Advances in Clinical Medicine 临床医学进展, 2026, 16(3), 3593-3607
Published Online March 2026 in Hans. https://www.hanspub.org/journal/acm
https://doi.org/10.12677/acm.2026.1631167

Hans汉斯

面向结直肠癌的几何感知结构稀疏特征选择(AW-MSGL)：以更少可解释基因实现稳定分类

韩君亚

河北工业大学理学院，天津

收稿日期：2026年2月18日；录用日期：2026年3月11日；发布日期：2026年3月20日

摘　要

结直肠癌(CRC)的早筛与分层诊断面临“样本少、维度高、冗余强”的基因表达数据挑战。本文在不改变流形稀疏组LASSO (MSGL)核心思想的前提下，提出自适应组权重的MSGL (AW-MSGL)，以更少且可解释的基因子集实现稳定分类。方法引入数据驱动的组权重 w_j 抑制共表达冗余；预处理采用F-score + KMeans自动构建模块；优化沿用加速近端梯度(APG)。在CRC微阵列数据(含独立测试集)上，AW-MSGL以显著更少的基因达到可比或更优的准确率，并在关键基因的生物学解释上保持一致性。该框架有望为CRC的轻量部署与可解释生物标志物发现提供数据驱动的工具。

关键词

特征选择，稀疏学习，流形正则化，组稀疏，自适应权重，基因表达，癌症分类

Geometry-Aware Structured Sparse Feature Selection for Colorectal Cancer (AW-MSGL): Stable Classification with Fewer Interpretable Genes

Junya Han

School of Science, Hebei University of Technology, Tianjin

Received: February 18, 2026; accepted: March 11, 2026; published: March 20, 2026

文章引用: 韩君亚. 面向结直肠癌的几何感知结构稀疏特征选择(AW-MSGL): 以更少可解释基因实现稳定分类[J]. 临床医学进展, 2026, 16(3): 3593-3607. DOI: 10.12677/acm.2026.1631167

Abstract

Early screening and stratified diagnosis of colorectal cancer (CRC) face challenges from gene expression data characterized by "small sample size, high dimensionality, and strong redundancy". Without altering the core philosophy of Manifold Sparse Group LASSO (MSGL), this paper proposes Adaptive Weighted MSGL (AW-MSGL) to achieve stable classification with a smaller, interpretable subset of genes. The method introduces data-driven group weights (w_j) to suppress co-expression redundancy; preprocessing employs F-score combined with KMeans to automatically construct modules; and optimization utilizes Accelerated Proximal Gradient (APG). On CRC microarray datasets (including an independent test set), AW-MSGL achieves comparable or superior accuracy with significantly fewer genes while maintaining consistency in the biological interpretation of key genes. This framework offers a data-driven tool for lightweight deployment and the discovery of interpretable biomarkers in CRC.

Keywords

Feature Selection, Sparse Learning, Manifold Regularization, Group Sparsity, Adaptive Weights, Gene Expression, Cancer Classification

1. 介绍

稀疏学习旨在剔除不重要特征、保留关键信息，特别适用于高维数据。将特征选择表述为带正则的优化问题：

$$\hat{\beta} = \arg\min_{\beta} \{L(y,\beta) + R(\lambda,\beta)\}, \tag{1}$$

其中，$L(y,\beta)$为损失项，$R(\lambda,\beta)$为惩罚(正则化)项；系数向量$\beta \in \mathbb{R}^p$描述模型。依据估计的系数向量$\hat{\beta}$进行选择：$\hat{\beta}$中非零项对应被保留的特征，其个数即为所选特征数。正则参数λ决定损失与惩罚的平衡；部分模型包含多个正则参数以协调不同惩罚。通过引入惩罚项，上式可有效抑制过拟合、提升泛化能力。

已有大量稀疏学习模型被提出，可大致分为“个体稀疏”和“组稀疏”两类。个体稀疏的特征选择模型又可细分为线性与非线性：线性模型的解路径呈分段线性，步进方向与跳跃幅度可闭式计算；非线性模型的解路径为曲线，需要迭代计算和更新方向，并判定每一段的端点，通常需多次遍历数据，因而较线性模型更慢。

考虑线性回归模型：

$$y = X\beta + \varepsilon, \tag{2}$$

其中，$\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n) \sim N(0, \sigma^2 I_n)$为误差向量，各分量独立同分布，均值为0、方差为$\sigma^2$。其预测响应为：

$$\hat{y} = X\hat{\beta} = \sum_{j=1}^{p} \hat{\beta}_j x^{(j)}, \tag{3}$$

其中，$\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)^{\mathrm{T}}$为由式(1)得到的估计系数向量。

在个体稀疏模型中，影响力最大的之一是 LASSO，由 Tibshirani [1]提出，采用L_1正则：

$$\hat{\beta}(\text{LASSO}) = \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\right\}, \tag{4}$$

其中 $\frac{1}{2}\|y - X\beta\|_2^2$ 为损失项， $\lambda\|\beta\|_1$ 为惩罚项；正则参数 $\lambda \geq 0$ 控制解 $\hat{\beta}$ 的稀疏性。绝对值系数和满足 $\sum_{j=1}^{p}|\beta_j| \leq t$ （$t \geq 0$ 为调节参数)。得益于正则化框架，LASSO 已成为稀疏回归的标准工具；当 $\lambda = 0$ 时退化为经典最小二乘。Bühlmann 与 van de Geer [2]讨论了 LASSO 在高维问题中的应用与理论性质；Liu 等[3]回顾了线性回归的正则化稀疏模型。

尽管 LASSO 应用广泛，其特征选择一致性仅在特定条件下成立[4]-[6]。为缓解该问题，Fan 与 Li [7]提出了 Smoothly Clipped Absolute Deviation (SCAD)罚函数，可降低估计偏差并产生连续解。其形式为：

$$\sum_{j=1}^{p} P_\lambda(\beta_j), \tag{5}$$

其中 $P_\lambda(\beta_j)$ 定义为：

$$P_\lambda(\beta_j) = \begin{cases} \lambda|\beta_j|, & \text{if } |\beta_j| \leq \lambda, \\ \dfrac{-\beta_j^2 + 2a\lambda|\beta_j| - \lambda^2}{2(a-1)}, & \text{if } \lambda < |\beta_j| \leq a\lambda, \\ \dfrac{(a+1)\lambda^2}{2}, & \text{if } |\beta_j| \geq a\lambda, \end{cases} \tag{6}$$

其中 $a > 2$ 且 $\lambda \geq 0$ 。该罚函数为在 λ 与 $a\lambda$ 处带节点的二次样条。基于贝叶斯论证与数值模拟，常推荐 $a = 3.7$ 。SCAD 在 $(-\infty,0)\cup(0,\infty)$ 上连续可导，但在 0 处奇异；其导数在区间 $[-a\lambda, a\lambda]$ 之外为 0。小系数被收缩到 0，而大系数得以保留，因而对大系数给出稀疏、连续且近乎无偏的估计。Bühlmann 与 Meier [8]提出了多步局部线性近似以增强稀疏性，并证明 SCAD 在高维场景下具备“oracle”性质。

