非平衡基因数据的差异表达基因选择算法研究

谢娟英[1] 王明钊[1),2)] 周 颖[1] 高红超[3] 许升全[2]

[1)](陕西师范大学计算机科学学院 西安 710119)

[2)](陕西师范大学生命科学学院 西安 710119)

[3)](中国科学院大学网络空间安全学院 北京 100093)

摘 要 针对准确率不适于评价不平衡数据特征子集性能的缺陷,提出了 $F2$-measure(简称 $F2$)准则.为避免 mRMR(minimal Redundancy-Maximal Relevance)的互信息方法倾向于选择多值特征,提出了归一化互信息 SU(Symmetrical Uncertainty).针对最大化 AUC(Area Under an ROC Curve)框架下,特征选择算法的特征与类标相关性、特征间相关性的取值范围(量纲)不一致问题,提出了归一化的特征权重.为加快特征选择过程,提出了结合 SU 和 AUC 的特征预选择,缩小特征搜索空间.提出动态加权顺序前向搜索 DWSFS(Dynamic Weighted Sequential Forward Search)和动态加权顺序前向浮动搜索 DWSFFS(Dynamic Weighted Sequential Forward Floating Search),以期得到分类性能更好的特征子集.基于最大化 AUC 和 mRMR 框架,结合上述创新点,设计出 16 种特征选择算法.7 个经典二类不平衡基因数据集、3 个多类不平衡(或近似平衡)基因数据集的 50 次重复实验表明:所提算法选择的基因子集具有非常好的分类识别能力;提出的 $F2$、SU、归一化基因权重、基因预选择,以及 DWSFS 和 DWSFFS 对选择非平衡基因数据集的差异表达基因非常有效.提出的 SU 在度量基因冗余性时优于斯皮尔曼等级相关系数 RCC(Rank Correlation Coefficient);基因选择过程中的权值度量采用基因与类标相关性减去基因间冗余性优于采用基因与类标相关性除以基因冗余性方案.与现有经典基因选择算法的实验比较表明:提出的基因选择算法的性能优于现有基因选择算法.

关键词 基因选择;AUC;互信息;mRMR;不平衡数据

中图法分类号 TP181 **DOI 号** 10.11897/SP.J.1016.2019.01232

Differential Expression Gene Selection Algorithms for Unbalanced Gene Datasets

XIE Juan-Ying[1] WANG Ming-Zhao[1),2)] ZHOU Ying[1] GAO Hong-Chao[3] XU Sheng-Quan[2]

[1)](*School of Computer Science, Shaanxi Normal University, Xi'an* 710119)

[2)](*College of Life Science, Shaanxi Normal University, Xi'an* 710119)

[3)](*School of Cyber Security, University of Chinese Academy of Sciences, Beijing* 100093)

Abstract To overcome the classification accuracy cannot evaluate the capacity of a selected feature subset for unbalanced gene datasets, $F2$-measure(referred to as $F2$) is proposed in this paper, so that the feature subset with much more capacity can be detected to recognize cancer patients. The normalized mutual information named SU(Symmetrical Uncertainty) is present to avoid the mutual information in mRMR(minimal Redundancy-Maximal Relevance) preferring to select those features with many values. To avoid the difference of score ranges of features to label and of between features when maximizing AUC(Area Under an ROC Curve) in feature selection process, a new normalized metric is present to unify the weights between feature and label and between features. To advance the efficiency of feature selection process, SU and AUC are linked

收稿日期:2017-06-28;在线出版日期:2018-01-28.本课题得到国家自然科学基金(61673251)、国家重点研发计划(2016YFC0901900)、科技成果转化培育项目(GK201806013)、中央高校基本科研业务费专项资金项目(GK201701006)、研究生培养创新基金资助项目(2015CXS028,2016CSY009)资助.**谢娟英**(通信作者),博士,教授,中国计算机学会高级会员(16704S),主要研究领域为机器学习、数据挖掘和生物医学大数据分析.E-mail:xiejuany@snnu.edu.cn.**王明钊**,博士研究生,研究方向为数据挖掘、生物信息学.**周 颖**,硕士,研究方向为数据挖掘、生物信息学.**高红超**,博士研究生,研究方向为深度学习和多媒体舆情分析.**许升全**(通信作者),博士,教授,主要研究领域为生物信息学.E-mail:xushengquan@snnu.edu.cn.

together to develop the feature preselection algorithm to reduce the number of candidate features. Dynamic Weighted Sequential Forward Search (DWSFS) and Dynamic Weighted Sequential Forward Floating Search (DWSFFS) are put forward to obtain the feature subset simultaneously with small size and strong recognition capacity when combining $F2$ and AUC as a criterion to evaluate the importance of a feature. We 16 feature subset selection algorithms based on the frame of mRMR with maximizing AUC while incorporating the aforementioned innovations. The mean results of 50 repeated experiments on 7 classical binary unbalanced gene expression datasets and 3 multi-class unbalanced or approximately balanced gene expression datasets proved that the developed algorithms in this paper can detect the gene subsets with superior classification power, and all the innovations proposed in this paper have got their superior capacities. Furthermore, the experimental results also demonstrate that the normalized mutual information SU is superior to the Spearman's rank correlation coefficient in evaluating the redundant of genes. At the same time that the weight of a gene by the difference of its correlation to labels and the redundant between genes is overwhelming the quotient of its correlation to labels divided by the redundant between genes. Our proposed gene subset selction algorithms defeat those available ones when compared to them.

Keywords　gene selection; AUC; mutual information; mRMR; unbalanced datasets

1　引　言

高维小样本癌症基因数据集为癌症预防、诊疗和致病机理研究提供了依据，但是该类数据经常包含大量与疾病无关或冗余的基因，因此，分析挖掘癌症患者与正常人的差异表达基因是生物医学研究的重点[1-4]，也是机器学习和数据挖掘领域的热点问题[1-7]. 然而该类数据中患者与正常人样本分布的不平衡给机器学习算法带来了严峻挑战[2,8].

癌症基因数据集中患者(正类)样本经常远少于正常人(负类)样本，类别分布偏斜. 分析该类数据时往往更关注异常样本，即患者，正确识别癌症患者是分析该类数据的主要目标. 传统机器学习算法倾向于学习大类而忽略小类，所得分类器具有较大偏向性，分类效果欠佳. 类别不平衡问题的常用处理方法包括样本重采样、代价敏感学习和特征选择等[9-10]. 特征选择从全部基因中发现最具识别能力的基因，可减轻类别不平衡对高维数据分析结果的影响[9]，有助于提高模型识别准确率，降低维数灾难、消除冗余基因、降低存储开销[11-12]. 因此，变量(特征)选择是不平衡癌症基因数据分析的首要任务. 样本类别分布均衡是以分类正确率为评价指标的基因选择算法的前提条件，但该条件在高维癌症基因表达数据经常不成立，因此，分类正确率不再适用于该类数据的特征子集性能评价. 现有算法层面和数据层面的特征选择研究，多针对不平衡文本数据[8,13-17].

为了解决样本分布不均衡癌症基因数据的特征选择问题，提出 $F2$-measure 准则(简称 $F2$)避免准确率评价的缺陷；提出归一化互信息 SU(Symmetrical Uncertainty)避免传统互信息偏爱多值特征的问题；将归一化互信息 SU 与 AUC(Area Under an ROC Curve)结合进行基因预选择，剔除部分与分类无关基因，缩小特征搜索空间；提出归一化特征权重，避免特征与类标相关性和特征冗余性的取值范围(量纲)不一致；提出 $F2$ 和 AUC 结合度量特征重要度的 DWSFS(Dynamic Weighted Sequential Forward Search)和 DWSFFS(Dynamic Weighted Sequential Forward Floating Search)，以发现分类能力更好的特征构成特征子集；基于 mRMR (minimal Redundancy-Maximal Relevance)[15]框架和最大化 AUC，以及上述创新点，提出 16 种针对癌症基因数据差异表达识别的基因选择算法.

以 SVM(Support Vector Machines)和 NB(Naive Bayes)为分类工具，采用 bootstrap[4,18]划分数据集为训练集和测试集，7 个经典二类不平衡基因数据集和 3 个多类不平衡(或近似平衡)基因数据集的 50 次实验测试，以及与现有经典基因选择算法的性能比较表明：提出的准则及基因选择算法能实现类别不平衡基因数据的差异表达基因识别，且提出算法的性能优于现有算法.

2 本文贡献

2.1 *F2*-measure 评价准则

AUC 和 *F*-measure 是癌症基因数据的差异表达基因分类能力的常用评价准则[9,14]. *F*-measure 的定义见式(1),是正类样本查准率和查全率的调和平均,其计算由表 1 所示混淆矩阵可得.

表 1 混淆矩阵

	分类为正类样本 P'	分类为负类样本 N'
实际为正类样本 P	TP	FN
实际为负类样本 N	FP	TN

$$\text{查准率}:precision=\frac{TP}{P'},$$

$$\text{查全率}:sensitivity=\frac{TP}{P},$$

$$F\text{-measure}=2\times\frac{precision\times sensitivity}{precision+sensitivity}=2\times\frac{\frac{TP}{P'}\times\frac{TP}{P}}{\frac{TP}{P'}+\frac{TP}{P}}=2\times\frac{TP}{P+P'} \tag{1}$$

由式(1)可见,*F*-measure 过分偏爱正类而忽略了负类信息,因此我们定义式(2)的 *F*2-measure(简称 *F*2),其中,$\sim precision=\frac{TN}{N'}$,表示分类器对负类的查准率.

$$F2\text{-measure}=2\times\frac{precision\times(\sim precision)}{precision+(\sim precision)}=2\times\frac{\frac{TP}{P'}\times\frac{TN}{N'}}{\frac{TP}{P'}+\frac{TN}{N'}}=\frac{2}{2+\frac{FN}{TN}+\frac{FP}{TP}} \tag{2}$$

由式(2)可知,*F*2 是正、负两类查准率的调和平均,因此 *F*2 更适合评价类别不平衡数据.

2.2 归一化互信息 *SU*

特征选择期望发现这样的特征:与类别标签相关性强,且其间没有冗余[15,19]. 互信息是衡量特征间相关(或冗余)程度的常用指标,其计算见式(3).

$$I(\boldsymbol{f}_i;\boldsymbol{f}_j)=\sum_{\boldsymbol{f}_j}\sum_{\boldsymbol{f}_i}p(\boldsymbol{f}_i;\boldsymbol{f}_j)\log\frac{p(\boldsymbol{f}_i;\boldsymbol{f}_j)}{p(\boldsymbol{f}_i)p(\boldsymbol{f}_j)} \tag{3}$$

式(3)的 $\boldsymbol{f}_i,\boldsymbol{f}_j$ 是任意两个特征向量,也可以是类标列 $\boldsymbol{Y}$,$p(\boldsymbol{f}_i;\boldsymbol{f}_j)$ 是特征向量 $\boldsymbol{f}_i,\boldsymbol{f}_j$ 的联合概率密度近似估计,$p(\boldsymbol{f}_i)$ 是特征向量 $\boldsymbol{f}_i$ 的概率密度近似估计. $I(\boldsymbol{f}_i;\boldsymbol{Y})$ 用于计算 $\boldsymbol{f}_i$ 与类标的相关性.

