

采用进化计算的高维多模态特征选择

丁转莲¹, 胡曦¹, 曹略¹, 孙登第², 张兴义³, 王晨旭⁴

(1. 安徽大学互联网学院, 230039, 合肥; 2. 安徽大学人工智能学院, 230601, 合肥;

3. 安徽大学计算机科学与技术学院, 230601, 合肥; 4. 西安交通大学软件学院, 710049, 西安)

摘要: 针对特征选择问题中的多模态特性和高维特性, 引入新的种群初始化策略对大规模多模态多目标优化算法进行改进, 提出了一种基于进化多目标优化的高维多模态特征选择算法。对原始的连续优化算法进行离散化处理, 用于评价离散优化问题中的个体, 并在6个高维特征选择数据集上进行验证。结果表明: 所提算法提升了初始种群的质量并加快了算法的收敛; 相比于其他同类算法, 所提算法获得了更优的帕累托前沿, 其超体积指标值整体最优, 并且在不影响分类精度的前提下可获得平均2.53个等效特征子集, 表明所提算法具有最好的分类精度和最多样化的等效特征子集。

关键词: 分类; 特征选择; 进化计算; 多模态多目标优化; 高维多模态特征选择

中图分类号: TP18 **文献标志码:** A

DOI: 10.7652/xjtub202409012 **文章编号:** 0253-987X(2024)09-0117-12

High-Dimensional Multimodal Feature Selection Based on Evolutionary Computation

DING Zhuanlian¹, HU Xi¹, CAO Lue¹, SUN Dengdi², ZHANG Xingyi³, WANG Chenxu⁴

(1. School of Internet, Anhui University, Hefei 230039, China; 2. School of Artificial Intelligence,

Anhui University, Hefei 230601, China; 3. School of Computer Science and Technology,

Anhui University, Hefei 230601, China; 4. School of Software Engineering,

Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Aiming at multimodal and high-dimensional characteristics in feature selection problems, a new population initialization strategy is introduced to improve the large-scale multimodal multiobjective optimization algorithm, and a high-dimensional multimodal feature selection algorithm based on evolutionary multiobjective optimization is proposed. The original continuous optimization algorithm is discretized to evaluate individuals in discrete optimization problems. This approach is validated across six high-dimensional feature selection datasets. Results demonstrate that the proposed algorithm improves the quality of the initial population and accelerates algorithm convergence. Compared to other algorithms, the proposed algorithm yields the superior Pareto front with the best overall hypervolume value. It can obtain an average of 2.53 equivalent feature subsets without compromising classification accuracy. These results show the proposed algorithm's ability to achieve optimal classification accuracy and the most diverse equivalent feature subsets.

收稿日期: 2024-02-25。 作者简介: 丁转莲(1988—), 女, 副教授, 硕士生导师; 孙登第(通信作者), 男, 教授, 博士生导师。

基金项目: 国家自然科学基金区域创新发展联合基金重点资助项目(U21A20512); 国家自然科学基金资助项目(62076005, 62272379)。

网络出版时间: 2024-06-04

网络出版地址: <https://link.cnki.net/urlid/61.1069.T.20240604.1030.002>

Keywords: classification; feature selection; evolutionary computation; multimodal multiobjective optimization; high-dimension multimodal feature selection

特征选择是机器学习、数据挖掘中的一项重要技术,是当前信息领域的研究热点之一^[1]。特征选择旨在从众多的候选特征中选出某些最有效的特征,其目标是得到令人满意的分类精度且使用的特征数量较少^[1]。在真实世界中,很多分类问题的原始特征集中无关或冗余的特征数量较多,这些特征在扰乱学习过程的同时严重影响分类精度,因此特征选择问题得到广泛的关注和研究。

经过多年研究,大量的特征选择方法被提出,包括传统特征选择方法^[2-3]和基于进化计算的特征选择方法^[4]。传统特征选择办法一般存在诸如嵌套效应、难以确定重要参数、容易陷入局部最优等问题。例如,序列前向选择^[2]和序列后向选择^[3]这类常用的封装式特征选择方法会受到嵌套效应的影响。进化计算模拟生物的演化过程,是一种随机全局优化方法。由于进化计算在全局搜索方面发挥出巨大作用,故而在特征选择问题中应用广泛^[5]。进化计算方法的优势主要在于:作为一种基于种群的方法,能够综合探索整个搜索空间;在搜索过程中对特征的选择没有限制,能克服嵌套效应和容易陷入局部最优问题;能同步搜索多个解,具有强大的并行搜索能力。

对于特征选择问题,决策者希望在获取较高分类精度的同时使用较少数量的特征,因此特征选择问题是一个多目标优化问题^[6]。同时,由于具有相同特征数量的不同特征子集往往可以得到相同或相似的分类精度,因此这种特征选择问题是一类多模态多目标优化问题。如果一个多目标优化问题至少有两个等效的全局帕累托最优解,它们对应帕累托前沿上同一点,则称该问题为多模态多目标优化问题,此类问题具有多模态特性^[7]。在一些现实应用中,经常会出现同时需要多个等效的全局或局部最优特征子集的情况,因为某个最优特征子集可能会包含过于昂贵的特征,这种特征子集在现实中不会被选择。为了给决策者提供更多的可供选择的方案,一些学者致力于研究特征选择中的多模态多目标优化问题并取得了较好的结果^[7-8]。然而,真实世界中问题的规模通常是不可控的。面对高维特征,传统的多模态优化方法无法获得令人满意的结果,因为它们不能有效应对高维数据所带来的“维数灾难”问题。近年来,一些求解高维特征选择问题的

规模多目标优化算法被提出,如 SparseEA^[9] 和 SLMEA 算法^[10]。然而遗憾的是,这些算法并不能处理特征选择问题中的多模态特性。由于目前这方面的工作还很少,因此研究高维多模态特征选择问题的求解具有重要意义。

1 相关工作

1.1 多模态特征选择问题

特征选择中有两个主要待优化目标,即分类精度和选择特征的数量。多目标特征选择问题^[6]通常在数学上定义为

$$\begin{cases} \min_x f_1(\mathbf{x}) = \frac{1}{D} \sum_{i=1}^D |x_i| \\ \min_x f_2(\mathbf{x}) = \text{ErrorRate}(\mathbf{x}) \end{cases} \quad (1)$$

式中:决策变量 $\mathbf{x}$ 表示一个特征子集; D 表示初始特征的个数; $f_1(\mathbf{x})$ 表示选择的特征数量占候选特征数的比例; $f_2(\mathbf{x})$ 表示每种特征组合对应的模型分类错误率。