另一项改进是 Adaptive LASSO (自适应 LASSO)，由 Zou [9]提出以实现一致的变量选择。其通过引入权重修正 LASSO 惩罚：

$$\lambda\sum_{j=1}^{p} w_j|\beta_j|, \tag{7}$$

其中 $w_j = \left|\hat{\beta}_j^{\text{ols}}\right|^{-\gamma}$ ，$\gamma > 0$ ，$\hat{\beta}_j^{\text{ols}}$ 为普通最小二乘估计。通过对较小系数施加更大惩罚，Adaptive LASSO 能自适应地选择特征。Lin 等[10]证明其在一定假设下具备 oracle 性质，Yuan 与 Lin [11]进一步证明了其一致性与分段线性解路径。

Zhang [12]提出了 Minimax Concave Penalty (MCP)，定义为：

$$\sum_{j=1}^{p} \phi_\lambda(\beta_j), \tag{8}$$

其中

$$\phi_\lambda(\beta_j) = \begin{cases} \lambda|\beta_j| - \dfrac{\beta_j^2}{2a}, & \text{当} |\beta_j| \leq a\lambda, \\ \dfrac{a\lambda^2}{2}, & \text{当} |\beta_j| > a\lambda, \end{cases} \tag{9}$$

其中 $a > 1$ 且 $\lambda \geq 0$ 。经验结果表明，MCP 常常优于 LASSO 与 SCAD [13]。SCAD 与 MCP 均利用凹或非凸惩罚以剔除不重要特征、保留重要特征，从而实现接近“oracle”的行为。

尽管线性模型计算高效，真实数据常呈现非线性关系，需用非线性模型刻画复杂结构。例如，L_1 正则化的逻辑回归与带 $L_{1/2}$ 罚的稀疏逻辑回归更契合此类模式。具体而言，逻辑回归刻画分类响应 Y 的后验概率：

$$\log\left(\frac{P(Y=k\mid x_i)}{P(Y=K\mid x_i)}\right)=\beta_0^{(k)}+x_i^{\mathrm{T}}\beta^{(k)},\ k=1,\cdots,K-1, \tag{10}$$

其类别概率为：

$$P(Y=k\mid x_i)=\frac{e^{\beta_0^{(k)}+x_i^{\mathrm{T}}\beta^{(k)}}}{1+\sum_{l=1}^{K}e^{\beta_0^{(l)}+x_i^{\mathrm{T}}\beta^{(l)}}},\ k=1,\cdots,K-1, \tag{11}$$

且 $P(Y=K\mid x_i)=1-\sum_{k=1}^{K-1}P(Y=k\mid x_i)$。Tibshirani 将 LASSO 推广至该设定，得到 L_1 正则的逻辑回归(LLR)，其对噪声具有鲁棒性并广泛适用。但在高维下，LLR 计算成本较高。Krishnapuram 等[14]提出多分类改进方法，Tian 等[15]引入二次下界策略。

Liang 等[16]针对癌症分类中的基因筛选，提出了带 $L_{1/2}$ 罚的稀疏逻辑回归(SLR-$L_{1/2}$)。其损失函数为：

$$-\sum_{i=1}^{n}\left\{y_i\log\left(f\left(x_i^{\mathrm{T}}\beta\right)\right)+(1-y_i)\log\left(1-f\left(x_i^{\mathrm{T}}\beta\right)\right)\right\}, \tag{12}$$

并配以罚项：

$$\lambda\sum_{j=1}^{p}\left|\beta_j\right|^{1/2}, \tag{13}$$

其中 $f(\tau)=e^{\tau}/(1+e^{\tau})$。对于 $0<q<1$ 的 L_q 罚，q 越小通常解越稀疏。Xu 等[17]表明，$L_{1/2}$ 在稀疏性与收敛性之间取得了更优平衡，相比 L_1 更稀疏、相比 L_0 更易求解，同时具备无偏性、稀疏性与接近“oracle”等良好性质。

除个体稀疏外，许多应用场景中特征天然呈现组结构，例如共表达基因构成功能通路。然而传统稀疏模型往往将特征视为相互独立。为此，引入具组效应的模型。Elastic Net 由 Zou 与 Hastie [18]提出，将 L_1 与平方 L_2 结合：

$$\lambda_2\|\beta\|_2^2+\lambda_1\|\beta\|_1, \tag{14}$$

其简化形式为：

$$\lambda\left((1-\alpha)\|\beta\|_2^2+\alpha\|\beta\|_1\right). \tag{15}$$

随后构造了自适应版本[19]：

$$\hat{\beta}(\mathrm{aen})=\left(1+\frac{\lambda_2}{n}\right)\arg\min_{\beta\in\mathbb{R}^p}\left\{L(y,\beta)+R(\lambda,\beta)\right\}, \tag{16}$$

其权重为 $\hat{w}_j=\left|\hat{\beta}_j(\mathrm{en})\right|^{-\gamma}$，可保证选择一致性与渐近正态性。

另一种方法 Fused LASSO [20]同时鼓励系数及其相邻差分的稀疏性：

$$\lambda_1\|\beta\|_1+\lambda_2\sum_{j=2}^{p}\left|\beta_j-\beta_{j-1}\right|. \tag{17}$$

针对结构化的组选择，Yuan 与 Lin [21]提出了 Group LASSO：

$$\min_{\beta\in\mathbb{R}^p}\left(\left\|y-\sum_{l=1}^{L}X_l\beta_l\right\|^2+\lambda\sum_{l=1}^{L}\sqrt{p_l}\left\|\beta_l\right\|\right), \tag{18}$$

该方法将 p 个变量划分为 L 个互不重叠的组，并对各组施加 L_2 范数惩罚，使整组可被同时剔除。但其缺乏组内稀疏性。为此，Friedman 等[22]提出了 Sparse Group LASSO (SGL)：

$$\min_{\beta\in\mathbb{R}^p}\left(\left\|y-\sum_{l=1}^{L}X_l\beta_l\right\|^2+\lambda_1\sum_{l=1}^{L}\left\|\beta_l\right\|+\lambda_2\left\|\beta\right\|_1\right), \tag{19}$$

其兼顾组层级与个体层级的稀疏性，实现双层选择。SGL 已广泛用于基因表达分析中的疾病预测。

然而，上述方法多数仅在特征空间中刻画线性关系，忽视了样本之间的几何结构与局部流形信息。实际上，高维数据(如基因表达、影像组学、用户行为)常嵌于低维流形之上；忽略该结构会扭曲语义相似性并削弱判别力。故越来越多的学者聚焦于流形正则化以致力于保存样本之间的集合结构。流形正则化已从分类扩展至嵌入、聚类、矩阵分解与生成建模，理论也从拉普拉斯嵌入发展到图扩散与谱学习。近期，Ma 等[23]提出一种流形正则模型以识别对阿尔茨海默病预测至关重要的血浆蛋白；通过保持邻域结构[24]，其准确率达到 97.5%。这表明几何学习与稀疏建模的协同效应日益增强。所提出的 MSGL 因而可视为该趋势的自然延伸：将流形正则与结构化稀疏相融合，以同时提升表征能力与生物学可解释性。