式(3)互信息方法倾向于选择多值特征[20],因此,提出式(4)的互信息归一化度量方法 *SU*.

$$SU(\boldsymbol{f}_i;\boldsymbol{f}_j)=0.5\times\left(\frac{I(\boldsymbol{f}_i;\boldsymbol{f}_j)}{H(\boldsymbol{f}_i)}+\frac{I(\boldsymbol{f}_i;\boldsymbol{f}_j)}{H(\boldsymbol{f}_j)}\right),$$

$$H(\boldsymbol{f}_i)=-\sum_{\boldsymbol{f}_i}p(\boldsymbol{f}_i)\log p(\boldsymbol{f}_i) \tag{4}$$

其中,$H(\boldsymbol{f}_i)$ 为特征向量 $\boldsymbol{f}_i$ 的熵.

2.3 特征预选择方法

ROC(Receiver Operating Characteristic)曲线[21]是描述分类模型性能的有效工具,曲线上一点是一个对应特定阈值的具体分类器,该分类器的"1-特异度"与"灵敏度"就是 ROC 曲线上的对应点坐标. 一个分类模型,通过调整阈值,得到一系列具体的分类器,将各分类器的"1-特异度"与"灵敏度"在以"1-特异度"为横坐标、"灵敏度"为纵坐标的 2 维空间散列出来并以线相连,就得到相应模型的 ROC 曲线. 为了定量描述分类模型的性能,ROC 曲线下面积 *AUC* 被用来度量相应分类模型的性能. *AUC* 的取值范围为[0,1],当 *AUC* 取最大值 1 时,所得分类模型最优,能识别所有样本,反之,当 *AUC* 取 0 值时,所得分类模型最差,将所有样本错分.

特征 $\boldsymbol{f}_j$ 对应的 *AUC* 值可用来度量该特征的类别区分能力. 此时 *AUC* 计算采用式(5)所示的简化计算[14].

$$AUC_{f_j}=\frac{\sum_{i=1}^{n}(r_i)-\frac{P\times(P+1)}{2}}{P\times N} \tag{5}$$

式(5)的 *P*、*N* 分别是正、负类样本数,$n=P+N$,r_i 为第 i 个样本按 $\boldsymbol{f}_j$ 的降序序号,最小序号为 1.

为了预先剔除部分不相关基因,减少候选基因数量,减少特征选择算法运行时间,并减少存储需求,我们分别采用式(4)和(5)计算特征的类别区分能力,并分别降序排序,从两个特征与类别相关性降序序列中依次交替取前 10%~20%的特征构成候选特征集合.

2.4 基因权重归一化方法

算法 ARCO[14]采用式(6)的特征 $\boldsymbol{f}_i$ 与类标 $\boldsymbol{Y}$ 的相关性和式(7)的特征 $\boldsymbol{f}_i$ 与已选特征 $\boldsymbol{f}_j$($\boldsymbol{f}_j\in\boldsymbol{S}$,$\boldsymbol{S}$ 表示被选特征集合)的冗余性之差,衡量特征 $\boldsymbol{f}_i$ 的分类性能,实现特征选择.

$$fscore\_Y=|AUC_{f_i}-0.5| \tag{6}$$

$$fscore\_S=\frac{1}{|\boldsymbol{S}|}\sum_{\boldsymbol{f}_j\in\boldsymbol{S}}RCC(\boldsymbol{f}_i,\boldsymbol{f}_j) \tag{7}$$

式(7)的 *RCC*(Spearman's Rank Correlation Coefficient)是斯皮尔曼等级相关系数[22],$|\boldsymbol{S}|$ 表示已选特征集合 $\boldsymbol{S}$ 包含的特征数. 分析式(6)、(7)可得:$fscore\_Y\in[0,0.5]$,$fscore\_S\in[0,1]$,取值范

围的不同将可能导致所选特征子集并非全局最优.为此提出归一化特征权重思想,计算方法见式(8).

$$\boldsymbol{X}'=\min x+\frac{(1-\min x)\times(\boldsymbol{X}-\min x)}{(\max x-\min x)} \tag{8}$$

其中,$\boldsymbol{X}$ 是待归一化特征权重构成的向量,$\max x$、$\min x$ 分别其最大、最小值,$\boldsymbol{X}'\in[\min x,1]$为归一化后结果.

2.5　基于 mRMR 框架的特征选择算法

以 mRMR 最小冗余最大相关为基本框架,结合提出的 SU 和经典互信息 I,提出表 2 的基因(特征)选择算法.其中,F 和 S 分别是预选择的候选基因集合和已选择基因集合.表 2 的 MID 算法和文献[15]的 mRMR 算法的区别在于,前者在特征选择之前进行了本文提出的基因预选择.

采用 AUC 度量基因与样本标签相关性,采用 RCC 和本文提出的 SU 计算基因与已选基因间的冗余性,基于 mRMR 最小冗余最大相关思想和式(8)的基因权重归一化,提出表 3 的特征选择算法.表 3 的 ARCD 算法与已有 ARCO 算法相比,增加了基因预选择与基因权重归一化步骤.表 3 的 F、S 与表 2 相同,分别表示预选择的候选基因集合和已选择基因集合.

表 2　基于互信息与 mRMR 框架的 4 种特征选择算法

算法缩写	算法全称	特征权值
MID	Mutual Information Difference	$\max\limits_{\boldsymbol{f}_i\in F-S}\left\{I(\boldsymbol{f}_i,\boldsymbol{Y})-\frac{1}{\lvert S\rvert}\sum\limits_{\boldsymbol{f}_j\in S}I(\boldsymbol{f}_i,\boldsymbol{f}_j)\right\}$
MIQ	Mutual Information Quotient	$\max\limits_{\boldsymbol{f}_i\in F-S}\left\{I(\boldsymbol{f}_i,\boldsymbol{Y})\Big/\frac{1}{\lvert S\rvert}\sum\limits_{\boldsymbol{f}_j\in S}I(\boldsymbol{f}_i,\boldsymbol{f}_j)\right\}$
nMID	normalized Mutual Information Difference	$\max\limits_{\boldsymbol{f}_i\in F-S}\left\{SU(\boldsymbol{f}_i,\boldsymbol{Y})-\frac{1}{\lvert S\rvert}\sum\limits_{\boldsymbol{f}_j\in S}SU(\boldsymbol{f}_i,\boldsymbol{f}_j)\right\}$
nMIQ	normalized Mutual Information Quotient	$\max\limits_{\boldsymbol{f}_i\in F-S}\left\{SU(\boldsymbol{f}_i,\boldsymbol{Y})\Big/\frac{1}{\lvert S\rvert}\sum\limits_{\boldsymbol{f}_j\in S}SU(\boldsymbol{f}_i,\boldsymbol{f}_j)\right\}$

表 3　基于最大化 AUC 与 mRMR 框架的 4 种特征选择算法

算法缩写	算法全称	特征权值
ARCD	AUC and Rank Correlation Coefficient Difference	$\max\limits_{\boldsymbol{f}_i\in F-S}\left\{(AUC(\boldsymbol{f}_i,\boldsymbol{Y})-0.5)-\frac{1}{\lvert S\rvert}\sum\limits_{\boldsymbol{f}_j\in S}RCC(\boldsymbol{f}_i,\boldsymbol{f}_j)\right\}$
ARCQ	AUC and Rank Correlation Coefficient Quotient	$\max\limits_{\boldsymbol{f}_i\in F-S}\left\{(AUC(\boldsymbol{f}_i,\boldsymbol{Y})-0.5)\Big/\frac{1}{\lvert S\rvert}\sum\limits_{\boldsymbol{f}_j\in S}RCC(\boldsymbol{f}_i,\boldsymbol{f}_j)\right\}$
AMID	AUC and Mutual Information Difference	$\max\limits_{\boldsymbol{f}_i\in F-S}\left\{(AUC(\boldsymbol{f}_i,\boldsymbol{Y})-0.5)-\frac{1}{\lvert S\rvert}\sum\limits_{\boldsymbol{f}_j\in S}SU(\boldsymbol{f}_i,\boldsymbol{f}_j)\right\}$
AMIQ	AUC and Mutual Information Quotient	$\max\limits_{\boldsymbol{f}_i\in F-S}\left\{(AUC(\boldsymbol{f}_i,\boldsymbol{Y})-0.5)\Big/\frac{1}{\lvert S\rvert}\sum\limits_{\boldsymbol{f}_j\in S}SU(\boldsymbol{f}_i,\boldsymbol{f}_j)\right\}$

2.6　特征搜索方法 DWSFS 和 DWSFFS

SFS(Sequential Forward Search)[23]不适于高维癌症基因数据的基因选择过程.为此,提出 DWSFS (Dynamic Weighted Sequential Forward Search),仅搜索权值较大的 top_k 个候选基因.结合 AUC 与提出的 $F2$,定义式(9)的特征子集性能评价准则 J,选择分类性能最强的基因.详细思想描述见算法 1.

$$J=\frac{AUC+F2}{2} \tag{9}$$

算法 1.　动态加权顺序前向搜索 DWSFS.

输入:$\boldsymbol{X}_{n\times m}$ 训练数据,样本标签 $\boldsymbol{Y}_{n\times 1}$,要选择基因数 k

输出:k 个被选择基因构成的集合 S

```
BEGIN
  S=∅; i=1;
  WHILE i≤m DO
    分别由式(4)、(5)计算基因 f_i 的 SU(f_i,Y)和 AUC;
  构造预选基因子集 ReT;
  max=0; i=1;
  WHILE i≤|ReT| DO
  BEGIN
    由式(6)计算基因 f_i 的 fscore_Y 并用式(8)归一化;
    IF fscore_Y(f_i)>max THEN
    BEGIN
      max=fscore_Y;
      sf=f_i;
    END //of IF
  END //of WHILE
  S=S∪{sf};
  MaxJ=J(S);
  WHILE |S|<k DO
  BEGIN
```

用表 3 和式(8)计算 ReT 中各基因的归一化权值，
并对基因降序排序；
$i=1$；
WHILE $i\leq top\_k$ DO
BEGIN
$J_i=J(S\cup\{\boldsymbol{f}_i\})$；
$find=$FALSE；
IF ($J_i>MaxJ$)&&(!$find$) TEHN
BEGIN
$S=S\cup\{\boldsymbol{f}_i\}$；
$find=$TRUE；
END //of IF
END //of WHILE
IF !$find$ THEN
BEGIN
$\boldsymbol{f}_i=\underset{\boldsymbol{f}_i\in ReT(top\_k)}{\arg\max}\ J_i$；
$S=S\cup\{\boldsymbol{f}_i\}$；
END //of IF；
$ReT=ReT-\{\boldsymbol{f}_i\}$；
$MaxJ=J(S)$；
END //of WHILE
END

算法 1 的 DWSFS 结合表 3 的 4 种特征选择算法，得到表 4 的 4 种基因选择算法.

