0254-6124/2023/43(5)-0927-11 *Chin. J. Space Sci.* 空间科学学报

ZHAO Yuwei, SU Ju. Satellite Anomaly Detection Method Based on Parameter Adaptive Optimization Clustering (in Chinese). *Chinese Journal of Space Science*, 2023, **43**(5): 927–937. DOI:10.11728/cjss2023.05.2022–0054

基于参数自适应优化聚类的卫星状态异常检测方法*

赵玉炜[1,2] 苏 举[1]

1(中国科学院国家空间科学中心 北京 100190)

2(中国科学院大学 北京 100049)

摘 要 对卫星在轨状态进行实时监测和异常检测，有利于保障卫星安全稳定运行。为解决在利用聚类分析检测卫星异常过程中，通过网格搜索选择最佳聚类超参数时精细度差、效率低的问题，本文将聚类超参数选择转化为单目标优化问题，并基于智能优化算法的启发式搜索能力，提出了超参数自适应优化的聚类算法UMOEAsII_BIRCH。为验证自适应搜索的有效性，在卫星遥测数据集和公开数据集上进行了测试。以网格搜索为基准，分别选取基于划分、基于密度和基于层次的聚类算法，对比自适应搜索和网格搜索两种方式下，异常检测的 F1-score 和算法执行时间。实验结果表明，自适应搜索克服了网格搜索中精细度与效率的矛盾，不受网格点的限制，异常检测效果优于基准方法，且在算法执行效率上具有显著优势。

关键词 遥测数据，异常检测，聚类分析，进化算法，网格搜索，参数自适应优化

中图分类号 V474

Satellite Anomaly Detection Method Based on Parameter Adaptive Optimization Clustering

ZHAO Yuwei[1,2] SU Ju[1]

1(*National Space Science Center, Chinese Academy of Sciences, Beijing* 100190)

2(*University of Chinese Academy of Sciences, Beijing* 100049)

Abstract Real-time monitoring and anomaly detection of satellites in orbit are conducive to ensuring the safe and stable operation of satellites. In order to solve the problems of poor fineness, low efficiency and limited grid points when selecting the optimal clustering hyper-parameters through grid search in the process of detecting satellite anomalies using cluster analysis, the selection of clustering hyper-parameters is transformed into a single-objective optimization problem. And based on the heuristic search abili-

* 中国科学院空间科学战略性先导科技专项项目资助（XDA15040100）

2022-09-21 收到原稿，2022-12-05 收到修定稿

E-mail：zhaoyuwei21@mails.ucas.ac.cn

ty of intelligent optimization algorithm, a hyper-parameter adaptive optimization clustering algorithm UMOEAsII_BIRCH is proposed. To verify the effectiveness of adaptive search, tests are conducted on a satellite telemetry data set and a public data set. Using grid search as the benchmark, the clustering algorithms based on partition, density and hierarchy are selected respectively to compare the F1-score of anomaly detection and the algorithm execution time in adaptive search and grid search. The experimental results show that the proposed adaptive search overcomes the contradiction between fineness and efficiency in grid search, and is not limited by grid points. Besides, adaptive search outperforms grid search in the F1-score of anomaly detection, and has a significant advantage in execution efficiency.

Key words Telemetry data, Anomaly detection, Clustering analysis, Evolutionary algorithm, Grid search, Parameter adaptive optimization

0 引言

随着硬件设施的发展和工程制造水平的提升,航天器仪器部件日益灵敏精细。同时,为满足更高目标任务的实施要求,航天器组成结构日益复杂。人造卫星是精密航天器的一种,在生态、经济、国防和国民生活等方面都发挥着重要作用。然而,外太空环境复杂,卫星长期在极端温度、空间大气、太阳风暴、强电磁辐射的恶劣环境中运行。此外,卫星由成千上万个元器件组成,元器件性能会随时间推移而逐渐退化,卫星在轨期间难免会发生状态异常。若能在卫星状态有异常倾向,但尚未发生严重故障时就检测出来,并采取有效干预措施进行修正,及时止损,将有利于保障卫星稳定、安全、可靠运行,从而延长卫星寿命,最大化任务收益。

异常检测是卫星故障诊断排查和实时健康监控的重要途径,常见的卫星状态异常检测方法有基于阈值、基于模型、基于规则和基于数据挖掘四大类。其中,基于阈值的异常检测,需要人力判别,工作量大,可扩展性差;基于模型的异常检测,建模过程复杂,对模型依赖度高;基于规则的异常检测,不能处理知识库中未涵盖的征兆;而基于数据挖掘的异常检测,不依靠先验知识,自动探寻隐藏在数据中的客观规律,通过归纳出的数据特征来检测异常,自动检测能力及可拓展性较强。

基于数据挖掘检测卫星状态,是近年来航天领域广为关注的研究热点。Pan 等[1] 结合核主成分分析和关联规则挖掘提出了一种传感器数据异常检测方法;Zheng 等[2] 基于卫星数据的波动特征,提出了一种基于序列概率比检验的方法,用于识别卫星状态;Zhao 等[3] 提出了基于 Petri 网的诊断方法,并应用于卫星导航接收系统的故障动态诊断;Zhang 等[4] 提出了一种基于深度学习的代表性特征自编码器,用于卫星电源系统的无监督异常检测;Li 等[5] 提出了一种基于 LightGBM 的卫星运行模式监测算法,用于卫星在轨运行模式的实时监控;Li 等[6] 提出了一种基于信息增益参数特征选择和集成学习的方法,能有效地用于载荷单机状态快速识别。

聚类分析是数据挖掘的方法之一,其广泛应用于车辆驾驶行为[7]、电力大数据[8]、核电站[9]、航空器飞行轨迹[10] 等的异常状态检测,并取得了不错的效果。然而应用聚类算法的主要不足是超参数选择不便,聚类超参数的微小差异可能导致全然不同的结果,不同数据集也对应不同的最佳超参数。目前常通过网格搜索的方式,选择聚类效果最优的超参数组合,但这个过程耗时耗力,如何能高效自适应地选择聚类超参数是亟待解决的问题。