在多目标特征选择中,通常难以通过单个解使其所有目标达到最优,对此可以通过进化算法来找到多个目标之间的最优解集。这些解集中可能包含多个不同的特征子集,但这些特征子集拥有的特征数量和实现的分类精度都是相同的,这便是多模态特征选择问题^[2]。如图 1 所示,两个待优化的目标 f_1 和 f_2 分别代表选择的特征数量和分类错误率,虚线表示的是这个多目标优化问题的帕累托前沿, $F_1 \sim F_5$ 是为决策者提供的 5 个可选特征,其中蓝色意味着该特征被选择,白色则表示该特征未被选择。从图 1 不难发现,在 $f_1 = 1, f_2 = 0.8$ 时存在 $\{F_1\}$ 、 $\{F_4\}$ 这两个等效的特征子集。类似地,在 $f_1 = 3, f_2 = 0.35$ 时有两个等效的特征子集 $\{F_2, F_3, F_4\}$ 、 $\{F_1, F_2, F_3\}$ 。一般的进化计算方法只会给出一个特征子集,而无法给出等效的特征子集。举个简单的例子,假设 $\{F_1, F_2, F_3\}$ 是一个细胞的形态特征(如体积、面积、质量等), $\{F_4\}$ 是这个细胞的光学特征(如吸收光谱的长度等),在现实世界中,由于获取细胞的形态特征会比获取细胞的光学特征更加容易,因此用户可以根据自己的偏好或实际情况进行选择。可见,求解多模态特征选择问题时,为决策者提供更多可选择的解决方案是必要的。

此外,真实世界中的高维度大规模数据集会引

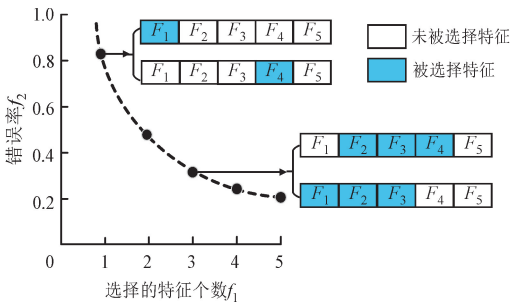


图 1 特征选择中的多模态多目标优化示意图

Fig.1 Illustration of multimodal multiobjective optimization in feature selection

起维度灾难。具有高维度大规模特征的多模态特征选择问题称之为高维多模态特征选择问题。由于要同时考虑高维特性和多模态特性,因此解决此类问题需要特殊的多模态多目标优化算法,这是一项具有挑战性的任务。

1.2 多模态多目标优化

不失一般性,多目标优化问题(MOP)^[7]在数学上可以定义为

$$\begin{cases} \min_x & f(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_m(\mathbf{x})) \\ \text{s. t.} & \mathbf{x} \in \Omega \end{cases} \quad (2)$$

式中: $\mathbf{x} = (x_1, x_2, \dots, x_D)$ 是一个在决策空间 Ω 中的 D 维决策向量; $f(\mathbf{x})$ 是一个包含 m 个冲突目标的目标函数。对于决策空间 Ω 中的两个不同的解 $\mathbf{x}_1$ 和 $\mathbf{x}_2$, 当且仅当满足式(3), 则称 $\mathbf{x}_1$ 支配 $\mathbf{x}_2$, 记为 $\mathbf{x}_1 < \mathbf{x}_2$

$$(\forall i: f_i(\mathbf{x}_1) \leq f_i(\mathbf{x}_2)) \wedge (\exists j: f_j(\mathbf{x}_1) < f_j(\mathbf{x}_2)) \quad (3)$$

式中: $i, j \in \{1, 2, \dots, m\}$ 。此外, 如果一个解没有被任何解所支配, 则称这个解是一个帕累托(Pareto)最优解。

多模态多目标优化问题(MMOP)是一类具有多模态特性的多目标优化问题, 求解 MMOP 的目的是为相同的帕累托前沿(PF)找到所有等效的帕累托解集(PS)。在实际应用中, 往往存在约束条件或环境变化, 而这些因素又难以在问题定义的公式中直接表示出来, 因此一些解在实际中可能是不可行的。可见, 有必要提供多个等效的 PS 作为决策者的替代选择。

近年来, 涌现出大量多模态多目标优化算法有效解决了 MMOP^[7]。其中, 一些算法通过决策空间中的拥挤距离来完成环境选择。例如, Omni-optimizer 算法^[11]同时引入决策空间和目标空间中每个解的拥挤距离, 以平衡两个空间中 Pareto 最优

解的多样性算法。除此之外, 小生境策略在解决 MMOP 问题时候也被广泛使用。例如, DN-NSGA-II 算法在交配池选择中采用了基于决策空间的小生境方法, 有助于为相同的 PF 找到多个等效的 PS^[12]; MO_Ring_PSO_SCD 算法采用了环形拓扑结构用于形成稳定的小生境并在进化过程中使用特殊拥挤距离^[13]; 类似地, RHEA 算法也采用了基于环形拓扑结构的方法, 在解决寻找等效 PS 时收敛性和多样性不平衡问题的同时能够保留优秀的局部 PS^[14]。另外, 基于分解的策略也常受到研究者的关注。例如, MOEA/D-AD 与 MMEA-DES 算法采用将一个或多个个体分配给同一个子问题的方法解决 MMOP 问题^[15-16]。DESS 算法是一种有效的子集选择框架^[17], 通过同时考虑决策空间和目标空间的多样性来选择解集, 用以提高各种 MMEAs 算法的性能。此外, 还有很多其他方法被用来解决 MMOP 问题。例如, CoMMEA 算法通过引入多样性存档和收敛性存档协同进化获得全局和局部 PS^[18]; CMMO 算法通过两个种群协同进化平衡目标空间和决策空间的多样性和收敛性^[19]; HREA 算法采用了层次排名的方法求解多模态问题。聚类方法也常被用于求解多模态问题^[20], 如 GSMPSO-MM 算法采用 K 均值聚类^[21], MMODE_CSCD 算法使用基于聚类的特殊拥挤距离^[22]。此外, 许多优秀的多模态多目标优化方法也相继被提出以提高算法性能。

实际应用中的一些多模态多目标优化问题含有非常多的决策变量, 这类问题称之为大规模多模态多目标优化问题。虽然采用上述多模态多目标优化方法求解 MMOP 均取得了较好的效果, 但是这些方法在应用于高维问题时还面临一些挑战。现有的多模态多目标优化方法由于难以在高维的搜索空间中度量种群多样性, 同时搜索能力变弱, 很难找到所有等效的 PS, 因此无法满足求解大规模多模态多目标优化问题的要求。此外, 现实世界中很多问题的 Pareto 最优解具有稀疏特性, 即大多数决策变量为 0。高维搜索空间和问题未知的稀疏特性给求解此类多目标求解问题带来了很大的挑战。总体而言, 这类问题的研究仍处在初级阶段。到目前为止, 仅 MP-MMEA^[23] 和 HHC-MMEA^[24] 算法专注于解决这类问题。

现实世界中存在很多大规模多模态多目标优化问题, 如神经网络权值训练、神经网络架构搜索及上述特征选择等问题。然而, 到目前为止, 针对高维多模态特征选择问题的研究非常少, 该问题的求解还

