面向微阵列基因表达数据的集成特征选择方法

王爱国†，陈桂林‡

†电子信息工程学院，佛山科学技术学院

‡计算机与信息工程学院，滁州学院

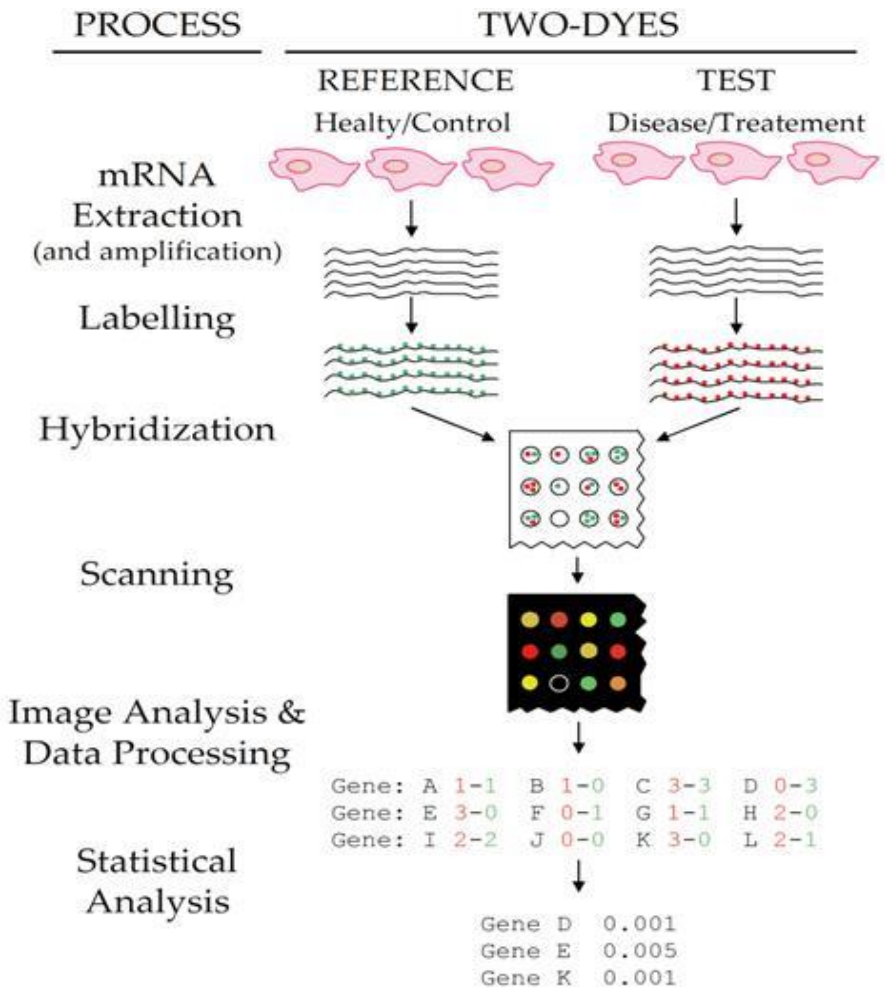
CCML 2021@长沙，中国

7th August, 2021

目录

- 背景与意义
- 集成特征选择框架
- 实验设置与结果分析
- 总结

微阵列技术



微阵列技术

https://en.wikipedia.org/wiki/DNA_microarray

基因*样本

基因表达谱

基因表达谱，服务于精准医学、个性化医疗

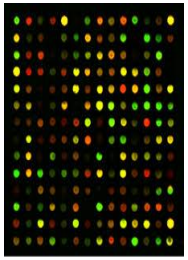
❑面向样本

-癌症预测

cancer vs. non-cancer

-肿瘤亚型识别

Small Round Blue Cell Tumor: EWS/NB/BL/RMS



vs.