本文将聚类分析应用于卫星状态异常检测,同时针对网格搜索中精细度与效率之间的矛盾,将聚类超参数选择转化为单目标优化问题,并基于智能优化算法的启发式搜索能力,提出了超参数自适应优化的聚类算法 UMOEAsII_BIRCH,经实验验证效果优于网格搜索。

1 卫星在轨异常状态检测

1.1 卫星遥测数据特征

卫星遥测数据的产生过程为:首先,星上安装的各传感器按照一定采样频率采集,并转换为电信号;电信号再利用调制编码技术,经无线电通信信道传输

至地面接收站; 最后通过信号解调, 还原为原始被测参量[11]。通常情况下, 在测控区内, 遥测数据实时下行传回; 而在测控区外, 遥测数据则是先存储在星上, 等到卫星过境、可以与地面建立通信时再下传。基于遥测数据监视卫星各分系统工作模式、运行状态, 是判断卫星是否正常运转的重要途径。

根据数据采集方式, 遥测数据分为模拟量和状态量两类。模拟量是连续值, 在一定范围内波动, 反映被测部件的性能状态; 状态量是离散值, 常在几个固定值间变化, 反映被测部件的功能模式。

卫星遥测数据为多维时序数据, 具有以下主要特征[12]。

(1)有噪声和异常值: 由于卫星长期处在太空中, 传感器受恶劣环境条件的影响和干扰, 采集过程可能有误; 此外, 信号在远距离传输过程中, 受无线条件影响, 也可能出错、丢失, 使得遥测数据信噪比低, 含有较多噪声和异常值。

(2)维度高、数据量大: 卫星物理结构复杂, 传感器数目众多, 被测参数可能有上千个, 使得遥测数据维度很高; 同时, 传感器具有较大的采样频率, 短时间内就会采集大量数据。

(3)各参数变化规律不尽相同: 有些参数随卫星的周期性运动也呈周期性变化趋势, 有些参数变化不显著, 在一段时间范围内波动轻微, 还有部分参数相互关联、共同变化。

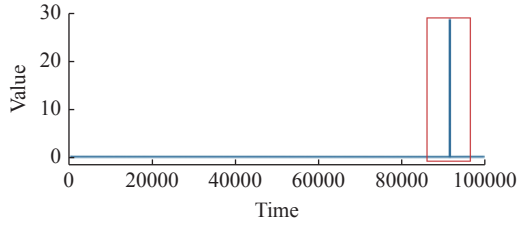

图 1 单点异常示例

Fig. 1 Example of point anomaly

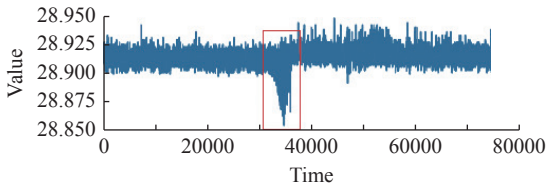

图 2 集体异常示例

Fig. 2 Example of collective anomalies

1.2 卫星在轨状态异常分类

根据异常表现形式, 卫星在轨状态异常分为点异常和序列异常。点异常是指在单一时间序列中, 遥测参数在某个时刻发生突变, 或一系列连续数据点呈现出与其他多数数据点不同的特征。点异常可以进一步分为单点异常和集体异常, 单点异常是指将遥测数据视为整体, 某个时刻的遥测参数与其他时刻相比存在明显差异, 如图 1 所示。

集体异常往往针对模拟量而言, 是指在一段连续时间内, 遥测参数趋势不正常变动, 超出门限范围波动, 或发生突变、缓变, 规律与大部分数据不符, 如图 2 所示。

序列异常是指某段时间序列呈现出与其他应有相似变化趋势的时间序列不同的波动规律, 在波形上有显著差异。序列异常不针对单一时间序列, 其数据对象是时间序列集合。单独分析异常的那一个时间序列, 其本身可能不具有点异常, 如图 3 所示。

2 参数自适应优化的聚类算法

2.1 BIRCH 算法

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)算法是基于层次的平衡迭代规约和聚类算法, 最早由 Zhang 等[13] 于 1996 年提出。BIRCH 算法首先将数据集以压缩形式存储, 再根据压缩后的数据进行聚类, 通过一次扫描即能得到较好的结果, 然后可以通过多次迭代改进聚类效果。该算法降低了 I/O 成本, 能在内存资源有限的情况下完成聚类, 适用于大数据集。

BIRCH 算法以聚类特征 (Cluster Feature) 和聚类特征树 (Cluster Feature Tree, CF 树, 定义符

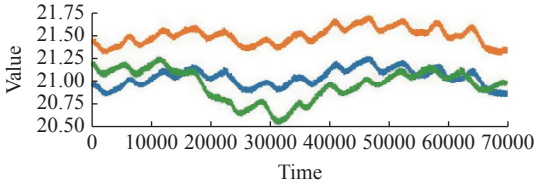

图 3 序列异常示例 (绿色曲线为异常序列, 黄色和蓝色曲线表示正常序列)

Fig. 3 Example of sequential anomalies (Green represents the abnormal sequence, yellow and blue represent normal sequence)

号 C) 为核心。聚类特征是描述分簇的元组,记录了分簇中数据点的信息。假定分簇包含 m 个数据对象 $\{x_1, x_2, x_3, \cdots, x_m\}$,每个数据对象具有 P 个属性特征(即 P 维),定义 $\boldsymbol{L}$ 为数据点的线性和,$\boldsymbol{S}$ 为数据点的平方和,则

$$\boldsymbol{L} = \sum_{i=1}^{m} X_i, \tag{1}$$

$$\boldsymbol{S} = \sum_{i=1}^{m} X_i^2. \tag{2}$$

聚类特征是一个三元组,假设分簇中有 N 个数据点,则聚类特征定义为 $\boldsymbol{C} = (\boldsymbol{N}, \boldsymbol{L}, \boldsymbol{S})$。其中,数据点的线性和反映了分簇的质心位置,数据点的平方和反映了分簇的半径大小。

