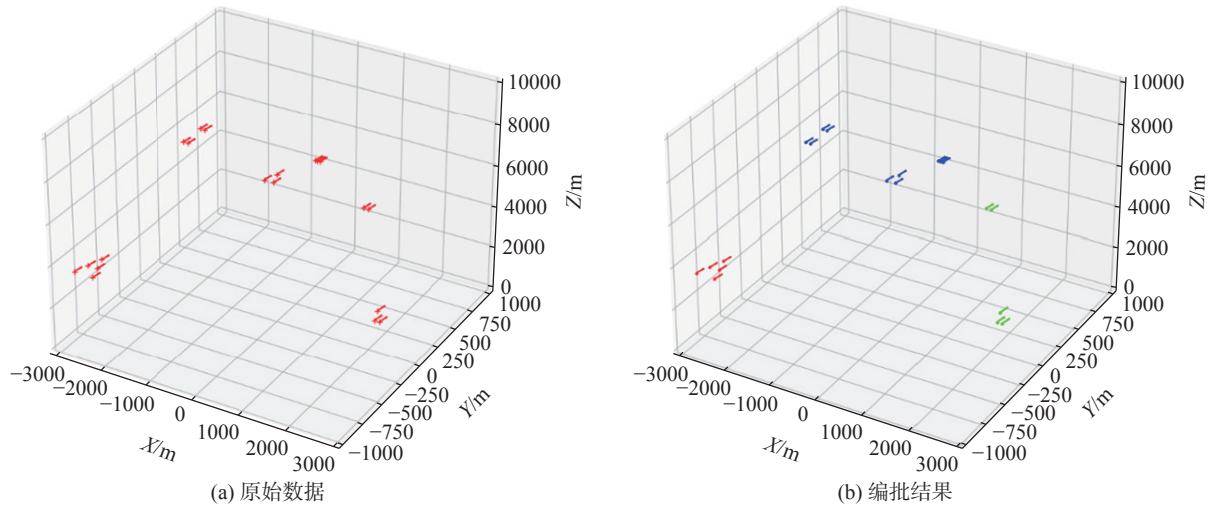
敌机编队状态从常规巡航状态变换到迎敌状态的过程中, 某时刻的状态如图 10(a) 所示, 根据本文算法将敌机编为 3 批, 如图 10(b) 所示.



(a) 原始数据　　(b) 编批结果

图 8　常规巡航状态下编队编批效果图. (a) 原始数据; (b) 编批结果

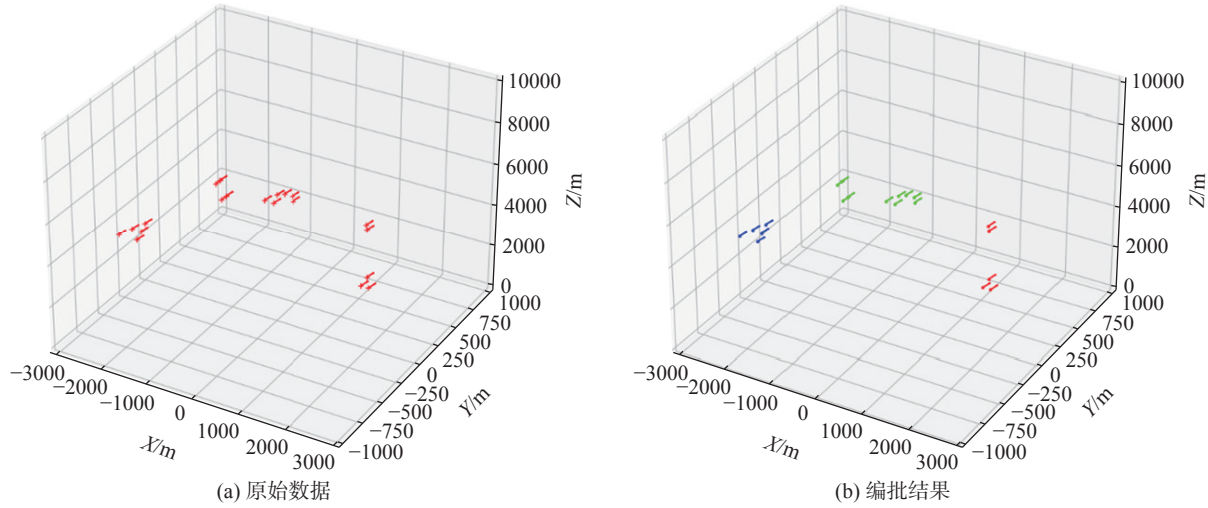
Fig.8　Illustration of formation and batch rendering under conventional cruise condition: (a) original data; (b) batching result



(a) 原始数据　　(b) 编批结果

图 9　迎敌状态下编队编批效果. (a) 原始数据; (b) 编批结果

Fig.9　Illustration of the effect of changing formations under fighting condition: (a) original data; (b) batching result



(a) 原始数据　　(b) 编批结果

图 10　编队状态变换过程中编批效果. (a) 原始数据; (b) 编批结果

Fig.10　Illustration of batch editing effect during formation condition change: (a) original data; (b) batching result

5　结论

本文针对多机协同对抗过程中的编批问题, 设计了一种基于改进自组织迭代聚类的多机协同编批方法. 研究分析与仿真结果表明:

(1) 本文选取的高维输入数据特征可以很好的表征多机协同态势信息.