表 4 基于 DWSFS 与不同最大化 *AUC* 方案的 4 种特征选择算法

算法缩写	算法全称	特征权值
ARCD-DWSFS	DWSFS using Dynamic AUC and Rank Correlation Coefficient Difference	$\max\limits_{\boldsymbol{f}_i\in F-S}\left\{(AUC(\boldsymbol{f}_i,\boldsymbol{Y})-0.5)-\frac{1}{\lvert S\rvert}\sum\limits_{\boldsymbol{f}_j\in S}RCC(\boldsymbol{f}_i,\boldsymbol{f}_j)\right\}$
ARCQ-DWSFS	DWSFS using Dynamic AUC and Rank Correlation Coefficient Quotient	$\max\limits_{\boldsymbol{f}_i\in F-S}\left\{(AUC(\boldsymbol{f}_i,\boldsymbol{Y})-0.5)\Big/\frac{1}{\lvert S\rvert}\sum\limits_{\boldsymbol{f}_j\in S}RCC(\boldsymbol{f}_i,\boldsymbol{f}_j)\right\}$
AMID-DWSFS	DWSFS using Dynamic AUC and Mutual Information Difference	$\max\limits_{\boldsymbol{f}_i\in F-S}\left\{(AUC(\boldsymbol{f}_i,\boldsymbol{Y})-0.5)-\frac{1}{\lvert S\rvert}\sum\limits_{\boldsymbol{f}_j\in S}SU(\boldsymbol{f}_i,\boldsymbol{f}_j)\right\}$
AMIQ-DWSFS	DWSFS using Dynamic AUC and Mutual Information Quotient	$\max\limits_{\boldsymbol{f}_i\in F-S}\left\{(AUC(\boldsymbol{f}_i,\boldsymbol{Y})-0.5)\Big/\frac{1}{\lvert S\rvert}\sum\limits_{\boldsymbol{f}_j\in S}SU(\boldsymbol{f}_i,\boldsymbol{f}_j)\right\}$

SFFS[24](Sequential Forward Floating Search)克服了 SFS 的子集嵌套包含缺陷. 故提出结合 SFFS 的 DWSFFS(Dynamic Weighted Sequential Forward Floating Search)对 DWSFS 进行改进，避免陷入局部最优解. DWSFFS 对加入特征子集的特征执行条件剔除：若 $\max\{J(S-\{\boldsymbol{f}_i\})\}$对应特征 $\boldsymbol{f}_i(i\leq|S|)$，满足 $\max(J(S-\{\boldsymbol{f}_i\}))>J(S)$，则从集合 S 中删除 $\boldsymbol{f}_i$. 算法 2 给出了 DWSFFS 的描述.

算法 2. 动态加权顺序前向浮动搜索 DWSFFS.
输入：$\boldsymbol{X}_{n\times m}$训练数据，样本标签 $\boldsymbol{Y}_{n\times 1}$，要选择基因数 k
输出：k 个被选择基因构成的集合 S
BEGIN
$S=\varnothing$；$i=1$；
WHILE $i\leq m$ DO
分别由式(4)和(5)计算基因 $\boldsymbol{f}_i$ 的 $SU(\boldsymbol{f}_i,\boldsymbol{Y})$和 AUC；
进行预选择获取候选特征子集 ReT；
$i=1$；
WHILE $i\leq|ReT|$ DO
由式(6)计算基因 $\boldsymbol{f}_i$ 的 $fscore\_Y$，用式(8)归一化；
$\boldsymbol{f}_i=\underset{\boldsymbol{f}_j\in ReT}{\arg\max}\{fscore\_Y(\boldsymbol{f}_j)\}$；
$S=S\cup\{\boldsymbol{f}_i\}$；
$MaxJ=J(S)$；
WHILE $|S|<k$ DO
BEGIN
用表 3 和式(8)计算 ReT 中各基因的归一化权值，
并对基因降序排序；
$i=1$；
WHILE $i\leq top\_k$ DO
BEGIN
$J_i=J(S\cup\{\boldsymbol{f}_i\})$；
$find=$FALSE；
IF ($J_i>MaxJ$)&&(!$find$) THEN
BEGIN
$S=S\cup\{\boldsymbol{f}_i\}$；
$find=$TRUE；
END //of IF
END //of WHILE
IF !$find$ THEN
BEGIN
$\boldsymbol{f}_i=\underset{\boldsymbol{f}_i\in ReT(top\_k)}{\arg\max}\ J_i$；
$S=S\cup\{\boldsymbol{f}_i\}$；
END //of IF；
$ReT=ReT-\{\boldsymbol{f}_i\}$；
$MaxJ=J(S)$；
$count=|S|$；
$Index=0$；
FOR $i=1$ to $count$ DO //浮动删除特征
BEGIN
IF $J(S-\{\boldsymbol{f}_i\})>MaxJ$ THEN
BEGIN
$MaxJ=J(S-\{\boldsymbol{f}_i\})$；

```
    Index=f_i;
  END //of IF
END //of FOR 浮动搜索
IF Index THEN
BEGIN
  S=S-{Index};
  MaxJ=J(S);
ELSE
  BREAK;
END //of IF
END //of WHILE
END
```

表 3 的特征选择算法与算法 2 的 DWSFFS 结合，得到表 5 的 4 种特征选择算法.

表 5　基于 DWSFFS 与不同最大化 *AUC* 方案的 4 种特征选择算法

算法缩写	算法全称	特征权值
ARCD-DWSFFS	DWSFFS using Dynamic AUC and Rank Correlation Coefficient Difference	$\max\limits_{f_i\in F-S}\left\{(AUC(\boldsymbol{f}_i,\boldsymbol{Y})-0.5)-\frac{1}{\lvert S\rvert}\sum\limits_{f_j\in S}RCC(\boldsymbol{f}_i,\boldsymbol{f}_j)\right\}$
ARCQ-DWSFFS	DWSFFS using Dynamic AUC and Rank Correlation Coefficient Quotient	$\max\limits_{f_i\in F-S}\left\{(AUC(\boldsymbol{f}_i,\boldsymbol{Y})-0.5)\Big/\frac{1}{\lvert S\rvert}\sum\limits_{f_j\in S}RCC(\boldsymbol{f}_i,\boldsymbol{f}_j)\right\}$
AMID-DWSFFS	DWSFFS using Dynamic AUC and Mutual Information Difference	$\max\limits_{f_i\in F-S}\left\{(AUC(\boldsymbol{f}_i,\boldsymbol{Y})-0.5)-\frac{1}{\lvert S\rvert}\sum\limits_{f_j\in S}SU(\boldsymbol{f}_i,\boldsymbol{f}_j)\right\}$
AMIQ-DWSFFS	DWSFFS using Dynamic AUC and Mutual Information Quotient	$\max\limits_{f_i\in F-S}\left\{(AUC(\boldsymbol{f}_i,\boldsymbol{Y})-0.5)\Big/\frac{1}{\lvert S\rvert}\sum\limits_{f_j\in S}SU(\boldsymbol{f}_i,\boldsymbol{f}_j)\right\}$

2.7　本文算法框架图

2.5～2.6 节在 2.1～2.4 节各创新点基础上，提出了 16 种特征选择算法，解决了非平衡基因数据的特征基因选择问题，并探讨了基因选择过程中的基因重要性度量的度量问题：采用基因与类标相关性减去还是除以基因之间冗余性. 图 1 给出了本文 16 种特征选择算法的逻辑框架. 这些算法的性能及本文各创新点的有效性将在第 4 节给予实验验证.

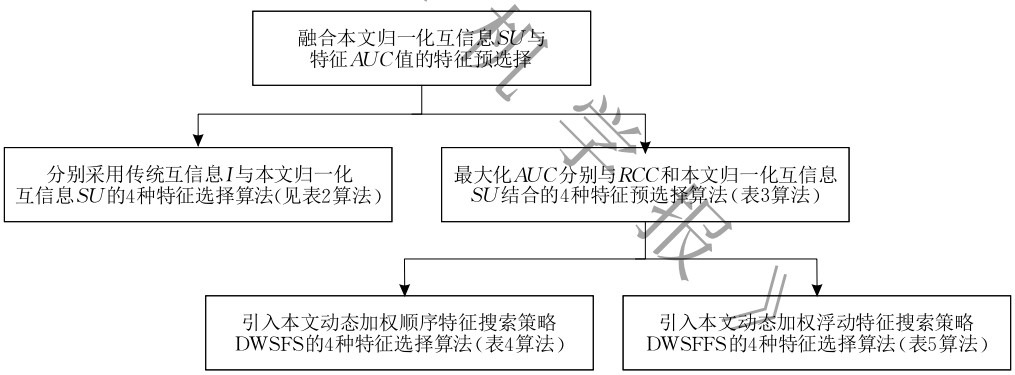


图 1　本文 16 种特征选择算法的逻辑关系

3　本文算法分析

设 m 和 n 分别表示训练集基因数和样本数. 本文算法首先使用 2.3 节提出的特征预选择剔除部分无关基因. 计算各基因的 SU 和 AUC 所需要时间分别是 $O(nm)$ 和 $O(n\log n)$；对基因降序排序的时间是 $O(m\log m)$. 由于本文基因数据的 $m\geq n$，因此预选择的时间复杂度为 $O(mn)$.

对表 2 算法，设预选基因数为 l，基因子集 S 的规模为 k，则计算待选基因与 S 中基因冗余性的时间复杂度是 $O(k(l-k)n)$；从预选基因中搜索满足最小冗余最大相关条件基因的时间复杂度是 $O(l)$，因此，表 2 算法的时间复杂度为 $O(k^2ln+kl)$. 通过记录各基因与已选基因相关性，使表 2 算法的时间复杂度降低为 $O(kln)$.

表 3 算法采用 AUC 度量基因的类别辨识能力，其中 ARCD 和 ARCQ 使用 RCC 方法计算待选基因与已选基因冗余性，AMID 和 AMIQ 采用提出的 SU 度量待选基因与已选基因的冗余性，因此，ARCD 和 ARCQ 的时间复杂度是 $O(kln\log n+kl)$；AMID 和 AMIQ 的时间复杂度是 $O(kln+kl)$.

表 4 各算法在表 3 算法基础上引入 DWSFS 策略，设 $top\_k=t$，表 4 前两算法和后两算法的基因类别辨识能力分别由 SVM 和 NB 分类器性能度量. SVM 和 NB 时间复杂度分别是 $O(kn^2)$ 和 $O(nk^2)$. 因此，算法 ARCD-DWSFS 和 ARCQ-DWSFS 的时间复杂度是 $O(kln\log n+kl+tk^2n^2)$，AMID-DWSFS

和 AMIQ-DWSFS 时间复杂度是 $O(kln+kl+tnk^3)$.

表 5 算法增加了对已选基因的有条件删除,使得使用 SVM 和 NB 分类器性能度量基因分类性能的时间复杂度分别变为 $O(n^2k^3)$ 和 $O(nk^4)$. 因此 ARCD-DWSFFS 和 ARCQ-DWSFFS 算法的时间复杂度是 $O(kln\log n+kl+k^3n^2)$, AMID-DWSFFS 和 AMIQ-DWSFFS 的时间复杂度是 $O(kln+kl+nk^4)$.