在聚类特征的基础上定义了 CF 树,CF 树的基本元素即为聚类特征。CF 树是一种平衡搜索树,每个节点包含了一个或多个聚类特征和指向其孩子节点的指针。CF 树包含两个重要参数,分支因子(Branching Factor)和阈值(Threshold)。分支因子规定了非叶节点与叶节点的最大分支树,阈值规定了分簇的最大半径或直径(通常以欧式距离为测度)。

BIRCH 算法的流程[14]可分为如下 4 个步骤。

步骤 1 扫描数据集,依次将样本点插入,建立一个初始的 CF 树。在这个过程中,将比较聚集的稠密点划分至子簇中,同时将比较离散的稀疏点标为噪声去除,以减少孤立点对聚类结果的影响。

步骤 2 根据需要调整 CF 树,以满足后续算法输入范围的需要。若内存占用较大,则可以增大阈值,建立一棵更小的树,使之达到速度和质量的要求。

步骤 3 使用全局或半全局算法对叶节点进行重聚类,以消除分裂导致的局部错位,使其更符合数据真实分布。

步骤 4 把步骤 3 产生的结果作为输入,重新将数据划分到最近的质心,保证重复数据分至同一个簇。

步骤 4 与步骤 2 均为非必需的,但步骤 4 得到的聚类结果往往比步骤 3 更精确。BIRCH 算法只需扫描一次数据集,时间复杂度相比于其他传统聚类算法更低,效率更高,利用 CF 树压缩数据也降低了空间复杂度,适用于处理大规模数据。

2.2 UMOEAs-II 算法

UMOEAs-II (United Multi-operator Evolutionary Algorithms-II)算法是改进的联合多算子进化算法,是 Elsayed 等[15]针对单目标优化问题于 2016 年提出的。该算法的设计基于产生的解决方案的质量和种群的多样性,将多种进化算法结合在一个单一框架中,每种进化算法均可以运行多个搜索算子。

Elsayed 等[16]于 2014 年将差分进化(DE)、遗传算法(GA)和协方差自适应矩阵(CMA-ES)结合在一起,形成了联合多算子进化算法 UMOEAs。UMOEAs-II 算法是在此基础上的进一步改进:利用高效多算子 DE 和 CMA-ES 的搜索能力,来进化两个不同的亚种群,达到预先设定的世代数后,再依据解决方案的质量和子种群的多样性更新每种算子在后续循环中应用的概率。

DE 是一种基于种群的随机进化算法,其模拟了自然界中优胜劣汰的进化过程,以不断提高个体的质量。该算法采用实数编码,按照贪婪选择策略,在决策空间中搜索最优解,通常目标是最小化适应度值。CMA-ES 是协方差矩阵自适应调整进化策略,利用协方差矩阵指导算法的进化。在该算法中,新个体由高斯分布抽样产生,考虑种群跨越世代的路径,而非单一的突变步骤。

UMOEAs-II 算法流程可分为以下 5 个步骤。

步骤 1 随机生成一个大小为 PS 的初始种群,并将其分为 PS_1 和 PS_2 两个大小不同的亚种群。其中,两个种群的解分别由 DE 和 CMA-ES 演化而来。定义 DE 和 CMA-ES 的应用概率分别为 $prob_1$ 和 $prob_2$,初始时均设为 1,并设定种群进化的代数 CS。

步骤 2 在 [0,1] 区间生成两个随机数,若第一个数小于 $prob_1$,则应用 DE 来进化个体;若第二个数小于 $prob_2$,则应用 CMA-ES 来进化个体。

步骤 3 当进化代数达到 CS 时,根据解决方案质量和子种群多样性更新 $prob_1$ 和 $prob_2$;当进化代数每达到 2×CS 时,进行信息共享,并将 $prob_1$ 和 $prob_2$ 重置为 1。

步骤 4 在后期阶段,应用内点法来寻找局部最优解,以发现迄今为止的最优个体。

步骤 5 循环执行步骤 2～4,直至评估次数达到最大评估次数 Max_FES。

在上述过程中,每一代时 DE 和 CMA-ES 均为并行的。在 UMOEAs-II 算法中,算子的结合、应用概率的动态变化、信息共享方案的采用和后期的局部搜索,使得算法在求解单目标优化问题时,能取得较高质量的解。

2.3 改进 UMOEAsII_BIRCH 算法原理

从优化的角度来看, 聚类是一类特殊的 NP 难分组问题。其以一定度量标准衡量聚类效果, 并寻找使得聚类效果最好的分簇方式。然而, 聚类算法对超参数高度敏感, 在缺乏先验知识的情况下超参数难以选择。为解决利用网格搜索选择超参数过程中精细度与效率冲突的问题, 本文将智能优化算法引入聚类分析中, 以实现聚类超参数的启发式自适应搜索。

进化算法是智能优化算法之一, 能够高效搜索近似最优解, 被证明对 NP 难问题有效, 已有学者尝试利用进化算法来优化聚类分析[17]。基于此, 本文将聚类超参数的选择转化为单目标优化问题, 并利用进化算法求解。在这个问题中, 要求的解向量为待选择的聚类超参数, 决策空间为超参数的取值范围, 目标函数值为选取的聚类效果评价指标。UMOEAs-II 算法对于单目标优化问题解决效果较好, BIRCH 算法适用于处理大规模数据集, 因此本文将 UMOEAs-II 算法与 BIRCH 算法相结合, 提出超参数自适应优化的聚类算法 UMOEAsII_BIRCH。

针对异常检测问题, 处理思路为: 选取异常检测中常用的效果评价指标 F1-score, 把 1/F1-score 作为进化算法目标函数值, 该值越小、越接近 1, 异常检测效果越好; 首先使用 UMOEAs-II 算法启发式搜索 BIRCH 聚类超参数取值空间中使得目标函数值最小的解向量, 最优解即对应 BIRCH 聚类最优超参数; 然后在搜索到的最优超参数下, 对数据集样本聚类, 正常数据将大量聚在一起形成密集的大簇, 而异常数据因分布规律不同, 将形成散落的小簇被区分出来。

