基于特征冗余分析的高维多任务多目标特征选择算法

朱苗苗[1],姚香娟[1*],巩敦卫[2],张 岩[3]
(1. 中国矿业大学数学学院,江苏徐州 221116;2. 青岛科技大学自动化与电子工程学院,山东青岛 266061;
3. 宿迁学院信息工程学院,江苏宿迁 223800)

摘 要: 在处理高维分类数据集时,多目标特征选择进化算法存在计算资源耗费高、收敛速度慢的问题. 多任务优化作为一种可以有效降低搜索维度、提升搜索效率的手段,已被引入该领域,但现有算法多仅考虑特征重要性,忽视了特征间的冗余关系. 针对这一不足,本文提出了一种基于特征冗余分析的多任务多目标特征选择算法 MTGA. 该算法首先基于特征冗余度对所有特征进行聚类,将高维特征划分为多个冗余度较低的特征簇. 随后从各个特征簇中选择少量的重要特征构建多个子任务,在保留关键信息的同时有效剔除大量冗余特征. 此外,针对各子任务,设计了基于特征重要性的繁殖算子,并通过知识迁移策略实现不同子任务间的重要特征共享,避免算法陷入局部最优. 在 14 个高维 UCI 数据集上的对比实验结果表明,所提算法优于多种经典特征选择方法,展现出良好性能.

关键词: 高维特征选择;特征冗余性;多目标进化算法;多任务优化;知识迁移机制
基金项目: 国家自然科学基金(No.62373357)
中图分类号: TP18;O29 **文献标识码:** A **文章编号:** 0372-2112(2025)07-2342-16
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250405

A High-Dimensional Multi-Task Multi-Objective Feature Selection Algorithm Based on Feature Redundancy Analysis

ZHU Miao-miao[1], YAO Xiang-juan[1*], GONG Dun-wei[2], ZHANG Yan[3]
(1. *School of Mathematics*, *China University of Mining and Technology*, *Xuzhou*, *Jiangsu* 221116, *China*;
2. *College of Automation and Electronic Engineering*, *Qingdao University of Science and Technology*, *Qingdao*, *Shandong* 266061, *China*;
3. *School of Communication Engineering*, *Suqian University*, *Suqian*, *Jiangsu* 223800, *China*)

Abstract: Evolutionary multi-objective feature selection algorithms face challenges such as high computational cost and slow convergence when addressing high-dimensional classification datasets. Multi-task optimization has emerged as an effective paradigm to reduce search dimensionality and improve efficiency, and has been increasingly applied to this domain. Nevertheless, existing approaches predominantly focus on feature importance while neglecting redundancy relationships among features, which may compromise optimization performance. To overcome this limitation, this study proposes a novel evolutionary multi-task multi-objective feature selection algorithm based on feature redundancy analysis, referred to as MTGA. The proposed method first clusters all features according to their redundancy metrics, dividing the high-dimensional space into low-redundancy clusters. Then, different features are selected from each cluster to construct multiple subtasks, thereby preserving key information while eliminating redundancy. For each subtask, a new reproduction operator is designed based on feature importance. Additionally, a knowledge transfer mechanism facilitates the sharing of important features across subtasks, mitigating the risk of premature convergence. To validate the proposed algorithm, extensive experiments are conducted on fourteen high-dimensional UCI benchmark datasets. The results demonstrate that MTGA outperforms multiple classical feature selection methods, exhibiting excellent performance.

Key words: high-dimensional feature selection; feature redundancy; multi-objective evolutionary algorithm; multi-task optimization; knowledge transfer mechanism
Foundation Item(s): National Natural Science Foundation of China (No.62373357)

收稿日期:2025-05-22;录用日期:2025-06-27;责任编辑:覃怀银
*通讯作者:姚香娟

1 引言

分类问题是机器学习和人工智能技术的重要组成部分,是模式识别、自然语言处理、计算机视觉等应用中实现更智能、更精确系统的关键[1,2]. 在大数据时代,各行各业数据规模呈日益增长之势,其中包含大量不相关、冗余的信息. 如何从高维数据中提取有效信息,降低训练数据维度,成为分类问题面临的新挑战[3]. 特征选择(Feature Selection,FS)作为一种有效的数据预处理技术,通过从原始特征中挑选特征组成特征子集,实现数据维度的降低,可以有效地减少空间存储,提高计算效率,是数据降维的有效手段[4]. 改进特征选择技术以应对超高维数据集,已成为当前信息社会发展的实际需求.

在特征选择过程中,评估一个特征子集的优劣通常会设定一些预定目标,如最大化分类精度、最小化特征数、最小化特征成本等,其中分类精度和特征数是最为常见的两个目标. 根据评估特征子集时与分类器的结合方式,可将特征选择方法划分为过滤式(Filter)、包装式(Wrapper)、嵌入式(Embedded)3类[5,6]. 过滤式方法独立于分类模型,运用统计指标或评分函数,如方差、相关系数、互信息、卡方检验等,来评估特征与目标的相关性,进而过滤掉无关或冗余的特征[7,8]. 包装式方法通过训练分类模型来评估子集效果,并依据评估结果采用特定算法迭代地选择或移除特征,例如前向选择、后向消除或遗传算法[9]. 嵌入式方法则将特征选择过程融入模型训练之中,借助模型自身的特性来评估每个特征的重要性,从而自动选择有用的特征,如基于正则化或基于树模型的特征重要性评估方法[10].

基于进化算法(Evolutionary Algorithm,EA)的特征选择方法作为一种有效的包装式方法,近年来受到了研究学者广泛的关注[11]. 进化算法通过交叉、变异、选择等操作生成多个候选特征子集,并通过训练分类器评估选择最优特征子集. 与传统的特征选择方法相比,进化算法具有全局搜索能力强、无须领域知识、搜索速度快等优点,在实际问题中展现出较好的性能和效果[11,12]. 在早期的研究中,研究学者通常将特征选择的多个目标通过权重参数聚合成为一个目标,再使用进化算法对其进行搜索,如遗传算法(Genetic Algorithm,GA)[13]、粒子群算法(Particle Swarm Optimization,PSO)[14]、蚁群算法(Ant Colony Optimization,ACO)[15]等. 这样设计的好处是简单便捷,但缺点是搜索到的最佳特征子集非常依赖权重参数的定义. 此外,分类错误率与所选特征数往往是相互冲突的,因此单目标特征选择的优化效果较为有限. 而多目标特征选择将分类错误率与特征数视为独立的两个目标,在搜索中平衡两个目标,最终提供多个解决方案,更符合实际应用的不同需求. 因此,多目标特征选择进化算法逐渐成为研究热点.

然而,在面对高维分类问题数据集时,多目标进化算法面临着Pareto前沿高度不连续、搜索空间庞大、变量交互关系复杂等问题[16]. 为克服上述困难,众多研究者针对性地提出了多种高效的改进算法. 如Nguyen等人[17]在基于分解的多目标进化算法(MOEA/D)框架下,提出了一种新的基于多参考点的动态与静态双机制分解方法(MOEAD-DYN). 其中静态机制缓解了分解对帕累托前沿形状的依赖和不连续性的影响,动态机制通过检测冲突区域有效分配计算资源. 为应对特征选择中冗余解频发与多样性下降的问题,Xu等人[18]提出了一种重复分析的方法(DAEA),在环境选择阶段主动识别并剔除冗余解,同时引入基于多样性的选择策略以增强解集在目标空间中的覆盖能力. 此外,Han等人[19]提出了一种基于自适应策略的多目标粒子群优化特征选择算法(MOPSO-ASFS),该方法通过引入投影距离指标的自适应调节机制,有效平衡种群进化过程中的探索与开发能力,提升了搜索效率与全局最优解的获取能力. 然而,以上算法主要聚焦于目标空间的解集分布与资源分配优化,并不直接涉及搜索空间的降维与特征交互分析. 对于高维特征选择问题,搜索空间无疑是巨大的,带来了搜索难度大、计算资源开销高等一系列问题. 而且,面对复杂的交互特征,种群在进化过程中的搜索方向趋于简单,这将导致种群很容易陷入局部最优,所获得的特征子集多样性差.

为进一步缓解上述高维特征选择问题中的搜索空间复杂性与变量交互挑战,近年来多任务优化(Multi-Tasking Optimization,MTO)逐渐被引入特征选择领域,并展现出良好的潜力[20]. 进化多任务是一种通过相关任务之间的知识迁移来增强全局搜索能力和加速整体收敛的有效范式[21]. 在特征选择问题中,通过自定义的任务生成策略,将一个高维的特征选择任务划分为多个相互关联的低维子任务,并在每个子任务中独立优化特征子集解[22]. 通过在低维子空间中并行搜索,并在任务之间实现有效的知识共享与迁移,不仅降低了整体搜索维度,也增强了搜索的多样性,从而有效避免陷入局部最优解. 其中任务生成机制、子任务搜索策略及知识迁移机制是多任务优化框架中的关键组成部分.

任务生成指的是从原始特征集中选择特征组成不同子任务的过程,当前的多数算法主要侧重于特征重要性的评估,却普遍忽视了特征间的冗余关系. 例如,Chen等人[20]最早提出的MTPSO算法基于ReliefF评分度量特征重要性,并设计了一种拐点划分策略,将所有特征划分为“有希望的特征集”与“剩余特征集”,进而据此构建低维子任务. Li等人[23]则采用了多种信息评估指标,从而提升任务间的多样性,并对竞争性群体优

化算法(Competitive Swarm Optimizer, CSO)进行了改进,引入知识迁移机制以实现跨任务协同优化. 针对以上方法仅采用过滤式信息指标的缺点,Feng等人[24]引入了加权聚类方法生成低维辅助任务,并与原始高维任务共同优化,在提升算法搜索效率方面取得了良好表现. 但以上算法仍然主要基于特征之间的相似性进行划分,进化搜索阶段也未建立有效的冗余去除机制,因此在任务间仍可能出现冗余特征的重复选择,限制了解的多样性和最终模型的泛化能力.

为了进一步提升多任务优化在特征选择中的效果,除了任务划分,合理设计知识迁移机制和搜索机制也至关重要. 如Lin等人[25]引入了一种将多个相互关联的特征选择任务的解合并到一个解中的形式,从而提高传递共同知识的有效性和效率. 此外,还设计了一种新的搜索方法来促进跨多个任务的种群进化,从而实现它们之间的有效知识转移. Xu等人[26]开发了适应空间分布的初始化策略,并且通过分析当前种群的状态动态调整多任务框架和繁殖策略,以更好地平衡收敛和多样性,取得了良好成效. 在高维特征空间中,特征之间关系复杂,合理利用这些信息指导子代生成,不仅能有效减少无效搜索,还能更快聚焦于关键特征. 基于此,本文设计了一种结合特征重要性和冗余度的繁殖算子,并嵌入NSGA-II的基本框架进行子任务的搜索,该繁殖算子能够产生较稀疏子代,相比于传统算子更适用于特征选择问题. 此外,针对设定的任务生成机制,开发了相应的知识迁移策略,以促进相关任务间的知识共享与协同进化.