需要深入研究。高维多模态特征选择的难点在于如何在高维的搜索空间中找到尽可能多的等效特征子集。传统的多目标优化方法只能找到一个特征子集,且在优化过程中找到等效的特征子集也会被丢弃,因此传统方法难以实现多模态特征选择在决策空间中解集分布具有多样性的需求。由于大规模多模态多目标优化方法具有较强的探索能力及能够在一次运行中找到多个合适的解,因此该类方法可以被认为是求解高维多模态特征选择问题的有力工具。然而,由于该类问题的困难与复杂,目前多模态多目标优化算法在该问题上应用的效果依旧有着巨大的提升空间。HHC-MMEA 算法中的多种群协同进化机制以及引入的混合层次聚类技术能够有效区分进化中的不同等效特征子集,提高算法的搜索能力,促使算法找到更多的等效特征子集。因此,本文对大规模多模态多目标优化算法 HHC-MMEA 进行改进,提出了一种基于进化多目标优化的高维多模态特征选择算法(FS-HHC-MMEA)。该算法采用 HHC-MMEA 算法框架,引入新的种群初始化策略来加快算法的收敛。另外,本文对连续优化算法 HHC-MMEA 进行了离散化处理用于求解特征选择问题。

2 基于进化多目标优化的高维多模态特征选择

2.1 算法描述

本文所提 FS-HHC-MMEA 算法使用改进的 HHC-MMEA 算法来搜索具有多模态特性的等效特征子集,算法总体流程如算法 1 所示。

算法 1 FS-HHC-MMEA(N)

输入:种群大小 N

输出:等效特征子集 P

```

1   $\mathbf{G} \leftarrow$  获取全局指导向量;
2   $P \leftarrow$  Initialization( $N, \mathbf{G}$ ); //算法 2
3   $K=1$ ; //初始子种群数量
4  while{未达到最大迭代次数} do
5      for  $i=1$  to  $K$  do
6           $\mathbf{I} \leftarrow$  更新第  $i$  个子种群的局部指导向量;
7          计算每个个体的特征个数和分类错误率;
8           $Q \leftarrow$  根据二元锦标赛选择  $2 \lfloor N/K \rfloor$  个父代;
9          offspring  $\leftarrow$  自适应交叉变异产生子代;
10          $\text{sub}P_i \leftarrow$  环境选择( $\text{sub}P_i \cup \text{offspring}$ );
11     end for
12     [ $K, \text{sub}P_1, \dots, \text{sub}P_k$ ]  $\leftarrow$  每隔一定代数进行

```

混合层次聚类;

```

13 end while

```

```

14  $P \leftarrow \text{sub}P_1 \cup \dots \cup \text{sub}P_k$ ;

```

```

15  $P \leftarrow$  根据  $P$  中非支配解集获得等效特征子集;

```

```

16 return  $P$ 

```

首先,将数据集划分成 80% 训练数据和 20% 测试数据^[9]。其次,采用决策变量分析方法^[9]获得全局指导向量(每个决策变量的得分),并利用新的种群初始化策略初始化种群个体。然后,计算每个个体的两个目标值,即选择的特征个数和利用 KNN 分类器得到的预测分类错误率。最后,采用 HHC-MMEA 算法中的子种群协同进化搜索多个等效特征子集。子种群的协同进化主要由以下几部分组成。首先,利用文献[23]提出的方法获得局部指导向量来确定每个子种群的搜索方向。然后,根据非支配前沿等级和解的拥挤度距离采取二元锦标赛选择策略选择交配父本。接着,使用基于全局指导向量和局部指导向量的自适应交叉及变异策略生成子代解,并利用改进的环境选择策略保留质量好的解。最后,使用混合层次聚类技术有效区分不同等效特征子集的分支从而将其尽可能地划分到不同子种群中,实现多个子种群间的高效合作以及协同进化。

2.2 新的种群初始化策略

相关研究表明,好的初始化策略对于提高算法的性能至关重要^[25]。对于高维多模态特征选择问题,多模态最优特征子集在整个决策空间中分布非常稀疏,即大部分决策变量位的值都是 0。在高维多模态特征选择中,随机初始化可能无法产生良好的效果,需要考虑初始解在多目标空间中的分布。原始 HHC-MMEA 算法中的拉丁超立方体采样种群初始化仅考虑初始解在特征数量上分布的均匀性,却忽略了特征的质量。新的种群初始化策略在原有的均匀采样基础上,基于全局指导向量,根据每个决策变量的得分采用类似二元锦标赛的方法来初始化种群,同时考虑了特征数量和特征质量,进而从解的多样性和解的质量方面提供更好的初始解。新的种群初始化策略的详细流程如算法 2 所示。

算法 2 Initialization($N, \mathbf{G}$)

输入:种群大小 N , 全局指导向量 $\mathbf{G}$

输出:初始种群 P

```

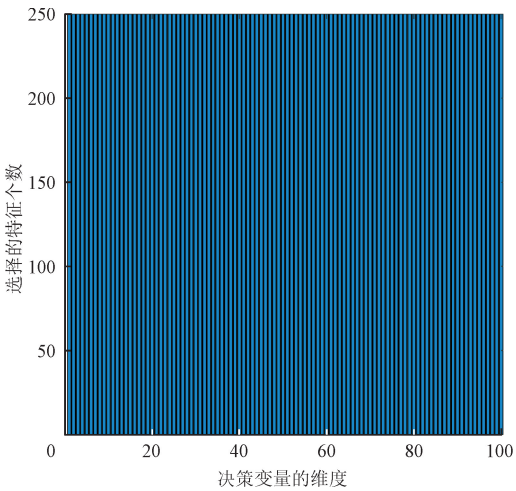
1   $D \leftarrow$  决策变量的数量;
2  [ $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$ ]  $\leftarrow$  随机生成  $N$  个实值向量;
3  [ $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_N$ ]  $\leftarrow$  使用拉丁超立方均匀采样生成  $N$  个二进制向量;

```

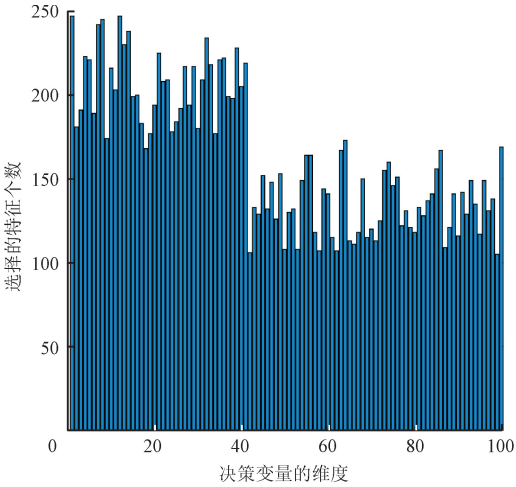
```
4   $r \leftarrow G$ ;  
5  for  $i=1$  to  $N$   
6    for  $j=1$  to  $R \times D$  //  $R$  表示介于 0 和 1 之间均匀分布的随机数  
7       $[m, n] \leftarrow$  随机挑选两个决策变量位;  
8      if  $r^m > r^n$  then  
9         $b_i^m \leftarrow 0$ ;  
10        $r^m \leftarrow r^m - 1$ ;  
11      else  
12         $b_i^n \leftarrow 0$ ;  
13        $r^n \leftarrow r^n - 1$ ;  
14      end if  
15    end for  
16  end for  
17   $P \leftarrow [r_1, r_2, \dots, r_N]$  和  $[b_1, b_2, \dots, b_N]$  对应位置的元素逐个相乘;  
18  return  $P$ 
```

