Advances in Applied Mathematics 应用数学进展, 2026, 15(5), 659-672
Published Online May 2026 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2026.155258

Hans汉斯

基于RFE引导PSO启发式特征选择的高维结直肠癌癌前病变分类

张雅洁，汪　颖*

大连交通大学基础部理学院，辽宁 大连

收稿日期：2026年4月28日；录用日期：2026年5月22日；发布日期：2026年5月29日

摘　要

针对结肠镜数据中存在的高维、小样本以及特征冗余严重等问题，构建一种RFE引导的PSO启发式特征选择与分类框架。首先，用F检验与相关性分析对原始特征进行初筛，以降低维度并减少冗余信息；引入RFE生成特征重要性评分引导搜索过程，并采用一种改进的PSO启发式优化策略实现对特征子集的全局搜索，结合局部搜索机制对候选解进一步细化，同时在适应度函数中引入特征规模约束与稳定性约束，增强搜索能力并提高最优解的鲁棒性。最后，采用RBF核的支持向量机作为分类器，在五折交叉验证下评估其分类性能。实验结果表明，其Recall、Accuracy、F1-score及G-mean等多种评价指标均优于对比方法。

关键词

特征选择，RFE引导，PSO启发式优化，高维数据，结直肠癌癌前病变分类

Classification of High Dimensional Colorectal Precancerous Lesion Based on RFE-Guided PSO Inspired Feature Selection

Yajie Zhang, Ying Wang*

School of Science, Department of Foundational Courses, Dalian Jiaotong University, Dalian Liaoning

Received: April 28, 2026; accepted: May 22, 2026; published: May 29, 2026

Abstract

To address the challenges of high dimensionality, small sample sizes, and severe feature redundancy in colonoscopy data, we propose an RFE-guided PSO heuristic framework for feature selection and

*通讯作者。

文章引用: 张雅洁, 汪颖. 基于 RFE 引导 PSO 启发式特征选择的高维结直肠癌癌前病变分类[J]. 应用数学进展, 2026, 15(5): 659-672. DOI: 10.12677/aam.2026.155258

classification. Firstly, the raw features are pre-screened using F-tests and correlation analysis to reduce dimensionality and eliminate redundant information. RFE is introduced to generate feature importance scores that guide the search process, whilst an improved PSO heuristic optimisation strategy is employed to perform a global search for feature subsets. This is combined with a local search mechanism to further refine candidate solutions. Additionally, feature scale and stability constraints are incorporated into the fitness function to enhance search capabilities and improve the robustness of the optimal solution. Finally, a radial basis function (RBF) kernel support vector machine is employed as the classifier, and its classification performance is evaluated under five-fold cross-validation. Experimental results demonstrate that various evaluation metrics, including Recall, Accuracy, F1-score and G-mean, outperform those of the comparison methods.

Keywords

Feature Selection, RFE Guidance, PSO Heuristic Optimisation, High-Dimensional Data, Classification of Colorectal Precancerous Lesions

1. 引言

结直肠癌是全球范围内发病率和死亡率较高的恶性肿瘤之一，其早期诊断对于改善预后和制定个体化治疗方案具有重要意义[1]。随着结肠镜检查及相关影像技术的发展，大量病灶特征数据被收集用于辅助诊断。然而，这类数据通常呈现高维、样本量有限、噪声多、特征冗余的特点，使传统统计方法和经典分类模型易出现过拟合及泛化能力不足的问题。

特征选择作为高维数据分析的核心环节，可从原始特征集合中筛选与分类高度相关且冗余最少的特征子集，从而降低维度与噪声影响，提升模型分类性能。常用方法包括过滤型(filter)、包装型(wrapper)和嵌入型(embedded)三类，分别侧重统计指标、分类器性能和模型训练过程[2]。与特征提取不同，特征选择保留了特征本身的属性，提高医学数据可解释性。在高维少样本数据中，合理的特征选择不仅可以降低模型过拟合风险，还可改善训练效率，减少计算成本，并在一定程度上缓解类别不平衡带来的影响。

茅婷等(2024)提出的 MMTS-AdaBoost [3]方法通过改进马田系统降维并结合 AdaBoost 分类，在结直肠癌癌前病变数据上取得了较高的准确率，优于多种对比算法。然而，该方法依赖固定降维策略，缺乏全局优化搜索与先验信息引导，在高维复杂场景下存在局限性。Deng 等(2024)的 U-RFE [4]方法通过多估计器递归消除与并集策略，在结直肠癌数据上取得 86.4%的准确率，但计算复杂、依赖规则迭代，且整体搜索能力有限，易陷入局部最优。Mohamed 等(2025)提出了一种结合 CNN 与 ACO-PSO [5]的结肠癌诊断方法：利用预训练模型提取特征并进行降维，采用 SVM、KNN 等分类器识别。实验在公开结肠癌组织病理图像上取得了较高的准确率与 F1 值。然而，ACO-PSO 主要用于特征选择，未对分类模型本身进行全局优化，整体框架较复杂。

结直肠病灶特征数据普遍存在高维特征、小样本及信息冗余等问题，因此有必要构建一种兼顾全局搜索能力与先验信息利用的特征选择方法，以提升病灶分类的准确性与稳定性[6]。在此背景下，采用多阶段特征筛选与引导策略：首先，采用过滤方法对特征进行初筛，以降低冗余信息干扰；随后，引入递归特征消除(Recursive Feature Elimination, RFE)评估特征重要性，并将其作为先验信息嵌入搜索过程。在优化阶段，构建 RFE 引导的粒子群启发式搜索机制，在借鉴粒子群优化(Particle Swarm Optimization, PSO)