综上所述,本文在经典非支配排序遗传算法NSGA-II的基础上,提出了一种基于特征冗余分析的多任务多目标特征选择算法MTGA,具体工作有以下几方面.

(1)提出一种基于特征冗余分析的子任务生成策略,使每一个子任务在尽量少地选择冗余特征的同时,尽量多地包含关键特征.

(2)在经典NSGA-II算法的基础上,设计了新颖的稀疏二进制交叉和变异算子,可以根据特征重要度及冗余度生成较低维度的子代,更适用于高维特征选择问题.

(3)为了提高种群多样性且加速收敛,设计了任务相似性驱动的知识迁移机制,在不同任务间迁移重要特征. 同时,基于以上策略,提出了面向高维分类数据的多任务多目标特征选择算法MTGA.

(4)在14个特征维度超过3000维的UCI公共数据集上对所有算法的有效性进行评估. 实验结果表明,MTGA在多个数据集上性能优于现有多种特征选择算法.

2 相关工作

2.1 高维特征选择问题

高维特征选择旨在从一个特征数量庞大的原始数据集中,挑选出一个最佳的特征子集,以替代原数据集训练分类器,进而提高分类性能并降低计算复杂度. 该过程可被建模为一个两目标优化问题,目标分别为分类精度和特征数,建模如下:

$$\min \begin{cases} F_1(S)=1-\text{Accuracy}(S) \\ F_2(S)=|S| \end{cases}, \ S\subseteq\Omega \tag{1}$$

其中,Ω为所有可能的特征子集形成的集合,S表示一个特征子集,Accuracy(S)是由该特征子集在训练集上所获得分类器在测试集上的分类准确率,$|S|$表示该特征子集中的特征数量.

多目标进化算法生成的Pareto最优解集具有较优的鲁棒性,能够适应数据分布的变化和模型的不确定性[27,28]. 因此,与将目标加权建模成单目标的算法相比,本文选择将特征选择问题的分类精度和特征数视为两个独立的目标,使用多目标进化算法来解决,通过寻找Pareto最优解集,在多个目标之间实现最佳权衡,以满足不同应用场景的需求.

2.2 对称不确定性衡量特征关系

在高维特征选择问题中,特征之间通常存在复杂的交互关系. 依据这些先验信息对特征进行划分,有助于指导特征选择.

根据特征与分类标签之间的相关性,特征可划分为3种:强相关特征、弱相关特征与不相关特征. 与此同时,特征之间还存在冗余关系,包括完全冗余、部分冗余与互不冗余. 特征选择的目标是在保持高精度的同时选择尽可能少的特征. 因此,从特征关系的角度出发,理想的特征子集应由低冗余强相关的特征组成. 然而,在已有研究中,学者多聚焦于保留强相关的特征,较少关注冗余特征的消除,这在高维数据场景下可能严重影响算法的性能和效率. 因此,本文在算法设计中多处使用了特征关系分析,旨在有效识别并剔除冗余特征,从而降低搜索维度,提高搜索效率,并提升最终选取特征子集的质量.

对称不确定性(Symmetrical Uncertainty, SU)是一种衡量两个变量之间依赖关系的对称度量[29],具有归一化的特性,值域在[0, 1]之间,其表达式为

$$\text{SU}(X;\ Y)=2.0\times\frac{I(X;\ Y)}{H(X)+H(Y)} \tag{2}$$

其中,$I(X;\ Y)$是X与Y的互信息(Mutual Information, MI),衡量变量间关系依赖的程度. $H(X)$与$H(Y)$是X与Y的信息熵,是对随机变量不确定性的衡量[30].

若两个特征之间的SU值较高,说明它们存在较强

的冗余关系;而当某一特征与分类标签之间的SU值较高时,则表明该特征对分类任务具有较强的判别能力.因此,本文采用SU值作为度量指标,分别用于评估特征间的冗余性以及特征与标签之间的相关性,从而选择更合理的特征组合.

2.3 进化多任务优化算法

假设现在有L个优化问题,用T_k表示第k个问题$(1\leqslant k\leqslant L)$,每一个问题的搜索空间为$X_k$,目标函数为$f_k$,则一个多任务优化问题可以表示为

$$\begin{aligned}&\min F\left(f_1(x_1),f_2(x_2),\cdots,f_L(x_L)\right)\\&\text{s.t. } x_1\in X_1,x_2\in X_2,\cdots,x_L\in X_L\end{aligned}\tag{3}$$

其中,F表示多任务优化的目标函数.与人工智能领域中的多任务学习概念相似,多任务优化旨在同时求解多个相似或相关的优化问题,并通过任务间的信息共享与协同进化,提升整体优化性能.

在特征选择问题中,可通过自定义的任务生成策略,将原始高维特征选择任务划分为多个并行的低维子任务,从而有效降低计算成本并加快收敛速度.此外,在优化过程中引入知识迁移机制,可促进不同子任务间的有效信息交流,增强种群多样性,提升搜索效率,并有助于避免算法陷入局部最优解.

因此,本文采用多任务多目标的算法框架来解决高维特征选择问题,所提出的算法能够在特征维度压缩与分类性能提高之间取得良好平衡.

3 基于冗余分析的多任务多目标特征选择算法

本节将详细介绍MTGA的核心组成模块,包括基于冗余分析的子任务生成策略、稀疏子代生成策略、任务相似性驱动的知识迁移策略,并在最后给出了算法整体框架.

3.1 基于冗余分析的子任务生成策略

采用多任务框架解决高维特征选择问题的首要步骤是根据原始问题生成多个平行子任务.要生成子任务,首先要确定任务的特征空间T,即一组经过筛选的重要特征集合,随后在该特征空间内进行搜索与优化.如前文所述,理想的特征子集应由低冗余强相干特征组成,因此子任务特征空间可通过删除大量不相关及相互冗余特征来构建.

本文使用SU作为特征重要性及冗余度的度量指标.$\mathrm{SU}(f_i,f_j)$表示特征f_i与特征f_j的SU值,定义f_i与f_j的距离$d_{i,j}$为

$$d_{i,j}=\frac{1}{\mathrm{SU}(f_i,f_j)}\tag{4}$$

在式(4)的距离定义下,两个特征之间距离越小表示相互冗余度越高.基于该距离定义,本文采用K近邻聚类算法(K Nearest Neighbor,KNN)对特征进行划分,将冗余度较高的特征分到相同的特征簇中.不同簇间特征冗余度较低,因此从每个特征簇中选择具有代表性的特征,可以获得一组较好的低冗余特征.在聚类过程中,本文算法的聚类中心采用文献[31]中提出的概率驱动初始化策略.该策略首先随机选择一个中心点,后续中心点优先选择距离已有中心较远的点作为新中心,这样操作有助于提升初始化的多样性,从而提高整体聚类质量.而聚类簇数的设定,考虑到过多的簇可能导致特征划分过于零散,而过少的簇又会导致划分失效,因此本文遵循常用的启发式设定规则,将簇数设为$K=\max\left(2,\left\lfloor\sqrt{D}/2+0.5\right\rfloor\right)$,其中$D$为特征总数.该设定综合考虑了特征维度与簇划分的平衡,在保证聚类有效性的同时避免过度分簇所导致的簇内特征稀疏问题,在聚类相关研究中被广泛采用[32].

在传统的基于聚类的特征选择方法中,通常假设不同特征簇之间互不冗余,因此仅从每个簇中选取一个代表特征.然而在实际的高维数据中,若过度压缩特征空间,可能会造成信息损失并影响分类性能.因此,本文根据特征的重要性评估,从每个簇中选取多个具有代表性的特征共同构建子任务的特征空间,以在保持低冗余的同时提升特征集的表达能力和分类效果.选择特征组成子任务的过程如下所述.

假设当前有K个特征簇,对于某一个特征簇G_k,G_k中的特征个数为Counts_k,则G_k的重要度可以用其包含的所有特征的平均SU值$I(G_k)$来衡量,如式(5)所示:

$$I(G_k)=\frac{\sum\limits_{f_j\in G_k}\mathrm{SU}(f_j)}{\mathrm{Counts}_k}\tag{5}$$

$I(G_k)$越高表示该簇特征与目标变量的平均相关性越强,则该簇特征被选择的概率也应该越高.根据$I(G_k)$设计的特征选择概率参数$P(G_j)$计算公式如下:

$$P(G_k)=\alpha\cdot\frac{\mathrm{e}^{I(G_k)}}{\sum\limits_{k=1}^{K}\mathrm{e}^{I(G_k)}}\tag{6}$$

其中,α是一个超参数,用于调节子任务特征空间的维度大小.通过这样的任务生成机制,从不同特征簇中选择较重要特征,减少选择冗余或无关的特征组,增强后续算法进化搜索阶段的效率.

图1展示了一次子任务生成过程:所有原始特征被划分为了3个特征簇,然后按设定概率从不同特征簇中选择特征组成多个子任务.该策略具体流程如算法1所示.

3.2 稀疏子代生成策略

在完成搜索空间的降维约简后,每个低维子任务的搜索至关重要.本文以NSGA-II作为基础优化算法,并针对高维特征选择问题的稀疏特性进行定制化改

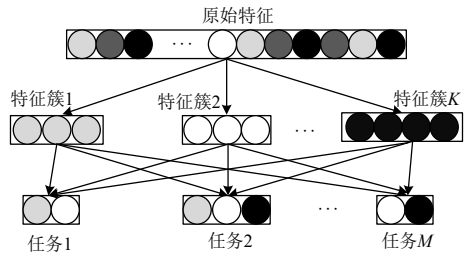


图1 子任务特征空间生成过程

算法1 子任务生成策略

输入：原始数据集$E=(X,Y)$，其中X为样本特征，Y为样本标签，任务数M，选择概率参数α

输出：M个子任务$\{T_1,T_2,\cdots,T_M\}$

1. 计算特征与标签之间以及特征两两之间的SU值
2. 根据式(4)计算特征之间距离$d_{i,j}$
3. 将距离输入K近邻聚类算法，得到K个特征簇$\{G_1,G_2,\cdots,G_K\}$，其中第k个特征簇包含$|G_k|$个特征
4. 根据式(6)~(7)计算每个特征簇G_k的选择概率$P(G_k)$
5. for $m=1:M$ do
6. for $k=1:K, j=1:|G_k|$ do
7. 如果随机数小于$P(G_k)$，则将G_k的第j个特征加入T_m
8. end for
9. end for
10. 返回生成的M个子任务

进. 算法使用二进制编码，假设原始数据集总特征数为D，则染色体长度等于D. 其中，每个位点取值为1表示选择对应特征，取值为0则表示不选择.

