基于 Fisher Score 和 RFE 的集成特征选择方法

盛煜轩[1] 严嘉浩[1] 苏志军[1]
SHENG Yuxuan YAN Jiahao SU Zhijun

摘 要 目前高维癌症基因数据普遍存在样本少、维度高、冗余性强等问题，文章提出了一种基于 Fisher Score 和递归特征消除法（RFE）的集成特征选择算法 FS-RFE。该算法旨在从高维癌症数据样本中挖掘出潜在的癌症基因标志物，为业界癌症诊断和积极治疗提供理论支持。FS-RFE 算法的流程为首先利用 Fisher Score 进行特征初筛以评估特征的重要性排序，然后筛选出初选特征子集，以达到快速降维的目的；之后为剔除冗余的特征，通过递归特征消除法对初选特征子集进行进一步特征筛选，从而提取出最优的特征子集。与传统的 4 种单一特征选择算法相比，FS-RFE 算法在全局搜索与局部搜索能力之间取得了较好的平衡，且显著提升了搜索效率。在 9 个公开肿瘤基因数据集上的实验结果表明，FS-RFE 算法在分类准确率（ACC）和 F_1-score 等评价指标上均优于对比的传统特征选择算法，表现出 FS-RFE 算法在处理高维小样本数据集时，能够筛选出具有较高质量的特征子集，从而提高分类性能，有效降低数据维度。所提研究为高维小样本癌症基因数据集的特征选择领域提供了一种高效的解决方案，具有较好的理论意义和应用价值。

关键词 机器学习；特征选择；高维数据；集成学习；生物信息学

doi：10.3969/j.issn.1672-9528.2025.06.040

0 引言

随着信息技术的快速发展，在各行各业中积累了海量的数据，呈现出“爆炸式”的增长趋势，特别是在生物医学、金融以及图像处理等领域，高维数据的处理已成为研究的热点问题。然而，往往在高维数据中包含着大量无关、冗余以及噪声特征，不仅增加了计算复杂度，还增加了模型过拟合的风险，从而影响着分类性能，这种现象在学术界被称为“维数灾难”[1]。为解决这一问题，特征选择作为一种重要的数据预处理技术逐渐体现出了重要性，并且广受关注。特征选择的原理是通过筛选出与目标变量最相关的特征子集，以达到有效降低数据维度的目的，从而显著提升了模型的泛化能力和计算效率[2]，在工业界以及学术界中有着较为重要的应用价值。

特征选择方法根据其与学习算法的关系，可以分为过滤式（Filter）、封装式（Wrapper）和嵌入式（Embedded）三类[3]。每种方法都有其优势和局限性，因此也适用于不同的应用场景之中。过滤式特征选择方法的核心思想是通过评估特征与目标变量之间的相关性来进行特征筛选，因此过滤式方法具有计算效率较高、适用性较强等优点，但过滤式方法的局限性是其过程与后续学习器无关，可能会导致筛选出的特征子集在特定任务中表现不佳。封装式特征选择方法的原理是通过结合分类器对特征子集进行评价，以分类器交叉验证的结果作为特征子集优秀的评估标准，但其缺陷是计算成本较高，特别在高维数据场景下，时间开销较大。嵌入式特征选择方法的原理是在分类器训练中自动进行特征选择，因此具有较快的速度。但嵌入式方法的缺点是其性能受限于模型的选择以及参数调优，且无法控制特征选择过程[4]。

近年来涌现了很多优秀的特征选择方法，例如基于信息论的特征选择算法（如 mRMR[5]）通过度量特征与目标变量之间的互信息进行筛选，但其缺陷是未充分考虑特征间的冗余问题；FCBF[6] 算法通过引入对称不确定性和近似马尔科夫毯理论，能够有效去除冗余特征，其缺点是忽略了多特征之间的交互影响；Lasso[7] 方法通过 L1 正则化实现特征选择，其缺陷是高冗余特征场景下可能误判冗余特征为重要特征。如何更好进行特征选择成为学术界共同追求的目标。近年来，集成特征选择方法备受关注。集成特征选择的核心思想是通过结合多种特征选择算法的结果，从而能够有效提升模型的稳定性和性能[8]，能够较好兼顾性能和时间开销，因此具有