全局最优引导思想的基础上，不再引入速度项与个体历史最优项，而是直接基于当前位置对解进行更新；同时将 RFE 获得的特征重要性作为先验引导信息，对粒子更新方向进行调节，使搜索更倾向于高重要性特征区域。引入局部搜索策略对候选解进行细化，以弥补未使用个体最优所带来的局部开发能力不足。通过全局最优驱动与先验引导的协同作用，实现全局探索与局部开发之间的有效平衡，从而获得更稳定、更具代表性的特征子集。在分类阶段，采用基于径向基函数核的支持向量机构建分类模型，以提升模型对复杂非线性关系的处理能力[7]；在降低特征维度的同时，有助于提高分类性能与结果稳定性，为结直肠癌病灶的计算机辅助诊断提供可行的技术路径。

2. 算法与理论

2.1. 方差分析(ANOVA) F 检验

方差分析(ANOVA) F 检验[8]是一种常用的单变量特征评价方法，通过比较类间差异与类内离散程度来衡量特征的类别可分性。当某一特征在不同类别之间具有较大的均值差异且类内波动较小时，其类别可分性更强，更有利于后续分类任务。

设数据集包含 C 个类别，则 F 统计量定义为：

$$F = \frac{\text{between-classvariance}}{\text{with-classvariance}} \tag{1}$$

其中，类间方差(between-class variance)用于刻画不同类别均值之间的差异程度，类内方差(within-class variance)用于衡量同一类别内部样本的离散程度。通常，F 值越大，表明该特征在不同类别之间的区分能力越强。

2.2. 相关性过滤

皮尔逊积距相关系数[9] (Pearson Product-Moment Correlation Coefficient, PPMCC)是一种衡量两个随机变量之间线性相关关系强度和方向的常用统计量。其基本思想是通过协方差与标准差的比值来标准化度量两个变量的线性依赖程度：当两个变量相对于各自均值的偏离方向一致时，相关系数为正且绝对值较大；偏离方向相反时，相关系数为负；若二者之间不存在显著线性关系，则相关系数接近于零。

设两个变量 X 和 Y 的样本分别为 $\{x_i\}_{i=1}^{n}$ 和 $\{y_i\}_{i=1}^{n}$，皮尔逊相关系数 r 定义为：

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{2}$$

其中，$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$ 分别表示两个变量的样本均值。n 为样本容量。分子表示 X 与 Y 的协方差，反映两个变量同步偏离各自均值的程度；分母是 X 与 Y 的标准差的乘积，对协方差进行归一化，使相关系数 r 取值落在 $[-1,1]$ 区间。$|r|$ 越接近 1，说明两个变量之间的线性关系越强；$r>0$ 表示正相关，$r<0$ 表示负相关，$r=0$ 表示不存在线性相关。

2.3. 递归特征消除法

递归特征消除法[10]是一种封装式特征选择算法，其核心思想是通过递归地移除最不重要的特征，从而提高分类模型的准确性。该方法使用一个基学习器(如随机森林)对每个特征进行重要性排序，然后反复剔除重要性较低的特征，最终通过交叉验证寻找最优特征子集。

设原始特征集为 $F=\{f_1,f_2,\cdots,f_m\}$，RFE 的迭代过程可描述如下：

(1) 使用第 t 次特征集 $F^{(t)}$ 训练基模型 $M^{(t)}$，获得每个特征的重要性评分 $S(f)=\text{importance}(f)$。

(2) 按照评分升序排序，剔除 k 个非重要特征(k 为步长)，得到剩余特征集 $F^{(t+1)}$。

(3) 重复步骤 1~2，直至剩余特征数达到预设的最小保留数 r。

(4) 采用交叉验证评估不同特征子集的性能，选择使分类精度最高的子集作为最优特征子集。

需要指出的是，RFE 的效果在较大程度上依赖于基学习器的特征评估能力，且由于需要多次训练模型，其计算开销相对较高。因此本文并未直接采用 RFE 进行特征子集的递归筛选，而是利用其生成的特征重要性评分作为先验信息，引导后续特征子集的搜索过程。一方面，该策略避免了 RFE 反复迭代带来的计算负担；另一方面，通过结合基于全局搜索的优化机制，有效弥补了 RFE 在特征选择过程中可能存在的局部性局限，从而在保证计算效率的同时提升特征选择的稳定性与性能。

2.4. 粒子群优化算法

粒子群优化算法[11]是 Kennedy 等人根据鸟群捕食行为中寻找最佳觅食区域的过程所提出的一种群智能算法，具有原理简单、参数少等优点。在粒子群优化算法中，鸟群中的每个个体都是一个粒子，每个粒子均记录自己所找到的最佳觅食位置(个体最优，P_{ij})，粒子群中所有粒子的最佳觅食位置可以看作全局最优解(P_{gj})，每个粒子的觅食位置拥有食物的可能性通过适应度刻画。

假设个体数为 N 的粒子群在 D 维空间中寻找最优解，在粒子群算法中对第 t 次迭代中第 i 个粒子在第 j 维上的位置速度分别表示为 x_{ij}^{t} 和 v_{ij}^{t}。在迭代过程中，粒子依据个体最优与全局最优不断更新自身状态，从而实现对最优解的搜索。

在第 t 次迭代时，第 i 个粒子在第 j 维上的速度与位置更新公式如下：

$$v_{ij}^{t+1}=\omega v_{ij}^{t}+c_1 r_1\left(p_{ij}-x_{ij}^{t}\right)+c_2 r_2\left(p_{gj}-x_{ij}^{t}\right) \tag{3}$$

$$x_{ij}^{t+1}=x_{ij}^{t}+v_{ij}^{t+1} \tag{4}$$

其中，ω 为惯性权重，用于控制粒子当前速度对历史速度的继承程度；c_1 和 c_2 分别为个体学习因子与群体学习因子；r_1 和 r_2 为区间 $[-1,1]$ 上的随机数。

根据等式(3) (4)可知，粒子群优化算法的搜索过程主要依赖个体最优与全局最优进行迭代更新。在迭代初期，群体多样性迅速下降，易出现早熟收敛，从而难以充分挖掘潜在有效特征子集；同时，在高维特征空间中缺乏有效引导，搜索过程具有一定盲目性，影响收敛效率[12]。