2. 实验设计与方法

2.1. 数据集说明

本研究仅使用 2 个公开的结直肠癌(CRC)高维微阵列数据集，均来源于 Gene Expression Omnibus (GEO)数据库，并设置训练与测试集合。各数据集的详细信息见表 1。

Table 1. Overview of cancer datasets
表 1. 癌症数据集概览

数据集(Dataset)	类型(Type)	样本数(Samples)	基因数(Genes)	类别(Classes)	正常/肿瘤 (Normal/Tumor)
GSE9348	结直肠(训练)	82	23521	肿瘤/正常	70/12
GSE8671	结直肠(测试)	64	23521	正常/肿瘤	32/32

如表 1 所示：

- 结直肠癌数据集由 GSE9348 (训练集)与 GSE8671 (测试集)构成。GSE9348 含 82 个样本(肿瘤 70、正常 12)，均使用 Affymetrix 平台检测 23,521 个基因；GSE8671 作为测试集，含 64 个样本，正常与肿瘤各 32 例，类别平衡性良好。

所用训练集均采用肿瘤与正常组织的配对设计，有效减少个体差异干扰并提升差异表达分析与模型训练的信噪比。数据基于 Affymetrix U133 Plus 2.0 芯片平台，确保 CRC 数据的可比性与稳定性。此外，设置独立的外部验证集以检验候选基因在新样本中的稳定性与一致性。

2.2. 基于 F-Score 的特征初选

高通量基因表达数据通常具有“维度极高、样本有限”的双重特征(动辄数万基因、样本数却不多)，这会放大过拟合风险并拖累计算效率。为此，我们采用两阶段的预处理流程：第一步基于监督的特征打分(F-score)筛选出判别性强的候选基因；第二步通过无监督聚类(KMeans)降低冗余、提升特征多样性，从

而为后续建模提供更加稳定、紧凑的输入。

第一步我们采用 F-score (单变量统计指标)对基因进行判别力排序(例如肿瘤 vs.正常)。对于第 i 个基因，F-score 定义为：

$$F(i)=\frac{\left(\mu_i^{(1)}-\mu_i\right)^2+\left(\mu_i^{(0)}-\mu_i\right)^2}{\sigma_i^{(1)^2}+\sigma_i^{(0)^2}}, \tag{20}$$

其中，$\mu_i^{(1)}$ 与 $\mu_i^{(0)}$ 分别表示基因 i 在癌症样本与正常样本中的均值，μ_i 为总体均值，$\sigma_i^{(1)^2}$ 、$\sigma_i^{(0)^2}$ 为类内方差。$F(i)$ 越大说明类间可分性越强、对分类任务越有价值。

图 1 展示了在结直肠癌数据集 GSE9348 上按 F-score 排序的前 20 个基因。其中前两名 *CA2* 与 *CLCA4* 得分显著偏高(分别为 1632.86 与 1243.85)，具有潜在的诊断或预后指征价值。

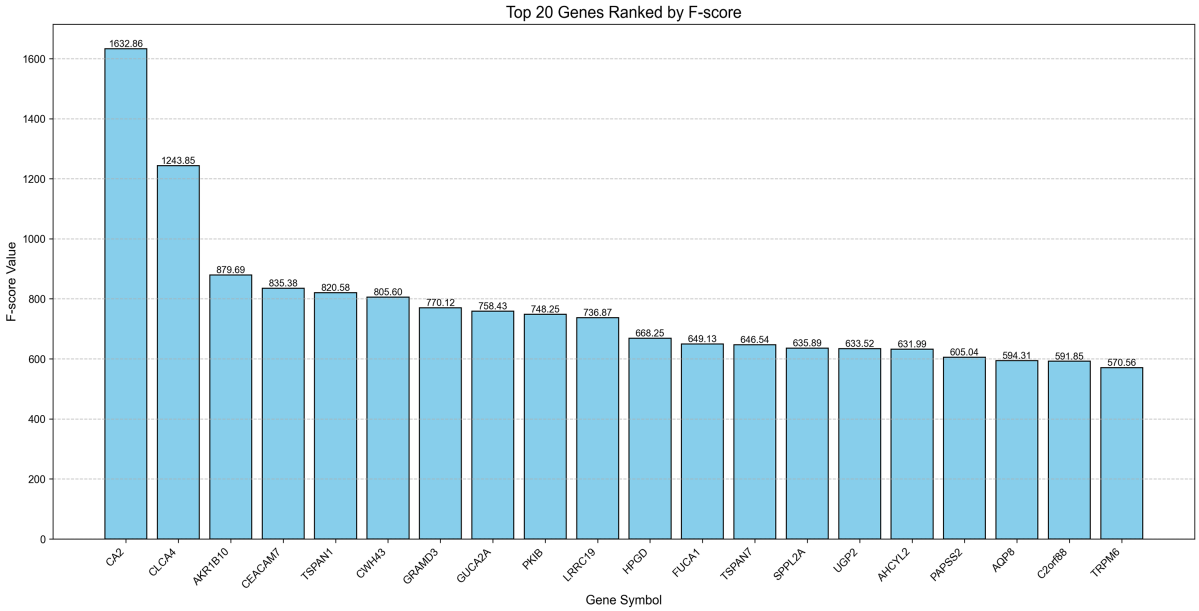

Figure 1. Bar chart of F-score gene ranking
图 1. F-score 基因排序柱状图

我们保留排名前 m 个基因(例如 $m=500$ 或 1000)用于后续分析。该初选步骤在显著降维的同时尽量保留生物学相关且具有类别判别性的特征，从而提升信噪比与计算效率。

2.3. KMeans 聚类促进特征多样性

尽管 F-score 能够有效识别信息量较高的基因，但它并未显式处理高相关基因间的冗余。为促进特征多样性、避免选择过多共表达基因，我们对排名靠前的候选基因执行 KMeans 聚类。

值得注意的是，为确保评估的泛化能力并避免数据泄露，本研究严格执行嵌套交叉验证策略。F-score 筛选和 KMeans 聚类步骤被完全封装在 10 折交叉验证的训练折(Training Fold)内部。即在每一次迭代中，仅利用 9 份训练数据计算 F-score 分值并构建 KMeans 聚类模型，确定特征子集和簇结构；随后将该映射规则应用于剩余 1 份测试数据进行验证。尽管这增加了计算量且每次迭代的组结构可能不同，但这保证了测试数据未参与任何特征选择或聚类过程。

设 $X\in\mathbb{R}^{N\times m}$ 为前 m 个基因的表达矩阵，其中每一列对应某个基因在 N 个样本上的表达向量。KMeans

通过最小化簇内平方和将这些基因划分为 K 个簇 $C_1, C_2, \cdots, C_K$：

$$\min_{C_1,\cdots,C_K} \sum_{k=1}^{K} \sum_{x_j \in C_k} \left\| x_j - \mu_k \right\|^2, \tag{21}$$