该算法采用双层编码的方式进行优化。首先,随机生成决策空间中的实值向量,利用拉丁超立方体均匀采样生成二进制向量。然后,在拉丁超立方体均匀采样的基础上根据决策变量的重要性程度对决策位进行翻转。辅助向量 r 首先初始化为全局指导向量,然后每次随机挑选两个决策变量位,比较其对应的辅助向量 r 上的得分值(决策变量得分值越高说明该特征的重要性越低),将重要性较低的决策变量位置为 0,同时将其对应的得分减 1。该策略在初始化阶段可以筛选掉一些不重要的特征,从而可以生成具有较高质量的初始解,加快算法的收敛。

为了验证新的种群初始化策略的有效性,图 2 给出了 HHC-MMEA 算法所采用的原始拉丁超立方体采样初始化策略与本文提出的 FS-HHC-MMEA 算法采用的新的种群初始化策略,在标准大规模多模态测试问题 SMMOP1^[23] 上产生的初始解的二进制变量的统计结果。可以看出,拉丁超立方体采样初始化策略产生的初始解在特征数量目标上分布非常均匀,但未考虑到特征的重要性差异。图中 SMMOP1 问题对应等效 PS 的个数为 4,决策变量数为 100,其中第 1 维到 40 维对应重要决策变量位。新的种群初始化策略不仅考虑了种群分布的多样性,也考虑了特征的重要性差异,筛选掉了一些不重要的特征,因此新的种群策略所包含的特征数量相较于原始策略会显著变少,加快了算法的收敛。



(a)HHC-MMEA 算法采用原始初始化策略的统计结果



(b)FS-HHC-MMEA 算法采用初始化策略的统计结果

图 2 两种初始化策略对应的算法在 SMMOP1 测试问题上初始化种群的二进制变量的统计结果

Fig. 2 Statistical results of the binary parts of the initial population obtained by two initialization strategies on SMMOP1 test problem

图 3 展示了两种初始化策略对应的算法在标准大规模多模态测试问题 SMMOP1、SMMOP6、SMMOP7 和 SMMOP8^[23] 上的决策空间反世代距离(IGDX)^[8] 收敛曲线。该问题对应等效 PS 的个数为 4,决策变量数为 100。从图 3 不难发现,与原始拉丁超立方体采样初始化策略相比,新的种群初始化策略在初始时拥有更低的反世代距离(值越小越好),表明其在进化初始阶段种群拥有更好的起点。新的种群初始化策略既考虑了初始解在特征数上分布的均匀性,又考虑了特征对目标函数值的贡献度,提高了所产生的初始解的质量,从而达到加快算法收敛的目的。

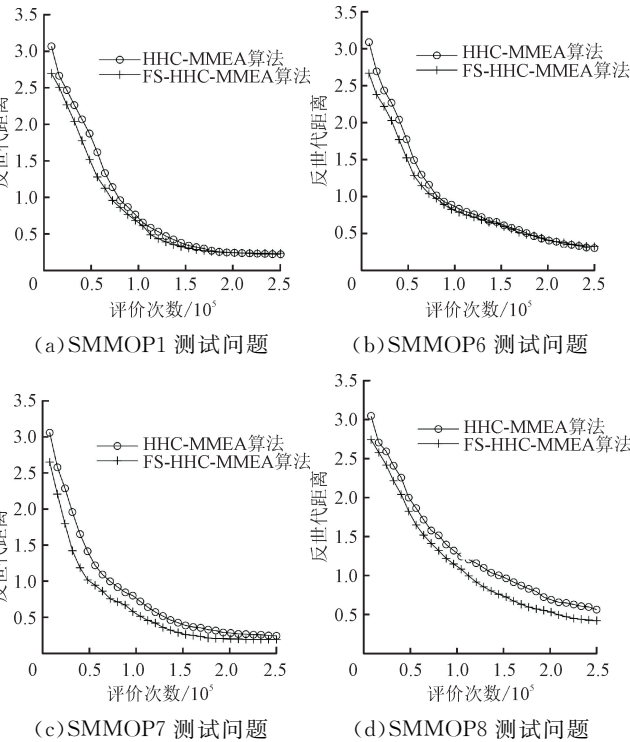


图 3 两种初始化策略对应的算法在 4 种测试问题上的反世代距离收敛曲线

Fig. 3 Convergence profiles of the IGDX values obtained by the algorithms with two initialization strategies on 4 test problems

2.3 个体评价

因为原始的 HHC-MMEA 算法是连续优化算法,不能直接用来求解特征选择这类离散组合优化问题,所以要对其进行特殊处理。图 4 展示了经处理的算法对每个个体目标值的计算方式。通过选定阈值 0.5,将实数映射为 0 或 1 二进制位,其中 0 表示该特征不被选择,1 表示该特征被选择。当某一维度的特征值大于 0.5 时,表示该维特征被选择,反之,表示该维特征被舍弃。例如,图 4 中选择的特征子集为 $\{F_1, F_4\}$,将选择的特征子集送入到分类器中得到分类精度,这里采用广泛使用的 KNN 分类器,其中 K 值被设置为 $3^{[9]}$ 。这样就得到了选择的

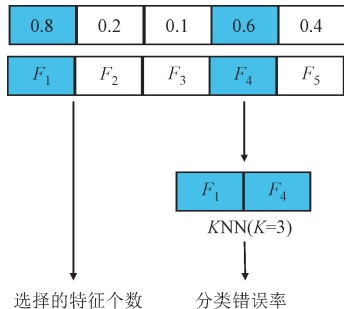


图 4 个体目标值计算示意图

Fig. 4 Illustration of fitness evaluation

特征数量以及分类错误率两个目标值,完成了一个个体的评价。

2.4 复杂度分析

对于 FS-HHC-MMEA 算法,时间复杂度主要由以下几个部分组成。假设种群的大小为 N ,目标空间和决策空间的维度分别为 M 和 D ,子种群的数量为 K ,种群初始化的时间复杂度为 $O(ND)$,所有个体非支配排序的时间复杂度为 $O(MN^2)$,子种群引导向量的更新和子代解的生成具有 $O(ND)$ 的时间复杂度,环境选择时间复杂度为 $O(ND)$,混合层次聚类的时间复杂度为 $O(ND)$,最终得到 FS-HHC-MMEA 算法的时间复杂度为 $O(MN^2 + ND)$ 。

3 实验结果与分析

为了验证新的种群初始化策略解决高维多模态特征选择问题的有效性,图 5 给出了新的种群初始化策略与拉丁超立方体采样初始化策略在 lung_discrete、Colon 数据集特征选择问题上的初始解。由图 5 可见,新的种群初始化策略得到的解支配了拉丁超立方体采样初始化策略所得到的多数解,且后者明显含有更少的特征数量。因此,新的种群初始化策略可以产生特征数量更少、质量更好的初始解。