2.5. 基于径向基函数核的支持向量机

径向基函数核支持向量机[13]是一种利用径向基核函数将数据映射到高维空间从而解决非线性分类问题的支持向量机模型。它通过最大化分类间隔寻找最优超平面，并使用核技巧避免显式计算高维映射。

RBF 核的定义为：

$$k\left(x_i,x_j\right)=\exp\left(-\frac{\left\|x_i-x_j\right\|^2}{2\sigma^2}\right) \tag{5}$$

其中 $\sigma>0$ 控制核函数的宽度。x_i 和 x_j 表示两个不同的输入样本(特征向量)，例如数据集中的第 i 个和第 j 个样本。

采用该核的 SVM 决策函数可写为：

$$y(x)=sign\left(\sum_{p=1}^{N}\alpha_p y_p k\left(x_p,x\right)+b\right) \tag{6}$$

式中 α_p 为拉格朗日乘子， y_p 为样本标签， b 为偏置， x_p 为支持向量(即 $\alpha_p \neq 0$ 的训练样本)。

RBF-SVM 通过调整参数 C 和 σ 来控制模型复杂度与分类精度，其中 C 是正则化参数，用于平衡模型复杂性与训练误差。

3. 基于 RFE 引导 PSO 启发式特征选择算法

群体智能优化算法(Swarm Intelligence Optimization, SIO)通过模拟自然界群体协同行为，实现对复杂搜索空间的高效探索，在特征选择问题中得到广泛应用[14]。典型算法如粒子群优化和灰狼优化算法等，具有易于理解、参数简单、搜索能力强等优点，能够有效地解决高维数据的问题。尽管基于群体智能的特征选择方法在高维数据处理中展现出较强的全局搜索能力，其在实际应用中仍面临一定挑战。一方面，传统粒子群优化方法缺乏有效的先验信息引导，在高维特征空间中容易出现搜索效率低及收敛不稳定的问题[15]；另一方面，单纯依赖随机初始化与经验更新机制，难以充分利用特征间潜在的重要性差异，从而影响特征子集的质量。相比之下，递归特征消除(RFE)能够基于模型对特征重要性进行评估，在一定程度上反映特征对分类任务的贡献，但其计算开销较高，且易受基学习器性能影响。

基于上述分析，有必要将基于模型的特征评估机制与群体智能优化方法相结合，以兼顾搜索效率与特征评估的有效性。为此，本文首先采用检验与相关性分析对高维原始数据集进行初筛，在降低维度的基础上构建一种 RFE 引导的 PSO 启发式特征选择方法，通过引入 RFE 生成的特征重要性信息对搜索过程进行引导，保持全局搜索能力的同时提高搜索的有效性，流程图如图 1 所示。

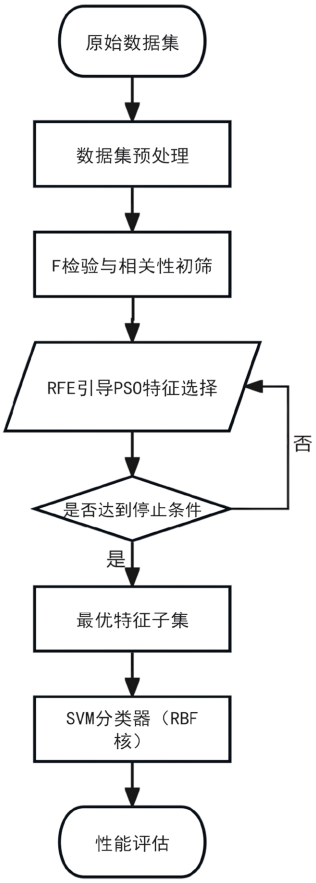

Figure 1. Flowchart of the RFE-guided PSO heuristic feature selection algorithm
图 1. 基于 RFE 引导的 PSO 启发式特征选择算法流程图

3.1. 预处理

本文对数据进行预处理，包括高缺失特征剔除与插补、近常数特征去除、异常值裁剪、以及标准化处理。随后利用 F 检验对特征进行重要性评估，根据检验得分大小对特征进行降序排序后选取前 k 个特征，其中 $k=150$，对选取的前 k 个特征进行相关性分析，去除高度相关的冗余特征，降低特征间冗余性并提升特征子集的紧凑性[16]。经过 F 检验与相关性过滤处理后，原始 698 维特征筛选为小于等于 150 维的候选特征子集，为后续 RFE 引导的 PSO 特征选择提供基础。

3.2. 基于 RFE 引导的 PSO 启发式特征选择算法

在获得候选特征子集后，进一步引入 RFE 引导的 PSO 启发式优化策略，对特征组合进行精细搜索，以获取最优特征子集。在该过程中，任一候选解均可被表示为一个二进制编码问题。向量中元素取值为 0 或 1，分别表示对应特征的舍弃与保留。

由于标准 PSO 适合针对连续优化问题而设计的，其速度–位置更新机制难以直接用二进制特征表示，通常需借助额外映射函数，可能导致信息偏差并影响搜索稳定性。PSO 在搜索过程中缺乏有效的先验引导信息，粒子更新主要依赖个体最优与全局最优，易在高维复杂空间中出现搜索方向盲目、收敛效率较低的问题。在迭代后期，种群多样性逐渐降低，容易陷入局部最优，从而影响最终解的质量与稳定性[17]。

针对上述问题，本文在 PSO 全局最优引导思想的基础上，引入 RFE 生成的特征重要性信息作为先验引导，并结合二值化映射机制与局部搜索策略，对搜索过程进行改进，从而提升特征选择的有效性与稳定性。改进如下：

3.2.1. 适应度函数的设计

为了能稳定地在分类性能与特征规模之间取得平衡，对候选解 X，本文构建如下适应度函数：

$$f(X)=F1_{mean}-\omega_1\cdot\frac{n.select}{X.shape[1]}-\omega_2\cdot\sigma(F1) \tag{7}$$