在种群初始化阶段，通过随机均匀的方式生成初始个体. 在迭代阶段，采用二进制锦标赛选择策略以生成父代个体. 针对子代的生成，本文设计了适用于高维特征选择场景的交叉与变异算子. 具体而言，在交叉操作中，对于某一个基因位，当两个父代取值相同时，意味着父代对该特征的重要度评价一致，所以子代在该基因位的取值与父代保持相同，不作改变. 为了生成稀疏的子代，对于取值不同的基因位，依据对应特征的SU值对其进行排序，将处于末位20%的特征视为不重要特征进行截断，即在子代中取值都赋0，以此减少特征数量. 最后，将剩余的特征随机分为不同两组，每个子代选择其中一组，这样就生成了两个不同的子代，并且兼顾保留重要特征与促进种群多样性. 图2展示了本文所定义的交叉操作产生子代的流程. 示例中，特征2、5、8为父代取值不同的特征，按照重要度排序后对特征2进行截断，即两个子代均不选择特征2，对于特征5与8子代随机选择.

在变异操作方面，结合高维特征选择问题所具有的稀疏特性，本文提出按位差异化的变异策略，即不同基因位具有不同的变异概率，且取值为1与取值为0的

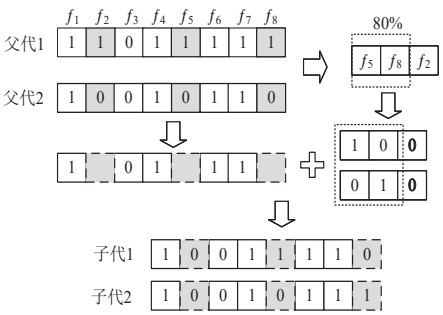


图2 交叉操作示例

基因位其变异规则亦不相同. 若某基因当前取值为1，则表示当前个体已选择该特征；若发生变异（即翻转为0），则意味着放弃该特征，因此其变异概率应与特征重要度成反比——SU值越高，变异概率越低. 反之，对于基因位当前取值为0的情况，若其翻转为1，则表示新增一个特征，因此其变异概率应与SU值成正比. 此外，为进一步强调选择稀疏性，在0翻转为1的情形下，总体变异概率需适当减小. 所以，变异概率设置如下：假设某个特征f_i^k属于特征簇G_k，并且取值为1，则其变异概率$P_{1\to0}$由式(7)计算，当随机数rand小于$P_{1\to0}^m$时进行变异操作翻转为0，这样高SU值特征更容易被选择，低SU值特征更容易被去除. 若f_i^k取值为0，则其变异概率$P_{0\to1}$由式(8)计算，其中$0\to1$变异参数β用于控制子代新增特征数. 在NSGA-II框架中嵌入以上交叉变异算子后，更新子任务种群的其他整体流程不变，如算法2所示.

$$P_{1\to0}=\frac{I(G_k)}{\mathrm{SU}(f_i^k)+I(G_k)} \tag{7}$$

$$P_{0\to1}=\beta\cdot\frac{\mathrm{SU}(f_i^k)}{\mathrm{SU}(f_i^k)+I(G_k)},\ (\beta\leqslant1) \tag{8}$$

算法2 子任务搜索

输入：子任务m的种群P_m及其种群规模N_m，变异概率参数β

输出：更新后的种群P_m及其精英存档A_m

1. 从P_m中根据竞标赛选择N_m个父代
2. 父代个体交叉、变异生成N_m个子代p_m
3. 合并种群$P'_m=P_m\cup p_m$
4. 对P'_m进行快速非支配排序并计算拥挤距离
5. 根据Pareto等级与拥挤度从P'_m中选择N_m个体更新P_m
6. A_m存储P_m中第一前沿个体
7. 返回P_m及A_m

3.3 知识迁移策略

在生成多个子任务之后，可以通过知识迁移机制在不同任务间实现信息共享，以提高算法效率并降低陷入局部最优的风险. 知识迁移分为4个关键组成部分：迁移发生、迁移对象、迁移内容与迁移方式，从这

4个部分可以系统地阐明知识迁移的整体过程.

(1)迁移发生:迁移应该在种群陷入局部最优时进行,否则会影响收敛过程. 对于多目标进化算法,可以通过检验当前Pareto前沿与上一代Pareto前沿的差异来判断种群是否处于局部最优. 计算当前代Pareto前沿解与上一代前沿解之间的欧氏距离,当连续改进距离为0时,说明种群进化停滞,未能进一步发现更优解. 在这种情况下需要跨任务学习以跳出局部最优.

(2)迁移对象:任务之间的相似性是决定从哪一个任务中迁移知识的关键依据,相似的任务间更有可能具备有用的知识. 本文采用Jaccard相似系数来衡量任务相似性,依据重叠特征的比例确定迁移对象,计算公式如式(9)所示,其中Fature_m与Fature_n分别是任务T_m与T_n搜索空间包含的特征.

$$\text{Sim}(T_m, T_n)=\frac{|\text{Fature}_m \cap \text{Fature}_n|}{|\text{Fature}_m \cup \text{Fature}_n|} \tag{9}$$

当触发迁移机制时,依据式(9)计算当前任务与其他任务之间的相似性,并选择相似度最高的任务作为知识迁移的来源. 图3展示了一次迁移对象选择的流程.

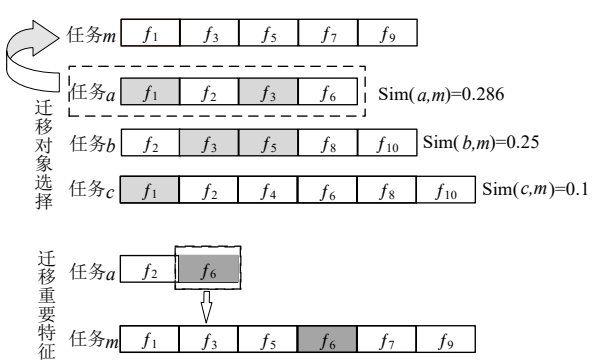


图3　选择迁移对象及迁移内容示例

(3)迁移内容:由此对所有特征进行降序排序. 其中,排名前20%的特征被认子任务生成机制可知,重要特征是否被包含于子任务的特征空间对最终优化结果具有决定性影响. 如果在任务构建阶段遗漏了关键特征,则对应子任务在后续进化中将无法获得高质量解,导致计算资源浪费. 因此,在进化过程中,需要动态调整各子任务的特征选择集合,知识迁移机制恰好可用于实现这一目标. 具体来说,在当前代,统计每个特征被选择的次数,并据定为重要特征,作为迁移内容在不同任务间共享. 在图3示例中,任务a特有的特征为特征2与特征6,对其进行排序后,选择更重要的特征6迁移到当前任务中.

(4)迁移方式:在知识迁移发生后,新引入的特征如何在种群个体中合理分配,是影响迁移效果的关键因素. 若所有个体均无条件选择新特征,随着进化进行,特征数量将不断膨胀,不利于以最小化特征子集为优化目标的特征选择任务. 为此,本文引入了存档机制:每一代将当前Pareto前沿的精英个体进行存档. 知识迁移发生后,非精英个体以一定概率选择新引入特征,而精英个体保持其原有特征子集不变. 此策略一方面保障了种群进化方向的稳定性,避免迁移引发的大幅度偏移;另一方面,也使得新增特征能够以合理比例分布于种群内部. 图4是迁移的一次示例,其中非精英个体以0.7的概率选择新特征,而精英个体保持不变.

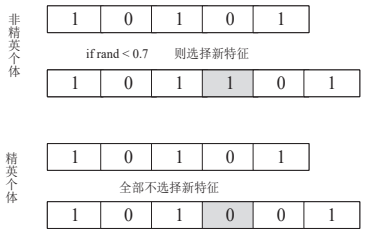


图4　知识迁移过程迁移方式示例

通过以上4个部分完整定义了一次知识迁移,其具体实现流程如算法3所示. 需要特别说明的是,尽管知识迁移过程中可能引入额外特征,但该操作是伴随进化逐步进行的,主要在进化后期触发,从而实现对解空间的精细化探索. 这种设计有助于在保持种群收敛性的同时,进一步提升最终解集的质量和多样性.

算法3　知识迁移策略

输入:当代种群$P=\{P_1,P_2,\cdots,P_M\}$,当代精英存档$A=\{A_1,A_2,\cdots,A_M\}$,上一代精英存档$A'=\{A'_1,A'_2,\cdots,A'_M\}$,子任务特征空间$T=\{T_1,T_2,\cdots,T_M\}$

输出:更新后的种群$P=\{P_1,P_2,\cdots,P_M\}$,更新后的子任务$T=\{T_1,T_2,\cdots,T_M\}$

1. 根据式(12)计算不同任务间相似度
2. for $m=1:M$ do
3. 计算A_m与A'_m之间的欧氏距离
4. 若精英存档改进欧氏距离连续为0,则认为T_m陷入局部最优,触发知识迁移,否则跳出循环
5. 选择相似度最高的任务$T_n(n\neq m)$
6. 计算T_n中每个特征被选择的次数,前20%特征F_{imp}迁移进T_m
7. 对于P_m中体,如果属于精英存档A_m,则保持不变,否则,根据概率选择F_{imp}中特征
8. end for
9. 返回更新后的P与T

3.4　多任务多目标特征选择算法框架

综合前述任务生成策略、子代更新策略及知识迁移策略,本文提出了一种基于特征冗余分析的进化多任务多目标特征选择算法MTGA. 该算法在充分利用特征间相关性与冗余信息的基础上,有效构建多任务

模型,并借助多目标优化框架实现特征数与分类性能的协同优化. 算法的整体流程如图5所示,算法的输入为包含全部原始特征及分类标签的样本数据集,输出为所有子任务搜索得到的多个特征子集进行非支配排序后组成的Pareto最优解集,可用于后续在测试集上进行模型性能评估. MTGA流程图对应的具体算法伪代码如算法4所示.