(a)基因表达谱　(b)组织形态学　(c) 临床数据

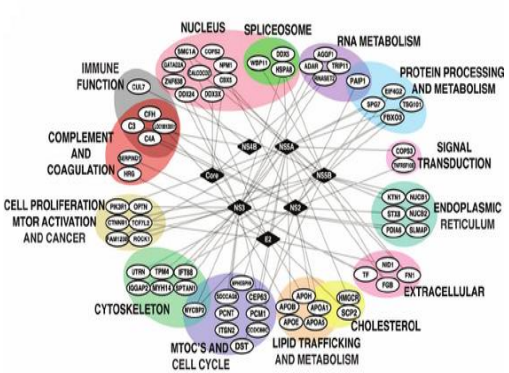
❑面向基因

-功能标注

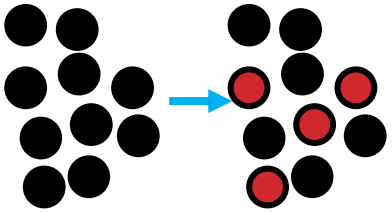
未知基因功能推测

-信息基因筛选

基因选择



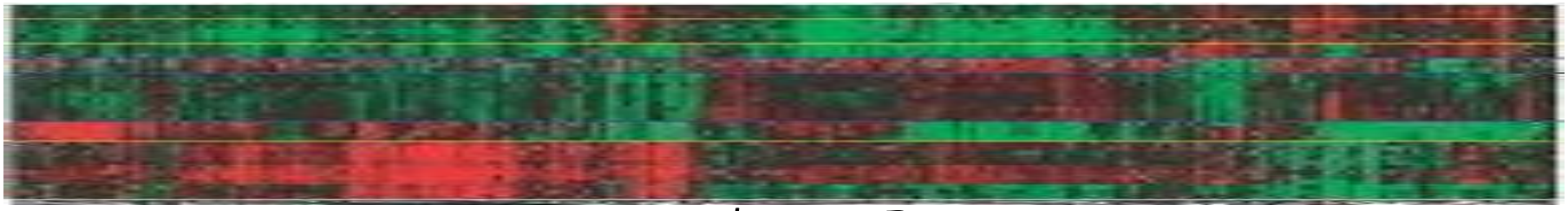
(a) 聚类



(b) 基因选择

Dolan et al. 2013. Identification and Comparative Analysis of Hepatitis C Virus-Host Cell Protein Interactions

样本数N



基因数P

- $P >> N$
- 易过拟合，泛化能力弱
- 原始基因空间中包含大量的***不相关***基因与***冗余***基因
  - 在微阵列环境下，称基因为特征
  - 少量的基因与研究的疾病相关
- 一些分类器对不相关特征/冗余特征敏感
  - 朴素贝叶斯vs. 冗余特征