其中 $x_j \in \mathbb{R}^N$ 为第 j 个基因的表达向量，μ_k 为第 k 个簇的质心。簇数 K 可依据经验或内部验证指标(如轮廓系数)选取。为提高稳定性，我们采用多次随机初始化并选择目标值最低的结果。

聚类后，选取每个簇质心最近的基因作为代表：

$$g_k^* = \arg\min_{g \in C_k} \left\| x_g - \mu_k \right\|. \tag{22}$$

该策略在低冗余的前提下，尽可能覆盖多样的共表达模式。

图 2 给出了对前 500 个 F-score 基因进行聚类($K = 154$)的热图示例。行表示基因，列表示样本，颜色表示表达水平(红：高，蓝：低)。竖线将肿瘤样本($n = 70$)与正常样本($n = 12$)分隔开。可以观察到清晰的表达差异——正常样本更为集中地出现红色区域，提示不同表型的表达模式差异显著，说明聚类较好地捕捉到了生物学上有意义的结构。

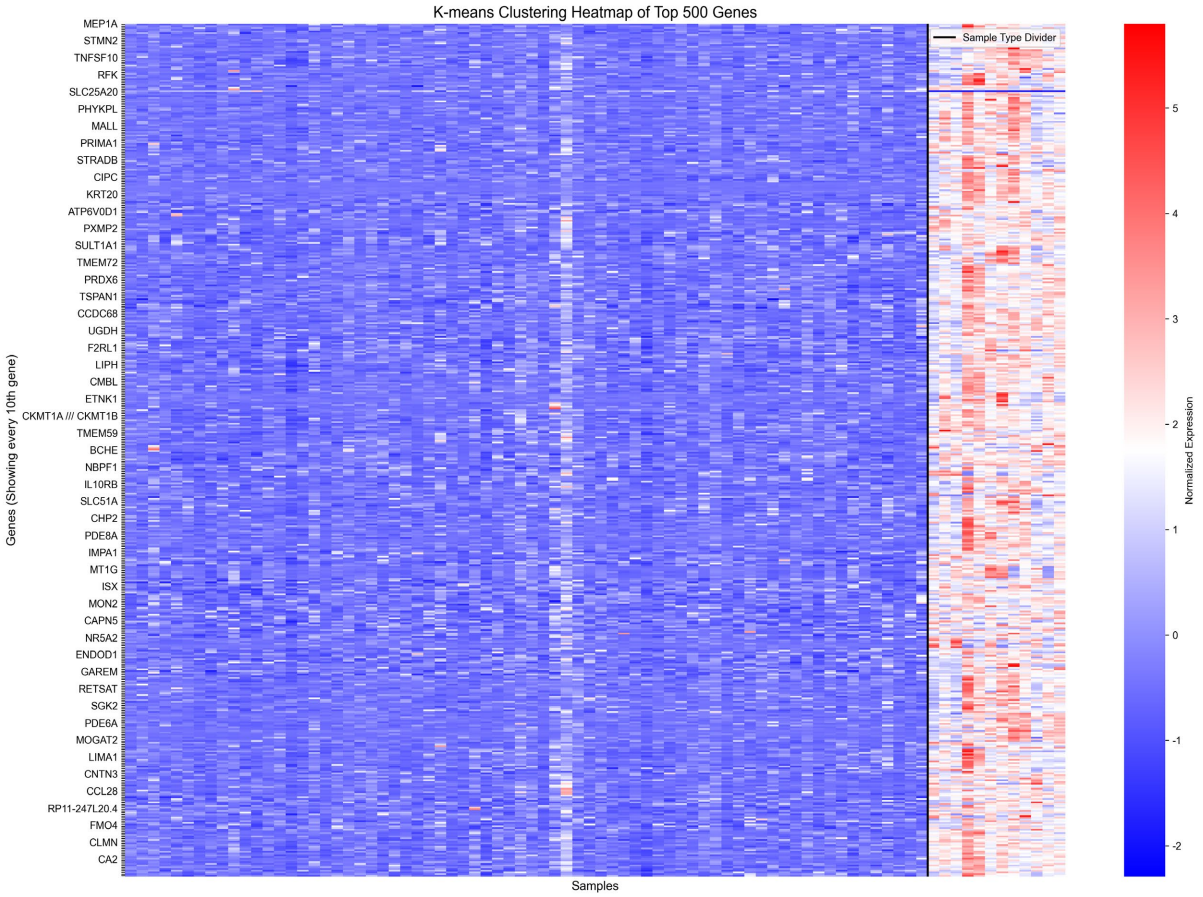

Figure 2. Gene expression heatmap of KMeans clustering
图 2. KMeans 聚类的基因表达热图

将 F-score 排序与 KMeans 聚类结合的预处理框架，在“判别性”与“多样性”之间取得了良好平衡。该降维后的基因集合作为 MSGL 的精炼输入，既提高了计算效率，也增强了生物学可解释性。

2.4. 流形稀疏组 LASSO (MSGL)模型

为有效从高维表达数据中识别具有判别力的基因，我们采用逐步融入结构先验的稀疏学习框架。方法设计遵循分层正则化策略：从基本稀疏性出发，逐步引入“组结构”与“几何/流形”约束。

(1) LASSO

在计算生物学场景中，从高维组学数据中定位与癌症相关的关键基因，本质上是一个特征选择问题。在众多稀疏建模方法中，LASSO (最小绝对收缩与选择算子)因兼具“变量选择 + 正则化”的特性而被广泛采用，其优化形式为：

$$\hat{w} = \arg\min_{w} \left\{ \|Y - Xw\|_2^2 + \lambda \|w\|_1 \right\}, \tag{23}$$

其中，$X \in \mathbb{R}^{M \times d}$ 为基因表达矩阵，行对应样本(如肿瘤或正常组织)，列对应具体基因的表达；响应向量 $Y \in \mathbb{R}^{M}$ 编码表型标签(如癌症亚型或生存状态)；参数向量 $w \in \mathbb{R}^{d}$ 为待学习的系数。正则参数 λ 控制稀疏程度：越大则非零系数越少，从而选择更精炼且更具预测性的基因子集。

需要强调的是，LASSO 倾向于“逐基因”选择，忽略了基因在致癌过程中常以通路或复合体的协同模块形式发挥作用的事实。这一局限促使我们在特征选择中引入“生物学组结构”。

(2) Group LASSO

为刻画基因之间的功能依赖关系，我们依据 KEGG、Reactome 或 GO 注释等先验知识，将基因组织为具有生物学意义的互斥组。考虑到致癌往往由协同功能的基因集群驱动，我们采用 Group LASSO 框架，在“组层面”施加稀疏约束。其目标函数为：

$$\hat{w} = \arg\min_{w} \left\{ \|Y - Xw\|_2^2 + \lambda_1 \sum_{j=1}^{N} \left\| w_{G_j} \right\|_2 \right\} \tag{24}$$