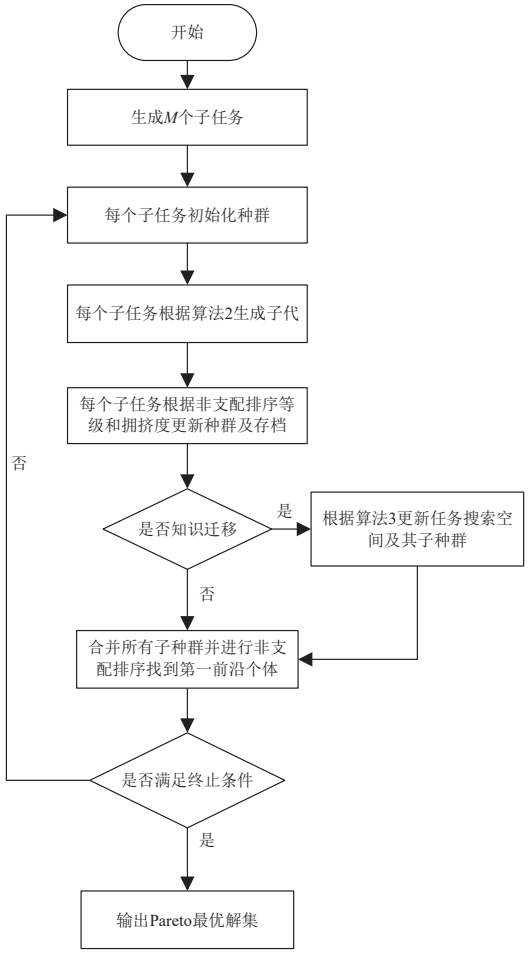


图5 MTGA流程图

算法4 MTGA流程

输入:样本数据集$E=\{X,Y\}$,其中X为样本特征,Y为样本标签,最大迭代次数iter,任务数M,种群总个体数N,选择概率参数α,变异概率参数β

输出:一组不同特征子集构成的Pareto解集

1. 生成子任务←算法1
2. 在每个子任务特征空间中随机初始化对应子种群,产生总种群$P=\{P_1,P_2,\cdots,P_M\}$,其中每个子种群P_m的个体数为N/M
3. 迭代计数$g\leftarrow 0$
4. while $g<$iter do
5. for $m=1{:}M$ do
6. 更新P_m和A_m←算法2
7. end for
8. 判断是否满足知识迁移条件,若满足则执行知识迁移策略←算法3
9. 对总种群所有个体执行非支配排序及拥挤距离计算
10. 更新全局Pareto解集
11. $g=g+1$
12. end while
13. 返回最终Pareto解集

4 实验及结果分析

为了验证所提方法的有效性,本节给出了MTGA的实验结果,主要包括4部分:(1)参数敏感性分析;(2)算法性能对比实验;(3)关键策略有效性分析;(4)计算复杂度分析. 最终实验结果验证了本文算法的有效性.

4.1 实验数据集及基本配置

为了全面评估MTGA在高维特征选择任务中的性能,本研究选用了来自UCI公共数据库的14个高维基准数据集进行实验. 这些数据集的特征规模处于3 312~12 533维之间,其余实例数、标签类别数等详细信息如表1所示.

表1 UCI公共数据集

序号	数据集	特征数	实例数	类别数
1	Lung	3 312	203	5
2	GLIOMA	4 434	50	4
3	DLBCL	5 469	77	2
4	Leukemia1	5 327	72	3
5	9Tumor	5 726	60	9
6	Brain1	5 920	90	5
7	drivface	6 400	573	2
8	Carcinom	9 182	174	11
9	nci9	9 712	60	9
10	acrene	10 000	100	2
11	Brain2	10 367	50	4
12	prostate	10 509	102	2
13	Leukemia2	11 225	72	3
14	11Tumor	12 533	174	11

本文采用6种不同类型的特征选择进化算法与MTGA进行对比实验. 这6种算法包括4种先进的多目标高维特征选择方法和2个经典多目标进化算法,具体为:(1)基于分解的多目标进化算法(MOEA/D)[33];(2)改进非支配排序遗传算法(NSGA-II)[34];(3)基于导向矩阵的多目标特征选择算法(SMMOEA)[35];(4)基于可变粒度搜索的多目标特征选择算法(VGSMOEA)[36];(5)基于问题重组和复制的多目标特征选择算法(PRDH)[37];(6)基于重复解分析的特征选择算法(DAEA)[18]. 除两种经典算法外,其余算法均为近年来提出的用于解决高维多目标特征选择问题的进化算

法. 通过与这6种算法进行性能对比实验,能够验证MTGA的有效性.

在实验过程中,为确保实验的公平性及结果的可对比性,所有进化算法的总种群数量和迭代次数均设定为100,其余参数均采用对应原文献或源代码中的取值. 不同算法参数设置的详细信息如下所示.

(1)MOEA/D:邻居数量$T=\lfloor 0.2N \rfloor$,N为种群规模,每代每个个体最多更新2个邻居个体,局部配对率0.9. 父代个体通过DE/rand/1算子生成子代解,缩放因子为0.5,交叉概率为0.9.

(2)NSGA-II:采用单点交叉和按位变异,交叉概率$P_c=1.0$,变异概率$P_m=1/D$,D为特征维度.

(3)SMMOEA:衰减因子$\gamma=0.1$.

(4)VGSMOEA:粒度转换参数$\alpha=10$,初始聚类参数$k=\min\{10\times\log_2 n, n\}$.

(5)PRDH:采用单点交叉和按位变异,交叉概率$P_c=1.0$,变异概率$P_m=1/D$,D为特征维度.

(6)DAEA:邻居数量$T=\max(4, \lfloor 0.2N \rfloor)$,局部配对率0.8,全局配对率0.2,修正变异率0.8,均匀变异率0.2.

对于多任务算法,由于每个任务都拥有独立的子种群,因此总的种群大小平均分配给所有任务. 例如本文算法MTGA,当任务数$M=4$时,则每个子任务种群数设置为25. 实验时,将数据集随机划分为测试集和训练集,划分比例分别为30%和70%. 测试集不参与进化算法的迭代评估过程,仅用于验证算法输出特征子集所训练模型的有效性;训练集则用于最终的模型训练以及进化过程中的子集评估. 为减少模型训练随机性对实验结果的影响,在进化过程中的子集评估采用5折交叉验证法. 此方法将训练集划分为5个子集,每轮使用其中4个子集训练模型,剩余1个用于验证,循环5次后取平均分类误差作为该子集的适应度指标. 通过交叉验证可有效减少偶然因素对特征选择结果的影响,提升优化算法的稳健性.

分类模型方面,实验选用KNN作为性能评估的分类器. KNN是一种非参数、基于实例的学习方法,实现简单、直观易懂,广泛应用于各种分类任务. KNN对噪声样本和冗余特征较为敏感,若包含大量无关或冗余特征,将显著降低分类性能. 因此,采用KNN作为分类器有助于直接反映特征选择对分类性能的提升作用,是验证特征子集质量的理想工具. 对于参数K,在特征选择进化算法研究中,K=5是被广泛采用的实验参数设置[36,37]. 并且在多数文献和实证研究中,K=5也被认为是能够在分类精度与计算效率之间取得良好权衡的经验值[38]. 因此,本实验采用K=5作为分类参数,以确保模型性能的稳健性与算法的可对比性.

在对比指标方面,本文从两个角度对算法性能进行考察:解集分布质量和特征子集质量. 在解集分布质量上,本文依据超体积指标(HyperVolume,HV)和反世代距离(Inverted Generational Distance,IGD)对算法性能进行评估[39,40]. 在HV计算过程中,将两个目标值归一化到[0,1]区间,并将参考点设置为[1,1]. 对于IGD指标,由于特征选择属于复杂的组合优化问题,难以获得真实的Pareto前沿,因此本文将所有方法运行20次所得的所有解组成联合种群,并对联合种群进行非支配排序,取所得最优前沿作为真实前沿用以计算IGD值. 在特征子集质量上,考虑到分类准确率是特征选择任务中的核心目标,本文以每组帕累托解在测试集上所能获得的最大分类准确率(Accuracy,ACC)作为主要评价依据,用以衡量所选特征子集的实际有效性.

4.2　参数敏感性分析

本节进行参数敏感性分析实验,系统考察不同参数对算法最终性能的影响. MTGA主要包含4个重要参数:(1)子任务数量参数M;(2)子任务选择概率参数α;(3)变异概率参数β;(4)特征截断阈值.

其中,α控制子任务搜索空间特征数量,β控制生成子代稀疏性. 为了确定α合适取值范围,首先在预实验中在[0.1,0.9]区间中进行粗粒度实验,结果发现当$\alpha<0.3$时,初始特征过少,表现普遍不稳定;当$\alpha>0.5$时,选入特征过多,任务间相似,反而降低算法性能. 基于以上观察,最终将α实验区间限定为[0.3,0.5],并在其中选取5个典型代表值,即$\alpha\in\{0.3, 0.35, 0.4, 0.45, 0.5\}$. 对于参数$\beta$进行了类似的预实验,最终设置了$\beta$的实验取值分别为$\beta\in\{0.005, 0.008, 0.01, 0.02, 0.03, 0.05\}$. 因此,共产生了2个参数共30种组合. 在4个不同维度的代表数据集(Lung、Leukemia1、drivface、Leukemia2)进行了实验,每组实验运行10次.