4 实验结果分析

我们首先在 7 个经典二类不平衡基因数据集上逐一测试提出的准则、策略与方法,并与已有同类研究进行比较;然后在 3 个多类基因数据集进一步测试比较本文算法与现有算法的性能. 7 个二类样本分布不平衡基因数据集①②③分别是:Colon[25]、CNS (Central Nervous System Embryonal Tumor data)[26]、ALL/AML Leukemia[27]、GLI_85[28]、DLBCL (Diffuse large B-cell lymphoma) Tumor[29]、Lung Cancer[30] 和 Lung Cancer(Michigan)[31],其中 Colon 数据集保留 500 个预选择基因,CNS、Leukemia、DLBCL、Lung Cancer(Michigan)4 个数据集预选择后保留 700 个基因,GLI_85 预选择后保留 2000 个基因,Lung Cancer 保留 1300 个基因. 该 7 个两类不平衡数据集的详细信息见表 6. 对多类数据集,采用一对其余或一对一转化为二类问题处理,本文算法即可适用,其中多类数据集的 *AUC* 计算采用文献[32]的 *MAUC*(Multi-class AUC), *F*2 值是对应多个二类问题的 *F*2 值平均. 实验采用一对一将 $M(>2)$ 类问题转化为 $M(M-1)/2$ 个二类问题,由混淆矩阵得到 $M(>2)$ 类问题的 *F*2 值计算公式如式(10)所示.

$$F2=\frac{2}{M(M-1)}\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\frac{2}{2+\frac{M_{ij}}{M_{jj}}+\frac{M_{ji}}{M_{ii}}} \tag{10}$$

表 6 二类非平衡数据集信息描述

数据集	特征数	样本数	样本分布
Colon	2000	62	22:40
CNS	7129	90	30:60
ALL/AML Leukemia	7129	72	25:47
GLI_85	22283	85	26:59
DLBCL Tumor	7129	77	19:58
Lung Cancer	12533	181	31:150
Lung Cancer(Michigan)	7129	96	10:86

本文使用的 3 个多类基因数据集的详细信息如表 7 所示④. 实验中 SRBCT[33] 保留 500 预选择基因, Lung Cancer[34] 和 Leukemia-MLL[35] 均保留 1260 个预选择基因. 采用 Libsvm 工具箱[36] 的惩罚因子 $C=20$ 的线性 SVM 分类器,以及 MATLAB 工具箱的 NB 分类器. 实验硬件环境为 8GB 内存的 PC,CPU 为 Intel(R)i5-6600@3.30GHz 3.31GHz. 所有实验均重复 50 次,比较各算法平均性能.

表 7 多类非平衡数据集信息描述

数据集	特征数	样本数	样本分布
SRBCT	2308	83	29:11:18:25
Lung Cancer	12600	203	139:20:6:21:17
Leukemia-MLL	12582	72	24:20:28

4.1 数据集划分与预处理

训练集和测试集使用 bootstrap[4,18] 划分得到. 数据集 ALL/AML Leukemia 已划分好训练集和测试集,但为了得到具有统计意义的实验结果,我们合并其训练集和测试集,然后对其随机打乱,再使用 bootstrap 划分得到训练集和测试集. 以最大最小标准化法预处理数据;采用算法 CAIM(Class-Attribute Interdependence Maximization)[37] 将数据离散化.

4.2 特征预选择方法与评价准则 *F*2 有效性验证

本小节将测试 2.3 节提出的特征预选择方法,并验证 2.1 节提出的 *F*2 准则. 我们对 CNS 数据集通过本文 2.3 节的方法选择前 700 个权值较大的基因构成预选基因子集,然后对包含该 700 个基因的 CNS 数据集和原始 CNS 数据集分别使用本文表 2 提出的 MID 和 MIQ 算法进行基因选择. 采用 bootstrap 方法获得训练集和测试集,图 2 给出了 MID 和 MIQ 算法 50 次重复运行选择到的基因子集对应 SVM 和 NB 分类器的平均 *AUC* 值和平均 *F*2 值. 表 8 展示了各算法 50 次重复运行的平均时间.

表 8 各算法在 CNS 数据集重复运行 50 次的平均时间

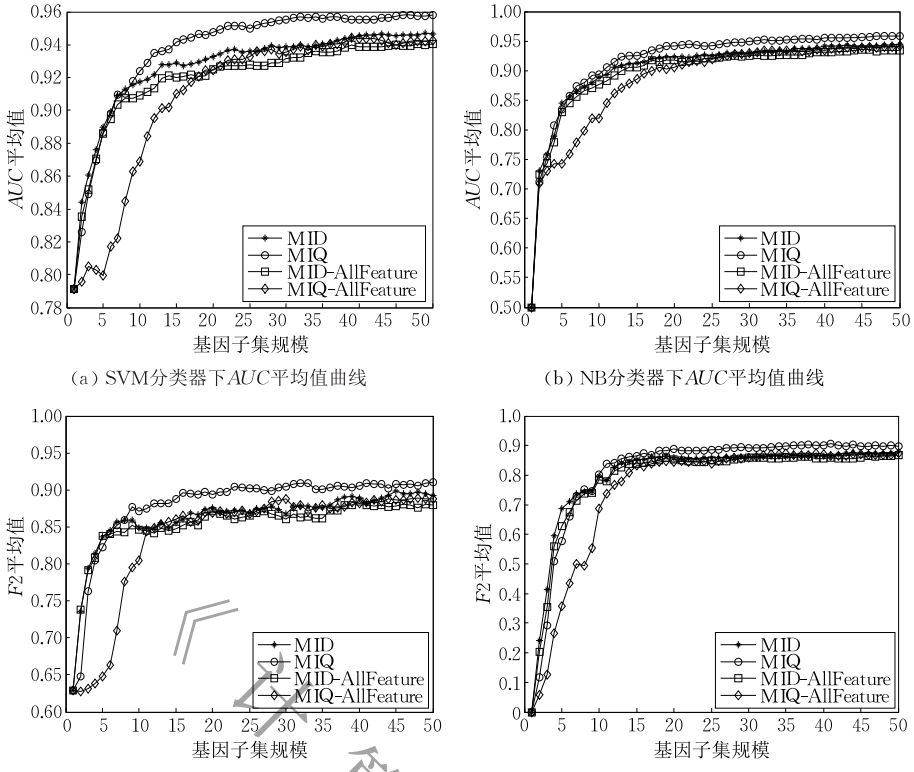
(单位:s)

MID	MIQ	MID-AllFeature	MIQ-AllFeature
5.52	5.42	57.98	56.66

从图 2 实验结果可以看出,算法 MID 和 MIQ 在包含 700 个预选基因的 CNS 数据选择的差异基因的分类性能优于在原始 CNS 数据选择的差异基因的分类性能,且 MIQ 选择的基因性能更优. 因此,本文提出的基因预选择方法可以有效剔除与分类任务无关或冗余基因,有助于基因选择算法选择到类别区分能力好的差异基因.

另外,图 2 实验结果还显示:*F*2 均值曲线和

① http://datam.i2r.a-star.edu.sg/datasets/krbd/
② http://featureselection.asu.edu/datasets.php
③ http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi
④ http://www.gems-system.org/



(a) SVM分类器下*AUC*平均值曲线　(b) NB分类器下*AUC*平均值曲线

(c) SVM分类器下*F*2平均值曲线　(d) NB分类器下*F*2平均值曲线

图 2　特征预选择及 *F*2 准则性能验证结果

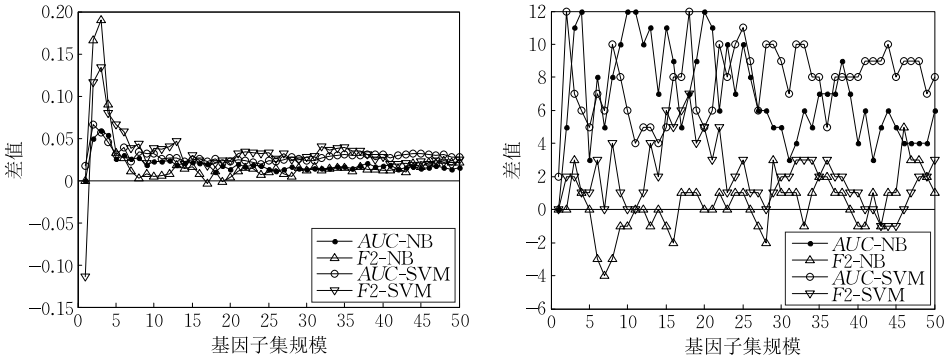
AUC 均值曲线的变化相似，但 *F*2 的取值低于 *AUC*. 这说明本文提出的 *F*2 准则可以有效评价基因子集的分类性能. 其值低于 *AUC* 的原因是：*F*2 度量的是基于被选基因子集的分类器对测试集正、负类样本的识别能力(即正类识别精度 *precision*(正类查准率)和负类识别精度～*precision*(负类查准率))的调和平均；而 *AUC* 度量了相应分类器对测试集正类样本识别能力(正类识别准确率(*Sensitivity*))和负类样本误识率(1－*Specificity*)之间的权衡.

表 8 的 MID 和 MIQ 算法分别在特征预选择前、后的 CNS 数据的 50 次重复运行的平均时间比较显示：在只含预选择基因的 CNS 数据进行特征选择的时间约是在原始 CNS 数据进行基因选择的 1/10. 这说明本文 2.3 节提出的特征预选择大大缩短了算法的时间需求.

综上可见，本文 2.3 节提出的基因预选择能识别并剔除分类能力较弱的基因，提升基因选择算法性能和效率；本文 2.1 节提出的 *F*2 准则能有效评估基因选择算法所选基因子集的性能.

4.3　特征权重归一化有效性验证

我们通过在 Colon 数据集进行实验，通过比较特征权重归一化算法 ARCD 和未归一化算法 ARCO 所选基因子集对应分类器的 *AUC* 值的差值，及其 *F*2 评价准则值的差值，来验证我们提出的归一化特征权重在解决基因权重计算过程中出现的量纲不一致问题的有效性. 图 3(a)～(b)分别为 ARCD 和



(a) *AUC*、*F*2平均值差值　(b) *AUC*、*F*2≥0.92次数的差值

图 3　归一化特征权重性能验证结果

ARCO 算法的平均 AUC、$F2$ 值差值及其 50 次重复运行中 AUC、$F2$ 均值大于 0.92 的次数的差值.

从图 3 实验结果可以看出,特征权值归一化的 ARCD 选择的基因子集对应分类器的 AUC 和 $F2$ 高于未进行特征权值归一化的 ARCO 选择的基因子集对应分类器的 AUC 和 $F2$,说明归一化基因权值对选择到分类性能更好的基因非常有效.

由此可见,归一化基因权值避免了基于 mRMR 框架并最大化 AUC 的基因选择算法的基因权值计算遇到的基因区分能力和冗余性取值范围(量纲)不一致问题,有助于选择到分类性能更好的基因.