其中，$F1_{mean}$ 表示在交叉验证下获得的加权 F1-score 均值，$n.select$ 为当前选择的特征数，$X.shape[1]$ 为总特征数，$\sigma(F1)$ 表示各折 F1-score 的标准差。$\omega_1=0.15$ 和 $\omega_2=0.15$ 分别为特征规模惩罚项与稳定性约束项的权重系数。

该适应度函数在保证分类性能的同时，对特征数量与结果波动进行约束，提高模型的泛化能力与鲁棒性。

3.2.2. RFE 引导机制

由于传统粒子群优化在搜索过程中粒子的更新依赖个体最优和全局最优，缺乏对种群其他有效信息的利用，易导致搜索方向单一[18]，尤其在高维特征空间中易陷入局部最优。为解决这一问题，本文引入递归特征消除(RFE)对候选特征子集进行重要性评估，并将其生成的评分作为先验知识引导粒子更新。

在 RFE 过程中，本文采用随机森林(Random Forest, RF)作为基学习器，其对高维数据具有鲁棒性强、能够刻画非线性特征关系等优点。具体参数设置为：树的数量(n_estimators)设为 50，其余参数采用默认配置。在特征递归消除过程中，特征保留比例设为当前特征数的一半，步长(step)为 0.1，即每轮迭代移除约 10%的低重要性特征。

基于 RFE 得到的特征排序结果，本文进一步构建归一化的重要性评分，将特征排名映射至区间[0,1]，排名越靠前的特征对应更高的引导分数，从而形成连续的特征引导向量。该引导向量嵌入粒子更新过程，为搜索提供方向性约束，使优化过程能够优先聚焦于高贡献特征区域。

需要指出的是，为降低 RFE 多轮迭代训练带来的计算开销，本文并未直接采用其进行特征子集筛选，而是仅利用其生成的特征评分作为搜索引导。该策略在降低计算复杂度的同时，保持了优化过程对高贡献特征的聚焦能力，从而提升了搜索效率与特征选择的稳定性。

3.2.3. RFE 引导的二进制映射机制

由于特征选择问题本质上属于离散优化问题，需要将粒子在连续空间中的位置映射为二进制特征选择向量[19]。为此，本文在传统 Sigmoid 阈值映射基础上，引入 RFE 生成的特征重要性分数，对二值化过程进行自适应调节。

具体而言，第 i 个粒子在第 t 次迭代后的位置为 $X_i^t=\left(x_{i1}^t,x_{i2}^t,\cdots,x_{iD}^t\right)$，其对应的二进制特征选择向量 S_i^t 定义为：

$$s_{ij}^t=\begin{cases}1, x_{ij}^t>\theta_j\\ 0, x_{ij}^t\le\theta_j\end{cases} \tag{8}$$

其中，阈值 θ_j 由 RFE 引导分数动态调整：

$$\theta_j=\theta_0\cdot\left(1-G_j\right)+\lambda\cdot G_j \tag{9}$$

在等式(9)中，G_j 为第 j 个特征的 $guidance_{RFE}$ 得分，θ_0 为初始阈值；λ 为调节系数。

本机制通过引入特征重要性先验信息，使重要特征对应的阈值降低，从而提高其被选中的概率，而低重要性特征则更易被剔除，实现了搜索过程对关键特征的自适应偏置。

3.2.4. 局部搜索机制

在得到二进制映射结果后，为进一步提升特征子集的精细程度并避免算法陷入局部最优，在粒子更新过程中，本文引入局部搜索机制，对当前候选解进行邻域微调优化。设当前特征子集掩码为 $m\in\{0,1\}^D$，D 为特征总数，集合 $S=\left\{j\in\{1,2,\cdots,D\}\mid m_j=1\right\}$ 表示当前选中的特征子集，集合 $T=\left\{j\in\{0,1,\cdots,D\}\mid m_j=0\right\}$ 表示未被选中的特征子集，则 $m_j=0$ 与 $m_j=1$ 分别表示第 j 个位置对应的特征未选中与选中的状态，那么特征的总集合 Δ 即可表示为 $\Delta=S+T$。在此次迭代中对当前种群中的部分候选解执行如下微调操作：

(1) 随机删除部分特征

若当前选中特征数量 $|S|$ 大于最小特征数 N_{min}，则随机删除一部分特征，删除数量定义为：

$$k_{drop}=\max\left(1,\left(|S|\cdot\mu\right)\right),\mu=0.1 \tag{10}$$

在(10)式中，μ 为需要删除的特征比例。即每次删除至少 1 个特征，最多删除约当前特征的 10%。被删除的特征索引从当前选中集合 S 中均匀随机选取，不依赖任何评分，从而保持扰动的随机性，增加跳出局部最优的机会。

(2) 引导添加部分特征

从未选中的特征集合 T 中，依据引导分数有放回地采样添加少量特征。添加数量为：

$$k_{add}=\max\left(1,\left(|T|\cdot\zeta\right)\right),\zeta=0.02 \tag{11}$$

即至少添加 1 个特征，每次最多添加约 2%的总特征数。在(11)中，ζ 表示需要添加的特征比例。$|T|$ 表示未被选中特征的数量，当我们决定要添加 k_{add} 个新特征时，特征集合 T 每个未选中特征 j 被选中的概率 p_j 与其引导分数成正比，即 p_j 为：

$$p_j=\frac{g_j}{\sum_{k\in T}g_k} \tag{12}$$

其中，g_j 是第 j 特征的引导分数 $guidance_{RFE}$，分母是 T 中所有特征的引导分数之和。这样，分数越高的特征，分子越大，被选中的概率就越大。

(3) 特征数量约束

经删除和添加后，若新特征子集规模 $|S'| > N_{\max}$，$N_{\max}$ 为设置最大特征数，则随机剔除多余特征，直至满足 $|S'| \le N_{\max}$，同时保证 $|S'| \ge N_{\min}$。

