CCML 2021 第十八届中国机器学习会议
The 18th China Conference on Machine Learning
2021年8月6-8日 中国·长沙
AUG 6-8, 2021 CHANGSHA, CHINA

面向微阵列基因表达数据的集成特征选择方法

背景与意义

■ 基因选择

从原始特征空间中剔除不相关的、冗余的特征，保留信息特征。准确的基因选择有助于推动癌症诊断、肿瘤分类、生物标志物识别和药物作用靶点寻找等研究

■ 稳定性

特征选择稳定性是指当训练样本集发生微小扰动时，特征选择算法能够选出相同或相似的特征子集。稳定的特征选择方法有助于获得可靠的特征子集，提高结果的可解释性；相反，稳定性较差的特征选择算法会降低生物医学研究人员使用该方法的信心

■ 如何设计稳定、准确的特征选择方法在实际应用中具有重要价值

集成特征选择框架

■ 在集成学习范式下设计特征选择框架

■ 利用resampling、bootstrap等采样技术产生M个训练子集，然后在每个子集上利用基特征选择算法得到一个特征子集；最后采用某种聚合策略合并M个子集，产生最终选择的特征子集

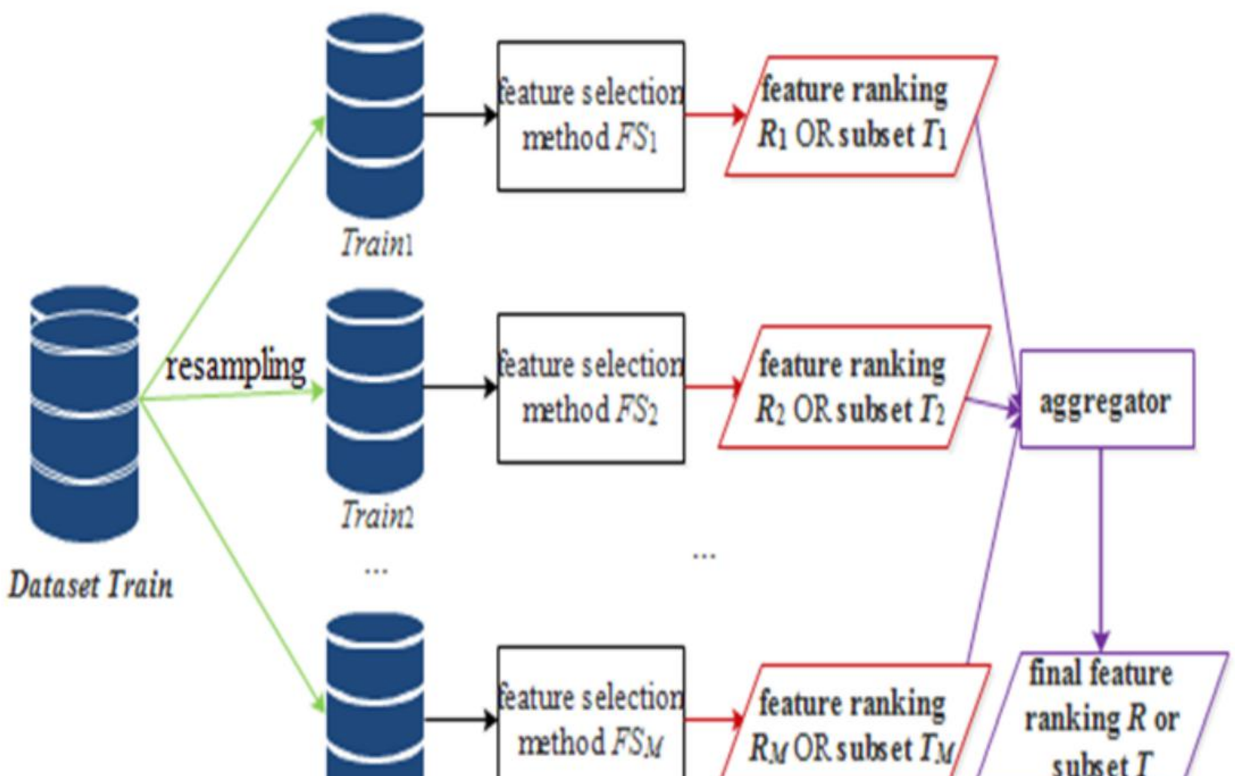


图1 集成特征选择框架

聚合策略

■ 聚合器是集成特征选择方法中的一个重要组件，其功能是根据一定的方式将若干个特征子集合并成一个特征子集S_F。本文以特征在特征子集集合中出现的频率作为筛选准则

■ 给定由基特征选择器输出的M个特征子集$\{T_1, T_2, \cdots, T_M\}$ $Q=\bigcup_{i=1}^{M} T_i$

■ p_f表示特征$f \in T$出现在M个集合中的频率，当p_f的值不小于给定的阈值γ时，将f加入到S_F中，即