4.4 归一化互信息 *SU* 有效性验证

本小节将通过表 2、表 3 各基因选择算法的性能测试和比较验证提出的归一化互信息 SU 用于判断基因区分能力大小和基因冗余性的有效性,并测试两种基因权值计算方法(基因的类别区分能力减去/除以基因冗余性)哪个更好.为节省篇幅,此处只给出 NB 分类器的实验结果.

4.4.1 归一化互信息 *SU* 与原始互信息 *I* 的比较

通过比较表 2 各算法对表 6 各数据所选基因子集的 NB 分类器的 AUC 和 $F2$ 均值来检测归一化互信息 SU 和原始互信息 I 的性能,验证了 2.2 节提出的归一化互信息 SU 的有效性.需要说明的是,根据审稿意见,我们补充了提出的归一化互信息 SU 优越于互信息 I 的理论证明,但因篇幅太长,此处省去了理论推导,只给出实验验证结果.图 4、图 5 展示了表 2 各算法在表 6 各数据集的实验结果.

从图 4 结果可以看出,提出的 SU 可有效度量基因的类别区分能力与冗余性.基于 SU 的 nMID 和

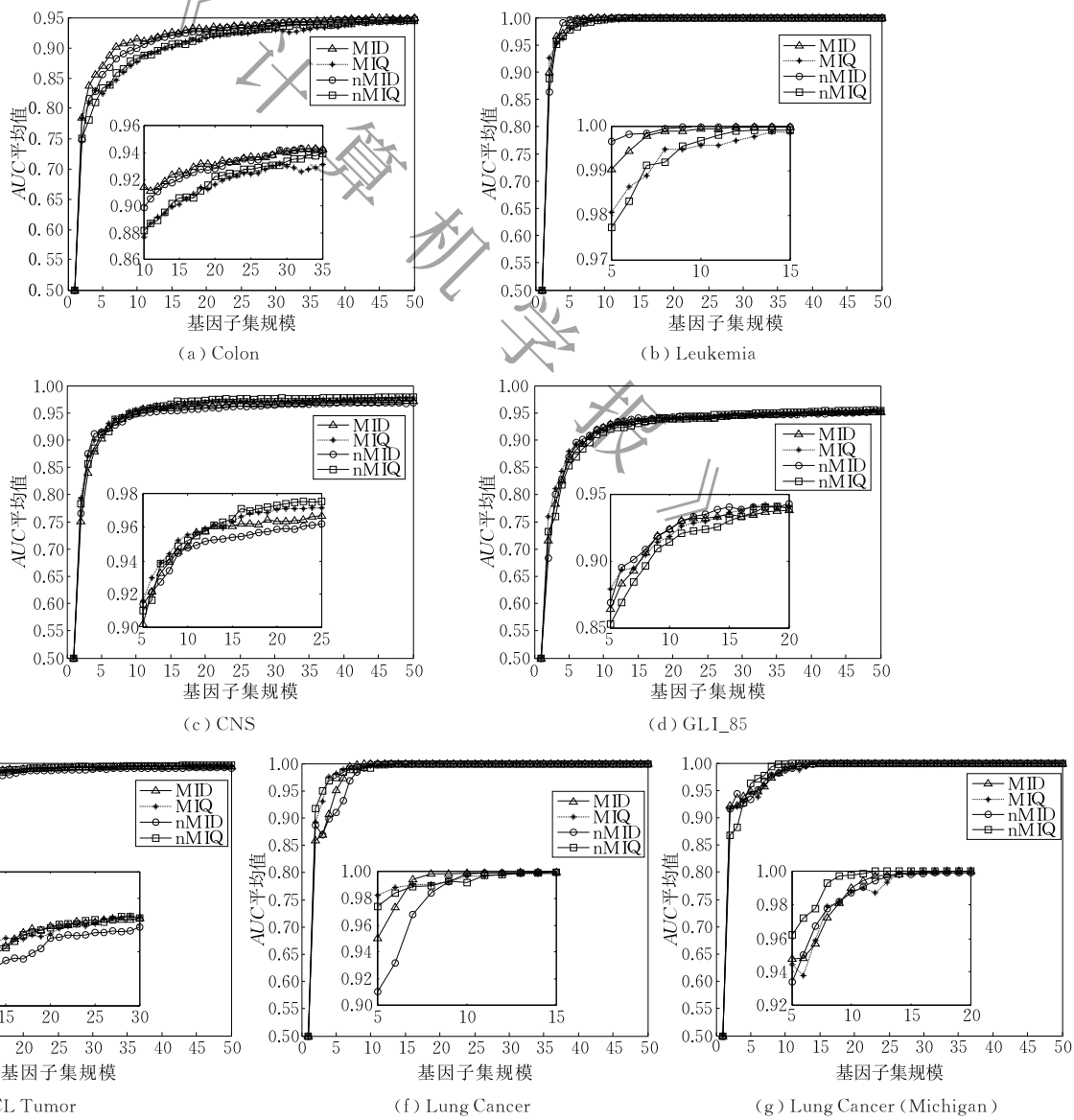


图 4 表 2 各算法在表 6 各数据集所选基因子集的 NB 分类器 AUC 均值曲线

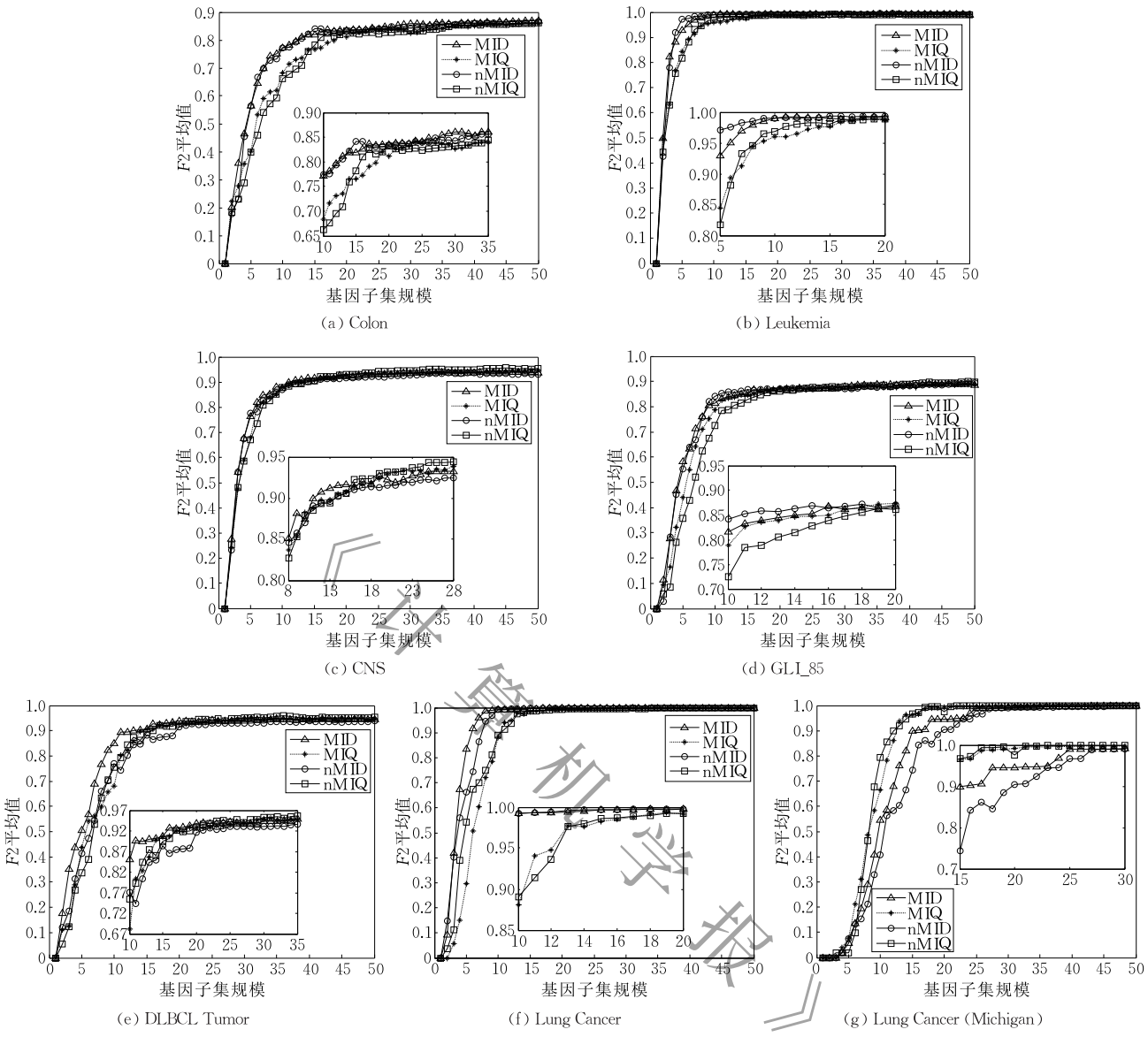


图 5　表 2 各算法在表 6 各数据集所选基因子集的 NB 分类器的 $F2$ 平均值曲线

nMIQ 算法能选择到区分能力非常好的差异表达基因. 在 Leukemia 数据集上，当基因子集规模大于 5，在 DLBCL、Lung Cancer 和 Lung Cancer(Michigan)数据集上，当基因子集规模大于 10 时，nMID 和 nMIQ 算法选择的基因子集的 NB 分类器的 AUC 值收敛到 1；在 Colon、CNS 和 GLI_85 数据集，nMID 和 nMIQ 算法选择的基因子集对应 NB 分类器的平均 AUC 值收敛到 0.95 的好结果.

图 5 实验结果揭示：基于 SU 的基因选择算法 nMID、nMIQ 选择的基因子集的平均 $F2$ 值较优. 在 Colon、Leukemia、GLI_85 和 Lung Cancer 数据集，nMID 算法选择的基因子集的平均 $F2$ 值最优. 在 Lung Cancer(Michigan)数据集，当基因子集规模较小时，nMIQ 算法的 $F2$ 值最优，但随着基因子集规模的增大，各算法的 $F2$ 值收敛到一致. 在 Leukemia、Lung Cancer 和 Lung Cancer(Michigan)数据集，各算法选择的基因子集的 NB 分类器的平均 $F2$ 值收敛到 1，意味着各算法选择的基因子集可以正确识别所有样本. 由此可见，2.2 节提出的归一化互信息 SU 的有效性.

综合图 4、图 5 关于表 2 各算法在表 6 各数据集的实验结果得出：我们 2.2 节提出的 SU 可以有效度量基因的类别辨识能力和基因冗余性，优于传统互信息方法 I，可选择到分类性能非常好的基因.

4.4.2　SU 与 RCC 的性能比较

本小节通过表 3 各算法在表 6 各数据集所选基因子集的 NB 分类器的 AUC 和 $F2$ 比较，比较 SU 和 RCC 度量基因冗余性的能力，从而验证 2.2 节提出的归一化互信息 SU 的有效性. 图 6、图 7 是表 3 各算法的实验结果.