1. 浙江同济科技职业学院 浙江杭州 311231
[基金项目] 高校一般科研项目“面向高维数据的集成特征选择研究”（ FRF24YB012）

较好的实际应用意义。

针对上述问题，本文提出了一种结合 Fisher Score 和递归特征消除法（RFE）的集成特征选择方法 FS-RFE。Fisher Score 是一种比较经典的过滤式方法，它能通过最大化类间差异和最小化类内差异来评估特征的重要性，具有计算简单、效率高等优点。但是 Fisher Score 无法充分地考虑特征间的组合关系，因此不能有效地处理冗余特征。SVM-RFE 是一种经典的封装式方法，通过递归地移除权重最小的特征，从而获得与分类任务高度相关的特征子集，但缺点是计算复杂度较高，在高维数据集中具有较高的时间开销。本文结合两种算法 Fisher Score 和 SVM-RFE 的优势，提出了针对高维癌症基因数的集成特征选择算法 FS-RFE，以此来实现特征的简约。

1 基础理论

1.1 Fisher Score

Fisher Score 算法是一种经典的过滤式特征选择方法，其核心思想是计算特征类间方差和类内方差的比值，来为每个特征打分。具体而言，类间方差反映了数据的不同类别之间的差别，而类内方差反映了数据同一类别的内部差异。因此将他们进行比值的目的是找到一个能够让不同类别的数据间隔最大化，同时让同类数据之间的间隔最小化的特征子集。Fisher Score 值的大小直接反映了特征是否符合上述所提要求，最终计算的分值越高，说明特征越符合目标。因此 Fisher Score 通过计算每个特征的权重分数，然后按照分数进行降序排序，从而筛选出最优的特征子集 [9]。

假设初始基因表达矩阵为 $\boldsymbol{X}\in\mathbb{R}^{m\times n}$，其中包含 c 个不同的类别。第 i 个特征的 Fisher Score 计算公式为：

$$S_b(f_m)=\sum_{k=1}^{c} d_k(\mu_m^{(k)}-\mu_m)^2 \tag{1}$$

$$S_t{}^{(k)}(f_m)=\sum_{n=1}^{n_k}(x_{mn}^{(k)}-\mu_m{}^{(k)})^2 \tag{2}$$

$$\mathrm{FS}(f_n)=\frac{S_b(f_m)}{\sum_{k=1}^{c} S_t{}^{(k)}(f_m)} \tag{3}$$

式中：$S_b(f_m)$ 表示第 m 个特征在不同类别之间的散度；d_k 表示第 k 个类别中的样本数量；$\mu_m^{(k)}$表示第 k 个类别中第 m 个特征的基因表达均值；μ_m 表示第 k 个类别中第 m 个基因的整体表达均值；$S_t{}^{(k)}(f_m)$表示第 m 个特征在第 k 个类别内的散度；$x_{mn}^{(k)}$表示第 k 个类别中第 n 个样本的第 m 个基因表达值。

Fisher Score 算法的优点是具有较高的分类准确率和较低的计算复杂度。但缺点是未考虑特征之间的相关性，导致所选特征子集中可能存在冗余特征。为解决上述问题，本文将 Fisher Score 与 SVM-RFE 算法结合使用，通过前者进行特征初筛，然后利用后者进一步剔除不重要的特征。通过结合两种算法的优势，能够更有效地筛选出高质量的特征子集，同时也能极大减少冗余特征的影响，为后续的分类任务打好基础。

1.2 SVM-RFE

在支持向量机（support vector machine, SVM）的理论框架下，Guyon 等人 [10] 提出了一种基于特征权重评估的特征选择方法，称为 SVM-RFE。该方法的核心思想是通过计算特征权重系数的绝对值，逐步剔除对分类贡献较小的特征，以达到特征降维的目的。