在 UMOEAsII_BIRCH 算法实现过程中, 通过不断迭代使适应度收敛至某一最小值。经算法多次测试, 找到最终适应度值最小的循环, 其求得的解向量即为问题近似最优解。最优解对应最佳聚类超参数, 通过上述过程, 实现了启发式自适应搜索聚类最优超参数的目的。其中, 最优聚类超参数在本文研究问题中意味着 F1-score 最大, 异常检测效果最好。

3 实验验证

3.1 异常检测效果评价指标

异常检测可以视作一个二分类问题, 将样本数据分为正常和异常两类。在这个问题中, 通常更关心异常的情况, 因此将异常样本设为正例(值为 1), 将正常样本设为负例(值为 0)。机器学习中常用混淆矩阵来衡量分类的质量, 各种情形列于表 1。

精确率 (Precision, 用 P 表示) 是指所有被预测为异常的样本中, 预测为异常、实际也为异常的样本所占的比例。其值越高, 说明误检率越低、异常检测效果越好, 有

$$P=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}}. \tag{3}$$

其中, N_{TP} 为正例的数量, N_{FP} 为负例的数量。

召回率 (Recall, 用 R 表示) 是指预测为异常、实际也为异常的样本, 占实际上所有异常的比例。其值越高, 说明漏检率越低、异常检测效果越好, 有

$$R=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}}. \tag{4}$$

其中, N_{FN} 为负例的数量。

但倘若只单独分析精确率或者召回率也是意义不大的: 例如为了不误检, 只将最可能是异常的几个样本点挑出来, 这样精确率很高, 但大量异常未被发现; 或是为了不漏检, 将所有疑似异常的都挑出来, 这样召回率很高, 但存在大量虚警。因此, 要综合考虑召回率和精确率, 才能反映真实效果。为此, 引入了 F1-score。

F1-score(用 F 表示)是精确率和召回率的加权调和平均, 取值范围为 [0,1], 其值越接近 1, 说明异常检测质量越高。在本文后续实验中, 选择 F1-score 作为主要评价指标, 衡量算法异常检测效果, 有

$$F=\frac{2PR}{P+R}. \tag{5}$$

3.2 卫星遥测数据集测试

3.2.1 实验数据

实验数据来自于中国某空间科学卫星, 采样周期为 1 s。选取 2020 年 12 月的部分延时遥测数据中的电源主要包进行算法验证, 共获得 58 维特征, 10 万条数据样本。

表 1 混淆矩阵

Table 1 Confusion matrix

		预测类别	
		正例	负例
真实类别	正例	TP	FN
	负例	FP	TN

在实际工程中，卫星状态异常并不常见，一段连续时间内发生多次异常的情况更少。此外，与其他设备不同，卫星的高造价和长期在太空环境中运行，就使得给卫星断电或人为模拟故障实验是不现实的。因此，本文在原始正常数据的基础上，根据各属性含义和取值范围等先验知识，结合专家经验，进行异常注入，得到模拟的含有点异常的卫星遥测数据（异常数据占比 0.33%）供后续实验。

所用遥测数据来自卫星电源分系统，包含电压、电流、开关状态、配电状态、温度等属性。针对原始数据，预处理主要包括特征选择、数据标准化、主成分分析等。

3.2.1.1 特征选择

本文所研究的卫星状态异常检测，针对点异常中的集体异常，研究对象是模拟量，即连续值。因此，首先删除遥测数据中的状态量，以及不具有实际意义的保留字段。剩余属性均属于研究范围，但各属性间并非独立，很多属性相互关联，有的还高度相关。对于相关性高的属性，所包含的信息往往极为相似。删除相似属性，可以减少特征冗余，提高效率。

皮尔森相关系数是常见的相关性度量指标之一，可以衡量线性相关的连续变量之间的相关程度。皮尔森系数 $\rho_{x,y}$ 介于 [–1, 1]，绝对值越接近 1，说明两个变量越相关。其中，负数代表负相关，正数代表正相关。通常认为，$|\rho_{x,y}| > 0.8$表示两个变量极强相关，$0.8 > |\rho_{x,y}| > 0.6$表示两个变量强相关。分别计算属性两两间的皮尔森相关系数，并画出属性相关性热力图，如图 4 所示。图 4 中颜色越浅代表属性正相关性越强，颜色越深代表属性负相关性越强。

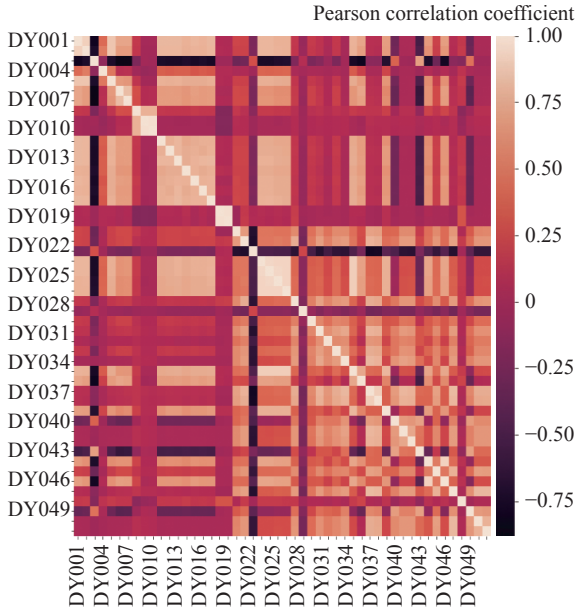

图 4 属性相关性热力图

Fig. 4 Attribute correlation heat map

3.2.1.2 数据标准化

使用 Z-Score 标准化（即标准差标准化），将数据转化为标准正态分布。该方法对于后续主成分分析和聚类的距离相似性度量有较好的效果。同时使用该方法在一定程度上可以避免 Min-Max 标准化等对异常值敏感的问题。