该局部搜索能够在全局搜索的基础上对特征组合进行邻域探索，避免陷入局部最优。

3.2.5. 引导式粒子更新策略

综合上述方法，在标准 PSO 的基础上，对粒子更新机制进行改进，本文引入全局最优引导与 RFE 先验引导的双驱动策略。第 i 个粒子位置更新迭代更新公式为：

$$X_i^{t+1} = X_i^t + r_1\left(g_{best} - X_i^t\right)\cdot F + r_2 \cdot guidance_{RFE} \tag{13}$$

其中，r_1 和 r_2 为 $[0,1]$ 的随机数，g_{best} 为全局最优位置，$guidance_{RFE}$ 为由 RFE 生成的特征先验得分，F 为动态收敛因子，其定义为：

$$F(t) = \exp\left(-\frac{t+1}{T}\right) \tag{14}$$

其中，t 为当前迭代次数，T 为最大迭代次数。

在基于 RFE 引导的 PSO 启发式特征选择算法中，每个候选解被编码为一个二进制向量，向量的每一位对应一个特征的取舍状态。适应度函数综合考虑分类性能、所选特征规模和结果稳定性，以全面评价候选解的质量。在全局搜索阶段，利用 RFE 生成的特征重要性分数作为先验知识，引导粒子的更新方向，从而提升高类别可区分能力特征被选中的概率；借助二进制映射机制将连续位置值转换为离散的特征子集。迭代过程中，引入局部搜索对粒子进行邻域微调：随机删除部分已选特征，依据 RFE 分数按概率添加特征，并对特征总数进行约束，以进一步优化特征组合。通过全局最优解与 RFE 先验信息双重驱动的更新策略，粒子在高维特征空间中既能保持足够的探索能力，又能聚焦于关键特征区域，显著提升特征选择的有效性与稳定性。

最后，采用基于径向基函数核的支持向量机作为分类器。该分类器能够有效处理高维、非线性特征空间中的分类问题，并具有较强的泛化能力，从而实现对优化后特征子集的稳定建模与分类。

综上所述，本文方法通过融合 F 检验初筛、RFE 先验引导、二进制 PSO 搜索机制以及局部搜索策略，在保持全局搜索能力的同时增强了特征选择过程的方向性与稳定性。

4. 实验结果与对比分析

4.1. 实验设置与分析

4.1.1. 实验参数设置

本文的实验数据集来源于 UCI 数据库中的结直肠癌癌前病变视频信息数据集，其特征维数为 698 [3]，样本数为 152。

实验在 Windows 11 64 位操作系统环境下进行，硬件配置为 AMD Ryzen 7 8745H 处理器及 16 GB 内存。算法基于 Python 实现，主要依托 NumPy、Pandas 以及 Scikit-learn 等开源机器学习库完成。为保证所提出的 RFE 引导 PSO 启发式特征选择方法具有良好的稳定性与可复现性，对算法中的关键参数进行了统一设置，如表 1 所示。

各参数的设定在综合考虑计算复杂度与模型分类性能的基础上确定，以在保证实验公平性与结果可

比性的同时，提高方法的稳定性与可重复性。

Table 1. Experimental parameter settings
表 1. 实验参数设置

参数名称	符号/变量	取值	说明
初筛特征数	k	150	初筛选择 Top-k 特征
树数量	n_estimators	50	随机森林规模
选择特征比例	—	50%	RFE 保留一半特征
步长	step	0.1	每轮递归删除比例
粒子数量	N	20	种群规模
最大迭代次数	T	30	停止条件
特征最小数量	—	2	子集下限约束
特征最大数量	—	50	子集上限约束
初始阈值	θ_0	0.45	基础选择阈值
调节系数	λ	0.3	RFE 引导调节
搜索次数	—	4-5	局部搜索每轮搜索次数

4.1.2. 实验参数分析

本文方法涉及多个关键参数，下面将对三个关键参数的合理设置包括初筛特征数 k 、粒子数量 N 、最大迭代次数 T 进行敏感性分析，并在固定其他参数的情况下从选择特征数量与六个评价指标 Precision、Specificity、Recall、Accuracy、F1-score、G-mean 展示其在一定范围内的变化对结果的影响。

(1) 初筛特征数 k

初筛阶段通过 F 检验选取 Top-k 特征。若 k 过小，可能丢失重要判别信息；若 k 过大，则会增加后续搜索空间复杂度。本文在 $k \in [100, 200]$ 范围内进行测试，结果表明当 $k = 150$ 时，模型在分类性能与计算效率之间取得较优平衡，因此本文采用 $k = 150$ 。

图 2 展示了在不同 k 值对多个分类评价指标的影响。可以看出， $k = 150$ 时所有指标均达到最高值，表现最佳， $k = 100$ 时各项指标相对较低， $k = 200$ 时部分指标略低于 $k = 150$ ，但整体仍优于 $k = 100$ 。

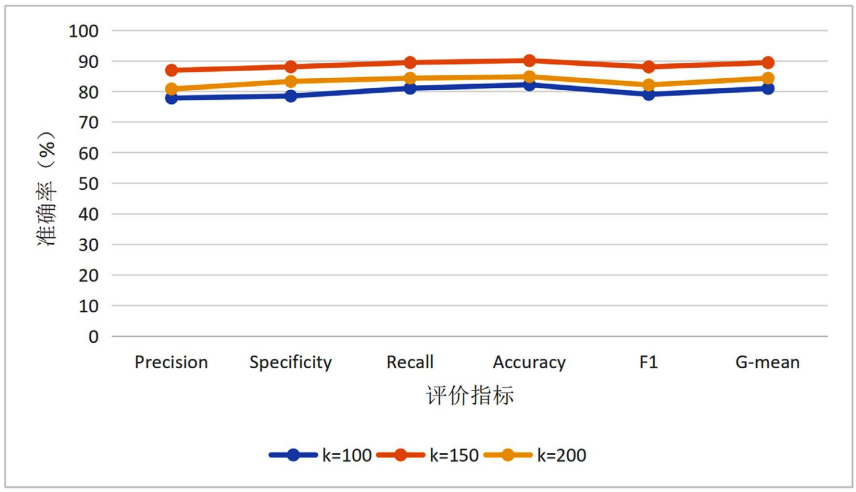

Figure 2. Effect of initial feature selection (k) on model performance and feature count
图 2. 不同初筛特征数 k 对模型性能与特征数影响