$$P(y_i|X) \propto P(y_i) * \prod_{k=1}^{p} P(a_k|y_i); X = (a_1, \dots, a_p)$$

  - k近邻分类器vs. 不相关特征

$$d(S,T) = \sum_{k=1}^{p} (S_k - T_k)^2$$

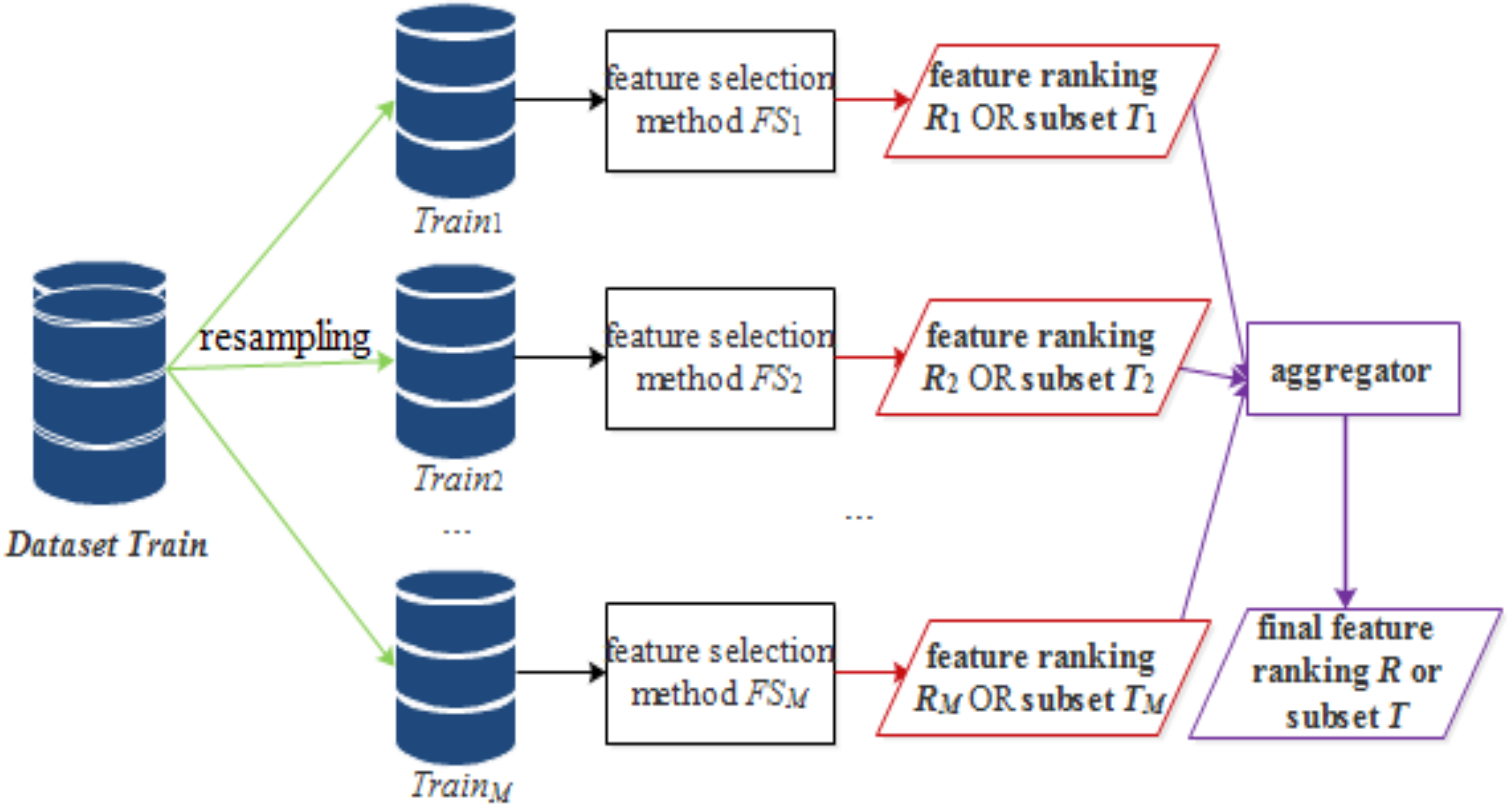
基因选择的稳定性问题

- 基因选择，其主要目标是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
- 在基因表达谱分析中，除了找出一组最具判别能力的特征子集外，特征选择的稳定性也是一个非常重要的方面
- 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，特别是当生物验证和分析基因功能的实验成本高昂时，导致基因选择算法难以广泛地应用
- 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

目录

- 背景与意义
- 集成特征选择框架
- 实验设置与结果分析
- 总结

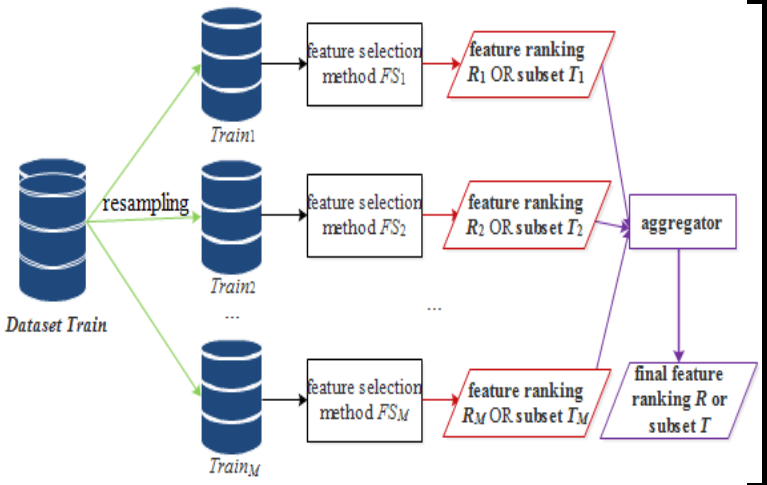
- 影响特征选择稳定性的因素主要包括
  - 数据层面(e.g.，高维小样本数据)
  - 算法层面(e.g.，所设计的算法未考虑稳定性、算法对超参数的初始值敏感)
  - 特定的应用领域(e.g.，存在多个数值和功能高度相关的基因)
- 集成特征选择



集成策略

- 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，*

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

- 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

目录

- 背景与意义
- 集成特征选择框架
- 实验设置与结果分析
- 总结

实验设置：基特征选择算法

- 基于相关性的特征选择算法(correlation-based feature selection, CFS)
- 返回一个特征子集，与特征排序方法相比，无需指定最终返回的特征个数
- 给定包含特征集*F*和类别变量*C*的数据集*Train*，CFS以式(1)为评价准则，采用最佳优先搜索(best first search)方式搜索特征空间

$$merit_S = \frac{k\overline{r}_{Cf}}{\sqrt{k + k(k-1)\overline{r}_{ff}}}, \qquad (1)$$