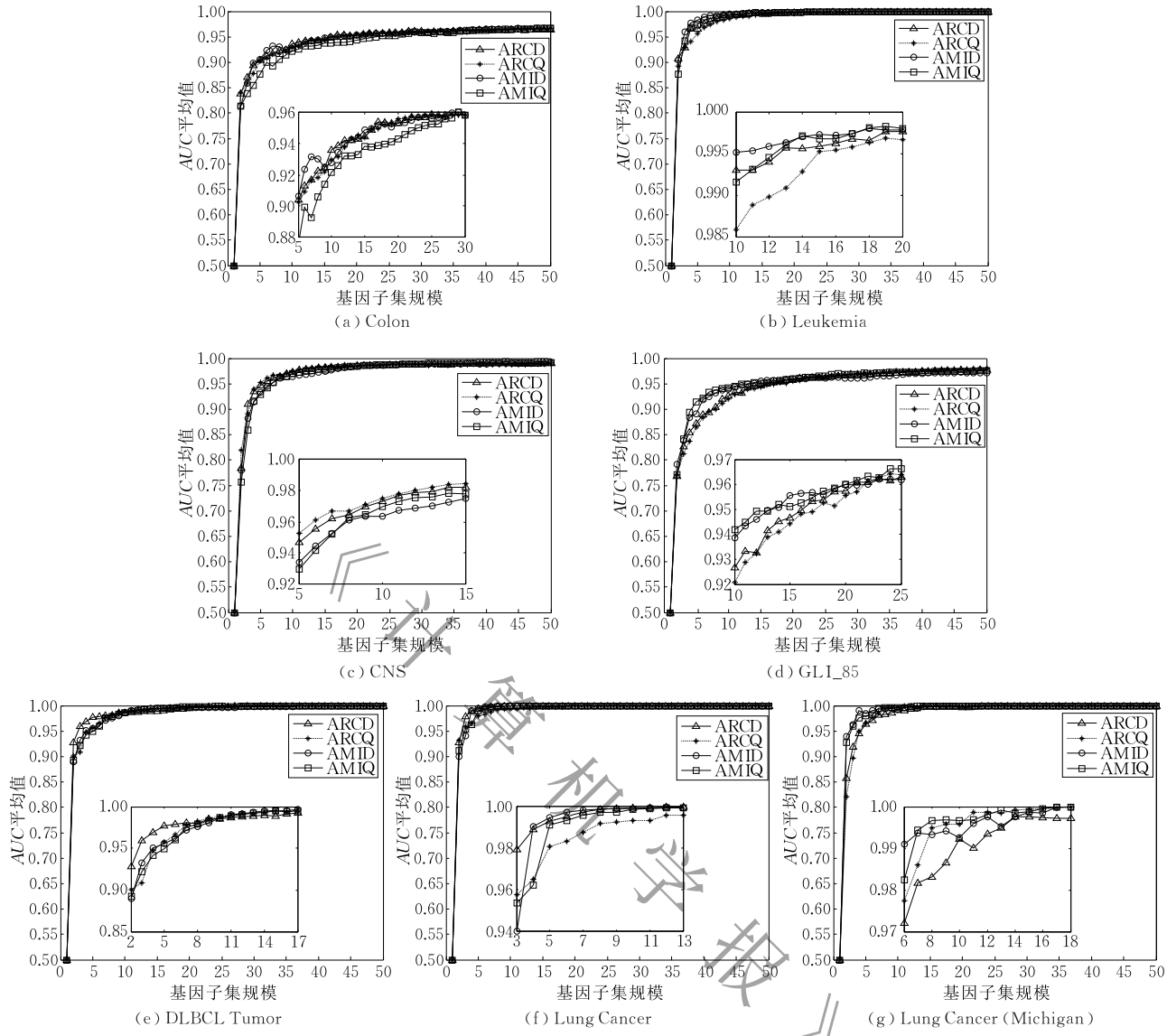


图 6 表 3 各算法在表 6 各数据集所选基因子集的 NB 分类器的 *AUC* 均值曲线

图 6 实验结果显示:表 3 各算法对表 6 各数据集均能选择到具有强分类能力的基因子集,建立在其上的分类器的平均 *AUC* 值均超过了 0.95,在 Leukemia、DLBCL Tumor、Lung Cancer 和 Lung Cancer(Michigan)数据集表现非常优,算法 ARCD、ARCQ、AMID 和 AMIQ 所选基因子集的平均 *AUC* 值均稳定收敛到 1,实现了对不同癌症患者的完全正确识别.

另外,图 6 实验结果还揭示:AMID 和 AMIQ 算法选择的基因子集对应分类器的平均 *AUC* 总体略高于 ARCD 和 ARCQ 算法的 *AUC*,但对 CNS 和 DLBCL Tumor 数据集,当基因子集规模较小时例外.

以上实验结果分析揭示:2.2 节提出的归一化互信息 *SU* 用于度量基因之间的相关性(基因冗余性)非常有效,优于采用 *RCC* 度量.

图 7 实验结果显示:表 3 各算法对表 6 数据集均能选择到具有强分类能力的基因子集,建立在其上的分类器的平均 *F*2 值稳定收敛到大于等于 0.8.其中 Colon 数据集最差,在 Leukemia、Lung Cancer 和 Lung Cancer(Michigan)数据集表现最优,各算法选择的基因子集的平均 *F*2 值均稳定收敛到 1.图 7 实验结果还揭示:ARCD、AMID 算法选择的基因子集对应分类器的平均 *F*2 值分别优于 ARCQ 和 AMIQ 算法选择的基因子集对应分类器的平均 *F*2.AMID 和 AMIQ 算法在多数情况下选择的基因子集对应分类器的 *F*2 值优于 ARCD 和 ARCQ 算法,除了在 Colon、Lung Cancer(Michigan)和 Lung Cancer 数据集,当选择较少的差异表达基因时,算法 ARCD 和 ARCQ 较优;特别是在 Lung Cancer