其中 G_j 表示第 j 个基因组，$\left\| w_{G_j} \right\|_2$ 为该组系数子向量的 ℓ_2-范数。通过惩罚组范数之和，上式鼓励整组同时被选择或剔除，体现了基因在生物系统中的“成组协作”特性。

需要注意的是，Group LASSO 不具备“组内稀疏”能力——一旦某组被激活，该组内所有基因往往被一并保留，即便其中只有少数真正有效，这可能导致过拟合并降低可解释性。

(3) Sparse Group LASSO

为同时实现“组间选择”与“组内选择”，我们采用 Sparse Group LASSO (SGL)，将组层面的 $\ell_{1,2}$ 罚与逐元素的 ℓ_1 罚相结合。其目标函数为：

$$\hat{w} = \arg\min_{w} \left\{ \|Y - Xw\|_2^2 + \lambda_1 \sum_{j=1}^{N} \left\| w_{G_j} \right\|_2 + \lambda_2 \|w\|_1 \right\} \tag{25}$$

该“混合正则化”使模型能够：

- 在组层面仅选择少量相关的功能模块(组间稀疏)；
- 在已选模块内进一步剔除冗余或判别力不足的基因(组内稀疏)。

这种双层选择机制契合生物系统的层级结构，有助于在癌症基因发现任务中同时提升预测性能与生物学合理性。

(4) 流形稀疏组 LASSO (MSGL)

传统的 SGL 方法通过线性映射 $f(X) = X\omega$ 将高维特征 $X \in \mathbb{R}^{N \times p}$ 映射到响应变量，并结合 ℓ_1 与组 ℓ_2 正则来进行特征选择。然而，此类方法仅关注“特征–标签”关系，忽略了输入空间中样本间固有的几何结构；结果可能使结构上相似的样本在投影空间中被拉得很远，进而影响模型的泛化与稳健性。

为克服上述不足，我们提出 Manifold Sparse Group LASSO (MSGL)，在 SGL 框架中引入流形正则以保留数据的局部几何结构。具体而言，MSGL 在样本层面构建 k -近邻(k -NN)图，并在邻近样本的模型输出上施加平滑性约束，从而保证相似样本获得相似预测。

考虑到基因表达数据通常缺乏预先定义的生物学分组，我们基于特征相似性提出“数据驱动”的分组策略。首先计算全部 p 个基因之间的余弦相似度矩阵 $S_f \in \mathbb{R}^{p\times p}$ 。对每个基因，仅保留其 k 个最相似的邻居以形成稀疏的 k -NN 图，从而得到稀疏邻接矩阵 A_f 。特征层面的图拉普拉斯算子定义为 $L_f = D_f - A_f$，其中 D_f 为对角度矩阵。该图既用于指导 Group LASSO 的分组结构，也提升了所选特征的生物学一致性。

此外，我们在样本层面构建图以刻画局部流形结构。样本 i 与 j 之间的相似度使用高斯核定义：

$$s_{ij} = \begin{cases} \exp\left(-\dfrac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right), & \text{if } x_j \in \mathcal{N}_k\left(x_i\right) \text{ or } x_i \in \mathcal{N}_k\left(x_j\right), \\ 0, & \text{otherwise}, \end{cases} \tag{26}$$

其中， $\mathcal{N}_k\left(x_i\right)$ 表示样本 i 的 k 个近邻， σ 为核带宽。样本层面的图拉普拉斯算子为 $L_s = D_s - A_s$ ，其中 $A_s = \left[s_{ij}\right]$， $D_s = \operatorname{diag}\left(\sum_j s_{ij}\right)$ 。

相应的流形正则项可写为：

$$\sum_{i,j} s_{ij}\left(\omega^{\mathrm{T}} x_i - \omega^{\mathrm{T}} x_j\right)^2 = 2\omega^{\mathrm{T}} X^{\mathrm{T}} L_s X \omega, \tag{27}$$

该项抑制相似样本之间的输出差异过大，从而维持局部结构一致性。

综合上述约束，MSGL 的整体目标函数为：

$$\min_{\omega} \frac{1}{2}\left\|Y - X\omega\right\|_2^2 + \lambda_1 \sum_{j=1}^{g} \left\|\omega_{G_j}\right\|_2 + \lambda_2 \left\|\omega\right\|_1 + \lambda_3 \omega^{\mathrm{T}} X^{\mathrm{T}} L_s X \omega, \tag{28}$$

其中， g 为通过特征聚类得到的分组数； ω_{G_j} 表示组 G_j 的系数子向量； $\lambda_1, \lambda_2, \lambda_3$ 分别控制组稀疏、逐元素稀疏与流形平滑强度。

2.5. 自适应组权重的 MSGL (AW-MSGL)

为增强对组内冗余的抑制，我们将组罚项改写为带权形式：

$$\min_{\omega} \frac{1}{2}\left\|Y - X\omega\right\|_2^2 + \lambda_1 \sum_{j=1}^{g} w_j \left\|\omega_{G_j}\right\|_2 + \lambda_2 \left\|\omega\right\|_1 + \lambda_3 \omega^{\mathrm{T}} X^{\mathrm{T}} L_s X \omega, \tag{29}$$

其中， $w_j > 0$ 为第 j 个组的自适应权重，取值由组内特征相似性给出。一个简便且稳健的选择是平均余弦相似度：

$$w_j = 1 + \eta \frac{2}{\left|G_j\right|\left(\left|G_j\right| - 1\right)} \sum_{u<v, u,v\in G_j} \frac{\left|x_u^{\mathrm{T}} x_v\right|}{\left\|x_u\right\|_2 \left\|x_v\right\|_2}, \ \eta \in [0,1], \tag{30}$$

并进行归一化以保持整体罚强度不变：

$$w_j \leftarrow \frac{w_j}{\frac{1}{g}\sum_{t=1}^{g} w_t}. \tag{31}$$

这样可让冗余度更高(相似性更强)的组受到更强约束，从而减少不必要的共表达特征被同时选入，提升模型紧凑性和稳定性。上述权重也可替换为平均绝对相关系数等相近度量而不影响整体框架。

该建模策略具备以下优势：

- **结构感知的特征选择**：在 SGL 框架中融合流形正则，同时兼顾判别能力与几何一致性。
- **稳定性与泛化提升**：保持样本的局部结构可降低过拟合，尤其适用于高维少样本场景。
- **稀疏性与精度兼顾**：在更少特征的前提下，实验显示分类准确率与 F1 指标更优。
- **可解释性增强**：被选基因往往构成紧密相连的子网络，便于生物学解读与通路分析。

2.6. 优化算法

为高效求解 MSGL 的非光滑凸目标，我们采用带 Nesterov 加速的近端梯度(APG)方法。目标函数可分解为光滑部分：

$$f(\omega)=\frac{1}{2}\|X\omega-Y\|_2^2+\lambda_3\omega^{\mathrm{T}}X^{\mathrm{T}}L_sX\omega, \tag{32}$$

以及非光滑部分：

$$g(\omega)=\lambda_1\sum_{j=1}^{g}\left\|\omega_{G_j}\right\|_2+\lambda_2\|\omega\|_1. \tag{33}$$