- $merit_S$表示包含k个特征的特征子集S的价值(*merit*)
- $\overline{r}_{Cf}$表示特征-类别平均相关性，$\overline{r}_{ff}$表示特征-特征平均相关性

实验设置：稳定性指标

- 给定K个特征子集$A=\{S_1, S_2, \cdots, S_K\}$
- 杰卡德指标

$$\psi_{Jaccard}(A)=\frac{2}{K(K-1)}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\frac{|s_i \cap s_j|}{|s_i \cup s_j|}$$

- 调整相似性指标

$$\psi_{Sim_L}(A)=\frac{2}{K(K-1)}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\frac{|s_i \cap s_j|-E(|r|)}{\max(|r|)-\min(|r|)}$$

  - r表示大小分别为$|s_i|$和$|s_j|$的两个集合的交集
  - $E(|r|)$、$\max(|r|)$、$\min(|r|)$分别表示$|r|$的期望大小、最大值以及最小值

实验设置：数据集&对比算法&其他设置

数据集

表 1 实验数据集

Dataset	*#Genes*	*#Samples*	*#Classes*	*#SGR*
Colon	2000	62 (40/22)	2	0.031
DLBCL	7129	77 (58/19)	2	0.011
Leukemia	5327	72 (38/9/25)	3	0.014
SRBCT	2308	83 (29/25/11/18)	4	0.036

对比算法

- 除了使用CFS和基于CFS的集成特征选择方法之外
- reliefF、互信息最大化 (mutual information maximization, MIM)、最小冗余最大相关性 (minimum redundancy maximum relevancy, MRMR)、条件互信息最大化(conditional mutual information maximization, CMIM)、联合互信息 (joint mutual information, JMI)、相关性快速过滤特征选择(fast correlation based filter, FCBF)等

其他

- 十折交叉验证
- 从选择的特征子集的大小、预测能力以及稳定性等方面评价特征选择方法

实验结果：特征子集的质量

加粗显示所获得的最好准确率；如果集成特征选择方法的结果不低于CFS的，下划线标识

表 2 朴素贝叶斯分类器下不同特征选择方法的实验结果

Dataset/Metric(%)		w/o	reliefF	MIM	MRMR	CMIM	JMI	FCBF	CFS	$enCFS_H$	$enCFS_Q$	$enCFS_U$	$enCFS_2$
Colon	Accuracy	58.07	**85.48**	82.26	82.26	83.87	82.26	80.65	82.26	**85.48**	83.87	**85.48**	82.26
	F-measure	61.02	84.01	81.37	81.37	83.39	81.37	80.71	81.37	84.81	83.39	84.81	81.37
DLBCL	Accuracy	79.22	92.21	89.61	90.91	**96.10**	89.61	89.61	90.91	92.21	92.21	92.21	93.51
	F-measure	71.09	89.52	87.28	88.01	94.88	87.28	85.41	87.58	89.24	89.52	89.52	91.44
Leukemia	Accuracy	**98.61**	91.67	94.44	95.83	94.44	97.22	95.83	93.06	95.83	95.83	97.22	93.06
	F-measure	97.70	86.06	90.38	92.90	91.75	95.42	92.90	89.25	92.90	92.90	95.42	89.25
SRBCT	Accuracy	**100.0**	91.57	98.80	98.80	96.39	97.59	98.80	98.80	**100.0**	98.80	98.80	98.80
	F-measure	100.0	90.55	98.20	98.20	96.34	97.24	99.09	99.09	100.0	99.09	98.44	99.09