SVM-RFE 的核心流程如下：

首先基于原始特征集合构建支持向量机模型（SVM），并根据特征权重系数对特征进行排序，之后移除权重系数最小的特征。在下一轮迭代中，基于剩余的特征子集重新训练支持向量机模型，并重复上述过程，直到所有特征都被移除。最后，根据特征被移除的顺序对所有特征进行相关性排序，最先移除的权重系数最小特征被认为对模型的影响较小，因此最后被移除的特征被认为是与类别标签最相关且排名最靠前的特征。

具体而言，SVM 的核心目标是通过构建最优超平面 $(w\cdot x+b)=0$ 作为决策边界，然后最大化分类间隔 $(2/\|\omega\|^2)$[11]。对于非线性或高维数据，SVM 利用核函数将数据映射到高维空间，使其在高维空间中线性可分。给定训练样本集 $\{(x_i,y_i),i=1,2,...,l\}$，$x_i\in\mathbb{R}^m$，$y_i\in\{-1,1\}$，在样本线性可分的情况下，目标函数需满足约束条件：

$$\begin{cases}\min J=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{k}\xi_i \\ s.t.\begin{cases}y_i(w\cdot x_i+b)\geqslant 1-\xi_i \\ \xi_i>0\end{cases} \quad ,i=1,2,\cdots,k\end{cases} \tag{4}$$

SVM 模型最终通过拉格朗日对偶定理和二次规划求得决策函数：

$$\begin{cases}f(x)=\mathrm{sgn}\,(w\cdot x+b) \\ w=\sum_{i=1}^{l}a_i y_i x_i\end{cases} \tag{5}$$

式中：权值 w 是拉格朗日乘子 α 不为零的支持向量的线性组合。特征权值越大，表明其包含的判别信息越丰富。

此外，当某个特征被移除时，其对目标函数的影响可通过泰勒展开近似计算：

$$\Delta J(i)=\frac{\partial J}{\partial \omega_i}\Delta\omega_i+\frac{\partial^2 J}{\partial \omega_i^2}(\Delta\omega_i)^2 \tag{6}$$

在目标函数 J 的最优解中，仅考虑二阶项时，近似计算为：

$$\Delta J(i)\approx(\Delta w_i)^2 \tag{7}$$

当第 i 个特征被移除时有 $\Delta\omega_i=\omega_i$，因此 SVM-RFE 以特征权值的平方 ω^2 作为排序准则，以此来衡量特征的重要性 [12]。

通过这种迭代式特征选择方法，SVM-RFE 能够有效筛

选出最具判别性的特征子集，从而提升模型的分类性能和泛化能力。

2 算法描述

2.1 算法核心思想

给定原始基因表达矩阵$\boldsymbol{M}=\{(X_i,C_i)\}_{i=1}^{n}$，假设每个样本有若干 m 个特征 $\{S_1,S_2,\cdots,S_m\}$，每个样本 C_k 且 $C_k\in\{0,1\}$。FS-RFE 算法的流程如图 1 所示。

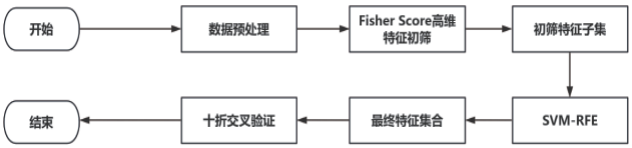

图 1 FS-RFE 算法流程

FS-RFE 算法由两阶段组成。首先在第一阶段中，计算特征在数据集上的类间方差s_B^k；接着计算特征在数据集上的类内方差s_ω^k；然后根据公式计算 Fisher Score，将计算出的特征权重进行降序排序，筛选初步的特征子集；最后在第二阶段中，使用 SVM-RFE 算法进行特征选择，得到最优特征子集。由此，提出一种针对高维小样本癌症数据集的集成特征选择算法 FS-RFE，算法详细伪代码描述如图 2 所示。