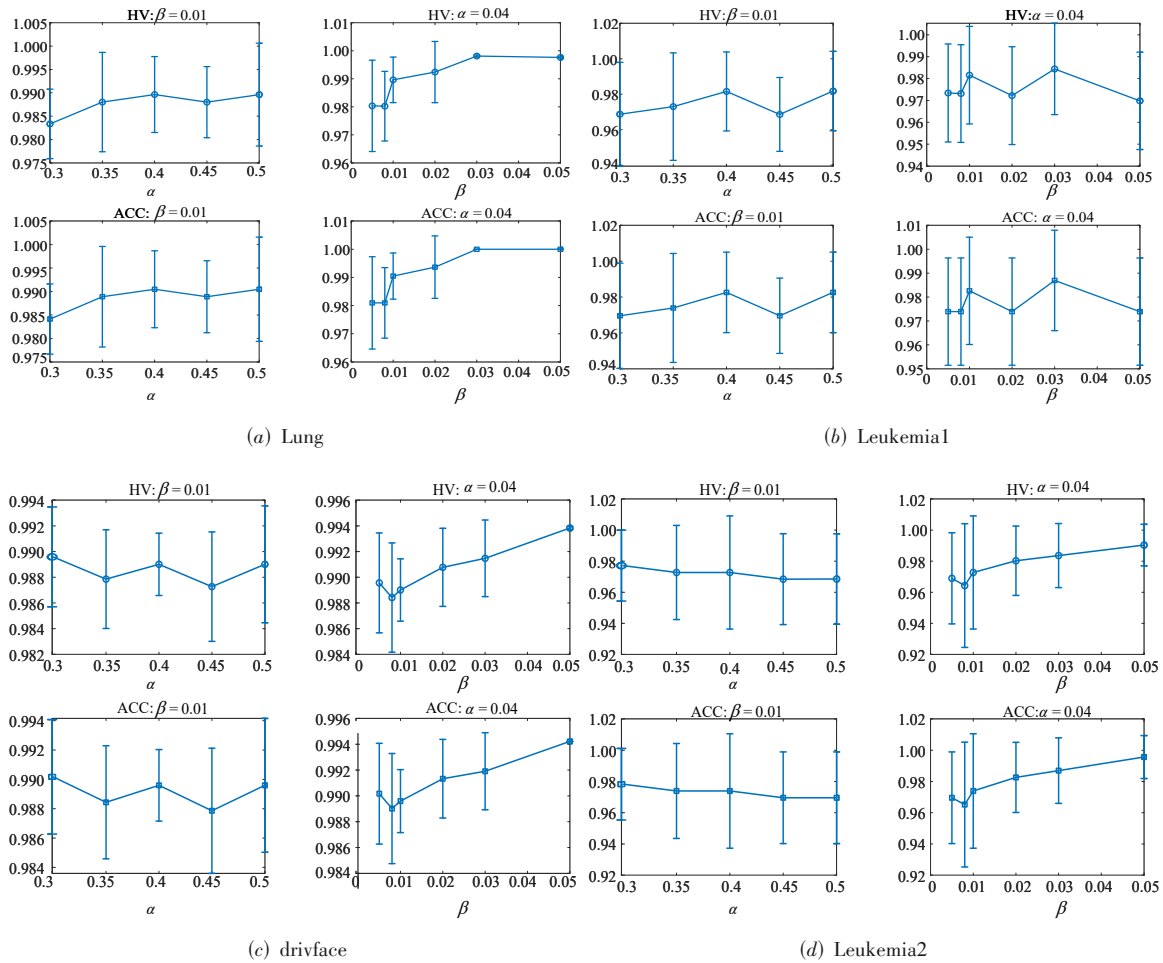
为了系统评估两个参数的敏感性,本文采用了两种可视化分析手段展示了实验结果. 首先,图6是每个数据集在固定一个参数取值下,另一个参数变化对算法性能的影响曲线,以HV与ACC作为评价指标,绘制了10次运行平均值及其标准差,揭示参数变化对收敛性与稳定性的影响. 实验结果表明,在设置范围内,数据集对α取值相对不敏感,性能变化平稳,表明算法具有良好鲁棒性. 但在Lung数据集上仍然能观察到较大的α能提升最终解集质量. 对于参数β,算法更为敏感,过小的β会导致关键特征被舍弃,表现为HV与ACC的显著下降,而较大的β在多数数据集上表现更为优越,能够在压缩特征维度的同时维持较好的解集质量.

图7通过箱线图展示了30组不同参数组合在4个数据集上10次运行中的表现,记录的每个数据点为算法在该参数组合下,在4个数据集中所获得的平均HV

与平均ACC. 从图7中可以观察到,不同组合参数在性能指标上存在显著差异,图中用红色标记出了平均表现最优的参数组合,即$\alpha=0.5,\beta=0.05$. 在这个组合下,所有数据集的平均HV与ACC中位数显著高于其他组合,并且上四分位数与中位数接近,未出现离群点,说明该参数组合具备一定鲁棒性与稳定性. 在最佳组合处,两参数相互补充,共同促进了算法在搜索空间中的高效搜索与稳定收敛.

在多任务特征选择框架中,子任务数量M取值往往与具体问题的特征维度、数据分布、种群规模以及可接受的计算开销等因素密切相关. 此外,已有研究表明,在中小规模多任务优化问题中,将子任务数量设为3~6往往能在性能与计算效率之间取得较好平衡[20]. 在本文的实验中,考虑到种群规模固定为100,为了保证实验的一致性,在所有数据集中统一将MTGA的子任务个数设定为4.

此外,在交叉算子和知识迁移策略中,本文使用了20%的截断阈值来筛选重要或不重要的特征,为了验证该阈值设定的合理性,在3个代表性数据集上分别对不同阈值(5%、10%、20%、30%、40%)进行了敏感性实验. 每组实验独立运行10次,并统计HV、ACC、ACC对应解的特征选择率(Ratio)3项指标. 如图8所示,每个点表示对应阈值下该指标的平均值,上下界则表示±1标准差,用以评估该设置下算法性能的稳定性. 图8中红色标注为对应指标的最优阈值位置,可以看到10%~20%为较优的截断阈值,因此本文后续实验中阈值取值为20%.



(a) Lung (b) Leukemia1

(c) drivface (d) Leukemia2

图6 4个数据集上α与β的参数敏感性分析

4.3 算法性能对比实验

表2和表3给出了MTGA与各对比算法在20次运行下的平均HV值及平均IGD值,其中阴影部分表示在对应数据集中的最优结果. 并在显著性水平为0.05下进行Wilcoxon秩和检验,“+”“=”和“−”分别表示所提算法相较于对比算法在统计意义上显著优于、无显著差异或显著劣于. 在HV指标对比中,MTGA在14个数据集中有11个取得了最佳的平均HV值,仅在Lung、9Tumor、11Tumor数据集上次于VGSMOEA和DAEA. 从Wilcoxon秩和检验结果来看,MTGA在14个数据集中

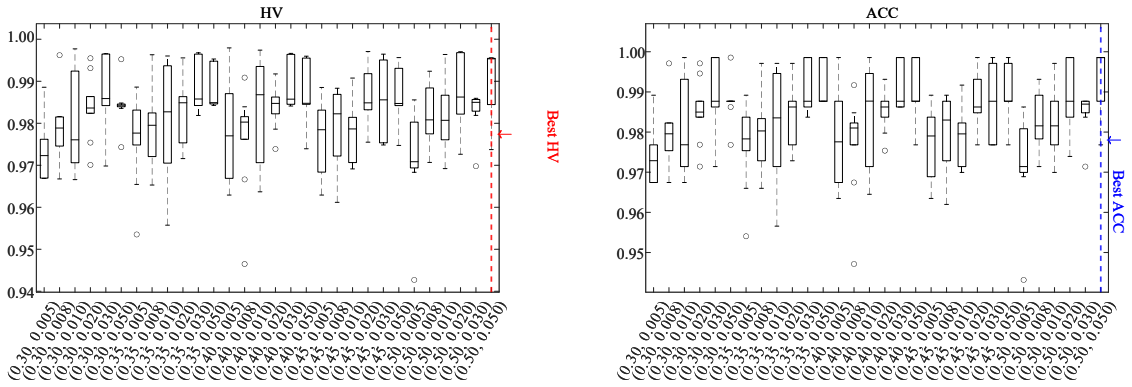
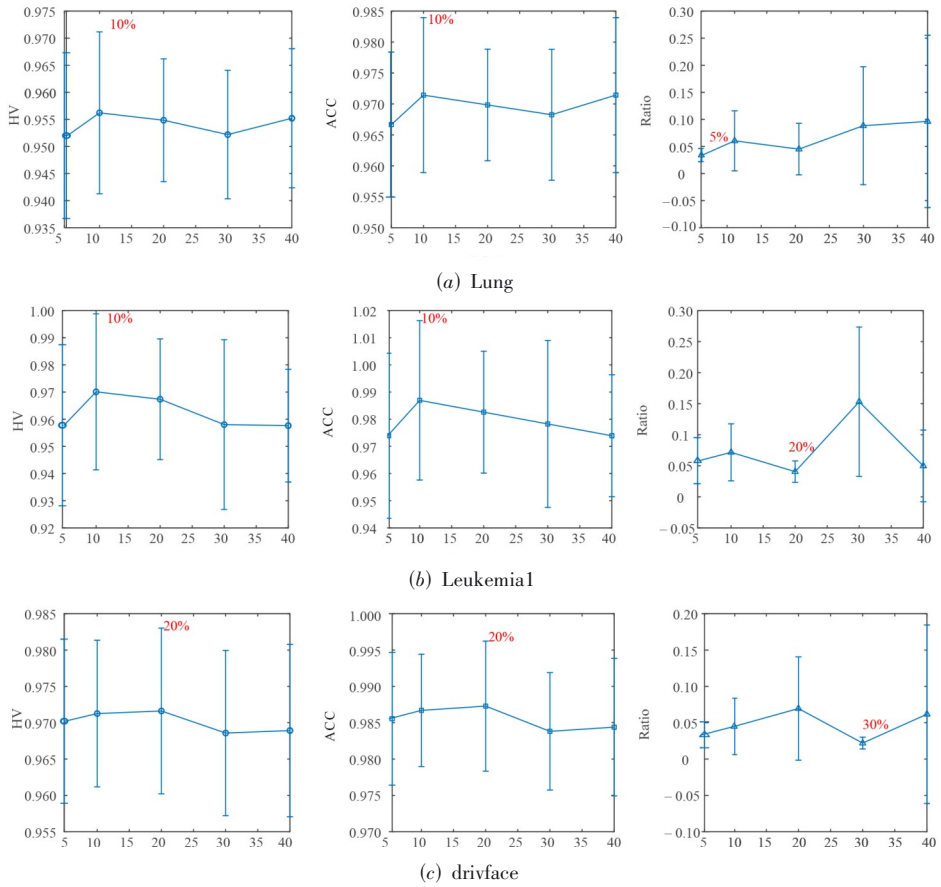


图7　(α,β)对平均HV与ACC的影响:箱线图分析



(*a*) Lung

(*b*) Leukemia1

(*c*) drivface

图8　特征截断阈值对算法性能的影响

分别在14、14、13、5、14、6和8个数据集上显著优于NSGA-II、MOEA/D、SMMOEA、VGSMOEA、PRDH和DAEA. 相比之下,只有VGSMOEA在9Tumor数据集上表现出显著优于MTGA的性能. 这可能是由于VGSMOEA在搜索过程中采用动态粒度调整策略,能够快速探索整个解空间,避免了搜索早期过快收敛的问题,从而在多个数据集上保持了较高的稳定性. 此外,通过Friedman检验对所有算法进行整体排序,MTGA的平均排名为1.21,明显优于其他对比算法,进一步表明MTGA整体性能优势突出.

为直观展示实验结果,本文通过图9绘制了各算法在部分数据集中所得的非支配前沿分布情况. 为了更清晰地呈现左侧帕累托前沿的细节,对图9中局部区域进行了放大. 从图9中可以观察到,NSGA-II和MOEA/D获得的解集虽然在分类错误率方面表现尚可,但选取的特征比例较高,导致其解集分布在目标空间中的局部区域. 相比之下,其余多目标特征选择算法在分类错误率与特征比例之间表现出更好的平衡性.