表 3 k 近邻分类器下不同特征选择方法的实验结果

Dataset/Metric(%)		w/o	reliefF	MIM	MRMR	CMIM	JMI	FCBF	CFS	$enCFS_H$	$enCFS_Q$	$enCFS_U$	$enCFS_2$
Colon	Accuracy	82.26	75.81	79.03	79.03	82.26	80.65	75.81	80.65	**85.48**	82.26	83.87	80.65
	F-measure	80.45	72.57	76.88	76.88	81.37	78.28	73.86	78.86	84.01	80.20	82.10	78.50
DLBCL	Accuracy	80.52	88.31	83.12	**98.70**	94.81	96.10	97.40	90.91	92.21	97.40	93.51	92.21
	F-measure	74.27	83.62	79.54	98.25	93.01	94.69	96.51	88.59	89.52	96.51	91.92	89.97
Leukemia	Accuracy	84.72	91.67	93.06	93.06	94.44	**95.83**	94.44	93.06	**95.83**	94.44	94.44	93.06
	F-measure	84.29	89.70	90.28	91.95	94.17	95.66	93.27	90.78	94.34	91.83	94.43	90.70
SRBCT	Accuracy	84.34	91.57	98.80	**100.0**	95.18	**100.0**	98.80	**100.0**	98.80	97.59	**100.0**	98.80
	F-measure	85.21	91.36	99.08	100.0	95.31	100.0	99.08	100.0	99.08	98.14	100.0	99.08

实验结果：特征子集的质量

加粗显示所获得的最好准确率；如果集成特征选择方法的结果不低于CFS的，下划线标识

表 2 朴素贝叶斯分类器下不同特征选择方法的实验结果

Dataset/Metric(%)		w/o	reliefF	MIM	MRMR	CMIM	JMI	FCBF	CFS	$enCFS_H$	$enCFS_O$	$enCFS_U$	$enCFS_2$
Colon	Accuracy	58.07	**85.48**	82.26	82.26	83.87	82.26	80.65	82.26	**85.48**	83.87	**85.48**	82.26
	F-measure	61.02	84.01	81.37	81.37	83.39	81.37	80.71	81.37	84.81	83.39	84.81	81.37
	Accuracy	79.22	92.21	89.61	90.91	**96.10**	89.61	89.61	90.91	92.21	92.21	92.21	93.51

表 4 支持向量机下不同特征选择方法的实验结果

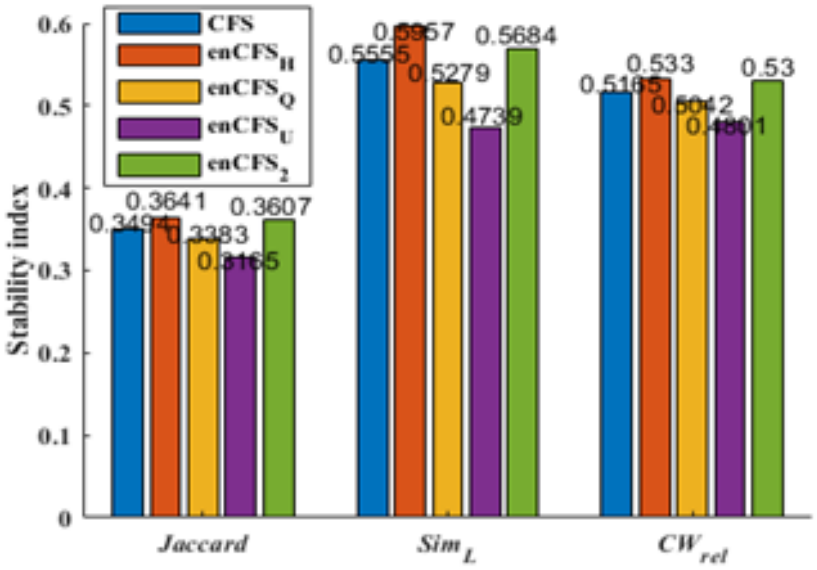
Dataset/Metric(%)		w/o	reliefF	MIM	MRMR	CMIM	JMI	FCBF	CFS	$enCFS_H$	$enCFS_O$	$enCFS_U$	$enCFS_2$
Colon	Accuracy	83.87	80.65	79.03	79.03	72.58	75.81	79.03	79.03	**88.71**	80.65	72.58	83.87
	F-measure	82.10	78.50	77.35	77.35	71.06	74.50	78.63	76.88	87.82	78.86	69.23	82.39
DLBCL	Accuracy	96.10	93.51	92.21	96.10	96.10	96.10	94.81	94.81	94.81	96.10	**97.40**	93.51
	F-measure	94.69	91.14	89.52	94.88	94.69	94.88	92.87	93.01	93.01	94.69	96.51	91.44
Leukemia	Accuracy	**95.83**	91.67	94.44	94.44	**95.83**	**95.83**	**95.83**	**95.83**	**95.83**	93.06	**95.83**	95.83
	F-measure	95.50	89.40	91.97	94.29	94.09	95.66	94.34	94.34	93.28	89.51	95.50	94.34
SRBCT	Accuracy	**100.0**	95.18	98.80	**100.0**	95.18	98.80	96.39	98.80	98.80	98.80	98.80	98.80
	F-measure	100.0	94.96	99.08	100.0	95.65	99.08	97.22	99.08	99.08	99.08	99.08	99.08
DLBCL	Accuracy	80.52	88.31	83.12	**98.70**	94.81	96.10	97.40	90.91	92.21	97.40	93.51	92.21
	F-measure	74.27	83.62	79.54	98.25	93.01	94.69	96.51	88.59	89.52	96.51	91.92	89.97
Leukemia	Accuracy	84.72	91.67	93.06	93.06	94.44	**95.83**	94.44	93.06	**95.83**	94.44	94.44	93.06
	F-measure	84.29	89.70	90.28	91.95	94.17	95.66	93.27	90.78	94.34	91.83	94.43	90.70
SRBCT	Accuracy	84.34	91.57	98.80	**100.0**	95.18	**100.0**	98.80	**100.0**	98.80	97.59	**100.0**	98.80
	F-measure	85.21	91.36	99.08	100.0	95.31	100.0	99.08	100.0	99.08	98.14	100.0	99.08