$$S_F = \{f \mid p_f \geq \gamma, f \in Q\}$$

■ 特别地，

$$\gamma = 0 \text{ 时，} S_F = \bigcup_{i=1}^{M} T_i \text{；} \gamma = 1 \text{ 时，} S_F = \bigcap_{i=1}^{M} T_i$$

■ 进一步地，通过上式获得S_F后，可以在S_F上再次使用特征选择算法FS以优化特征空间

实验与结果

■ 数据集

Dataset	*#Samples*	*#Classes*	*#SGR*
Colon	62 (40/22)	2	0.031
DLBCL	77 (58/19)	2	0.011
Leukemia	72 (38/9/25)	3	0.014
SRBCT	83 (29/25/11/18)	4	0.036

■ 对比方法

- 选择Correlation-based feature selection (CFS)作为基特征选择算法
- ReliefF、Mutual Information Maximization (MIM)、Min-Redundancy Max-Relevance (MRMR)、Joint Mutual Information (JMI)、Fast Correlation Based Filter (FCBF)、conditional mutual information maximization (CMIM)
- 对于聚合策略，选择不同的阈值γ。将其值分别设置为1、0.75、0.5、0.25和0，并记对应的集成特征选择方法为$enCFS_I$、$enCFS_T$、$enCFS_U$、$enCFS_H$、$enCFS_Q$、$enCFS_U$