表2 不同算法的HV值对比

HV	MOEA/D	NSGA-II	SMMOEA	VGSMOEA	PRDH	DAEA	MTGA
Lung	0.620 1(+)	0.719 0(+)	0.938 8(+)	0.982 8(=)	0.952 1(+)	0.991 7(=)	0.985 0
GLIOMA	0.505 9(+)	0.597 1(+)	0.760 8(+)	0.868 5(+)	0.724 3(+)	0.917 3(=)	0.928 4
DLBCL	0.581 7(+)	0.583 1(+)	0.930 8(+)	0.990 8(+)	0.945 4(+)	0.944 9(+)	0.992 9
Leukemia1	0.548 4(+)	0.556 1(+)	0.873 4(+)	0.953 4(=)	0.893 2(+)	0.873 7(+)	0.960 3
9Tumor	0.351 3 +)	0.412 6(+)	0.451 2(=)	0.532 9(=)	0.422 4(+)	0.520 4(=)	0.530 8
Brain1	0.460 9(+)	0.535 4(+)	0.721 2(+)	0.770 6(+)	0.731 9(+)	0.773 0(=)	0.794 0
drivface	0.596 6(+)	0.694 0(+)	0.969 1(+)	0.981 3(−)	0.970 6(+)	0.968 3(=)	0.983 5
Carcinom	0.496 7(+)	0.580 0(+)	0.824 6(+)	0.872 1(=)	0.738 9(+)	0.802 8(+)	0.873 5
nci9	0.308 9(+)	0.308 5(+)	0.479 2(+)	0.632 5(=)	0.481 5(+)	0.619 2(+)	0.646 4
acrene	0.559 0(+)	0.658 6(+)	0.868 7(+)	0.937 1(=)	0.849 0(+)	0.936 8(=)	0.948 8
Brain2	0.460 9(+)	0.535 4(+)	0.721 2(+)	0.770 6(+)	0.731 9(+)	0.773 0(+)	0.794 0
prostate	0.565 8(+)	0.562 0(+)	0.881 7(+)	0.933 0(=)	0.914 3(+)	0.897 1(+)	0.944 2
Leukemia2	0.486 7(+)	0.623 3(+)	0.818 5(+)	0.955 1(=)	0.792 0(+)	0.946 9(=)	0.962 6
11Tumor	0.473 7(+)	0.476 8(+)	0.748 3(+)	0.756 4(+)	0.702 2(+)	0.790 9(=)	0.789 0
+/−/=	14/0/0	14/0/0	13/0/1	5/1/8	14/0/0	6/0/8	—
rank	6.86	6.14	4.43	2.29	4.14	2.93	1.21

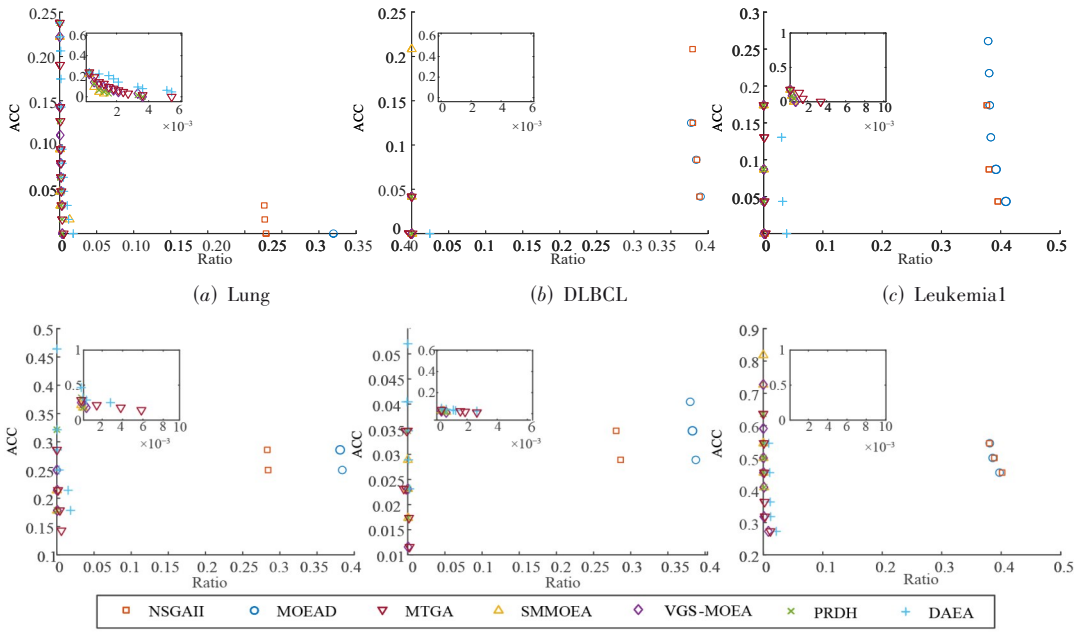
表3 不同算法的IGD值对比

IGD	MOEA/D	NSGA-II	SMMOEA	VGSMOEA	PRDH	DAEA	MTGA
Lung	0.390 3(+)	0.293 5(+)	0.035 8(+)	0.012 3(−)	0.030 1(=)	0.025 3(=)	0.025 1
GLIOMA	0.400 5(+)	0.302 0(+)	0.131 3(+)	0.056 8(+)	0.167 4(+)	0.038 1(=)	0.038 4
DLBCL	0.394 0(+)	0.394 9(+)	0.058 5(+)	0.008 6(−)	0.044 0(+)	0.043 8(+)	0.011 4
Leukemia1	0.397 9(+)	0.395 6(+)	0.061 9(+)	0.028 6(=)	0.047 9(+)	0.077 6(+)	0.027 4
9Tumor	0.394 1(+)	0.297 2(+)	0.120 0(+)	0.057 4(=)	0.128 2(+)	0.069 8(=)	0.062 8
Brain1	0.395 8(+)	0.297 9(+)	0.101 3(+)	0.052 1(+)	0.090 6(+)	0.057 4(+)	0.034 8
drivface	0.386 0(+)	0.286 0(+)	0.013 6(+)	0.004 0(−)	0.012 2(+)	0.016 2(+)	0.009 1
Carcinom	0.419 5(+)	0.330 7(+)	0.117 1(+)	0.049 1(=)	0.147 9(+)	0.057 7(=)	0.059 6
nci9	0.407 5(+)	0.408 2(+)	0.103 2(+)	0.052 8(=)	0.105 1(+)	0.076 1(+)	0.042 4
acrene	0.391 6(+)	0.294 0(+)	0.072 1(+)	0.033 1(=)	0.078 5(+)	0.032 0(=)	0.027 3
Brain2	0.395 8(+)	0.297 9(+)	0.101 3(+)	0.052 1(+)	0.090 6(+)	0.057 4(+)	0.034 8
prostate	0.389 3(+)	0.390 7(+)	0.072 2(+)	0.040 8(=)	0.054 6(+)	0.059 3(+)	0.031 1
Leukemia2	0.418 1(+)	0.310 9(+)	0.136 1(+)	0.034 5(=)	0.133 6(+)	0.037 7(+)	0.028 3
11Tumor	0.403 8(+)	0.404 0(+)	0.066 2(+)	0.043 9(=)	0.068 6(+)	0.088 7(+)	0.039 1
+/−/=	14/0/0	14/0/0	14/0/0	3/3/8	11/0/1	9/0/5	—
rank	6.71	6.29	4.36	1.79	4.14	3.21	1.50

VGSMOEA、SMMOEA与PRDH在特征压缩方面具有一定优势，能够有效筛除冗余特征. 但解集分布呈现出不均匀的态势，集中在帕累托前沿一侧，更倾向于搜索具有较少特征数的子集，而忽略了高精度的子集，在搜索过程中容易陷入局部最优解，无法在整个目标空间内进行全面且有效的搜索. 而DAEA所获得的前沿在形状上分布较为均匀且广泛，覆盖了目标空间的较大范围，但该算法在收敛性方面存在不足，这在一定程度上限制了其在实际场景中的应用效果. 相比之下，MTGA能够找到一组在收敛性与多样性之间更具均衡性的近似帕累托解集，在保证特征压缩能力的同时，更全面地逼近真实的帕累托前沿.

在IGD指标上，MTGA在14个数据集中的8个数据集优于其他算法，平均排名为1.5，优于其他算法. 但在较低维数据集上的IGD表现却略逊于VGSMOEA，如Lung、drivface、DLBCL等，但差距非常有限，大多落在±0.01以内. 这是因为MTGA是通过多任务并行全局搜索生成分布更广的解，但可能造成部分解偏离真实前沿，导致IGD较高. 但在高维空间中，维度灾难影响更大，此时MTGA的优势更明显.

表4给出了不同算法在分类准确率方面的具体表现. 从整体结果来看，MTGA在所有数据集上均表现出



(a) Lung　(b) DLBCL　(c) Leukemia1

(d) Brain1　(e) drivface　(f) nci9

图9　不同算法得到的帕累托前沿分布

表4　不同算法的分类正确率对比

ACC	MOEA/D	NSGA-II	SMMOEA	VGSMOEA	PRDH	DAEA	MTGA
Lung	0.947 7(+)	0.892 7(+)	0.939 7(+)	0.890 1(=)	0.953 2(+)	0.997 6(–)	0.989 7
GLIOMA	0.819 4(+)	0.804 2(+)	0.761 1(+)	0.748 1(+)	0.725 0(+)	0.922 2(+)	0.947 2
DLBCL	0.943 7(+)	0.894 1(+)	0.931 3(+)	0.991 7(=)	0.945 8(+)	0.979 2(+)	0.997 9
Leukemia1	0.891 3(+)	0.826 3(+)	0.873 9(+)	0.954 3(+)	0.893 5(+)	0.910 9(+)	0.980 4
9Tumor	0.572 7(+)	0.481 8(+)	0.452 3(+)	0.408 8(+)	0.422 7(+)	0.525 0(+)	0.579 5
Brain1	0.750 0(+)	0.734 8(+)	0.721 4(+)	0.771 4(+)	0.732 1(+)	0.776 8(+)	0.801 8
drivface	0.971 1(+)	0.965 8(+)	0.969 4(+)	0.981 8(=)	0.970 8(+)	0.975 4(+)	0.982 6
Carcinom	0.810 3(+)	0.784 4(+)	0.826 7(+)	0.875 0(+)	0.760 3(+)	0.806 0(+)	0.889 7
nci9	0.504 5(+)	0.479 5(+)	0.479 5(+)	0.634 1(+)	0.481 8(+)	0.629 5(+)	0.663 6
acrene	0.912 9(+)	0.889 5(+)	0.869 4(+)	0.938 7(+)	0.869 4(+)	0.938 7(+)	0.956 5
Brain2	0.750 0(+)	0.734 8(+)	0.721 4(+)	0.771 4(+)	0.732 1(+)	0.776 8(+)	0.801 8
prostate	0.925 8(+)	0.882 0(+)	0.882 3(+)	0.935 5(+)	0.914 5(+)	0.911 3(+)	0.961 3
Leukemia2	0.795 7(+)	0.785 4(+)	0.819 6(+)	0.956 5(=)	0.813 0(+)	0.952 2(+)	0.980 4
11Tumor	0.775 9(+)	0.742 6(+)	0.751 7(+)	0.759 5(=)	0.702 6(+)	0.806 0(=)	0.783 6
+/–/=	14/0/0	14/0/0	14/0/0	9/0/5	14/0/0	12/1/1	—
rank	3.93	5.86	5.57	3.29	5.43	2.79	1.14

比较优越的分类能力. 在14个数据集的实验中,MTGA仅在Lung与11Tumor两个规模最小与最大的数据集上劣于DAEA,在其他数据集上所获结果均优于其他算法,显示出其在多目标特征选择任务中的稳定性与有效性,特别是在如GLIOMA、DLBCL、Carcinom等中型规模数据集中,MTGA相较于其他算法获得了明显更高的准确率. 进一步从平均排名来看,MTGA以1.14的最低平均秩值排名第一,显著优于其他方法. 其中,DAEA排名第二(2.79),VGSMOEA排名第三(3.29),而NSGA-II等传统算法则排名靠后. 相较于其他算法,MTGA能在分类准确率指标上取得较优结果,这是因为MTGA多任务多种群的结构可以更充分地搜索决策空间,并且自适应扩大搜索空间的知识迁移机制也更侧重于分类性能的提高.