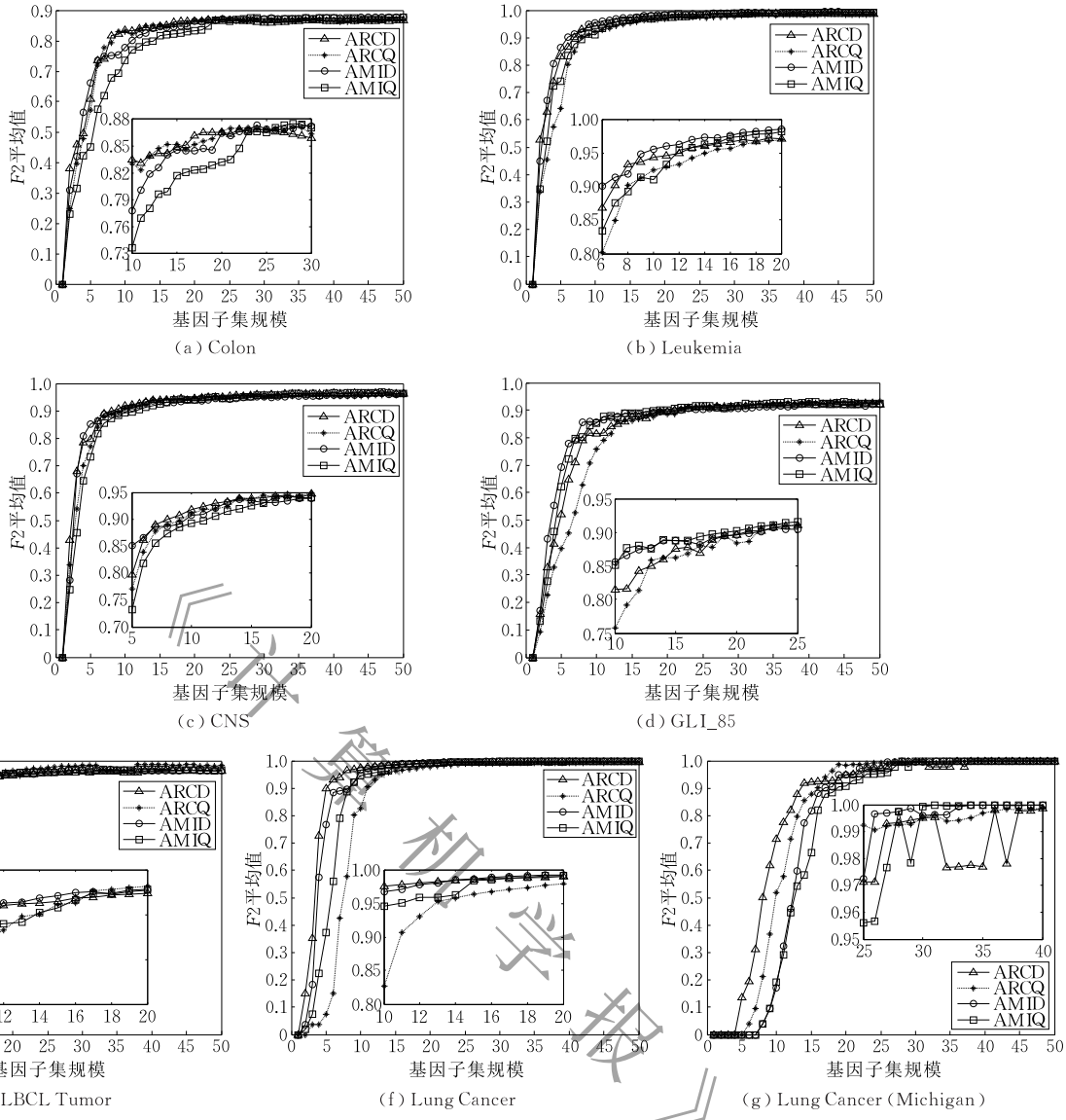


图 7　表 3 各算法在表 6 各数据集所选基因子集的 NB 分类器的 $F2$ 平均值曲线

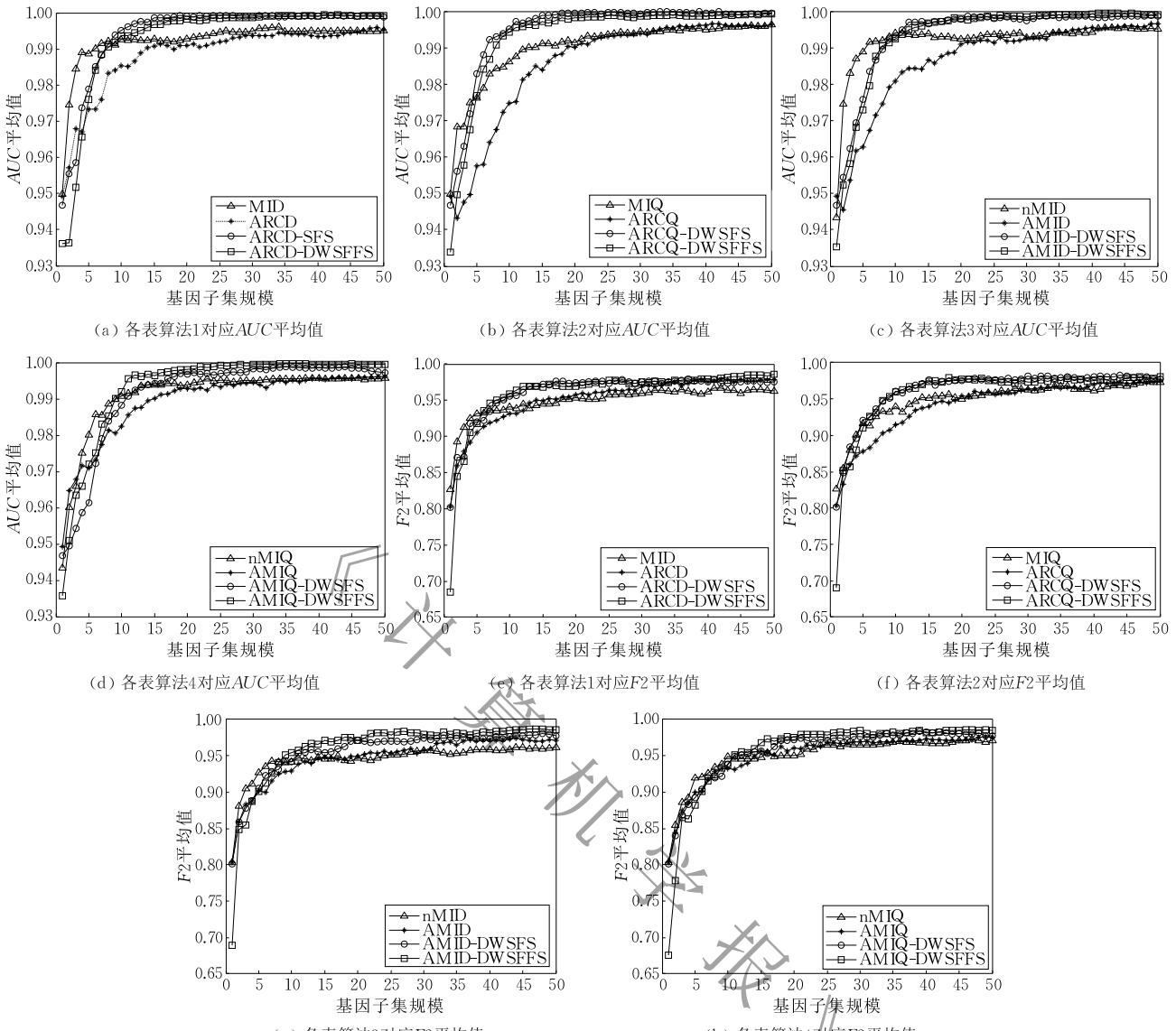
(Michigan)数据集，当选择的基因子集规模较小时，ARCD 最优，其次是 ARCQ，AMID 和 AMIQ 则相对略弱，但随着基因子集规模的增大，AMID 非常好，最终 AMIQ 也表现出很好性能.

综合图 6、图 7 关于表 3 各算法在表 6 各数据集所选基因子集的分类器的 AUC 和 $F2$ 平均值实验结果分析得出：本文 2.2 节提出的归一化互信息 SU 是一种度量变量之间相关性的有效方法，以此度量基因之间的冗余性比采用 RCC 度量基因之间冗余性更有效. 同时，基因权值计算采用基因与类标的相关性减去基因之间的相关性更好.

4.5　DWSFS 和 DWSFFS 搜索策略有效性验证

为了验证提出的 DWSFS 和 DWSFFS 的有效性，我们以表 6 的 ALL/AML Leukemia 数据集为例，选用 SVM 分类器，横向比较表 2～表 5 中各算法的性能. 在这里，各表的第 1～4 个算法依次命名为算法 1～算法 4. 本文提出的 16 种算法在 ALL/AML Leukemia 数据集的实验结果如图 8 所示. 图 8(a)和(e)、图 8(b)和(f)、图 8(c)和(g)、图 8(d)和(h)分别展示了表 2～表 5 的第 1～4 个算法选择的基因子集对应 SVM 分类器的平均 AUC 值和平均 $F2$ 值.

图 8 实验结果揭示：当基因子集规模非常小时，表 2 各算法所选基因子集分类性能最好；随着基因子集规模增大，表 4、表 5 的分别融合搜索策略 DWSFS 和 DWSFFS 的基因选择算法优于表 2、表 3 的算法，选择的基因子集分类能力更强，且收敛速度也远快于表 2、表 3 各算法. 由此可见，本文 2.6 节提出的特征搜索策略 DWSFS 和 DWSFFS 在选择类别不平衡基因数据集的差异表达基因时非常有效.



(a) 各表算法1对应*AUC*平均值 (b) 各表算法2对应*AUC*平均值 (c) 各表算法3对应*AUC*平均值

(d) 各表算法4对应*AUC*平均值 (e) 各表算法1对应*F*2平均值 (f) 各表算法2对应*F*2平均值

(g) 各表算法3对应*F*2平均值 (h) 各表算法4对应*F*2平均值

图 8 表 2～表 5 各基因选择算法在 Leukemia 数据集所选基因子集对应 SVM 分类器的平均 *AUC* 值和 *F*2 值曲线

4.6 本文算法与现有基因选择算法性能比较

由前面实验分析得知：2.2 节提出的归一化互信息 *SU* 是非常好的基因相关性度量准则，基因权值计算采用基因与类标相关性减去基因之间冗余性更优，2.6 节的特征搜索策略能加快算法搜索到更优特征子集，因此，本小节以基于归一化互信息 *SU* 的 nMID 和 AMID 算法，以及分别结合特征搜索策略 DWSFS、DWSFFS 的 AMID-DWSFS 和 AMID-DWSFFS 算法为例，即表 2～表 5 的算法 3，与现有经典基因选择算法 ARCO[14]、mRMR[15]、WAC Roulette wheel[4]、LLEScore[38]、Relief[39]、SVM-RFE[40]、Lasso[41] 等进行对比，以验证本文提出的基因选择算法的有效性，以及本文提出的归一化互信息 *SU*、基因权重归一化方法、特征搜索策略 DWSFS 与 DWSFFS、基因预选择方法等的有效性. 分类器采用 NB 分类器. 图 9、图 10 分别展示了本文基因选择算法与 ARCO、mRMR、WAC Roulette wheel、LLEScore、Relief、SVM-RFE 算法在表 6 各数据集所选基因子集的 NB 分类器的平均 *AUC* 值和平均 *F*2 值. 表 9 给出了本文算法与 Lasso 算法在表 6 各数据集所选基因子集的 NB 分类器的 *AUC* 和 *F*2 指标及对应基因子集规模比较. 本文 nMID、AMID、AMID-DWSFS 和 AMID-DWSFFS 算法的实验结果是 50 次实验所得相应指标平均值的最优值及其对应基因子集规模；Lasso 算法是 50 次实验中，每次自动停止时得到的基因子集对应 NB 分类器的平均 *AUC* 值和平均 *F*2 值，及自动停止时所得基因子集规模的平均值.

从图 9 实验结果可以看出：基于 2.2 节归一化互信息 *SU* 的基因选择算法 nMID 和 AMID、结合

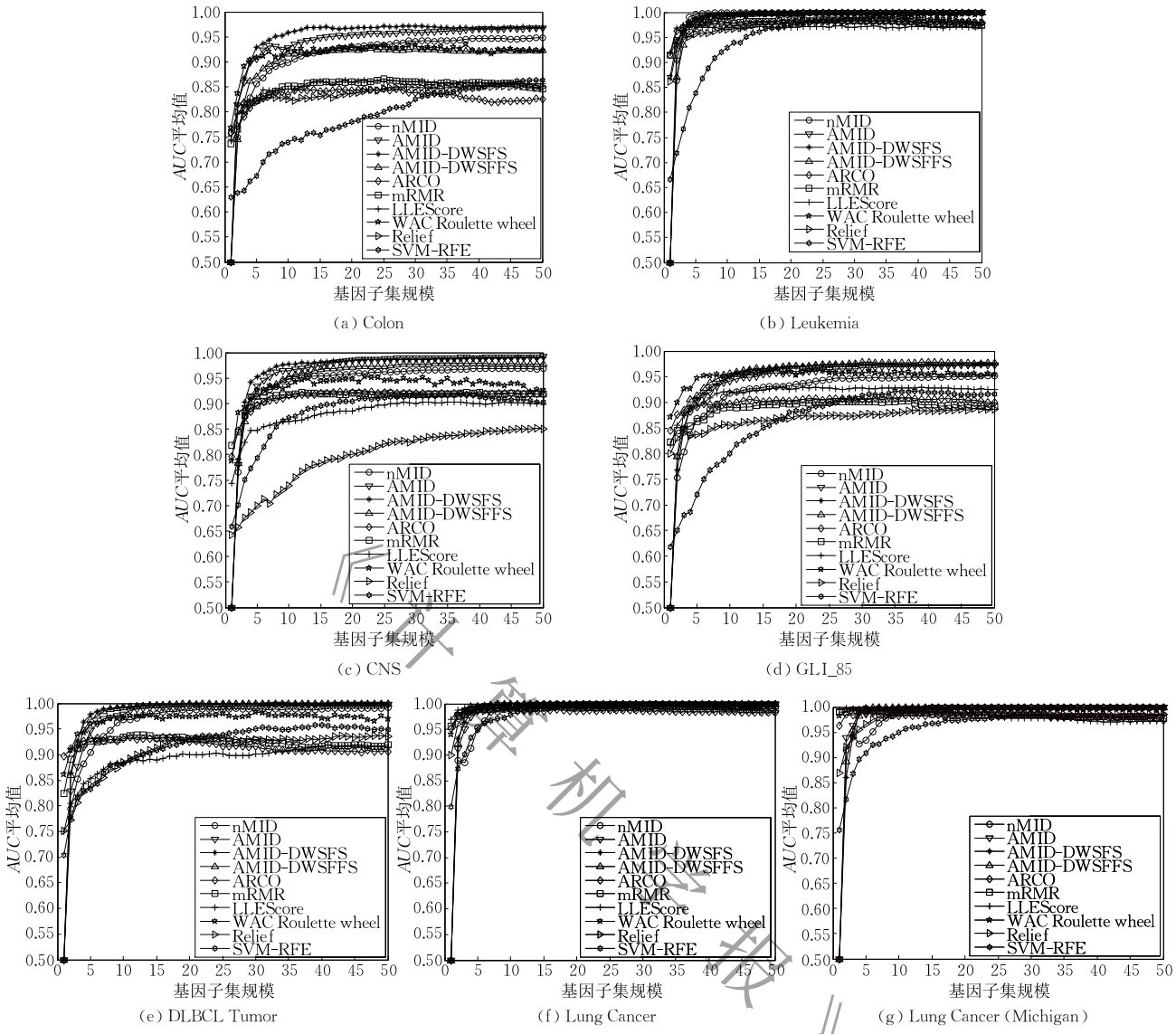


图 9　10 个特征选择算法在表 6 各数据集所得基因子集的 NB 分类器的 *AUC* 均值曲线

本文 2.6 节特征搜索策略 DWSFS、DWSFFS 的基因选择算法 AMID-DWSFS 和 AMID-DWSFFS 所选基因子集的分类性能要明显优于其他对比算法，对比算法 WAC Roulette wheel 选择的基因子集的平均 *AUC* 值位居第五，仅次于本文 4 种基因选择算法 nMID、AMID、AMID-DWSFS 和 AMID-DWSFFS.