实验结果：特征子集的大小

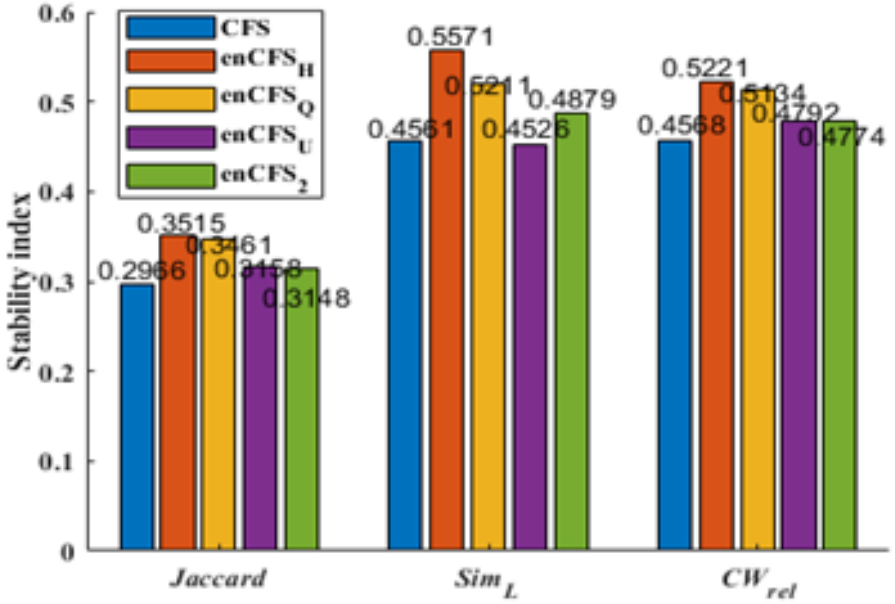
表 5 不同特征选择方法选择的特征个数(均值±标准差)

Dataset	CFS		$enCFS_U$		$enCFS_O$		$enCFS_U$		$enCFS_2$	
	#avg	#std	#avg	#std	#avg	#std	#avg	#std	#avg	#std
Colon	23..8	4.2	11.7	2.8	24.3	2.9	74.9	7.9	21.4	3.7
DLBCL	42.7	5.2	10.4	1.9	28.7	3.2	179.6	13.0	34.7	4.4
Leukemia	91.1	6.0	45.3	3.8	82.9	6.9	231.4	8.3	77.9	5.6
SRBCT	116.7	8.4	72.8	4.3	120.6	9.7	241.6	9.7	105.2	7.4