(2) 粒子数量 N

粒子数量决定搜索空间的覆盖能力。较小的 N 可能导致搜索不足，而过大的 N 会显著增加计算开销。通过实验对比 $N \in \{10,20,30\}$，发现当 $N=20$ 时，算法在性能与运行时间之间表现较为稳定，选取该值。

图 3 展示了在不同样本量 N 下各分类评价指标的变化情况。可以看到，$N=20$ 时所有指标均达到最高，表现最佳。$N=10$ 时部分指标相对较低，但 Specificity 和 Accuracy 尚可。$N=30$ 时多数指标略低于 $N=20$，但接近 $N=10$ 的水平。总体来看，样本量 $N=20$ 在该特征数下最有利于模型性能。

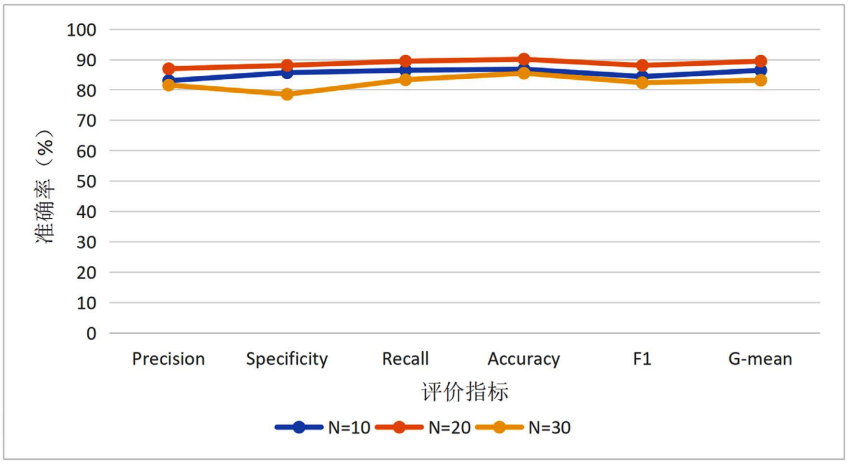

Figure 3. Effect of particle number (N) on model performance and selected feature count
图 3. 不同粒子数 N 对模型性能与特征数影响

(3) 最大迭代次数 T

迭代次数影响算法收敛程度。本文测试了 $T \in \{20,30,40\}$，结果显示，当 $T=30$ 时算法已基本收敛，继续增加迭代次数对性能提升有限，但会显著增加计算时间，因此最终设置 $T=30$。

图 4 展示了在不同参数 T 下各分类评价指标的变化情况。$T=30$ 时绝大部分指标均达到最高值，表现最佳。$T=40$ 时指标普遍处于中等水平，Recall 与 G-mean 接近于 $T=30$。$T=20$ 时多数指标明显下降。总体来看，$T=30$ 是最优参数设置。

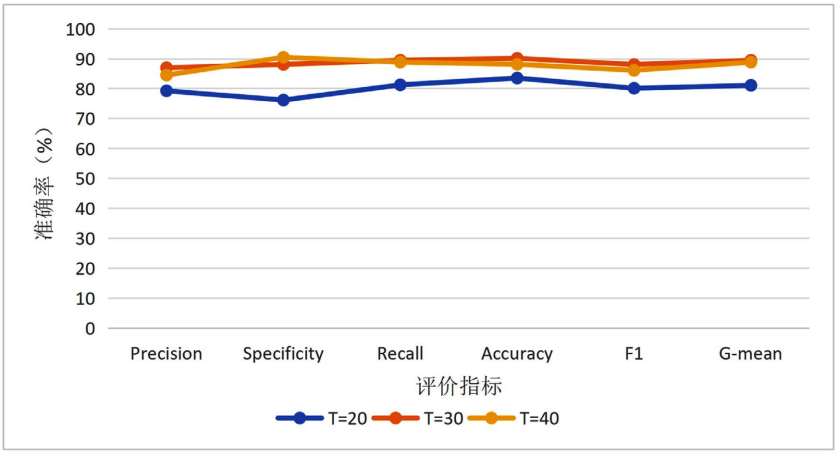

Figure 4. Effect of maximum iteration number (T) on model performance and selected feature count
图 4. 不同迭代次数 T 对模型性能与特征数的影响

综合上述分析，本文所采用的参数设置在性能与计算成本之间取得了良好的平衡，能够保证算法在高维小样本数据上的稳定性与有效性。

4.2. 评价指标

本文选取查准率(Precision)、特异性(Specificity)、召回率(Recall)、准确率(Accuracy)、F1 值及 G-mean 等多种评价指标，对模型的分类性能进行全面评估。

在二类问题的混淆矩阵中，对于二分类问题，可将样例根据其真实的类别与学习器预测的类别组合划分为真正例(TP)、假正例(FP)、真负例(TN)、假负例(FN)四种情形[20]。其中，查准率用于衡量被模型预测为正类的样本中，实际为正类的比例，其定义为：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

特异性衡量实际为负类的样本中被正确预测为负类的比例，可定义为：

$$\text{Secificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{16}$$

召回率衡量实际为正类的样本中被正确预测为正类的比例，定义为：

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{17}$$

准确率是模型预测正确的样本数占总样本数的比例。其计算公式为：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{18}$$