算法 FS-RFE 思想
输入：基因表达矩阵 $M=\left\{(X_i,C_i)\right\}_{i=1}^{n}$ 一共有 m 个特征：$\left\{S_1,S_2,...,S_m\right\}$
每个样本 X_i 具有一个类别标签 C_k 且 $C_k\in\left\{0,1\right\}$ 第一轮 Fisher Score 保留的特征个数 NumF
输出：特征子集 R

```
function Step 1 数据预处理 (Gene Matrix:M)
    应用 RMA 算法进行背景矫正，对数据进行归一化处理，转化为 log2 以降低运算复杂度
    return 预处理过的基因表达矩阵 M^1
end function
function Step 2 使用 Fisher Score 进行第一轮特征筛选 (Gene Matrix:M^1)
    for i = 0 → m − 1 do
        计算第 i 个特征在数据集上的类间方差 S_B^i
        计算第 i 个特征在数据集上的类内方差 S_ω^i
        计算第 n 个特征的 Fisher Score 为 S_n = S_B^i / S_ω^i
    end for
    根据 S_n 的值对基因矩阵 M 进行降序排序
    留下前 NumF 个特征组成特征子集 M^2 = M^1[0:NumF]
    return M^2
end function
function STEP3 使用递归特征消除筛选最终特征 (GeneMatrix:M^2)
    使用所有特征训练分类器 Classifier
    计算特征权重系数绝对值排序 F_rank
    while |R_cur|>NumF do
        //R_cur 为当前特征子集
        R_cur ← 根据 F_rank 移除权重系数绝对值最小的特征
        使用 R_cur 训练分类器 Classifier
        重新计算当前的特征权重系数绝对值排序 F_rank
    end while
    M^3 ← R_cur
    return M^3
end function
```

图 2 FS-RFE 算法流程的伪代码描述

2.2 时间复杂度分析

FS-RFE 算法的时间复杂度分析如下：

假设输入数据集的样本个数为 N，特征个数为 M，Step1 计算个特征在数据集上的类间方差和类内方差并求得第 n 个特征的 Fisher Score，其时间复杂度为 $O(M)$，Step2 中假设每次从特征列表中只删除一个特征，故 SVM-RFE 算法的时间复杂度为 $O(N\cdot M^2)$。由此可以得到 FS-RFE 算法总的时间复杂度为 $O(\text{FS-RFE})=O(M+N\cdot M^2)$[13]。

3 实验与分析

3.1 实验环境及数据集

本文的实验环境如下：操作系统为 Windows11，处理器为 Intel(R) Core(TM) Ultra 5，用于训练和测试的软件为 Python3.6，开发工具使用 PyCharm。

为验证 FS-RFE 的分类性能，将其应用于 9 个癌症微阵列基因表达数据集（GSE5564、GSE5764、GSE5788、GSE8671、GSE9844、GSE15471、GSE16088、GSE18842、GSE23400）。表 1 总结了这些关于不同癌症的公共基因表达数据集的属性。所有数据集均从名为基因表达综合数据库（GEO）的公共存储库中检索[14]。

表 1 癌症微阵列基因表达数据集的属性

编号	数据集	特征维数	癌症样本	正常样本
GSE5563	Vulva	20 862	9	10
GSE5764	Breast	20 862	10	20
GSE5788	Leukemia	12 645	6	8
GSE8671	Colon	20 862	32	32
GSE9844	Head–neck	20 862	26	12
GSE15471	Pancreas	20 862	39	39
GSE16088	Osteosarcoma	12 645	14	6
GSE18842	Lung	20 862	46	45
GSE23400	Esophageal	12 645	53	53