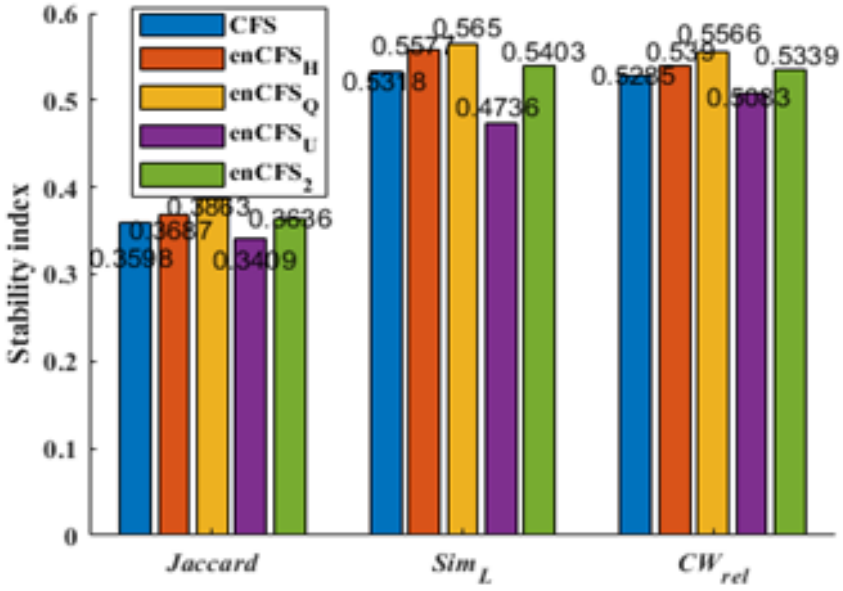
实验结果：特征子集的稳定性



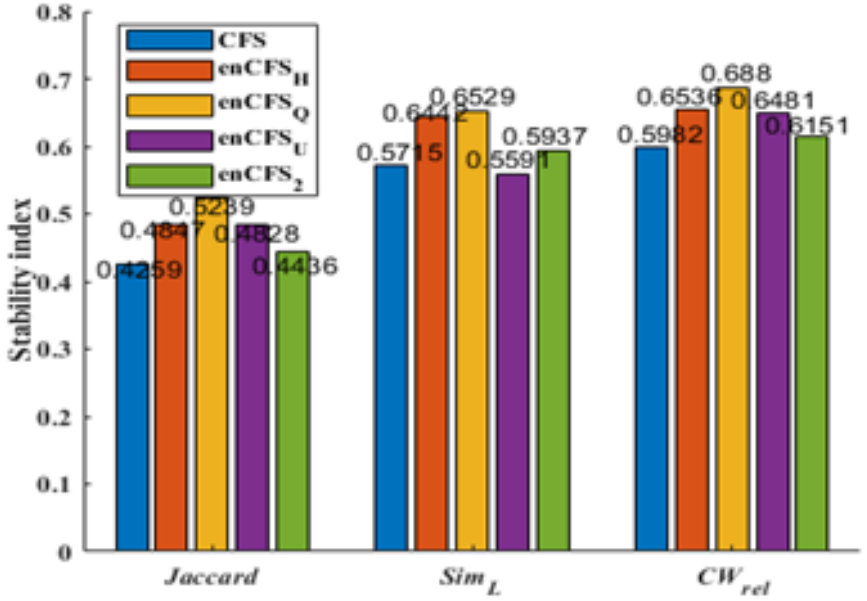
(a) *Colon*



(b) *DLBCL*

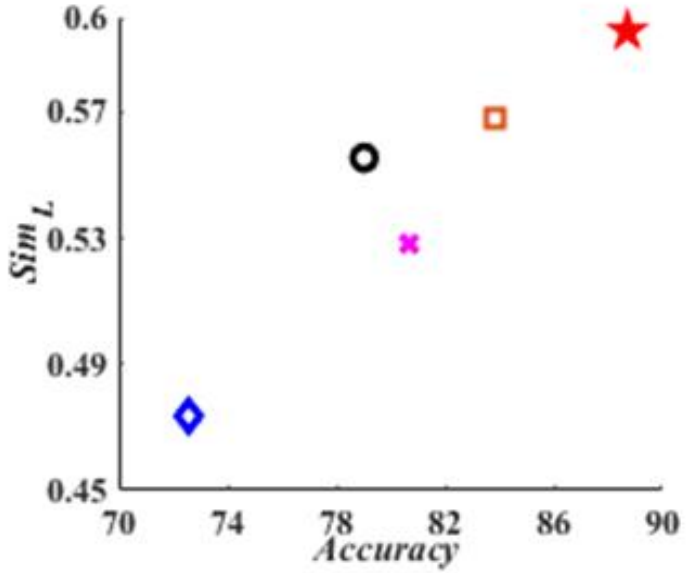


(c) *Leukemia*

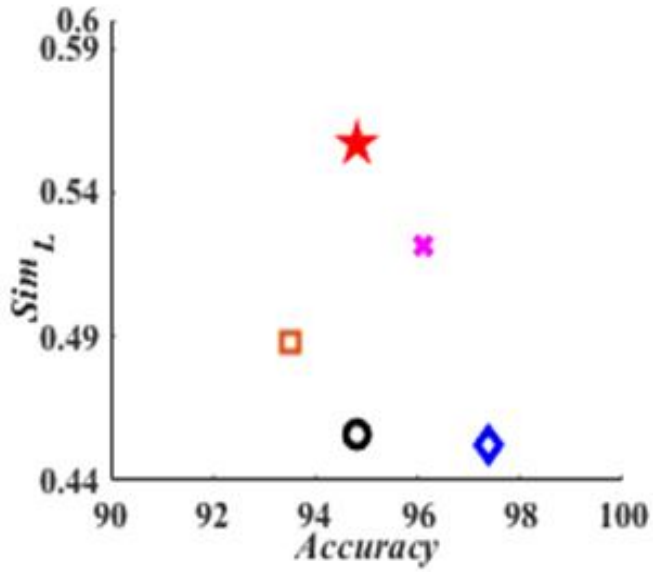


(d) *SRBCT*

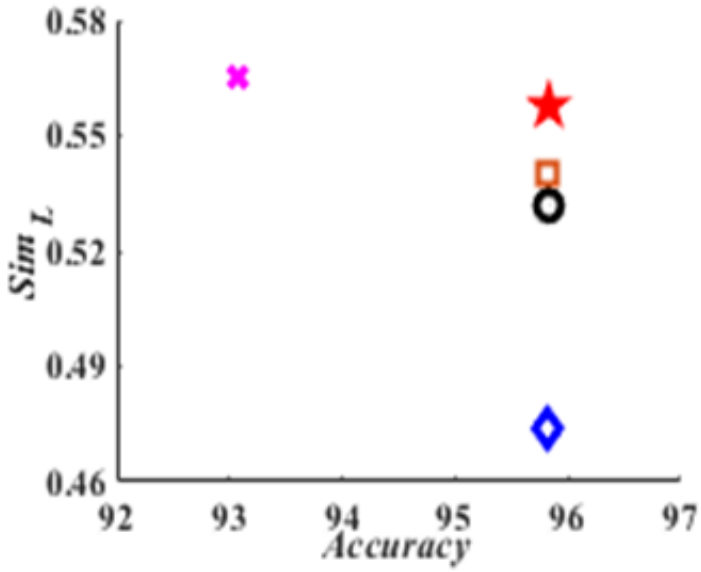