F1 值是查准率与召回率的调和平均，用于综合衡量模型的分类性能[21]，其定义为：

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{19}$$

几何均值用于衡量模型在正负类上的综合分类能力，尤其适用于类别不平衡问题[22]，其定义为：

$$\text{G-mean} = \sqrt{\text{Specificity} \times \text{Recall}} \tag{20}$$

4.3. 消融实验

为了验证本文提出的关键模块对最终分类性能的贡献，我们进行了消融实验(Ablation Study)。具体评估了以下三个模块的作用：

(1) RFE 引导机制；

(2) 局部搜索机制；

(3) F 检验初筛特征。

通过逐步移除或禁用每个模块，并观察分类性能和特征选择稳定性变化，可分析各模块的有效性与必要性。

4.3.1. 实验设计

本研究在固定其他参数的条件下，分别对 RFE 引导机制、局部搜索机制和 *F* 检验初筛特征进行消融实验。

具体做法为：

(1) 移除 RFE 引导分数以评估其对特征搜索的指导作用；

(2) 禁用局部搜索机制以考察其对特征子集微调与稳定性的影响；

(3) 移除 F 检验初筛以观察其对计算效率和性能的影响。

通过 5 折交叉验证记录六个分类指标的结果及最终选择的特征数量，从而分析各模块在模型中的作用。

4.3.2. 消融实验结果

表 2 总结了不同消融设置下模型在六个指标上的性能，其中标黑部分为最优结果。

Table 2. Ablation study results comparison
表 2. 消融实验结果对比

消融模块	Precision	Specificity	Recall	Accuracy	F1	G-mean
完整方法	86.98	**88.1**	**89.5**	90.13	88.08	**89.49**
移除 RFE 引导分数	82.32	83.33	85.3	86.18	83.53	85.28
禁用局部搜索机制	81.73	76.19	82.64	85.53	82.16	82.39
去除 F 检验初筛	**89.53**	83.33	88.94	**91.45**	**89.23**	88.76

在消融实验中，完整方法、移除 RFE 引导分数、禁用局部搜索机制、去除 F 检验初筛特征数四个方法最终取得的特征数量分别为 28、29、29、249。结合表 2 可以看出：

完整方法在整体性能上表现最佳，同时特征数量最少，仅 28 个，说明该方法能够实现高效而准确的分类性能。

移除 RFE 引导分数导致 Precision、Specificity、Recall、以及 G-mean 都有不同程度下降，说明 RFE 引导在特征选择中对模型性能具有明显贡献。

禁用局部搜索机制使 Specificity 及 G-mean 下降更明显，表明局部搜索对优化特征子集的稳定性和分类平衡性起到重要作用。

去除 F 检验初筛虽然在 Precision、F1-score 和 Accuracy 上略有提升，但特征数为 249 个，特征数量大幅增加，表明初筛对于提高计算效率和控制特征冗余仍然重要。

4.4. 结果分析

图 5 中展示了本文提出方法在六个评价指标上的性能分布情况，包括 Precision、Specificity、Recall、Accuracy、F1-score 以及 G-mean。各指标以极坐标形式均匀分布，数值范围为 0~100。由图可见，各项指标均处于较高水平，约 86%~91%之间，整体轮廓接近规则六边形，表明模型在不同评价指标之间表现均衡。

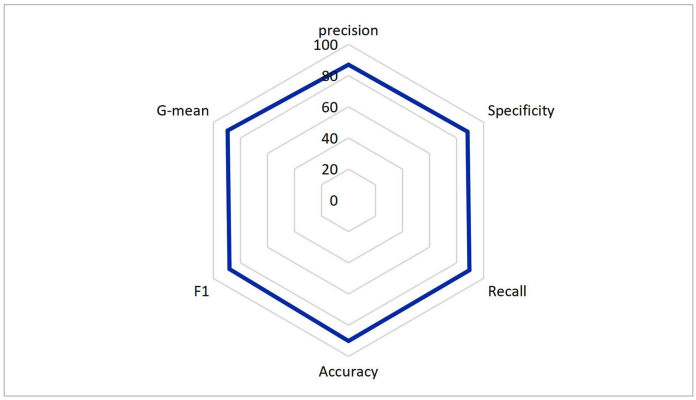

Figure 5. Classification performance radar chart
图 5. 分类性能雷达图

为验证所提出方法的有效性，选取 MMTS-AdaBoost、mRMR-AdaBoost、Chi-square-AdaBoost 以及未进行特征选择的 AdaBoost、BP 神经网络(Back Propagation Neural Network, BP NN)、朴素贝叶斯(Naive Bayes, NB)和支持向量机(Support Vector Machine, SVM)等方法进行对比分析。表 2 中加粗结果表示各评价指标下的最优值。

Table 3. A comparison of classification performance among different methods on high-dimensional data of colorectal lesions
表 3. 不同方法在高维结直肠病灶数据上的分类性能比较

算法	维度	precision /%	Specificity /%	Recall /%	Accuracy /%	F1 /%	G-mean /%
F 检验 + RFE 引导的 PSO	28	86.98	88.1	**89.5**	**90.13**	**88.08**	**89.49**
MMTS-AdaBoost	**10**	**100.00**	**100.00**	60.00	84.62	75.00	77.46
mRMR-AdaBoost	30	83.33	93.75	50.00	76.92	62.50	68.47
Chi-square-AdaBoost	30	83.33	93.75	50.00	76.92	62.50	68.47
AdaBoost	415	83.33	94.12	55.56	80.77	66.67	72.31
BP NN	415	83.33	94.12	55.56	80.77	66.67	72.31
NB	415	50.00	82.35	33.33	65.38	40.00	52.39
SVM	415	**100.00**	**100.00**	54.55	80.77	70.59	73.85

如表 3 所示，本文提出的 *F* 检验 + RFE 引导的 PSO 启发式特征选择方法在大多数评价指标上表现优异。具体而言，本方法在 Recall、Accuracy、F1-score 以及 G-mean 上分别达到了 89.5%、90.13%、88.08%和 89.49%，在保证较少特征数量的情况下取得了整体最佳的性能，显示出模型在有限特征条件下仍保持了稳定的综合性能。相比之下，MMTS-AdaBoost 虽然在 Precision 和 Specificity 上达到 100%，但其 Recall 仅为 60.00%，说明其对正类样本识别能力较弱，存在明显偏置。mRMR 和 Chi-square 方法整体表现较低，尤其在 Recall 和 F1-score 指标上下降明显。未进行特征选择的分类器(如 AdaBoost、BP NN、NB 和 SVM)虽然在部分指标上表现较好，但整体性能不稳定，且在 Recall 和 G-mean 方面存在明显不足。