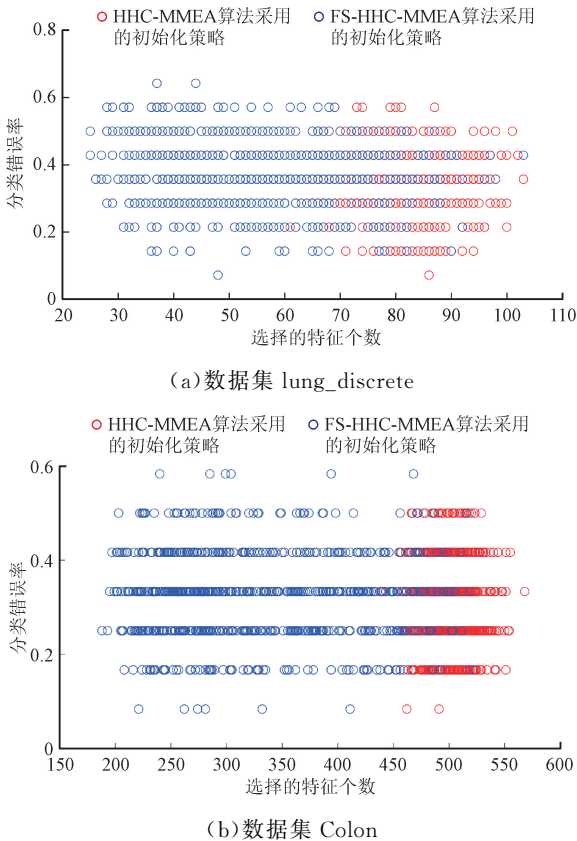


图 5 两种种群初始化策略所产生的解

Fig. 5 Solutions generated by two initialization strategies

为了验证算法解决高维多模态特征选择的有效性,本文在来自亚利桑那州立大学(ASU)的 6 个基准数据集上进行了实验,数据集的信息如表 1 所示。每个数据集被划分成训练集(80%)和测试集(20%),实验采用常用的 KNN 分类器预测分类精度,其中 K 设置为 3。由于实际问题的真实帕累托前沿未知,因此本文采用广泛使用的超体积指标^[26]作为算法的性能指标。超体积指标估计了解集中的点与参考点所围成区域的超体积,值越大表示解集的多样性和收敛性越好。关于超体积指标中参考点的选取可参考文献[27]。

表 1 数据集信息
Table 1 Information of datasets

数据集	样本数	特征数	类别数
lung_discrete	73	325	7
Colon	62	2 000	2
TOX_171	171	5 748	4
Prostate_GE	102	5 966	2
leukemia	72	7 070	2
CLL_SUB_111	111	11 340	3

本文对比了 MO_Ring_PSO_SCD^[13]、MP-MMEA^[23]、HREA^[20]和 SLMEA^[10]4 种算法以证明 FS-HHC-MMEA 算法在求解高维多模态特征选择问题上的有效性,其中 MO_Ring_PSO_SCD 和 HREA

算法是具有代表性的多模态多目标优化算法,MP-MMEA算法是一种先进的大规模多模态多目标进化算法,SLMEA 算法代表稀疏大规模多目标优化算法。为了保证公平,对于所有算法,种群大小设置为 800,评价次数设置为 500 000。除特别说明外,每个算法的参数采用其在相关文献中的推荐设置。MO_Ring_PSO_SCD 算法基于粒子群优化生成子代^[13],其中参数 C_1 、 C_2 、 W 分别设置为 0.5、0.5 和 0.1。对于其他算法,采用单点交叉和位变异产生二进制变量的子代解,采用模拟二进制交叉和多项式变异生成实数变量的子代解,其中交叉概率设置为 1,变异概率设置为 $1/D$ (D 为决策变量维数),交叉和变异的分布指数均设置为 20。所有算法的最终结果只输出非支配解。

表 2 列出了所有算法在 6 个数据集上分别进行 30 次独立运行得到的平均超体积指标。图 6、图 7 和图 8 展示了所有算法在不同数据集上得到的帕累托前沿。从表 2 可以看出,除了 MO_Ring_PSO_SCD 与 HREA 算法外,其他 3 种算法在 6 个数据集上都取得了较好的超体积指标。从图 6、图 7 和图 8 可以看出,除了 MO_Ring_PSO_SCD 与 HREA 算法外,其他 3 种算法在目标空间都能够获得较优的帕累托前沿。

表 2 5 种算法在 6 个数据集上的超体积指标值
Table 2 Hypervolume values obtained by five algorithms on six datasets

数据集	特征数	超体积指标				
		MO_Ring_PSO_SCD	MP-MMEA	SLMEA	HREA	FS-HHC-MMEA
		算法	算法	算法	算法	算法
lung_discrete	325	0.000 0±0.000 0	0.321 1±0.000 0	0.315 2±0.008 4	0.321 1±0.000 0	0.321 1±0.000 0
Colon	2 000	0.000 0±0.000 0	0.648 8±0.000 0	0.641 9±0.000 0	0.000 0±0.000 0	0.648 8±0.000 0
TOX_171	5 748	0.000 0±0.000 0	0.129 3±0.004 4	0.116 4±0.012 0		0.128 0±0.013 1
Prostate_GE	5 966	0.000 0±0.000 0	0.168 1±0.002 8	0.136 4±0.017 5		0.169 0±0.013 1
leukemia	7 070	0.000 0±0.000 0	0.157 0±0.004 7	0.138 1±0.016 7		0.161 8±0.000 0
CLL_SUB_111	11 340	0.000 0±0.000 0	0.151 0±0.005 0	0.137 9±0.008 0		0.154 8±0.000 0

注:表中超体积指标使用平均值±标准差表示;每个数据集上的最优解用加粗字体标出。

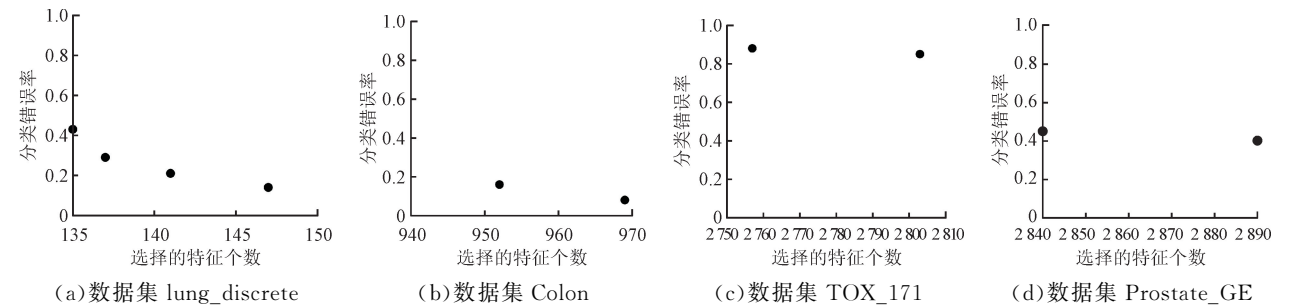
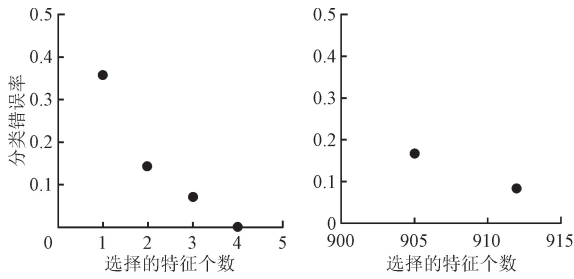