由于癌症数据集的样本数有限，为提高模型评估的准确性和可靠性，采用 K 折交叉验证（K-fold cross-validation）方法对数据进行划分，其中的 K 值可以根据数据集的维度进行设定。在此采取十折交叉验证法，将原始数据集均匀地划分为 10 个子集。在下一步的验证过程中，将依次从 10 个子集中选取其中的 1 个子集作为测试集，然后将剩下的 9 个子集作为训练集来训练模型和构建分类器。通过这种方式，能够充分利用数据集进行模型型验证，提高评估结果的稳定性和可靠性。十折交叉验证方法如图 3 所示。

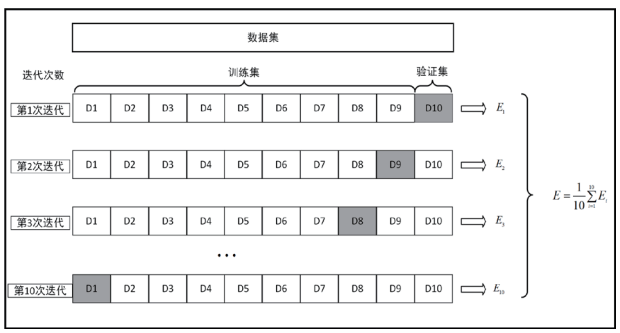

图 3 十折交叉验证示意图

3.2 数据的预处理

从公共数据库 GEO 中下载了原始基因芯片数据之后，首先要进行预处理步骤。分为背景校正、标准化和汇总，预处理后的基因表达矩阵用于后续的进一步分析。为了消除背景噪声，本文采用了稳健多阵列平均值（RMA）方法进行背景校正。与 MAS5 等其他方法相比，RMA 更能准确地反映出真实生物学的变化 [15]。通过图 4 所述步骤，获得了高质量的基因表达矩阵，为后续的特征选择奠定了可靠的基础。

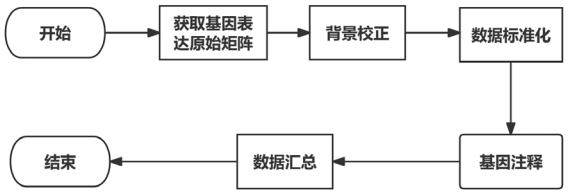

图 4 基因表达数据的预处理流程

3.3 评估标准

基因特征数据集通常在特征选择后使用分类器进行分类实验。本文的癌症基因数据集属于二分类数据集，使用如下指标作为评估标准；准确率（Accuracy）、特异性（SP）、敏感性（SN）和精准度（Precision）。

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{8}$$

$$F_1\text{-Score} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}} \tag{9}$$

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

式中：TP 指真正类，即样本为正类且被正确预测为正类的数量；TN 指真负类，即样本为负类且被正确预测为负类的数量；FP 指假正类，即样本为负类但被错误预测为正类的数量；FN 指假负类，即样本为正类但被错误预测为负类的数量。

3.4 实验与分析

在特征降维阶段，本实验基于上述癌症基因数据集，使用全部癌症数据集的样本进行测试。本次实验将 FS-RFE 与 4 种常用的特征选择方法（ANOVA、XGBoost、ReliefF 和 Random Forest）进行了比较。众所周知，在一种癌症中，只有几十个甚至几个基因与其生成和发展高度相关。在此前提下，我们最初将基因子集的大小限制在 100 以下，因此在癌症基因数据集中，最终筛选的基因子集大小设置为 60 以达到最佳分类性能 [16]。

表 2 FS-RFE 与其他传统特征选择算法性能对比

编号	FS-RFE（Ours）		ANOVA		XGBoost		ReliefF		RandomForest	
	ACC	F_1-score	ACC	F_1-score	ACC	F_1-score	ACC	F_1-score	ACC	F_1-score
GSE5563	100	100	100	100	94.74	94.11	100	100	100	100
GSE5764	96.67	94.74	93.33	88.89	73.33	33.33	93.33	88.89	90	82.35
GSE5788	100	100	100	100	71.43	100	100	100	57.14	100
GSE8671	100	100	100	100	100	100	100	100	100	100
GSE9844	100	100	100	100	76.32	83.02	97.37	98.04	92.11	94.12
GSE15471	91.03	91.57	92.31	92.86	89.74	90	88.46	88.89	91.03	91.76
GSE16088	100	100	100	100	100	100	100	100	100	100
GSE18842	100	100	100	100	100	100	100	100	100	100
GSE23400	94.34	95.33	93.4	93.58	95.28	95.24	93.4	93.33	94.34	94.44
Mean	**98**	**97.96**	97.67	97.26	88.98	88.41	96.95	96.57	91.62	95.85