3.2.1.3 主成分分析

经特征选择，数据属性已大幅减少，但仍具有较高维度，影响聚类效率。因此，通过主成分分析进一步将数据维度降至三维，从而能可视化显示，便于直观观察。在三维坐标系中，可视化显示出经预处理后的数据点，如图 5 所示。可以看出，绝大多数正常点密集分布在一起，而异常点形成几个小簇，比较分散，符合利用聚类进行异常检测的认知规律。

3.2.2 参数设置

为验证提出的聚类超参数自适应搜索的效果，以网格搜索为基准，选取基于划分的聚类 K-Means，基于密度的聚类 MeanShift、DBSCAN 和基于层次的聚类 BIRCH 算法进行对比测试。对于各聚类算法，分别采用网格搜索的方式和自适应搜索的方式，选择异常检测效果最好的聚类超参数。

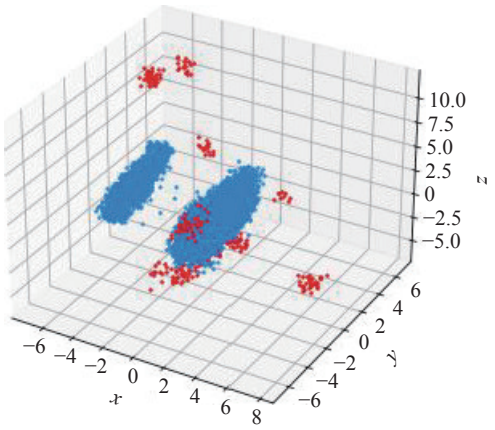

图 5 异常点分布（蓝色表示正常点，红色表示异常点）

Fig. 5 Distribution of anomalies (Blue represents normal points, and red represents abnormal points)

对于智能优化算法的测试, 为避免偶然情况的干扰, 需要多次运行, 求得最优值、最差值、均值和方差, 来综合评定算法性能。因此, 自适应搜索方式中, 对于 UMOEAs-II 算法, 本文设定的运行次数为 10 次, 各算法具体参数设置见表 2。

自适应搜索中, 每次评价执行一次聚类算法, 计算适应度值。网格搜索中, 每个网格点对应一组超参数取值, 执行一次聚类算法。因此, 聚类算法运行总次数分别对应自适应搜索中的最大评估次数和网格搜索中的网格点数。为使两种搜索方式处在可比的规模, 在网格搜索中, 设定网格点数 NUM 为自适应搜索中 UMOEAsII 算法的最大评估次数 Max_FES, 保证两种方式测试了同等数量的聚类超参数组合。

相应地, 为避免实验的偶然性, 在保证划分精细度不变的前提下, 随机网格的具体划分方式, 并运行 10 次, 与自适应搜索保持一致。每种聚类算法测试的参数组合列于表 3, 统计 F1-score 的最优值、最差值、均值和方差。同时, 为比较搜索效率, 记录算法执行时间。

3.2.3 实验结果

分别统计网格搜索(MeanShift, K-Means, DBSCAN, BIRCH)和自适应搜索(UMOEAsII_MeanShift, UMOEAsII_K-Means, UMOEAsII_DBSCAN, UMOEAsII_BIRCH)两种方式下, 算法运行 10 次的结果, 得到的统计量列于表 4。

比较传统的聚类算法 K-Means, MeanShift, DBSCAN 和 BIRCH, 可以发现, 基于层次的 BIRCH 聚类效果最好, 在最优超参数下, 利用 BIRCH 算法进行异常检测的 F1-score 优于其他聚类。同时比较各聚类算法平均执行一次的时间可得, BIRCH 算法的执行效率高于 K-Means 和 DBSCAN, 仅次于 MeanShift, 但 BIRCH 算法异常检测的性能远优于 MeanShift。这也印证了 BIRCH 算法通过只扫描一次数据集, 然后迭代改进的方式聚类, 时间复杂度更低、效率更高, 适用于处理大规模数据。

分别比较各聚类算法两种搜索方式下 10 次运行的结果, 由表 4 可以看出, 自适应搜索方式得到的最

表 2 UMOEAs-II 算法参数设置

Table 2 Parameter settings of UMOEAs-II

	最大评估次数	种群大小	问题维度	超参数名称	超参数取值范围
UMOEAsII_MeanShift	120	5	2	quantile n_samples	[0.01, 1.0] [500, 2000]
UMOEAsII_K-Means	200	8	1	n_clusters	[2, 1000]
UMOEAsII_DBSCAN	1600	22	2	eps min_samples	[0.2, 1.0] [2, 160]
UMOEAsII_BIRCH	2500	36	2	threshold branching_factor	[0.001, 2.0] [2, 200]

注 问题维度, 代表聚类算法要寻优的超参数个数。

表 3 网格搜索测试参数组合

Table 3 Parameter combinations of grid search

算法名称	超参数名称	超参数取值范围	搜索步长	网格点数
MeanShift	quantile n_samples	(0.001, 1.0) (500, 2000)	0.025 500	120
K-Means	n_clusters	(2, 1000)	5	200
DBSCAN	eps min_samples	(0.2, 1.0) (2, 160)	0.02 4	1600
BIRCH	threshold branching_factor	(0.001, 2.0) (2, 200)	0.04 4	2500

优值、最差值和均值都高于或与网格搜索方式的结果持平。同时,对比算法运行时间可得,自适应搜索在最好、最坏和平均情况下均能以更高效率得到比网格搜索更优或等优的解。此外,自适应搜索的方差相较于网格搜索更小、稳定性更高。

为观察算法收敛过程,依据记录点给出 F1-score 最优值演化过程折线图(见图 6)。折线反映了利用进化算法搜索最优聚类超参数过程中 F1-score 的演化过程。可以看出,算法在运行结束前数十次至上百次评价时就已达到收敛,具有较好的收敛性能。这说明自适应搜索实际找到最优解的时间小于执行时间,效率高于网格搜索。