图 6 MO_Ring_PSO_SCD 算法在部分数据集上得到的帕累托前沿
Fig. 6 PF obtained by MO_Ring_PSO_SCD algorithm on partial datasets



(a) 数据集 lung_discrete

(b) 数据集 Colon

图 7 HREA 算法在 2 个数据集上得到的帕累托前沿

Fig. 7 PF obtained by HREA algorithm on two datasets

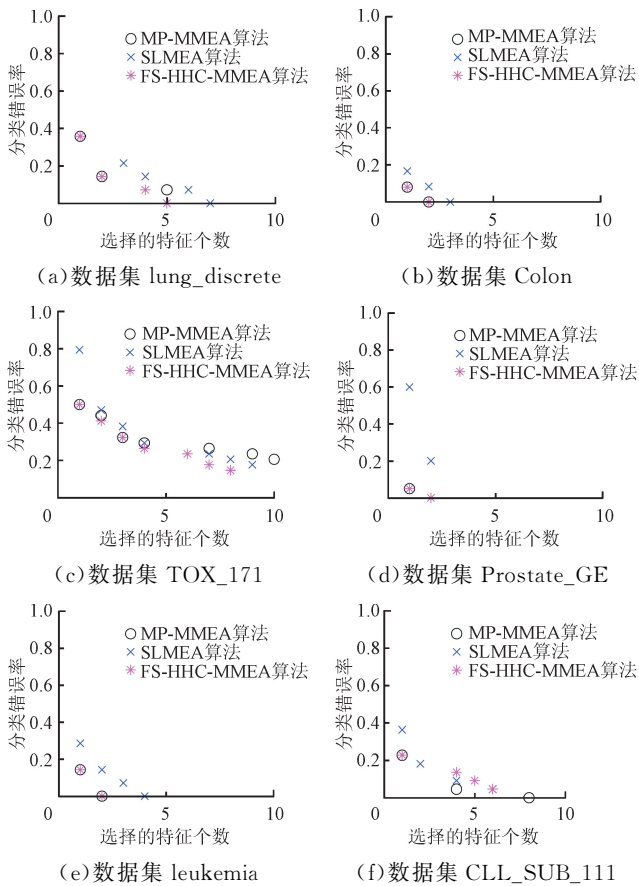


图 8 MP-MMEA、SLMEA 和 FS-HHC-MMEA 算法在 6 个数据集上获得的帕累托前沿比较

Fig. 8 Comparisons of PFs obtained by MP-MMEA, SLMEA and FS-HHC-MMEA algorithms on six datasets

由于 MP-MMEA、SLMEA 和 FS-HHC-MMEA 这 3 种算法都是大规模多目标优化算法,因此在解决高维问题时都获得了较高的超体积指标值与较优的帕累托前沿。FS-HHC-MMEA 算法的超体积指标值在 6 个数据集上获得了 5 个具有竞争力的结果,整体上取得了最好的结果,说明目标空间得到多样性和收敛性更好的解集。从图 8 的帕累托前

沿可看出,FS-HHC-MMEA 算法得到的特征子集具有最好的分类精度和最多多样化的特征数量。

MO_Ring_PSO_SCD 多模态优化算法在面对高维问题时很难收敛到最优的帕累托前沿,从而可看出 MO_Ring_PSO_SCD 算法不适用于高维问题,因此接下来的实验部分将不再展示 MO_Ring_PSO_SCD 算法的性能。HREA 多模态优化算法在 325 维的 lung_discrete 数据集上能够收敛到帕累托前沿,而在 2 000 维的 Colon 数据集上无法收敛到帕累托前沿,且在维度数量更多的其他数据集上, HREA 算法运行都超时(超过 24 h),无法运行出实验结果,表 2 中用空白表示这种情形。HREA 同样难以面对高维问题所带来的挑战,因此下面将不再展示 HREA 算法性能。

图 9 展示了 MP-MMEA、SLMEA 和 FS-HHC-MMEA 这 3 种算法在 6 个数据集上获得的等效特征子集个数。图 10 展示了各算法在每个数据集上获得的平均等效特征子集个数。从图 9 和图 10 可以看出,SLMEA 算法并不存在等效特征子集,因为 SLMEA 算法没有能力保存等效特征解,无法求解多模态多目标优化问题。在不影响分类精度的前提下,多模态多目标优化算法 MP-MMEA 能给出一些等效特征子集(所有数据集上的等效特征子集平均个数是 1.50),而 FS-HHC-MMEA 算法能找到更多的等效特征子集(所有数据集上的等效特征子集平均个数是 2.53)。例如,在 CLL_SUB_111 数据集上,当特征个数为 1 和 6 时,FS-HHC-MMEA 算法分别找到 2 个等效特征子集,而 MP-MMEA 算法只能找到 1 个特征子集。

表 3 展示了 FS-HHC-MMEA 算法在 6 个数据集上寻找到的帕累托最优解集。由表 3 可以看出,算法找到的帕累托最优解集不仅具有多模态特性,而且具有稀疏特性。具体来说,在 Colon 数据集上, $\{F_{317}\}$ 、 $\{F_{302}\}$ 、 $\{F_{842}\}$ 和 $\{F_{1623}\}$ 4 个特征子集所包含的特征个数相同,而且分类错误率也相等,都为 0.083 3。因此,该结果为特征选择问题提供了多个可选的最优特征子集,决策者可根据问题领域的成本和限制来选择合适的特征子集。FS-HHC-MMEA 算法在其他数据集上也找到了多个等效的特征子集。结合图 8 和图 9 及表 3 来看,在保证较理想的稀疏多模态解搜索能力的先决条件下,FS-HHC-MMEA 算法能同时得到较优的帕累托前沿。综合来看,FS-HHC-MMEA 算法是一种有效求解高维多模态特征选择问题的算法。