实验结果：特征子集的质量与稳定性之间的tradeoff



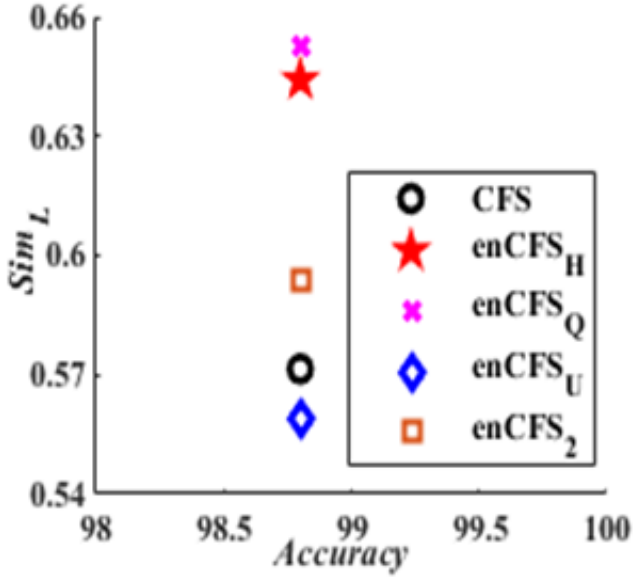
(a) *Colon*



(b) *DLBCL*



(c) *Leukemia*



(d) *SRBCT*

总结

- 在集成学习范式下设计了一个以基于相关性的特征选择算法为组件的稳定特征选择方法
- 给出了两种不同的聚合策略
- 以分类准确性和三个稳定性指数为评价指标，在四个公共的微阵列数据集上进行了对比实验
- 与对比方法相比，本文提出的方法可以获得相当甚至更好的分类准确率，并且能够提高选择特征结果的稳定性

谢谢各位聆听！

Q&A