表 4 算法测试结果对比

Table 4 Comparison of algorithm test results

	F1-score最优值	F1-score最差值	F1-score均值	F1-score方差	算法执行时间/s
MeanShift	0.6922	0.6922	0.6922	0.00000	4780
UMOEAsII_MeanShift	0.6922	0.6922	0.6922	0.00000	4087
K-Means	0.8006	0.7545	0.7871	0.01994	156639
UMOEAsII_K-Means	0.8006	0.7650	0.7969	0.01066	104939
DBSCAN	0.8297	0.8277	0.8286	0.00066	502272
UMOEAsII_DBSCAN	0.8305	0.8294	0.8298	0.00039	394553
BIRCH	0.8519	0.7734	0.8191	0.02580	172106
UMOEAsII_BIRCH	**0.8610**	**0.8542**	**0.8568**	0.00205	86930

注 加粗数据表示最优结果。

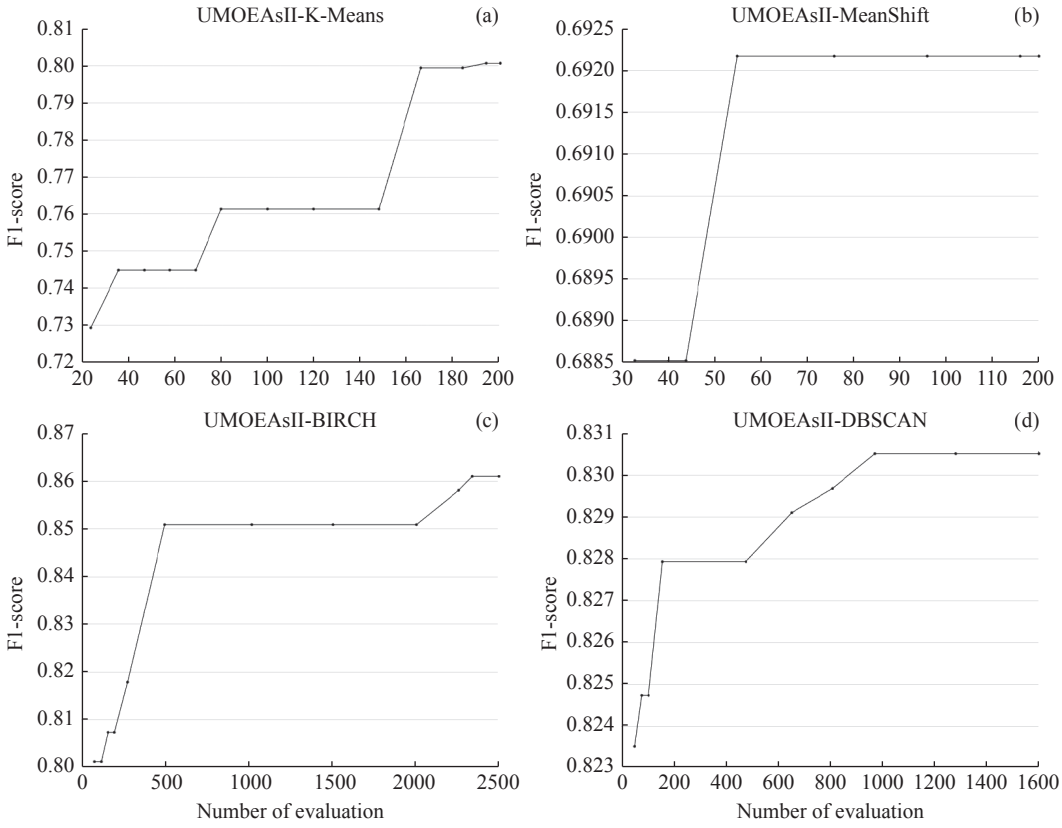

图 6 进化算法演化过程曲线

Fig. 6 Evolution process curve of evolutionary algorithm

为展示网格搜索方式寻找最优参数的过程, 依据聚类超参数组合和对应的 F1-score 绘制曲线图与曲面图(见图 7)。从图 7 可以看出, 当 F1-score 最大时, 异常检测效果最好, 则寻优过程可以理解为搜索图中最高点, 最高点处对应的超参数即可近似视为最优聚类参数组合。

由图 7 还可以看出, F1-score 随超参数变化剧烈, 无明显规律, 趋势难以预知。若设置的步长过大, 则搜索不够精细, 可能错过更优的取值; 若设置的步长过小, 则需测试很多种组合, 增大时间开销。而在固定网格划分精度的前提下, 无论选择哪种具体划分方式, 都只能搜索到网格点上的参数组合, 无法测试不在网格上的点, 约束性强。

综上所述, 本文提出的 UMOEAsII_BIRCH 算法通过自适应搜索的方式, 在决策空间内搜寻, 克服了网格搜索方式中搜索精细程度与效率之间的平衡问题, 能够以更高效率发现更加优异的解, 且超参数个数越多, 优势越显著。同时, 该方法需要的人工干预较少, 不受先验知识的限制, 达到了改进的预期效果。此外, 改进的 UMOEAsII_BIRCH 算法适用于大数据集, 在卫星遥测数据集上测试有效, 异常检测 F1-score 可达 0.86。

3.3 公开数据集测试

为验证参数自适应聚类的可拓展性, 选取异常检测公开数据集 Thyroid 进行测试。该数据集包括 3772 条样本, 每条样本含 6 维属性, 共分为正常和异常两个类别, 其中异常样本有 93 条, 占比 2.466%。分别选取基于划分的聚类 K-Means、基于密度的聚类 DBSCAN 和基于层次的聚类 BIRCH 算法进行测试, 设定算法运行次数为 10 次, 自适应搜索和网格搜索