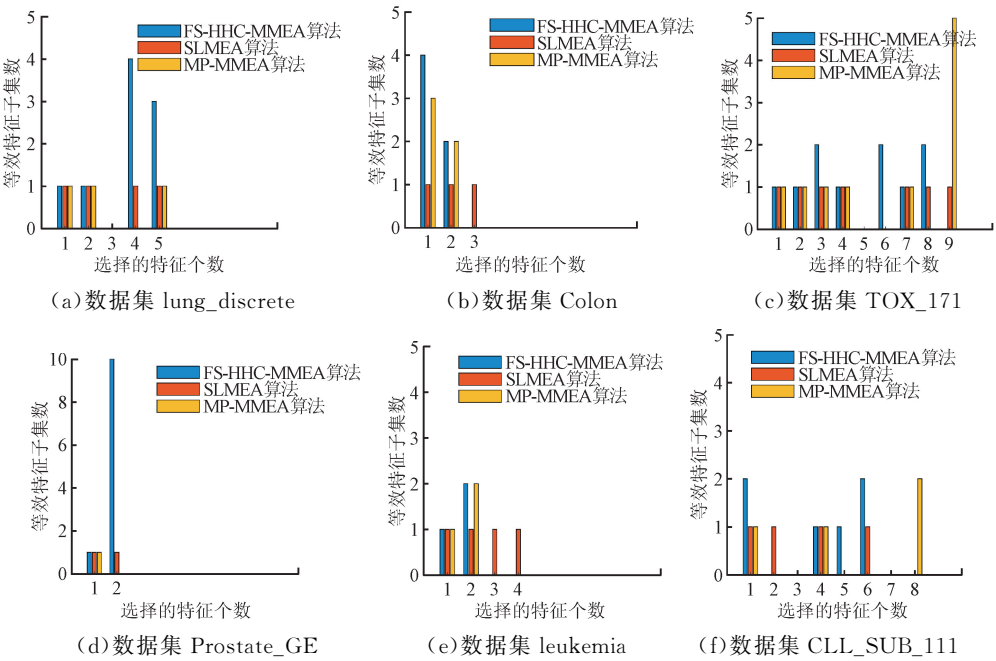


图 9 不同算法得到的等效特征子集数量对比

Fig. 9 Comparisons of the number of equivalent feature subsets obtained by different algorithms

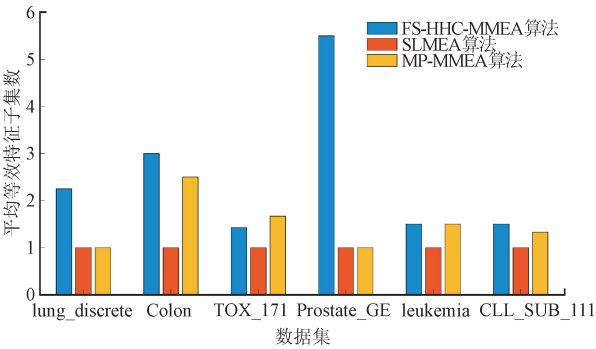


图 10 不同算法得到的平均等效特征子集数量对比

Fig. 10 Comparisons of the number of average equivalent feature subsets obtained by different algorithms

表 3 FS-HHC-MMEA 算法选出的特征子集

Table 3 Feature subsets obtained by FS-HHC-MMEA algorithm

数据集	每个特征子集包含的特征数	分类错误率	选择的特征	等效特征子集数
lung_discrete	1	0.357 1	$\{F_{68}\}$	1
	2	0.142 9	$\{F_{68}, F_{99}\}$	1
	4	0.071 4	$\{F_{53}, F_{68}, F_{113}, F_{211}\}$	4
			$\{F_{42}, F_{45}, F_{96}, F_{205}\}$	
			$\{F_{53}, F_{81}, F_{219}, F_{261}\}$	
			$\{F_{68}, F_{156}, F_{160}, F_{211}\}$	
	5	0	$\{F_{23}, F_{76}, F_{111}, F_{205}, F_{261}\}$	3
			$\{F_{55}, F_{99}, F_{105}, F_{128}, F_{211}\}$	
			$\{F_{53}, F_{68}, F_{113}, F_{211}, F_{286}\}$	

续表

数据集	每个特征子集包含的特征数	分类错误率	选择的特征	等效特征子集数
Colon	1	0.083 3	$\{F_{317}\}$	4
			$\{F_{302}\}$	
			$\{F_{842}\}$	
			$\{F_{1623}\}$	
TOX_171	2	0	$\{F_{302}, F_{1872}\}$ $\{F_{302}, F_{556}\}$	2
	1	0.500 0	$\{F_{1496}\}$	1
	2	0.411 8	$\{F_{2039}, F_{4430}\}$	1
	3	0.323 5	$\{F_{1084}, F_{1121}, F_{4430}\}$ $\{F_{2039}, F_{3277}, F_{4430}\}$	2
	4	0.264 7	$\{F_{1084}, F_{1121}, F_{1814}, F_{4430}\}$	1
	6	0.235 3	$\{F_{141}, F_{1121}, F_{1766}, F_{3318}, F_{4570}, F_{4691}\}$ $\{F_{141}, F_{1121}, F_{1766}, F_{4257}, F_{4570}, F_{4691}\}$	2
	7	0.176 5	$\{F_{141}, F_{1121}, F_{1766}, F_{2242}, F_{4039}, F_{4691}, F_{5655}\}$	1
	8	0.147 1	$F_{141}, F_{1121}, F_{1766}, F_{2242}, F_{3323}, F_{4039}, F_{4691}, F_{5655}$ $F_{141}, F_{1121}, F_{1766}, F_{2242}, F_{2298}, F_{4039}, F_{4691}, F_{5655}$	2
Prostate_GE	1	0.050 0	$\{F_{117}\}$	1
	2	0	$\{F_{4284}, F_{5939}\}$	10
			$\{F_{3845}, F_{4284}\}$	
			$\{F_{117}, F_{3960}\}$	
			$\{F_{117}, F_{1590}\}$	
			$\{F_{117}, F_{3543}\}$	
			$\{F_{3039}, F_{4284}\}$	
			$\{F_{1522}, F_{4284}\}$	
			$\{F_{2240}, F_{4284}\}$	
			$\{F_{3975}, F_{4284}\}$	
leukemia	1	0.142 9	$\{F_{1869}\}$	1
	2	0	$\{F_{1720}, F_{3197}\}$	2
			$\{F_{2382}, F_{4307}\}$	
	4	0.136 4	$\{F_{300}, F_{2799}, F_{7495}, F_{8963}\}$	1
CLL_SUB_111	1	0.227 3	$\{F_{5763}\}$ $\{F_{7893}\}$	2
	5	0.090 9	$\{F_{300}, F_{1577}, F_{7495}, F_{8963}, F_{9972}\}$	1
	6	0.045 5	$\{F_{300}, F_{1577}, F_{7227}, F_{7495}, F_{8963}, F_{9972}\}$	2
			$\{F_{300}, F_{1577}, F_{7495}, F_{8963}, F_{9972}, F_{10637}\}$	

4 结 论

针对特征选择问题中的多模态特性和高维特性带来的挑战,本文基于改进的大规模多模态进化算法对高维多模态特征选择问题进行有效求解,结合实验验证,得出如下结论。

(1)引入新的种群初始化策略对大规模多模态多目标优化算法进行改进,实验结果表明新的种群初始化策略可以产生质量更好的初始解,加速了算法的收敛。

(2)对原始的多模态多目标优化算法进行离散化处理,用于评价特征选择问题中的个体,并使用离散化的多模态多目标优化算法搜索具有多模态特性的特征子集。6 个高维特征选择数据集上的实验结果表明,本文算法能够找到更多的多模态特征子集并且获得更优的帕累托前沿。

值得注意的是,本文算法中的编码并不能反映特征之间的相关性和相似性,因此一种更加有效的编码方法有待进一步研究。另外,本文仅考虑搜索高维多模态特征选择问题中的多个最优特征子集,而忽略了次优特征子集的优化,最优和次优最优特征子集共存的高维多模态特征选择问题的研究也是将来的努力方向之一。

参考文献:

[1] 商立群,李洪波,侯亚东,等. 基于特征选择和优化极限学习机的短期电力负荷预测 [J]. 西安交通大学学报, 2022, 56(4): 165-175.
SHANG Liqun, LI Hongbo, HOU Yadong, et al. Short-term power load forecasting based on feature selection and optimized extreme learning machine [J]. Journal of Xi'an Jiaotong University, 2022, 56(4): 165-175.