光滑部分的梯度为：

$$\nabla f(\omega)=X^{\mathrm{T}}(X\omega-Y)+\lambda_3X^{\mathrm{T}}L_sX\omega. \tag{34}$$

为加速收敛，我们采用 Nesterov 动量。在第 i 次迭代，先计算外推搜索点：

$$q^i=\omega^i+\alpha_i\left(\omega^i-\omega^{i-1}\right), \tag{35}$$

其中，$\alpha_i=\frac{1-\rho_{i-1}}{2}$，$\rho_i=\frac{2}{i+3}$，并令 $\omega^{-1}=\omega^0$。

步长 l 通过回溯线搜索确定，以满足下降条件。随后通过求解以下近端子问题得到下一次迭代：

$$\omega^{i+1}=\arg\min_{\omega}\left\{\left\langle\omega-q^i,\nabla f\left(q^i\right)\right\rangle+\frac{l}{2}\left\|\omega-q^i\right\|^2+g(\omega)\right\}. \tag{36}$$

该子问题可通过“分组软阈”获得闭式解。对每个组 G_j，定义：

$$\omega_{G_j}^{i+1}=\begin{cases}\left(1-\dfrac{\lambda_1w_j}{l\left\|\hat{\omega}_{G_j}\right\|_2}\right)\left(\hat{\omega}_{G_j}-\dfrac{\lambda_2}{l}\operatorname{sgn}\left(\hat{\omega}_{G_j}\right)\right), & \left\|\hat{\omega}_{G_j}\right\|_2>\dfrac{\lambda_1w_j}{l},\\ 0, & \text{otherwise.}\end{cases} \tag{37}$$

算法迭代直至权重 ω 的相对变化低于设定的容差(例如 10^{-6})，或达到最大迭代次数为止。该优化方案高效地结合了流形保持、组稀疏与元素级稀疏，使其能够在高维基因组数据上实现可扩展且稳定的特征选择。

3. 结果

为系统检验 MSGL 的实际效能，我们在结直肠癌(CRC)微阵列数据上开展对照实验：先按方法部分的流程进行预筛，再以各方法选出的特征训练 SVM 分类器，分别用 10 折交叉验证与独立测试集进行评估。对比方法涵盖 LASSO、GL、SGL 与 FL。本文的表格与实证分析仅报告 CRC 结果。

从图 3 可见，MSGL 在结直肠癌(CRC)数据上以更小的特征集取得可比甚至更优的性能。以 CRC 为

例，MSGL 仅需 11 个基因，而 LASSO 与 FL 往往超过 30 个；更紧凑的特征集提升了模型的可解释性与稳健性。

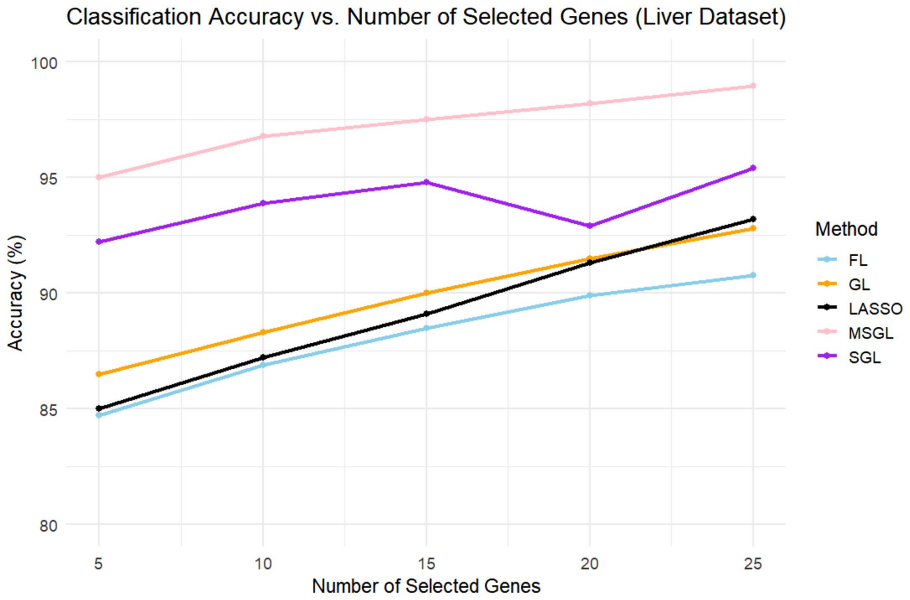

Figure 3. Trend of classification accuracy under varying numbers of features
图 3. 分类准确率随特征数量变化的趋势

3.1. 结直肠癌(CRC)：核心应用案例

依照方法部分的同一预处理与评估流程(F-score + K-means, $k \approx 154$)，我们在结直肠癌数据上进行对比；指标采用准确率、F1 值、MCC、AUC 与 PRC。对比方法包括 LASSO、GL、SGL 与 FL。

采用 5 × 10 折交叉验证策略以确保评价的公平与稳健。报告结果取 10 次独立运行的平均值。如表 2 所示，MSGL 在结直肠癌数据集的所有指标上均达到了完美表现，Accuracy、F1-score、MCC、AUC 与 PRC 均为 100.00%。这说明 MSGL 不仅能识别高度具辨识性的基因，还能构建具有极佳泛化与稳定性的结直肠癌检测模型。相比之下，表现最好的对比方法 SGL 的准确率为 94.10%、PRC 为 94.20%，显著低于 MSGL。这些结果突出显示了 MSGL 在处理高维、相对小样本的基因组数据时的优越特征选择能力。且经 Wilcoxon 符号秩检验，MSGL 与 SGL 之间的性能差异具有统计学意义($P < 0.05$)，证明性能提升并非随机误差。

在散发性 CRC 数据集中，特征选择还捕获到若干具生物学意义的候选基因；代表性结果及其可能作用见表 3。

碳酸酐酶 VII (CA7)被鉴定为结直肠癌(CRC)中的肿瘤抑制因子，其下调与疾病侵袭性增强及不良临床结局相关。qPCR、Western blot 与免疫组化分析显示，CRC 组织中 CA7 的 mRNA 与蛋白表达较邻近正常黏膜显著降低[25]。机制上，碳酸酐酶家族成员(包括 CA7)参与细胞内 pH 稳态维持，并可能防止肿瘤微环境的胞外酸化——这一状态已被证实会促进侵袭、上皮–间质转化(EMT)与免疫逃逸[26]。低表达 CA7 与肿瘤分期进展、淋巴结转移、差的组织学分化相关，并能独立预测更短的无病生存与总体生存，确立其在 CRC 中的预后价值。在胃癌中也观察到类似发现：CA7 下调与肿瘤去分化及不良预后相关，提示 CA7 在胃肠道恶性肿瘤中具有保守的肿瘤抑制作用[27]。因此，CA7 的缺失可能通过破坏 pH 调控并增强促肿瘤信号通路来推动肿瘤进展。这些发现将 CA7 定位为胃肠道肿瘤发生的关键调节因子，以及