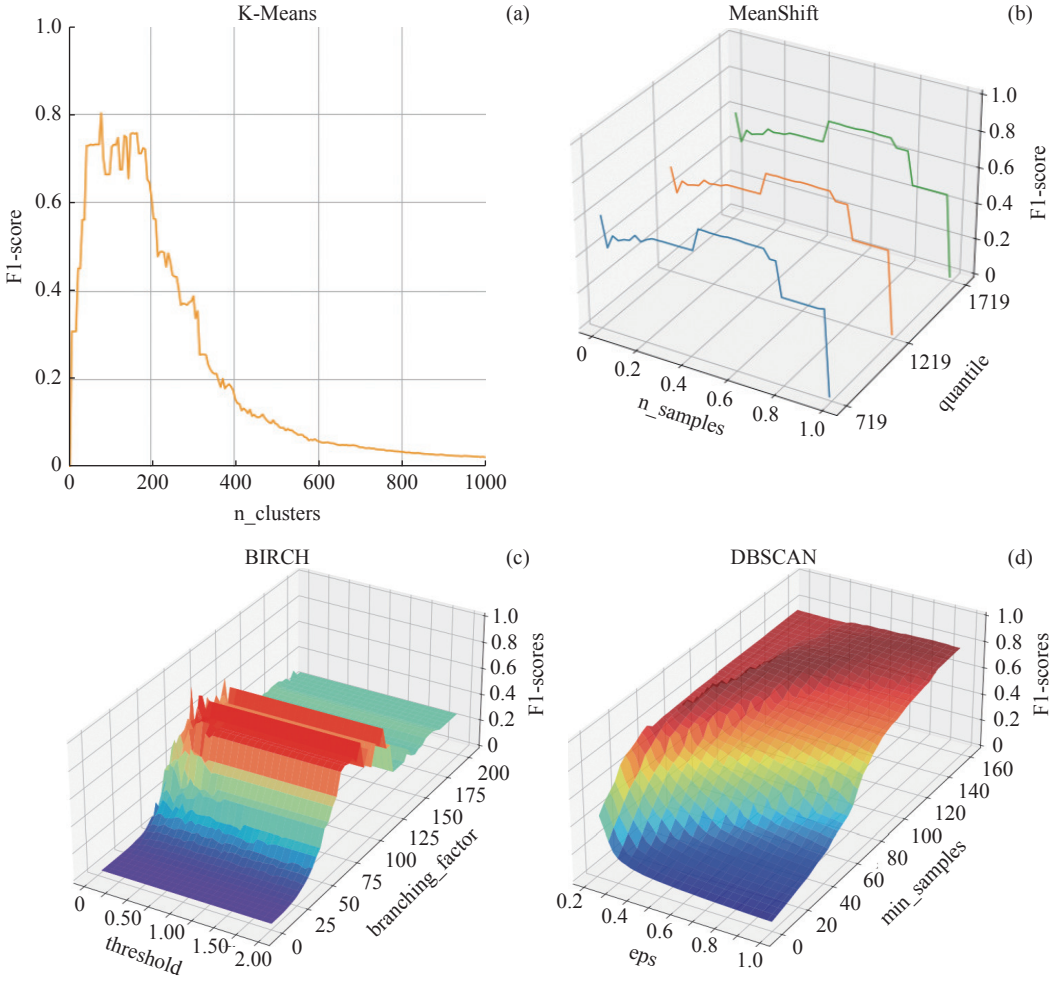

图 7 参数网格搜索过程

Fig. 7 Process of parameter grid searching

方式的具体参数设置列于表 5 和表 6。

分别统计网格搜索和自适应搜索两种方式下算法运行 10 次的结果，记录 F1-score 的最优值、最差值、均值、方差以及算法的执行时间，结果列于表 7。

由表 7 可得，基于层次的 BIRCH 聚类异常检测效果较其他聚类更好。对于各聚类算法，自适应搜索在最坏情况下的解优于网格搜索在最好情况下的解，说明自适应搜索能以更高效率搜索到更优的聚类超参数。同时，自适应搜索得到的 F1-score 的方差很小，说明该搜索方式有较强的稳定性。实验结果表明，本文提出的 UMOEAsII_BIRCH 算法在公开数据集 Thyroid 上测试有效，可拓展至不同数据集。

4 结语

为解决传统聚类算法对超参数高度敏感，网格搜索过程繁琐、结果不精细、效率低的问题，基于智能优化算法的启发式搜索能力，提出了超参数自适应优化的聚类算法 UMOEAsII_BIRCH，并在卫星遥测数据集和公开数据集上进行了验证。以网格搜索为基准，分别选取基于划分、基于层次和基于密度的聚类算法，对比本文提出的自适应搜索和网格搜索两种方式在等规模聚类次数下，算法 10 次运行中 F1-score 的最优值、最差值、均值、方差以及算法的执行时间。实验结果表明，自适应搜索克服了网格搜索中

表 5 UMOEAs-II 算法参数设置

Table 5 Parameter settings of UMOEAs-II

	最大评估次数	种群大小	问题维度	超参数名称	超参数取值范围
UMOEAsII_K-Means	200	8	1	n_clusters	[2, 1000]
UMOEAsII_DBSCAN	1200	22	2	eps	[0.1, 1.0]
				min_samples	[2, 160]
UMOEAsII_BIRCH	1600	24	2	threshold	[0.001, 2.0]
				branching_factor	[2, 200]

表 6 网格搜索测试参数组合

Table 6 Parameter combinations of grid search

算法名称	超参数名称	超参数取值范围	搜索步长	网格点数
K-Means	n_clusters	(2, 1000)	5	200
DBSCAN	eps	(0.1, 1.0)	0.03	1200
	min_samples	(2, 160)	4	
BIRCH	threshold	(0.001, 2.0)	0.05	1600
	branching_factor	(2, 200)	5	

表 7 算法测试结果

Table 7 Algorithm test results

	F1-score最优值	F1-score最差值	F1-score均值	F1-score方差	算法执行时间/s
K-Means	0.5965	0.5348	0.5674	2.6505×10^{-2}	1091.75
UMOEAsII_K-Means	0.5965	0.5965	0.5965	2.3406×10^{-16}	681.86
DBSCAN	0.4639	0.4470	0.4571	4.7600×10^{-3}	1443.58
UMOEAsII_DBSCAN	0.4671	0.4671	0.4671	4.6800×10^{-16}	1299.00
BIRCH	0.6139	0.4865	0.5379	3.8275×10^{-2}	2944.44
UMOEAsII_BIRCH	**0.6415**	**0.6415**	**0.6415**	0.0000	1351.90