结果如表 2 所示，其中加粗表示取得最佳性能，下划线表示次优性能。从 ACC 指标上来看，FS-RFE 在 7 个数据集中取得了最高的性能，在 GSE15471 数据集中也取得了较优的性能，在 9 个癌症数据集上的准确度平均值表现最佳。从 F_1-score 指标上来看，FS-RFE 在 6 个数据集中表现出了最佳性能，在 3 个数据集上表现出了次优性能，在 9 个癌症数据集上的 F_1-score 平均值表现最佳。

从不同数据集取得的分类结果上来看，FS-RFE 在大部分数据集中都取得了最好的分类性能，尤其是在如 GSE5764（乳腺癌）、GSE5788（白血病）和 GSE9844（口腔癌）质量较低的数据集中，FS-RFE 表现了较为优异的分类性能，而传统特征选择方法在部分数据集中表现欠佳，例如在 GSE5764 数据集中，XGBoost 的 F_1-score 仅为 33.33%，在 GSE5788 数据集中，Random Forest 的 ACC 仅为 57.14%，可以推测出传统的数据选择方法在个别癌症数据集中无法建立出稳定的分类模型，可能与数据及标签不平衡以及样本高噪声等因素有关。因此可以得出结论，相较于传统的特征选择方法，FS-RFE 采用了以集成学习为核心的多阶段特征选择方法具有较好稳定性，对于一些质量欠佳的同样有良好的表现，这体现出了 FS-RFE 在执行癌症基因筛选任务中具有较好的泛化能力和鲁棒性。从图 5 和图 6 中可以看到更直观的实验结果。

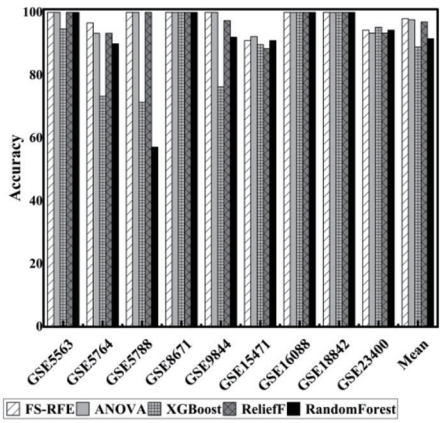

图 5 FS-RFE 算法和其他 4 种特征选择算法准确率对比

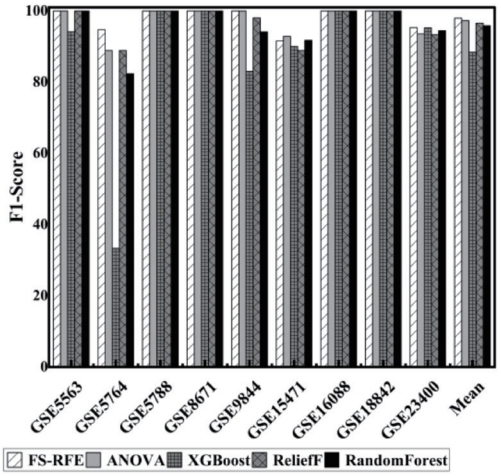

图 6 FS-RFE 算法和其他 4 种特征选择算法 F_1-score 指标对比

4 结论

本文基于集成学习提出了一种针对癌症基因筛选的两阶段特征选择方法 FS-RFE，该方法的核心思想是，首先计算特征的 Fisher Score，然后计算权重系数从而得出各个特征的重要性排序，筛选出初步的特征子集，最后使用 RFE 算法对特征子集的基因进行重要性排序，从而得到最终特征子集。经过 9 个癌症数据集上测试，使用分类的准确率和 F_1-score 指进行对比，结果显示 FS-RFE 算法具有较好的分类效果，特别是在一些低质量数据集中，与传统的特征选择算法表现相比，FS-RFE 算法具有更好的性能，但是与其他算法对比时，本文的算法表现并不是在所有集合中均表现最优，因此在未来的研究工作中，将结合最新的集成学习研究成果，进一步优化针对高维数据集的特征选择方法。