CRC 风险分层的有前景生物标志物。

Table 2. Performance comparison of feature selection methods on cancer datasets
表 2. 各癌症数据集特征选择方法的性能对比

方法	准确率	F1	MCC	AUC	PRC	P-value (vs MSGL)
			GSE9348			
MSGL	100.00%	100.00%	100.00%	100.00%	100.00%	-
GL	92.35%	91.50%	89.20%	94.50%	93.00%	0.008
SGL	94.10%	92.75%	91.50%	95.10%	94.20%	0.032
FL	89.85%	89.00%	87.80%	91.75%	91.00%	0.004
LASSO	90.50%	90.30%	88.90%	93.20%	92.80%	0.015

注：P 值通过 Wilcoxon 符号秩检验计算得出，用于比较各方法与 MSGL 的性能差异。$P < 0.05$ 表示差异具有统计学意义。

Table 3. Representative genes identified by feature selection in sporadic colorectal cancer datasets
表 3. 散发性结直肠癌数据集中通过特征选择识别的代表性基因

基因	全称(缩写)	在结直肠癌(CRC)中的关键功能作用(Key Functional Role in CRC)
KLF4	Krüppel-like factor 4	调控分化、EMT 与 STAT3 信号的转录因子；预后标志物。
SEMA3E	Semaphorin 3E	PLXND1 的配体；通过 PI3K/AKT 通路促进 EMT 与转移。
NHEJ1	非同源末端连接因子 1 (XLF)	DNA 修复因子；通过 NHEJ 通路促进化疗耐药。

Krüppel-like factor 4 (KLF4)在结直肠癌(CRC)中发挥肿瘤抑制作用。与邻近正常上皮相比，肿瘤组织中的 KLF4 表达常见下调，且低表达与分化差、TNM 分期更晚、总体生存更短密切相关，凸显其预后意义[28]。KLF4 通过多种机制发挥抗肿瘤效应，包括直接结合 β-catenin 抑制 Wnt 信号活性，以及抑制 cyclin D1 表达从而诱导细胞周期阻滞、降低增殖。此外，KLF4 可激活膜结合原钙黏蛋白的表达，使 β-catenin 锚定于质膜，进一步削弱致癌性 Wnt 通路的活化。最新研究显示，KLF4 还可通过靶向 RAB26 并抑制自噬来增强 CRC 的化疗敏感性，自噬是 5-氟尿嘧啶(5-FU)耐药的关键机制。体内外实验表明，KLF4 过表达可降低 LC3-II 的积累与自噬体形成，并在 5-FU 处理下增加细胞凋亡[29]。此外，肠上皮细胞中 KLF4 的缺失会导致细胞迁移增加与促癌信号通路的激活，强调其在维持上皮稳态中的作用。上述结果共同确立了 KLF4 在 CRC 中的肿瘤抑制、分化与治疗反应调控中的关键地位。

五聚蛋白家族成员 3 (PTX3)在结直肠癌中上调，并与侵袭性肿瘤行为及不良预后相关。血浆 PTX3 水平升高与更晚的 TNM 分期、淋巴结转移以及更短的总体生存显著相关，确立其作为 CRC 患者独立预后生物标志物的作用[30]。在功能层面，PTX3 通过增强肿瘤微环境中 M2 样巨噬细胞极化与免疫抑制活性来促进肿瘤进展，从而推动免疫逃逸[31]。机制研究显示，抑制 PTX3 可下调 IL-10、TGF-β 等关键免疫抑制因子的表达，并减少调节性 T 细胞(Tregs)的募集，从而增强抗肿瘤免疫。此外，在结肠癌细胞中敲低 PTX3 能通过下调基质金属蛋白酶(MMP2、MMP9)以及抑制 NF-κB 信号通路活化来抑制细胞增殖、迁移与侵袭[32]。体内实验亦证实，沉默 PTX3 能降低肿瘤生长与转移潜能，凸显其在 CRC 进展中的功能重要性。上述发现将 PTX3 定位为肿瘤免疫调控的关键介质以及 CRC 的潜在治疗靶点。

3.2. 比较与讨论

如表 4 所示，MSGL 在结直肠癌(CRC)的独立测试集上整体表现更优。尽管在更贴近真实测试条件下，各方法均出现一定程度的性能下降，MSGL 在 Accuracy、F1-score、MCC、AUC 与 PRC 指标上持续

取得最佳或近最佳结果，体现出较强的泛化能力与鲁棒性。在 GSE8671 数据集上，MSGL 显著优于 GL、SGL、FL 与 LASSO。相比之下，GL 与 FL 在多数指标上的表现相对较弱；LASSO 在准确率上具有一定竞争性，但在更均衡的指标(如 MCC 与 AUC)上不及 MSGL。总体而言，MSGL 在独立 CRC 测试条件下展现出更优的分类性能与特征选择效果，验证了其识别高判别性、可泛化基因特征的能力。

Table 4. Performance of feature selection methods on independent test sets
表 4. 独立测试集上特征选择方法的性能

方法	准确率	F1	MCC	AUC	PRC	P-value (vs MSGL)
			GSE8671			
MSGL	93.50%	93.80%	90.20%	95.00%	95.30%	-
GL	86.10%	85.60%	80.30%	87.50%	86.80%	0.006
SGL	88.60%	87.40%	83.80%	89.70%	89.00%	0.028
FL	83.40%	83.00%	77.90%	85.10%	84.30%	0.003
LASSO	85.00%	84.80%	80.50%	86.90%	86.20%	0.012

注：P 值通过 Wilcoxon 符号秩检验计算得出，用于比较各方法与 MSGL 的性能差异。$P < 0.05$ 表示差异具有统计学意义。

4. 结论与未来工作

本研究围绕“以更少且可解释的基因子集实现稳定分类”的应用导向目标，提出几何感知的结构化特征学习框架(MSGL)。我们以结直肠癌(CRC)为核心应用场景展开验证。方法将流形正则与组稀疏惩罚结合，刻画基因表达的内在几何结构并施加具生物学意义的模块化约束。在缺乏预定义生物通路的情况下，基于余弦相似度的聚类自动构建基因模块，提升跨癌种适用性。综合微阵列实验显示，MSGL 以显著更少的基因实现更优或相当的分类性能，并保留较强的泛化与解释能力，适合向临床转化场景落地。

尽管本框架表现出色，当前流程在预处理阶段仍依赖经验性参数选择(如基因簇数量)。未来工作将致力于开发用于自动结构发现的自适应策略，包括基于数据的簇数确定以及多组学先验的融合。此外，将该框架拓展至深度稀疏模型并应用于单细胞 RNA-seq 数据，有望进一步提升其在精准肿瘤学中的可扩展性与生物学可解释性。

参考文献

[1] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[2] Bühlmann, P. and Van De Geer, S. (2011) Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Science Business Media.