(2) 该方法不依赖传统自组织迭代聚类算法中进行合并和分裂操作时的复杂超参数, 能够在给定少数直观超参数条件下, 对人工构造数据给出合理的聚类结果.

(3) 该方法适用于实际的巡航无人机编队想定场景, 在不同的任务状态下可以完成对应的编批结果, 满足真实对抗场景中的编批要求.

参　考　文　献

[1] James G, Witten D, Hastie T, et al. *An Introduction to Statistical Learning: with Applications in Python*. Cham: Springer, 2023

[2] Li F J, Wang J T, Qian Y H, et al. Fuzzy ensemble clustering based on self-coassociation and prototype propagation. *IEEE Trans Fuzzy Syst*, 2023, 31(10): 3610

[3] Li Y, Fan B, Guo J, et al. Attribute reduction method based on k-prototypes clustering and rough sets. *Comput Sci*, 2021, 48(Suppl 1): 342
(李艳, 范斌, 郭劼, 等. 基于 k-原型聚类和粗糙集的属性约简方法. 计算机科学, 2021, 48(增刊 1): 342)

[4] Alguwaizani A. Degeneracy on K-means clustering. *Electron Notes Discrete Math*, 2012, 39: 13

[5] Campello R J G B, Kröger P, Sander J, et al. Density-based clustering. *WIREs Data Min Knowl*, 2020, 10(2): 1343

[6] Fahim A. A varied density-based clustering algorithm. *J Comput*

Sci, 2023, 66: 101925

[7] An X Y, Wang Z M, Wang D, et al. STRP-DBSCAN: A parallel DBSCAN algorithm based on spatial-temporal random partitioning for clustering trajectory data. *Appl Sci*, 2023, 13(20): 11122

[8] Huang Q R, Gao R, Akhavan H. An ensemble hierarchical clustering algorithm based on merits at cluster and partition levels. *Pattern Recognit*, 2023, 136: 109255

[9] Dutta A K, Elhoseny M, Dahiya V, et al. An efficient hierarchical clustering protocol for multihop Internet of vehicles communication. *Trans Emerging Tel Tech*, 2020, 31(5): 3690

[10] Karypis G, Han E H, Kumar V. Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 1999, 32(8): 68

[11] Hennig C. An empirical comparison and characterisation of nine popular clustering methods. *Adv Data Anal Classif*, 2022, 16(1): 201

[12] Li S, Wang Y F. A distributed multi-sensor track association algorithm based on K-means clustering. *Telecommun Eng*, 2018, 58(3): 295
(李素, 王运锋. 应用 K-means 聚类的分布式多传感器航迹关联算法. 电讯技术, 2018, 58(3): 295)

[13] Xia D L, Ning F F, He W N. Research on parallel adaptive canopy-K-means clustering algorithm for big data mining based on cloud platform. *J Grid Comput*, 2020, 18(2): 263

[14] Arai K. Improved ISODATA clustering method with parameter estimation based on genetic algorithm. *Int J Adv Comput Sci Appl*, 2022, 13(5)

[15] Liu Q J, Zhao Z M, Li Y X. Application of fuzzy clustering algorithm on feature selection. *J Nanjing Univ Aeronaut Astronaut*, 2012, 44(6): 881
(刘全金, 赵志敏, 李颖新. 模糊聚类在特征选取中的应用. 南京航空航天大学学报, 2012, 44(6): 881)

[16] Lemenkova P. Evaluating land cover types from Landsat TM using SAGA GIS for vegetation mapping based on ISODATA and K-means clustering. *Acta Agriculturae Serbica*, 2021, 26(56): 159

[17] Lv Z Z, Liu Q C. Imbalanced data over-sampling method based on ISODATA clustering. *IEICE Trans Inf Syst*, 2023, E106D(9): 1528

[18] Tang Y M, Huang J J, Pedrycz W, et al. A fuzzy clustering validity index induced by triple center relation. *IEEE Trans Cybern*, 2023, 53(8): 5024

[19] Wang P L. *A Kind of Efficient Clustering Validity Index and Its Application* [Dissertation]. Tianjin: Tianjin University, 2014
(王鹏龙. 一类高效的聚类有效性指标及应用[学位论文]. 天津: 天津大学, 2014)

[20] Duan X J, Ma Y, Zhou Y Q, et al. A novel cluster validity index based on augmented non-shared nearest neighbors. *Expert Syst Appl*, 2023, 223: 119784

[21] Mittal H, Saraswat M. A new fuzzy cluster validity index for hyperellipsoid or hyperspherical shape close clusters with distant centroids. *IEEE Trans Fuzzy Syst*, 2020, 29(11): 3249

[22] Vergara V M, Salman M, Abrol A, et al. Determining the number of states in dynamic functional connectivity using cluster validity indexes. *J Neurosci Methods*, 2020, 337: 108651

[23] Akhanli S E, Hennig C. Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes. *Stat Comput*, 2020, 30(5): 1523

[24] Chowdhury K, Chaudhuri D, Pal A K. An entropy-based initialization method of *K*-means clustering on the optimal number of clusters. *Neural Comput Appl*, 2021, 33(12): 6965

[25] Wu C H, Ouyang C S, Chen L W, et al. A new fuzzy clustering validity index with a Median factor for centroid-based clustering. *IEEE Trans Fuzzy Syst*, 2015, 23(3): 701

[26] Liu Y, Jiang Y F, Hou T, et al. A new robust fuzzy clustering validity index for imbalanced data sets. *Inf Sci*, 2021, 547: 579