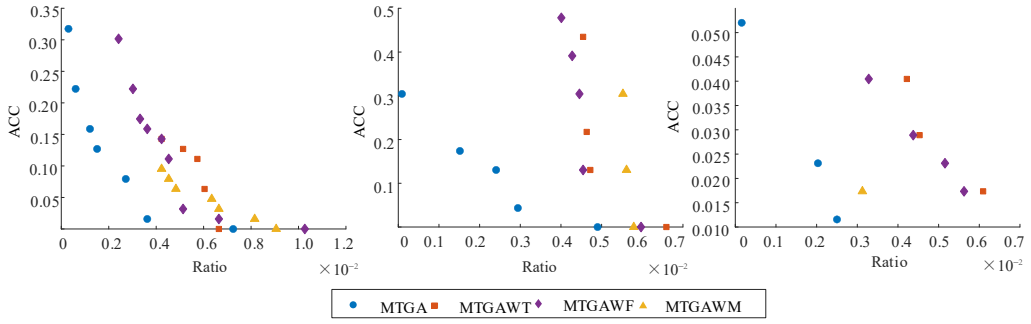
通过在14个高维分类数据集上的对比实验可以看出,MTGA在多数数据集上取得了较优结果,验证了其在解集多样性与收敛性方面的综合优势. 特别是在高维数据场景下,MTGA凭借基于冗余分析的任务生成策

略，有效压缩了搜索空间，提升了搜索效率；同时，通过知识迁移机制实现跨任务的信息共享，使得算法能跳出局部最优，进一步提升了解集质量．综合来看，实验结果充分证明了MTGA在高维多目标特征选择问题中的有效性与创新性．

4.4 算法关键策略有效性分析

为了验证MTGA的3个重要组成部分，即任务生成策略、子任务搜索策略及知识迁移策略在提升算法性能方面的作用，本节进行了消融实验，对比方案包括完整的MTGA算法及其3种变体．各变体算法在除关键策略替换外，其余部分与标准MTGA保持一致，具体如下：(1)采用随机任务生成的MTGAWT；(2)采用均匀交叉和随机变异的MTGAWM；(3)不引入知识迁移机制的MTGAWF.

图10展示了标准MTGA及3种变体在Lung、drivface和Leukemia2这3个数据集上，运行20次中所得最佳HV值对应Pareto前沿分布情况．结果表明，标准MTGA在各数据集上均获得更优质的特征子集，相比之下，去除任一策略后的算法变体均在解集的多样性或整体目标值方面表现不如标准MTGA，表明各组件在整体框架中发挥了相互补充、协同增强的关键作用，充分验证了所提出的任务生成、交叉变异及知识迁移策略在整体框架中的协同有效性．



(a) Lung (b) drivface (c) Leukemia2

图10 MTGA及其变体算法性能对比

此外，为了进一步验证知识迁移机制对算法收敛性能的影响，本文设计了独立对比实验，比较引入知识迁移机制的MTGA和不引入迁移机制的MTGAWF的HV收敛表现．为减少随机因素对评价指标的干扰，采用滑动平均对每代的HV值进行平滑处理，以更直观地展示算法的收敛趋势．图11绘制了在6个不同规模数据集上，MTGA及其变体在20次独立运行下滑动平均后的HV收敛曲线．如图11所示，整体而言MTGA展现出更快的HV增长趋势，显著早于MTGA-WF收敛至较优水平，说明知识迁移机制能够有效引导搜索方向，加快算法趋近最优解．在Leukemia1与Leukemia2的实验中，MTGA在迭代早期的性能略低于MTGAWF，这可能是因为在算法早期，迁移特征的引入导致搜索空间发生扰动，种群内特征子集结构较为不稳定，短期内表现出一定的性能下降，导致其在早期略逊于无迁移机制的MTGAWF. 但随着进化过程的推进，任务特征空间趋于稳定，种群逐渐适应和优化迁移特征，知识迁移的正面作用逐步显现，帮助算法跳出局部最优并取得更优的收敛表现．这一趋势表明，知识迁移在复杂高维问题中虽可能带来初期波动，但从长期来看可以在降低维度的同时保留重要特征，因此仍具备显著的全局优化优势．综上所述，不同任务间知识迁移机制在多目标特征选择中具备有效性，能够显著提升算法的空间探索效率与收敛表现．

4.5 计算复杂度分析

为进一步评估MTGA算法的运行效率，本节对其主要组成模块进行计算复杂度分析，并讨论其在高维乃至超高维特征空间下的可拓展性．

在初始化阶段，MTGA基于对称不确定性(SU)进行特征冗余聚类，用于构建子任务，该过程需要计算所有特征对之间的SU值．设原始特征维度为D，则计算复杂度为$O(D^2)$. 在迭代过程中，假设总种群规模为N，划分为多个子种群后，每个任务在其子空间中独立进化．假设子任务平均维度为$\bar{d}$，则子代生成整体复杂度为$O(N\bar{d})$. 在环境选择阶段，MTGA采用NSGA-II框架的快速非支配排序与拥挤度计算机制，该过程计算复杂度为$O(N^2)$. 对于知识迁移机制，其触发频率低且开销小．具体而言，迁移是否发生与迁移对象的选择所引入的计算开销均为常数级别，在整体复杂度中可以忽略．迁移特征选择基于少量精英个体所选特征频次统计排序，因此复杂度约为$O(\bar{d}\log\bar{d})$，迁移后仅更新非精英个体的少数特征位，复杂度约为$O(N)$.

综上所述，MTGA的初始化阶段计算复杂度为$O(d^2)$，仅在算法开始时执行一次．在迭代阶段，每代计算复杂度为$O(N\bar{d}+N^2+\bar{d}\log\bar{d})$，与现有基于NSGA-II

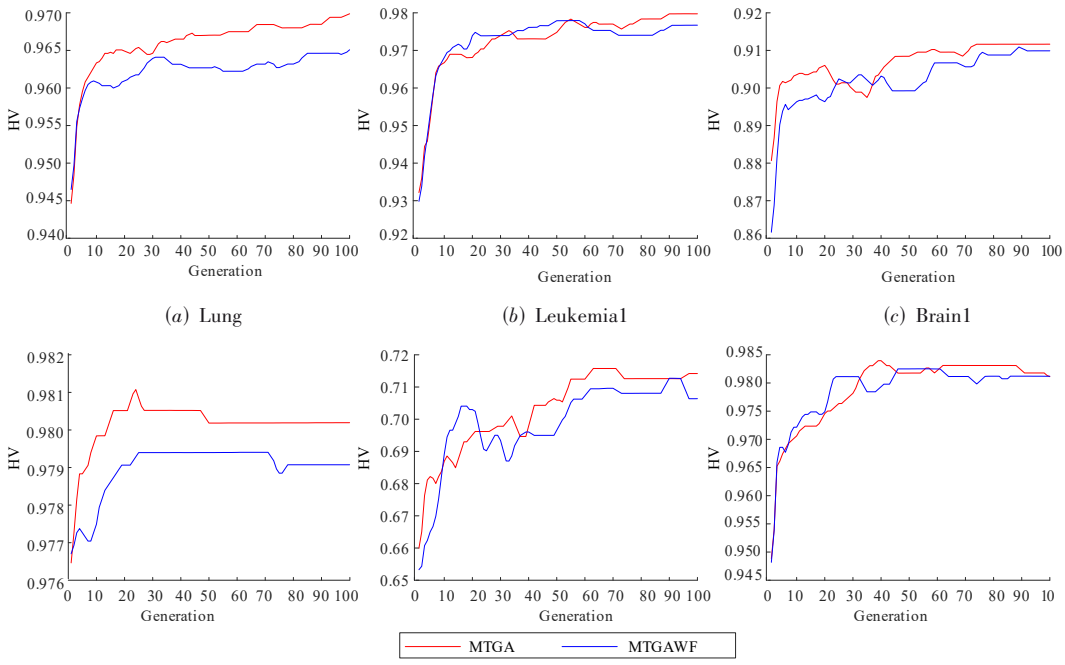


图 11　MTGA 与 MTGA-WF 在 HV 收敛性能上的对比

的多目标算法计算量相当. 但初始阶段的 $O(D^2)$ 级别的 SU 计算,在面对百万级甚至更高维度的特征空间时,仍然可能成为整体效率的瓶颈. 尽管如此,引入 SU 作为任务划分依据可以显著提升子任务的结构合理性与搜索效率,避免关键特征被遗漏或冗余特征干扰,从而增强整体优化性能. 因此,SU 计算虽代价较高,但其对算法收敛速度与解集质量的正向贡献是值得的.

针对超高维数据下的可拓展性问题,未来可进一步探索具有更高计算效率的相关性估计方法(如稀疏矩阵近似、基于采样的 SU 估计、局部特征筛选策略等),或开发更轻量的统计指标替代 SU,用于子任务生成与特征重要性评估,以在保持性能的同时降低预处理计算负担.