注 加粗数据表示最优结果。

精细程度与效率之间矛盾的问题, 不受网格点的限制, 能在更短时间内寻得更精细的聚类超参数取值, 无需人工干预, 证明了将智能优化算法与聚类结合以实现聚类超参数自适应搜索的有效性。此外, 本文提出的算法结合了 BIRCH 算法和 UMOEAs-II 算法的优势, 实现了基于超参数自适应优化聚类的异常检测, 综合性能较高。不过, 该方法目前只针对离线检测进行了实验, 即在大量历史数据中发现异常样本。未来如拓展至实时在线检测将具有更重要的意义。

参考文献

[1] PAN D W, LIU D T, ZHOU J, *et al*. Anomaly detection for satellite power subsystem with associated rules based on Kernel Principal Component Analysis[J]. *Microelectronics Reliability*, 2015, **55**(9/10): 2082-2086

[2] ZHENG L, GUANG J, TANG S H. Fluctuation feature extraction of satellite telemetry data and on-orbit anomaly detection[C]//2016 Prognostics and System Health Management Conference. Chengdu: IEEE, 2017: 1-5

[3] ZHAO Jia, LV Hong, LIU Bao, *et al*. Fault diagnosis for GNSS receiver based on fuzzy Petri Net[J]. *Journal of Test and Measurement Technology*, 2017, **31**(5): 438-442 (赵佳, 吕弘, 刘宝, 等. 基于模糊Petri网的卫星导航接收系统故障诊断[J]. 测试技术学报, 2017, **31**(5): 438-442)

[4] ZHANG Huaifeng, JIANG Jing, ZHANG Xiangyan, *et al*. Novel anomaly detection method for satellite power system[J]. *Journal of Astronautics*, 2019, **40**(12): 1468-1477 (张怀峰, 江婧, 张香燕, 等. 面向卫星电源系统的一种新颖异常检测方法[J]. 宇航学报, 2019, **40**(12): 1468-1477)

[5] LI Yukui, LI Hu, HU Tai. In-orbit operational pattern monitoring algorithms based on LightGBM for Hard X-ray Modulation Telescope Satellite[J]. *Chinese Journal of Space Science*, 2020, **40**(1): 109-116 (李钰骅, 李虎, 胡钛. 基于LightGBM的HXMT在轨运行模式监测算法[J]. 空间科学学报, 2020, **40**(1): 109-116)

[6] LI Hu, GUO Guohang, HU Tai, *et al*. Ensemble learning for state recognition of payload from telemetry data[J]. *Journal of National University of Defense Technology*, 2021, **43**(6): 33-40 (李虎, 郭国航, 胡钛, 等. 遥测参数数据载荷状态判别集成学习方法[J]. 国防科技大学学报, 2021, **43**(6): 33-40)

[7] CASAS P, MAZEL J, OWEZARSKI P. UNADA: unsupervised network anomaly detection using sub-space outliers ranking[C]//Proceedings of the 10 th International Conference on Research in Networking. Valencia: Springer, 2011: 40-51

[8] LU Chunguang, YE Fangbin, ZHAO Ling, *et al*. Abnormal value detection of large power data based on density peak clustering[J]. *Science Technology and Engineering*, 2020, **20**(2): 654-658 (陆春光, 叶方彬, 赵羚, 等. 基于密度峰值聚类的电力大数据异常值检测算法[J]. 科学技术与工程, 2020, **20**(2): 654-658)

[9] WANG H, PENG M J, YU Y, *et al*. Fault identification and diagnosis based on KPCA and similarity clustering for nuclear power plants[J]. *Annals of Nuclear Energy*, 2021, **150**: 107786

[10] LI Nan, QIANG Yigeng, FAN Rui, *et al*. On the abnormal detection of the aircraft flight trajectory based on the abnormal factor statistics[J]. *Journal of Safety and Environment*, 2021, **21**(2): 643-648 (李楠, 强懿耕, 樊瑞, 等. 基于异常因子的航空器飞行轨迹异常检测研究[J]. 安全与环境学报, 2021, **21**(2): 643-648)

[11] PENG Xiyuan, PANG Jingyue, PENG Yu, *et al*. Review on anomaly detection of spacecraft telemetry data[J]. *Chinese Journal of Scientific Instrument*, 2016, **37**(9): 1929-1945 (彭喜元, 庞景月, 彭宇, 等. 航天器遥测数据异常检测综述[J]. 仪器仪表学报, 2016, **37**(9): 1929-1945)

[12] KANG Xu. Anomaly Detection Method for Sequential Telemetry Data[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2018 (康旭. 时序遥测数据异常检测方法研究[D]. 南京: 南京航空航天大学, 2018)

[13] ZHANG T, RAMAKRISHNAN R, LIVNY M. BIRCH: An efficient data clustering method for very large databases[J]. *ACM SIGMOD Record*, 1999, **25**(2): 103-114

[14] CHEN Jingwen. Research and Application of an Improved BIRCH Algorithm Based on Link[D]. Changchun: Jilin University, 2019 (陈婧文. 基于链接改进的BIRCH算法的研究与应用[D]. 长春: 吉林大学, 2019)

[15] ELSAYED S, HAMZA N, SARKER R. Testing united multi-operator evolutionary algorithms-II on single objective optimization problems[C]//2016 IEEE Congress on Evolutionary Computation (CEC). Vancouver: IEEE, 2016: 2966-2973

[16] ELSAYED S M, SARKER R A, ESSAM D L, *et al*. Testing united multi-operator evolutionary algorithms on the CEC2014 real-parameter numerical optimization[C]//2014 IEEE Congress on Evolutionary Computation (CEC). Beijing: IEEE, 2014: 1650-1657

[17] HRUSCHKA E R, CAMPELLO R J G B, FREITAS A A, *et al*. A survey of evolutionary algorithms for clustering [J]. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 2009, **39**(2): 133-155