参考文献：

[1] 翟峻仁 , 殷柯欣 , 谢爱锋 , 等 . 互信息集成学习特征选择 [J]. 长春工业大学学报 , 2023, 44(2): 183-188.

[2] 程凤伟 , 王文剑 , 张珍珍 . 面向高维小样本数据的层次子空间 ReliefF 特征选择算法 [J]. 南京大学学报 (自然科学), 2023, 59(6): 928-936.

[3] KHAIRE U M, DHANALAKSHMI R. Stability of feature selection algorithm: a review[J]. Journal of king saud university - computer and information sciences, 2022,34(4):1060-1073.

[4] 孟圣洁 , 于万钧 , 陈颖 . 最大相关和最大差异的高维数据特征选择算法 [J]. 计算机应用 ,2024,44(3):767-771.

[5] 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.

[6] YU L, LIU H. Feature selection for high-dimensional data: a fast correlation-based filter solution[C]//ICML'03: Proceedings of the Twentieth International Conference on International Conference on Machine Learning.Palo Alto: AAAI Press, 2003: 856-863.

[7] TIBSHIRANI R. Regression shrinkage and selection via the Lasso[J]. Journal of the royal statistical society: series b (methodological), 1996,58(1), 267-288.

[8] 周钢 , 郭福亮 . 基于特征选择的高维数据集成学习方法研究 [J]. 计算机科学 , 2021,48(S1):250-254.

[9] 孙林 , 张起峰 , 徐久成 . 基于互信息的 Fisher Score 多标记特征选择 [J]. 南京大学学报 (自然科学),2023,59(1):55-66.

[10] GUYON I, WESTON J, BARNHILL S, et al. Genese-lection for cancer classification using support vector machines[J]. Machine learning, 2002, 46: 389-422.

[11] 刘嘉欣 , 王宏伟 , 王佳 . 基于 Lasso-RFE 的乳腺癌预后仿真 [J]. 计算机仿真 , 2022, 39(12): 330-335.

[12] 林俊 , 许露 , 刘龙 . 基于 SVM-RFE-BPSO 算法的特征选择方法 [J]. 小型微型计算机系统 , 2015, 36(8):1865-1868.

[13] 穆晓霞 , 郑李婧 . 基于 F-score 和二进制灰狼优化的肿瘤基因选择方法 [J]. 南京师大学报 (自然科学版), 2024, 47(1): 111-120.

[14] GE C Y, LUO L Q, ZHANG J L, et al. FRL: an integrative feature selection algorithm based on the fisher score, recursive feature elimination, and logistic regression to identify potential genomic biomarkers[J]. Biomed research international, 2021: 4312850.

[15] IRIZARRY R A, HOBBS B, COLLIN F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data[J]. Biostatistics, 2003, 4(2): 249-264.

[16] ZHANG J L, XU D, HAO K J, et al. FS-GBDT: identification multicancer-risk module via a feature selection algorithm by integrating Fisher score and GBDT[J]. Briefings in bioinformatics, 2021, 22(3): 189.

【作者简介】

盛煜轩（1996—），男，硕士研究生，助教，研究方向：数据挖掘，email: nero0802@163.com。

严嘉浩（1993—），男，硕士研究生，讲师，研究方向：计算机技术，email: 328521516@qq.com 。

苏志军（1979—），男，本科，副教授，研究方向：数据分析，email: 35372963@qq.com。

（收稿日期：2025-01-26 修回日期：2025-06-05）