5　结束语

本文针对高维特征选择问题的特点,提出了一种多任务多目标特征选择进化算法 MTGA. 针对高维数据集带来的维度灾难问题,采用多任务并行搜索机制,将原高维问题转换为多个低维问题,从而有效降低搜索复杂度. 同时,针对多任务框架的特性,设计了适用于特征选择任务的知识迁移机制,促进不同子任务间的知识共享与传递,加速收敛过程. 此外,本文对传统交叉变异算子进行了针对性改进,使其更契合特征选择问题的优化需求. 在实验验证方面,本文在 14 个高维数据集上系统评估了 MTGA 的性能,并通过消融实验验证了各关键策略的有效性. 实验结果表明,与各变体算法相比,完整的 MTGA 能够获得更优质的特征子集,充分证明了各组成策略在提升搜索效率与分类性能方面的积极贡献,以及其组合带来的整体性能优势.

本文所提出的方法在高维特征选择任务中表现出不错的竞争力,但未来仍有进一步改进的空间. 首先,高维数据中存在复杂的特征交互关系,因此开发更加高效且鲁棒的特征评价指标具有重要意义. 其次,初始化策略在高维特征选择中发挥着关键作用,未来可以探索适用于多任务特征选择的新型初始化机制,以更好地覆盖搜索空间和目标空间. 最后,目前在部分数据集中仍存在特征数量偏多的问题,后续研究可进一步结合特征压缩与选择策略,以在提升分类精度的同时有效降低特征数量.

参考文献

[1] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[2] YOUNG T, HAZARIKA D, PORIA S, et al. Recent trends in deep learning based natural language processing [review article] [J]. IEEE Computational Intelligence Magazine, 2018, 13(3): 55-75.

[3] 黄铉. 特征降维技术的研究与进展[J]. 计算机科学, 2018, 45(B06): 16-21, 53.
HUANG X. Research and development of feature dimensionality reduction[J]. Computer Science, 2018, 45(B06):

16-21, 53. (in Chinese)

[4] LI J D, CHENG K W, WANG S H, et al. Feature selection: A data perspective[J]. ACM Computing Surveys, 2018, 50(6): 1-45.

[5] 刘晓明, 李丞正旭, 吴少聪, 等. 文本分类算法及其应用场景研究综述[J]. 计算机学报, 2024, 47(6): 1244-1287.
LIU X M, LI C Z X, WU S C, et al. A survey of text classification algorithms and application scenarios[J]. Chinese Journal of Computers, 2024, 47(6): 1244-1287. (in Chinese)

[6] GUYON I, ELISSEEFF A. An introduction to variable and feature selection[J]. Journal of Machine Learning Research, 2003(3): 1157-1182.

[7] PENG H C, LONG F H, DING C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238.

[8] HANCER E, XUE B, ZHANG M. Differential evolution for filter feature selection based on information theory and feature ranking[J]. Knowledge-Based Systems, 2018, 140: 103-119.

[9] JIANG L X, KONG G G, LI C Q. Wrapper framework for test-cost-sensitive feature selection[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 51(3): 1747-1756.

[10] GUYON I, WESTON J, BARNHILL S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2002, 46(1): 389-422.

[11] 王艳丽, 梁静, 薛冰, 等. 基于进化计算的特征选择方法研究概述[J]. 郑州大学学报(工学版), 2020, 41(1): 49-57.
WANG Y L, LIANG J, XUE B, et al. Research on evolutionary computation for feature selection[J]. Journal of Zhengzhou University (Engineering Science), 2020, 41(1): 49-57. (in Chinese)

[12] 高慧敏, 王云鹤, 卞闯, 等. 基于混合进化算法的特征选择方法研究[J]. 电子学报, 2023, 51(6): 1619-1636.
GAO H M, WANG Y H, BIAN C, et al. Research on feature selection based on hybrid evolutionary algorithm[J]. Acta Electronica Sinica, 2023, 51(6): 1619-1636. (in Chinese)

[13] TAN F, FU X Z, ZHANG Y Q, et al. A genetic algorithm-based method for feature subset selection[J]. Soft Computing, 2008, 12(2): 111-120.

[14] XUE B, ZHANG M J, BROWNE W N. Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms[J]. Applied Soft Computing, 2014, 18: 261-276.

[15] KASHEF S, NEZAMABADI-POUR H. An advanced ACO algorithm for feature subset selection[J]. Neurocomputing, 2015, 147: 271-279.

[16] JIAO R W, NGUYEN B H, XUE B, et al. A survey on evolutionary multiobjective feature selection in classification: Approaches, applications, and challenges[J]. IEEE Transactions on Evolutionary Computation, 2024, 28(4): 1156-1176.

[17] NGUYEN B H, XUE B, ANDREAE P, et al. Multiple reference points-based decomposition for multiobjective feature selection in classification: Static and dynamic mechanisms[J]. IEEE Transactions on Evolutionary Computation, 2020, 24(1): 170-184.

[18] XU H, XUE B, ZHANG M J. A duplication analysis-based evolutionary algorithm for biobjective feature selection[J]. IEEE Transactions on Evolutionary Computation, 2021, 25(2): 205-218.

[19] HAN F, CHEN W T, LING Q H, et al. Multi-objective particle swarm optimization with adaptive strategies for feature selection[J]. Swarm and Evolutionary Computation, 2021, 62: 100847.

[20] CHEN K, XUE B, ZHANG M, et al. Evolutionary multitasking for feature selection in high-dimensional classification via particle swarm optimization[J]. IEEE Transactions on Evolutionary Computation, 2021, 26(3): 446-460.

[21] 李豪, 汪磊, 张元侨, 等. 演化多任务优化研究综述[J]. 软件学报, 2023, 34(2): 509-538.
LI H, WANG L, ZHANG Y Q, et al. Survey of evolutionary multitasking optimization[J]. Journal of Software, 2023, 34(2): 509-538. (in Chinese)

[22] 林炜星, 王宇嘉, 陈万芬, 等. 基于多因子粒子群的高维数据特征选择算法[J]. 计算机工程与应用, 2021, 57(22): 199-207.
LIN W X, WANG Y J, CHEN W F, et al. High-dimensional data feature selection algorithm based on multifactor particle swarm optimization[J]. Computer Engineering and Applications, 2021, 57(22): 199-207. (in Chinese)

[23] LI L J, XUAN M L, LIN Q Z, et al. An evolutionary multitasking algorithm with multiple filtering for high-dimensional feature selection[J]. IEEE Transactions on Evolutionary Computation, 2023, 27(4): 802-816.

[24] FENG Y, FENG L, LIU S, et al. Towards multi-objective high-dimensional feature selection via evolutionary multitasking[J]. Swarm and Evolutionary Computation, 2024, 89: 101618.

[25] LIN J B, CHEN Q, XUE B, et al. Evolutionary multitasking for multiobjective feature selection in classification[J].

IEEE Transactions on Evolutionary Computation, 2024, 28(6): 1852-1866.
[26] XU H, XUE B, ZHANG M J. An adaptive initialization and multitasking based evolutionary algorithm for bi-objective feature selection in classification[J]. Complex & Intelligent Systems, 2025, 11(7): 310.
[27] 张梦婷, 杜建强, 罗计根, 等. 多目标优化特征选择研究综述[J]. 计算机工程与应用, 2023, 59(3): 23-32.
ZHANG M T, DU J Q, LUO J G, et al. Research on feature selection of multi-objective optimization[J]. Computer Engineering and Applications, 2023, 59(3): 23-32. (in Chinese)
[28] 王朝, 黄慧涛, 张晶, 等. 基于解空间降维的大规模约束多目标进化算法[J]. 电子学报, 2023, 51(11): 3120-3127.
WANG C, HUANG H T, ZHANG J, et al. A large-scale constrained multi-objective optimization algorithm based on solution space reduction[J]. Acta Electronica Sinica, 2023, 51(11): 3120-3127. (in Chinese)
[29] YU L, LIU H. Feature selection for high-dimensional data: A fast correlation-based filter solution[C]//Proceedings of the 20th International Conference on Machine Learning (ICML-03). California: AAAI, 2003: 856-863.
[30] COVER T M. Elements of Information Theory[M]. New Jersey: John Wiley & Sons, 1999.
[31] ARTHUR D, VASSILVITSKII S. K-means++: The Advantages of Careful Seeding[R]. California: Stanford University, 2008.
[32] JAIN A K. Data clustering: 50 years beyond K-means[J]. Pattern Recognition Letters, 2010, 31(8): 651-666.
[33] ZHANG Q F, LI H. MOEA/D: A multiobjective evolutionary algorithm based on decomposition[J]. IEEE Transactions on Evolutionary Computation, 2007, 11(6): 712-731.
[34] DEB K, PRATAP A, AGARWAL S, et al. A fast and elitist multiobjective genetic algorithm: NSGA-II[J]. IEEE Transactions on Evolutionary Computation, 2002, 6(2): 182-197.
[35] CHENG F, CHU F X, XU Y, et al. A steering-matrix-based multiobjective evolutionary algorithm for high-dimensional feature selection[J]. IEEE Transactions on Cybernetics, 2022, 52(9): 9695-9708.
[36] CHENG F, CUI J, WANG Q, et al. A variable granularity search-based multiobjective feature selection algorithm for high-dimensional data classification[J]. IEEE Transactions on Evolutionary Computation, 2022, 27(2): 266-280.
[37] JIAO R W, XUE B, ZHANG M J. Solving multiobjective feature selection problems in classification via problem reformulation and duplication handling[J]. IEEE Transactions on Evolutionary Computation, 2024, 28(4): 846-860.
[38] KOHAVI R, JOHN G H. Wrappers for feature subset selection[J]. Artificial Intelligence, 1997, 97(1/2): 273-324.
[39] ZITZLER E. Evolutionary Algorithms for Multi-Objective Optimization: Methods and Applications[M]. Ithaca: Shaker, 1999.
[40] BOSMAN P A N, THIERENS D. The balance between proximity and diversity in multiobjective evolutionary algorithms[J]. IEEE Transactions on Evolutionary Computation, 2003, 7(2): 174-188.

作者简介

朱苗苗　女,1999年4月出生于江苏省连云港市. 现为中国矿业大学数学学院硕士研究生. 主要研究方向为智能优化算法及其应用.
E-mail:TS23080041A31LD@cumt.edu.cn

巩敦卫　男,1970年出生于江苏省徐州市. 现为青岛科技大学自动化与电子工程学院教授、博士生导师. 主要研究方向为多目标优化、智能软件工程等.
E-mail: dwgong@vip.163.com

姚香娟　女,1975年3月出生于河北省赵县. 现为中国矿业大学数学学院教授、博士生导师. 主要研究方向为进化测试、运筹优化.
Email: yaoxj@cumt.edu.cn

张　岩　女,1972年出生于黑龙江省集贤县. 现为宿迁学院信息工程学院教授、硕士生导师. 主要研究方向为智能软件工程、家禽智慧养殖等.
E-mail: zhangyan@suq.edu.cn