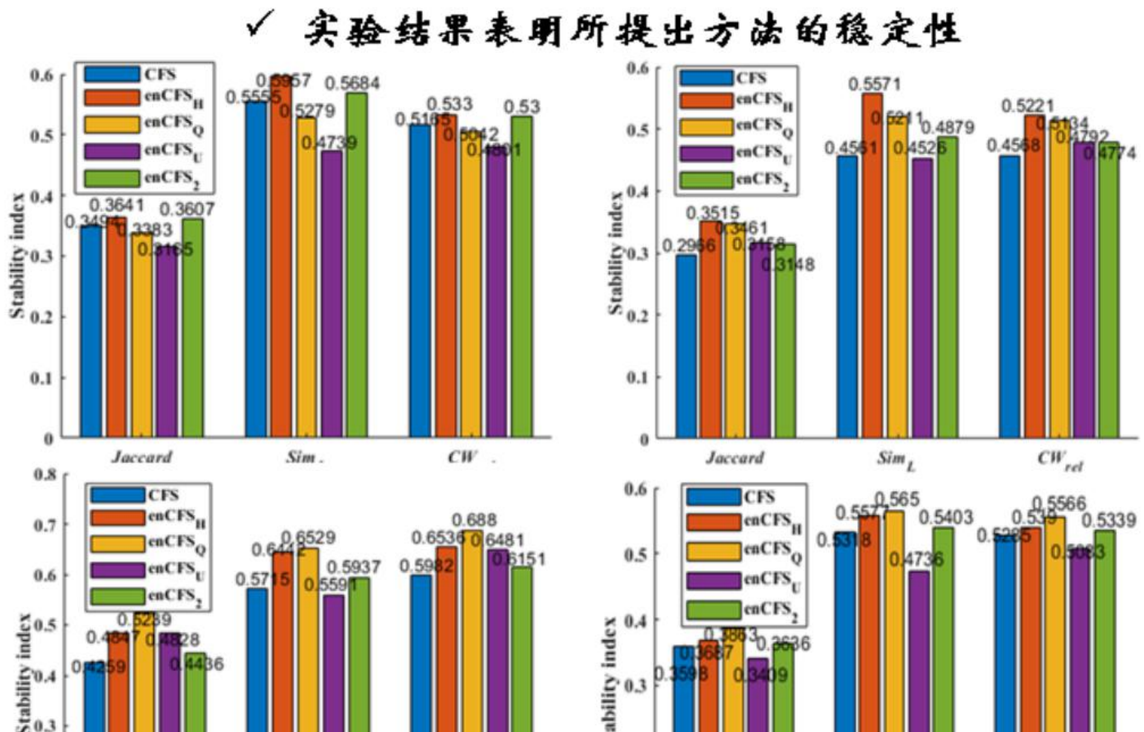
■ 评价指标

- 特征子集质量：Accuracy和F-measure
- 特征子集大小
- 选择的特征子集的稳定性：杰拉德指数、调整相似性指数

结果

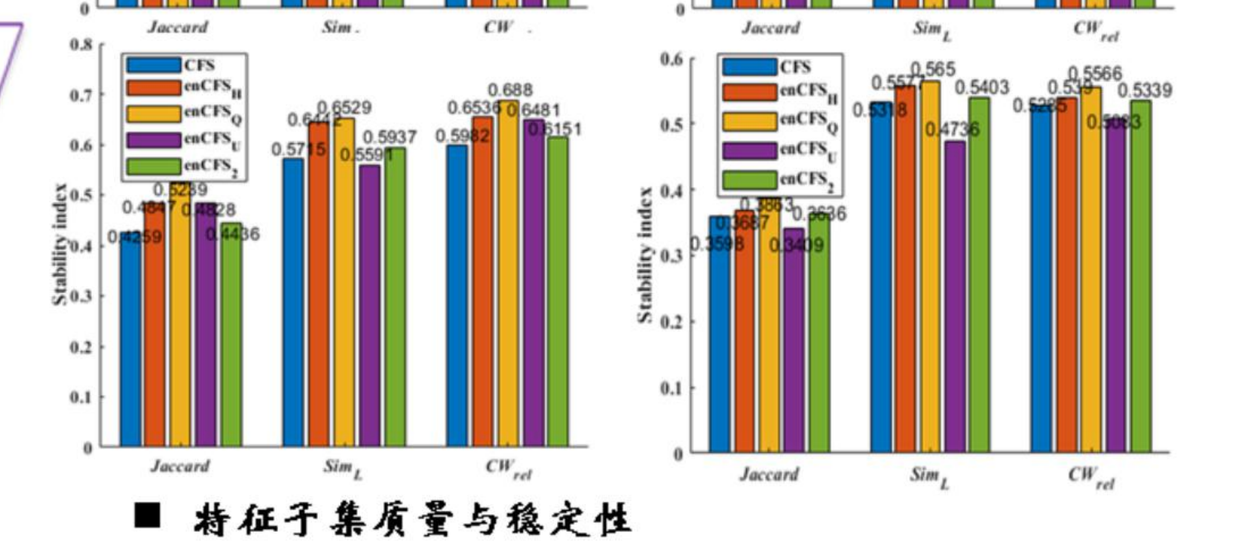
■ 在不同数据集上比较特征选择方法的稳定性（从左往右，从上往下对应实验数据集）

✓ 实验结果表明所提出方法的稳定性



■ 特征子集质量与稳定性

✓ 表明$enCFS_H$获得了较高的准确率和稳定性



投稿人 姓名：王爱国、陈桂林
单位： 佛山科学技术学院，滁州学院