此外，从特征维度角度来看，本文方法仅选取 28 个特征即可获得最优性能，相比原始 415 维特征显著降低了特征规模，仅次于 MMTS-AdaBoost 选择的特征数量，但也有效减少了冗余信息，提高了模型的计算效率与泛化能力。

综上所述，所提出方法在保证较低特征维度的同时，实现了分类性能与稳定性的综合提升，具有较强的特征筛选能力和良好的实际应用价值。

5. 结语

本文针对结直肠癌癌前病变数据中存在的高维、小样本及特征冗余问题，提出了一种基于 RFE 引导的 PSO 启发式特征选择方法。通过融合 *F* 检验与相关性分析实现初步降维，引入 RFE 生成特征重要性作为先验信息引导搜索过程，并结合局部搜索机制与改进的适应度函数，在保证特征规模可控的同时提升了分类性能与结果稳定性。实验结果表明，所提出方法在 Accuracy、Recall、F1-score 及 G-mean 等多个指标上均优于对比方法，且在显著降低特征维度的情况下仍保持较高的分类准确率，验证了其在高维医学数据分析中的有效性与优越性。但 RFE 与 PSO 结合后整体计算开销相对较高，在大规模高维数据场景下的运行效率仍有进一步优化空间。

参考文献

[1] 景凯. 基于生物信息学分析技术筛选结直肠癌相关基因及其功能[D]: [硕士学位论文]. 济南: 山东大学, 2024.

[2] 孙丽芹. 基于智能优化的高维数据特征选择算法研究[D]: [博士学位论文]. 西安: 西安电子科技大学, 2023.

[3] 茅婷, 张月义, 孙叶芳, 虞岚婷. 基于 MMTS-AdaBoost 的高维结直肠癌癌前病变分类[J]. 计算机应用与软件, 2024, 41(1): 291-296.

[4] Deng, F., Zhao, L., Yu, N., Lin, Y. and Zhang, L. (2024) Union with Recursive Feature Elimination: A Feature Selection Framework to Improve the Classification Performance of Multicategory Causes of Death in Colorectal Cancer. *Laboratory Investigation*, **104**, Article ID: 100320. https://doi.org/10.1016/j.labinv.2023.100320

[5] Ali A. Mohamed, A., Rahebi, M., Hançerlioğulları, A. and Rahebi, J. (2025) An Approach Based on Convolutional Neural Network and ACO-PSO for Colon Cancer Disease Diagnosis. *Politeknik Dergisi*, **28**, 649-659. https://doi.org/10.2339/politeknik.1419744

[6] 廖南清, 张祁新. 结直肠癌病理类型的多模态融合分类模型[J]. 生物医学, 2025, 15(5): 1012-1023.

[7] 曹君杰, 冯爱芬, 常芳欣, 杨双杨, 蒋智涵, 王世杰. 网格搜索的支持向量机方法在乳腺癌诊断中的应用[J]. 应用数学进展, 2025, 14(5): 238-243.

[8] Rayarao, S.R. (205) F-Tests: A Comprehensive Review of Theory, Applications, and Statistical Inference. Authorea.

[9] 徐维超. 相关系数研究综述[J]. 广东工业大学学报, 2012, 29(3): 12-17.

[10] 林小棋, 任超, 李毅, 等. 基于 Relief F-RFE 特征优选的桉树人工林提取[J]. 测绘科学, 2023, 48(10): 107-115.

[11] Gad, A.G. (2022) Particle Swarm Optimization Algorithm and Its Applications: A Systematic Review. *Archives of Computational Methods in Engineering*, **29**, 2531-2561. https://doi.org/10.1007/s11831-021-09694-4

[12] 陈垂丽. 基于多目标进化优化的特征选择理论与方法[D]: [硕士学位论文]. 北京: 中国矿业大学, 2025.

[13] Du, K., Jiang, B., Lu, J., Hua, J. and Swamy, M.N.S. (2024) Exploring Kernel Machines and Support Vector Machines: Principles, Techniques, and Future Directions. *Mathematics*, **12**, Article 3935. https://doi.org/10.3390/math12243935

[14] 高岳林, 杨钦文, 王晓峰, 等. 新型群体智能优化算法综述[J]. 郑州大学学报(工学版), 2022, 43(3): 21-30.

[15] 冯茜, 李擎, 全威, 裴轩墨. 多目标粒子群优化算法研究综述[J]. 工程科学学报, 2021, 43(6): 745-753.

[16] Xie, S., Zhang, Y., Lv, D., Chen, X., Lu, J. and Liu, J. (2022) A New Improved Maximal Relevance and Minimal Redundancy Method Based on Feature Subset. *The Journal of Supercomputing*, **79**, 3157-3180. https://doi.org/10.1007/s11227-022-04763-2

[17] 孙会岳. 基于粒子群优化的高维特征选择方法研究[D]: [硕士学位论文]. 大连: 大连理工大学, 2025.

[18] 彭建新, 詹志辉. 全局信息引导的改进粒子群优化算法[J]. 小型微型计算机系统, 2016, 37(7): 1518-1521.

[19] 肖胤喆. 基于改进粒子群优化算法的特征选择方法研究[D]: [硕士学位论文]. 长春: 吉林大学, 2022.

[20] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

[21] Takahashi, K., Yamamoto, K., Kuchiba, A. and Koyama, T. (2022) Confidence Interval for Micro-Averaged F_1 and Macro-Averaged F_1 Scores. *Applied Intelligence*, **52**, 4961-4972. https://doi.org/10.1007/s10489-021-02635-5

[22] de la Cruz Huayanay, A., Bazán, J.L. and Russo, C.M. (2024) Performance of Evaluation Metrics for Classification in Imbalanced Data. *Computational Statistics*, **40**, 1447-1473. https://doi.org/10.1007/s00180-024-01539-5