[3] Li, Q. (2023) A Comprehensive Survey of Sparse Regularization: Fundamental, State-of-the-Art Methodologies and Applications on Fault Diagnosis. *Expert Systems with Applications*, **229**, Article ID: 120517. https://doi.org/10.1016/j.eswa.2023.120517

[4] Frank, L.E. and Friedman, J.H. (1993) A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, 109-135. https://doi.org/10.1080/00401706.1993.10485033

[5] Meinshausen, N. and Bühlmann, P. (2006) High-Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics*, **34**, 1436-1462. https://doi.org/10.1214/009053606000000281

[6] Xu, J. and Ying, Z. (2008) Simultaneous Estimation and Variable Selection in Median Regression Using Lasso-Type Penalty. *Annals of the Institute of Statistical Mathematics*, **62**, 487-514. https://doi.org/10.1007/s10463-008-0184-2

[7] Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. https://doi.org/10.1198/016214501753382273

[8] Bühlmann, P., Meier, L. and Zou, H. (2008) Discussion of "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models" by H. Zou and R. Li. *The Annals of Statistics*, **36**, 1534-1541.

[9] Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429. https://doi.org/10.1198/016214506000000735

[10] Lin, Z., Xiang, Y. and Zhang, C. (2009) Adaptive Lasso in High-Dimensional Settings. *Journal of Nonparametric Statistics*, **21**, 683-696. https://doi.org/10.1080/10485250902984875

[11] Yuan, M. and Lin, Y. (2007) Model Selection and Estimation in the Gaussian Graphical Model. *Biometrika*, **94**, 19-35. https://doi.org/10.1093/biomet/asm018

[12] Zhang, C. (2010) Nearly Unbiased Variable Selection under Minimax Concave Penalty. *The Annals of Statistics*, **38**, 894-942. https://doi.org/10.1214/09-aos729

[13] Breheny, P. and Huang, J. (2011) Coordinate Descent Algorithms for Nonconvex Penalized Regression, with Applications to Biological Feature Selection. *The Annals of Applied Statistics*, **5**, Article No. 232. https://doi.org/10.1214/10-aoas388

[14] Tian, G.L., Tang, M.L., Fang, H.B., *et al*. (2008) Efficient Methods for Estimating Constrained Parameters with Applications to Regularized (Lasso) Logistic Regression. *Computational Statistics & Data Analysis*, **52**, 3528-3542. https://doi.org/10.1016/j.csda.2007.11.007

[15] Adeli, E., Li, X., Kwon, D., Zhang, Y. and Pohl, K.M. (2020) Logistic Regression Confined by Cardinality-Constrained Sample and Feature Selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, 1713-1728. https://doi.org/10.1109/tpami.2019.2901688

[16] Liang, Y., Liu, C., Luan, X., Leung, K., Chan, T., Xu, Z., *et al*. (2013) Sparse Logistic Regression with a L1/2 Penalty for Gene Selection in Cancer Classification. *BMC Bioinformatics*, **14**, Article No. 198. https://doi.org/10.1186/1471-2105-14-198

[17] Xu, Z., Zhang, H., Wang, Y., Chang, X. and Liang, Y. (2010) L 1/2 Regularization. *Science China Information Sciences*, **53**, 1159-1169. https://doi.org/10.1007/s11432-010-0090-0

[18] Zou, H. and Hastie, T. (2005) Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **67**, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[19] Zou, H. and Zhang, H.H. (2009) On the Adaptive Elastic-Net with a Diverging Number of Parameters. *The Annals of Statistics*, **37**, Article No. 1733. https://doi.org/10.1214/08-aos625

[20] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2004) Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **67**, 91-108. https://doi.org/10.1111/j.1467-9868.2005.00490.x

[21] Yuan, M. and Lin, Y. (2005) Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **68**, 49-67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

[22] Friedman, J., Hastie, T. and Tibshirani, R. (2010) A Note on the Group Lasso and a Sparse Group Lasso.

[23] Ma, Z., Guan, X., Liu, Y. and Shao, W. (2024) Identification of Essential Plasma Protein Using Manifold Regularized Sparse Group-Lasso for Prediction of Alzheimer's Disease. *Displays*, **81**, Article ID: 102578. https://doi.org/10.1016/j.displa.2023.102578

[24] Chen, X., Pan, W., Kwok, J.T. and Carbonell, J.G. (2009) Accelerated Gradient Method for Multi-Task Sparse Learning Problem. 2009 9*th IEEE International Conference on Data Mining*, Miami Beach, 6-9 December 2009, 746-751. https://doi.org/10.1109/icdm.2009.128

[25] Yang, G.-Z., Hu, L., Cai, J., *et al*. (2022) Prognostic Value of Carbonic Anhydrase VII Expression in Colorectal Carcinoma. *Frontiers in Immunology*, **13**, Article ID: 1051353.

[26] Svastová, E., Hulíková, A., Rafajová, M., *et al*. (2004) Carbonic Anhydrase XII Is a Membrane-Bound Hypoxia-Inducible Protein beyond Carbonic Anhydrase IX. *Journal of Biological Chemistry*, **279**, 23433-23441.

[27] Kondo, H., Yamada, D., Fujii, S., *et al*. (2018) Reduced Expression of Carbonic Anhydrase VII in Gastric Cancer: Its Association with Differentiation and Prognosis. *Histopathology*, **72**, 987-997.

[28] Parenti, S., Montorsi, L., Fantini, S., Mammoli, F., Gemelli, C., Atene, C.G., *et al*. (2018) KLF4 Mediates the Effect of 5-ASA on the B-Catenin Pathway in Colon Cancer Cells. *Cancer Prevention Research*, **11**, 503-510. https://doi.org/10.1158/1940-6207.capr-17-0382

[29] Zheng, Y., Wu, J., Chen, H., Lin, D., Chen, H., Zheng, J., *et al*. (2023) KLF4 Targets RAB26 and Decreases 5-FU Resistance through Inhibiting Autophagy in Colon Cancer. *Cancer Biology & Therapy*, **24**, Article ID: 2205253. https://doi.org/10.1080/15384047.2023.2226353

[30] Zhang, J., Wang, T. and Niu, X. (2016) Increased Plasma Levels of Pentraxin 3 Are Associated with Poor Prognosis of Colorectal Carcinoma Patients. *The Tohoku Journal of Experimental Medicine*, **240**, 39-46.

https://doi.org/10.1620/tjem.240.39

[31] Chen, F.W., Wu, Y.L., Cheng, C.C., Hsiao, Y., Chi, J., Hung, L., *et al*. (2024) Inactivation of Pentraxin 3 Suppresses M2-Like Macrophage Activity and Immunosuppression in Colon Cancer. *Journal of Biomedical Science*, **31**, Article No. 10. https://doi.org/10.1186/s12929-023-00991-7

[32] Li, M., Hu, Y., Wang, J., Xu, Y., Hong, Y., Zhang, L., *et al*. (2023) The Dual HDAC and PI3K Inhibitor, CUDC-907, Inhibits Tumor Growth and Stem-Like Properties by Suppressing PTX3 in Neuroblastoma. *International Journal of Oncology*, **64**, Article No. 14. https://doi.org/10.3892/ijo.2023.5602