[2] 刘彪,刘金长. 基于用户画像分析预测电费敏感型客户的建模实践 [J]. 电力大数据, 2017, 20(8): 20-24.
LIU Biao, LIU Jinchang. Modeling practice of forecasting electricity fees sensitive customers based on user profile analysis [J]. Power Systems and Big Data, 2017, 20(8): 20-24.

[3] PUDIL P, NOVOTICOVÁ J, KITTLER J. Floating search methods in feature selection [J]. Pattern Recognition Letters, 1994, 15(11): 1119-1125.

[4] 王艳丽. 基于多模态进化计算的特征选择算法研究 [D]. 郑州: 郑州大学, 2020.

[5] 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 Yanli, LIANG Jing, XUE Bing, et al. Research on evolutionary computation for feature selection [J]. Journal of Zhengzhou University(Engineering Science), 2020, 41(1): 49-57.

[6] 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 Mengting, DU Jianqiang, LUO Jigen, et al. Research on feature selection of multi-objective optimization [J]. Computer Engineering and Applications, 2023, 59(3): 23-32.

[7] 岳彩通,梁静,瞿博阳,等. 多模态多目标优化综述 [J]. 控制与决策, 2021, 36(11): 2577-2588.
YUE Caitong, LIANG Jing, QU Boyang, et al. A survey on multimodal multiobjective optimization [J]. Control and Decision, 2021, 36(11): 2577-2588.

[8] ZHOU Aimin, ZHANG Qingfu, JIN Yaochu. Approximating the set of Pareto-optimal solutions in both the decision and objective spaces by an estimation of distribution algorithm [J]. IEEE Transactions on Evolutionary Computation, 2009, 13(5): 1167-1189.

[9] TIAN Ye, ZHANG Xingyi, WANG Chao, et al. An evolutionary algorithm for large-scale sparse multiobjective optimization problems [J]. IEEE Transactions on Evolutionary Computation, 2020, 24(2): 380-393.

[10] TIAN Ye, FENG Yuandong, ZHANG Xingyi, et al. A fast clustering based evolutionary algorithm for super-large-scale sparse multi-objective optimization [J]. IEEE/CAA Journal of Automatica Sinica, 2023, 10(4): 1048-1063.

[11] DEB K, TIWARI S. Omni-optimizer: a procedure for single and multi-objective optimization [C]//Evolutionary Multi-Criterion Optimization. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005: 47-61.

[12] LIANG J J, YUE C T, QU B Y. Multimodal multi-objective optimization: a preliminary study [C]//2016 IEEE Congress on Evolutionary Computation (CEC). Piscataway, NJ, USA: IEEE, 2016: 2454-2461.

[13] YUE Caitong, QU Boyang, LIANG Jing. A multiobjective particle swarm optimizer using ring topology for solving multimodal multiobjective problems [J]. IEEE Transactions on Evolutionary Computation, 2018, 22(5): 805-817.

[14] LI Guoqing, SUN Mengyan, WANG Yirui, et al. A ring-hierarchy-based evolutionary algorithm for multimodal multi-objective optimization [J]. Swarm and Evolutionary Computation, 2023, 81: 101352.

- [15] TANABE R, ISHIBUCHI H. A decomposition-based evolutionary algorithm for multi-modal multi-objective optimization [C]//Parallel Problem Solving from Nature: PPSN XV. Cham, Switzerland: Springer International Publishing, 2018: 249-261.
- [16] SUN Yu, ZHANG Shen. A decomposition and dynamic niching distance-based dual elite subpopulation evolutionary algorithm for multimodal multiobjective optimization [J]. Expert Systems with Applications, 2023, 231: 120738.
- [17] PENG Yiming, ISHIBUCHI H. A diversity-enhanced subset selection framework for multimodal multiobjective optimization [J]. IEEE Transactions on Evolutionary Computation, 2022, 26(5): 886-900.
- [18] LI Wenhua, YAO Xingyi, LI Kaiwen, et al. Coevolutionary framework for generalized multimodal multi-objective optimization [J]. IEEE/CAA Journal of Automatica Sinica, 2023, 10(7): 1544-1556.
- [19] MING Fei, GONG Wenyin, WANG Ling, et al. Balancing convergence and diversity in objective and decision spaces for multimodal multi-objective optimization [J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2023, 7(2): 474-486.
- [20] LI Wenhua, YAO Xingyi, ZHANG Tao, et al. Hierarchy ranking method for multimodal multiobjective optimization with local Pareto fronts [J]. IEEE Transactions on Evolutionary Computation, 2023, 27(1): 98-110.
- [21] LI Guoqing, WANG Wanliang, ZHANG Weiwei, et al. Grid search based multi-population particle swarm optimization algorithm for multimodal multi-objective optimization [J]. Swarm and Evolutionary Computation, 2021, 62: 100843.
- [22] LIANG Jing, QIAO Kangjia, YUE Caitong, et al. A clustering-based differential evolution algorithm for solving multimodal multi-objective optimization problems [J]. Swarm and Evolutionary Computation, 2021, 60: 100788.
- [23] TIAN Ye, LIU Ruchen, ZHANG Xingyi, et al. A multipopulation evolutionary algorithm for solving large-scale multimodal multiobjective optimization problems [J]. IEEE Transactions on Evolutionary Computation, 2021, 25(3): 405-418.
- [24] DING Zhuanlian, CAO Lve, CHEN Lei, et al. Large-scale multimodal multiobjective evolutionary optimization based on hybrid hierarchical clustering [J]. Knowledge-Based Systems, 2023, 266: 110398.
- [25] BANGYAL W H, NISAR K, AG IBRAHIM A A B, et al. Comparative analysis of low discrepancy sequence-based initialization approaches using population-based algorithms for solving the global optimization problems [J]. Applied Sciences, 2021, 11(16): 7591.
- [26] ISHIBUCHI H, IMADA R, MASUYAMA N, et al. Comparison of hypervolume, IGD and IGD⁺ from the viewpoint of optimal distributions of solutions [C]//Evolutionary Multi-Criterion Optimization. Cham, Switzerland: Springer International Publishing, 2019: 332-345.
- [27] CHEN Huangke, TIAN Ye, PEDRYCZ W, et al. Hyperplane assisted evolutionary algorithm for many-objective optimization problems [J]. IEEE Transactions on Cybernetics, 2020, 50(7): 3367-3380.

(编辑 亢列梅)