图 10 各算法选择的基因子集对应 NB 分类器的平均 *F*2 值比较显示：在基因子集规模较小时，对比算法选择的基因子集的平均 *F*2 值较优，而本文 nMID、AMID、AMID-DWSFS 和 AMID-DWSFFS 算法在基因子集规模增大时，选择的基因子集对应分类器的平均 *F*2 值优于对比算法. 对比算法 WAC Roulette wheel 选择的基因子集对应分类器的平均 *F*2 值和 *AUC* 值一样稳定且较优. 分析原因是：WAC Roulette wheel 算法[4] 采用 *K*-means 算法对基因进行聚类，对各类簇分别训练 SVM 分类器，分别计算基因的辨识能力，然后使用轮盘赌策略从各类簇选取辨识能力最强的基因组成基因子集，因此所得基因子集的分类性能很稳定. 本文算法首先预选择与类标高度相关的基因，然后基于 mRMR 思想逐步选择与类标相关度高且与已选择基因尽可能不相关的基因构成基因子集.

从图 9、图 10 的实验结果分析可见：(1) 提出的算法选择的基因子集的分类识别能力很好；(2) 提出的归一化互信息 *SU* 在度量基因相关性时非常有效；(3) 提出的特征搜索策略 DWSFS、DWSFFS 在选择不平衡数据集的差异表达基因时非常有效；(4) 提出的归一化基因权值方法很有效；(5) 基因权值计算采用基因区分能力减去其冗余性更有效；(6) 提出的基因预选择方法非常有效.

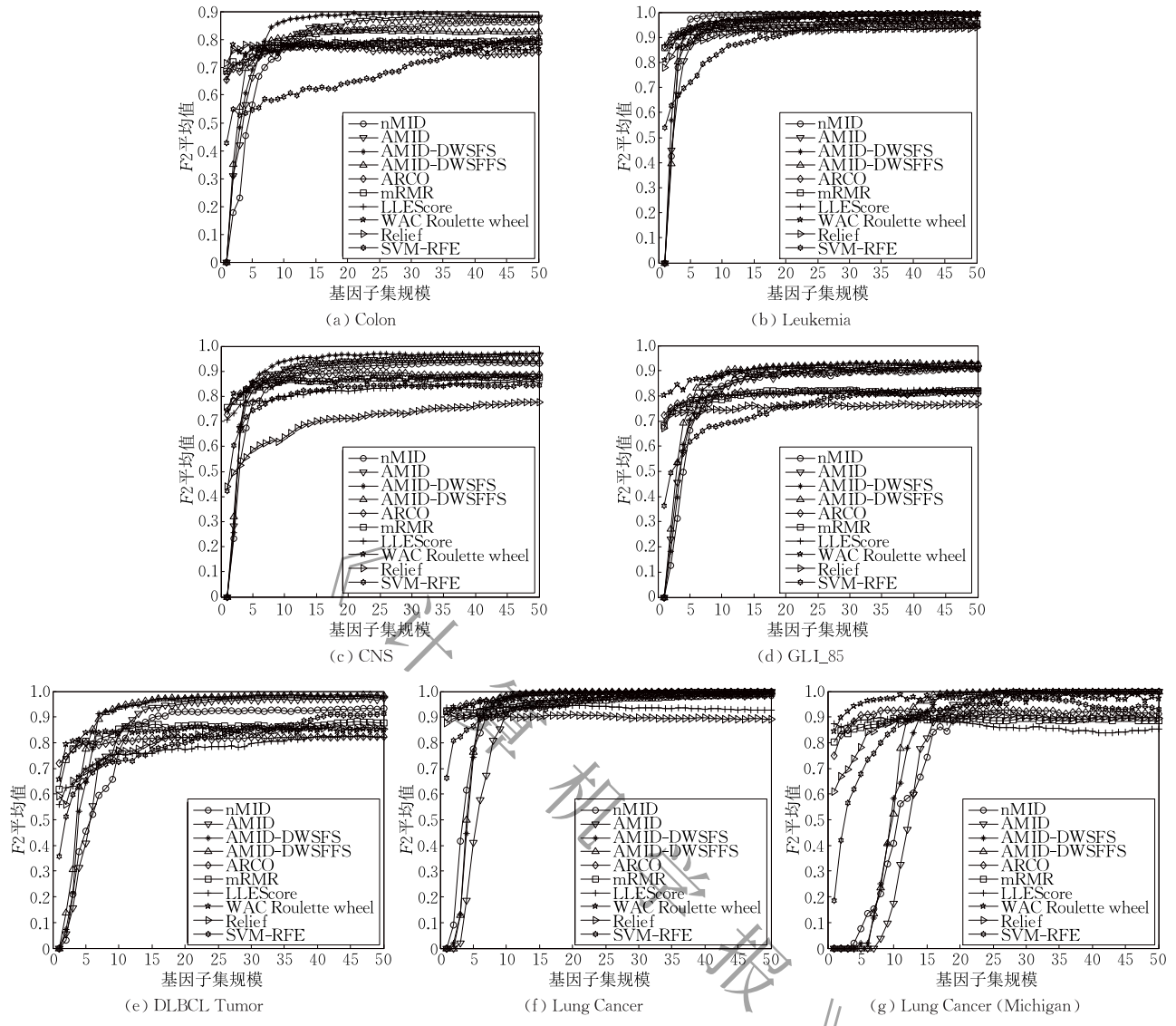


图 10 10 个特征选择算法在表 6 各数据集所得基因子集的 NB 分类器的 $F2$ 平均值曲线

表 9 本文 4 种基因选择算法与 Lasso 算法的实验结果比较

算法	Lasso		nMID		AMID		AMID-DWSFS		AMID-DWSFFS	
指标/基因子集规模	*AUC*/规模	*F2*/规模	*AUC*/规模	*F2*/规模	*AUC*/规模	*F2*/规模	*AUC*/规模	*F2*/规模	*AUC*/规模	*F2*/规模
Colon	0.86/13	0.80/13	0.95/50	0.87/50	0.97/50	0.88/49	**0.97/28**	**0.89/21**	0.93/23	0.83/26
CNS	0.92/29	0.84/29	0.97/46	0.94/36	0.99/50	0.96/47	**0.99/48**	**0.97/50**	0.98/33	0.95/42
ALL/AML Leukemia	0.98/24	0.93/24	1.00/36	**0.99/36**	1.00/41	0.99/43	1.00/46	0.99/49	**1.00/28**	0.99/48
GLI_85	0.90/26	0.78/26	0.95/50	0.91/48	0.97/48	0.91/43	0.97/49	0.93/46	**0.98/39**	**0.93/37**
DLBCL Tumor	0.95/28	0.86/28	0.99/50	0.93/47	1.00/46	0.97/50	1.00/48	0.99/43	**1.00/36**	**0.99/36**
Lung Cancer	0.99/37	0.93/37	1.00/43	**1.00/31**	1.00/31	1.00/50	**1.00/16**	1.00/49	1.00/23	1.00/49
Lung Cancer(Michigan)	0.99/15	0.89/15	1.00/21	1.00/50	1.00/18	**1.00/34**	1.00/15	1.00/44	**1.00/12**	1.00/38

注:加粗和下划线字体为同一数据集中 *AUC*、*F2* 的最大值(最大值相同的情况下,基因子集规模小的优先).

从表 9 可见:提出的算法优于对比算法 Lasso. 尽管 Lasso 算法在各数据集所选基因子集的 NB 分类器的 *AUC* 和 *F2* 值也很不错,但是没有在任何一个数据集的性能超过本文算法. 本文算法中,AMID-DWSFFS 性能最优,AMID-DWSFS 位居第二,nMID 排名第三,AMID 排第四;AMID 算法只在 Lung Cancer(Michigan)数据集的 *F2* 指标值和对应基因子集规模最优. 由此可见,本文 4 种算法能选择到性能很好的基因子集;也验证了本文提出的系列创新点的有效性.

4.7 本文各算法时间性能比较

前面第 3 小节对提出的各算法的时间复杂度进

行了理论分析，这里我们进一步比较各算法的实际运行效率. 图 11 是各算法在 CNS 数据集重复运行 50 次的平均时间比较.

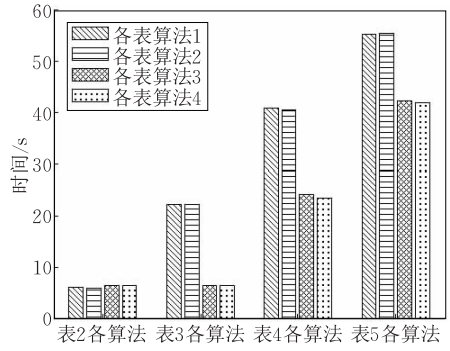


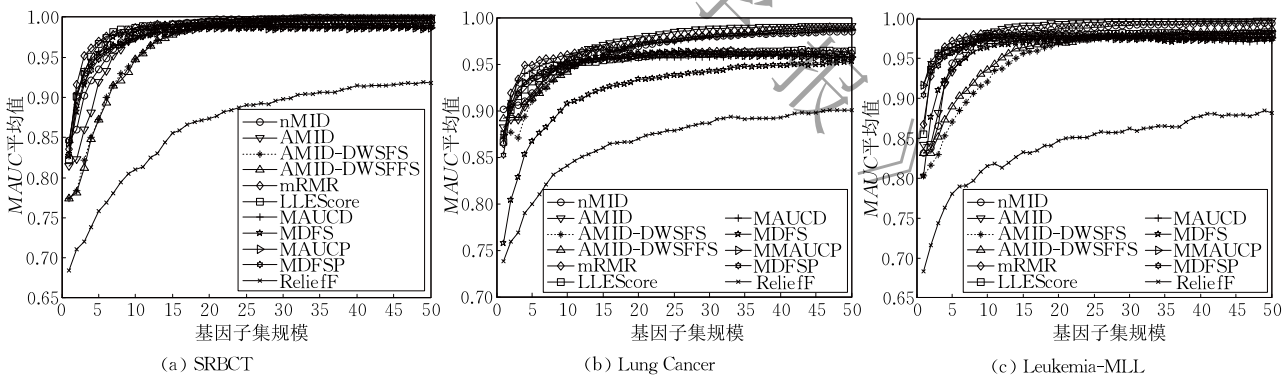
图 11　表 2～表 5 各基因选择算法运行时间对比

图 11 各算法的平均运行时间比较显示：表 2 各算法效率最高，表 5 各算法效率最差. 表 3 算法 AMID 和 AMIQ(表 3 算法 3、算法 4)和表 2 算法 nMID 和 nMIQ(表 2 算法 3、算法 4)的时间性能相当，远优于表 3 算法 ARCD 和 ARCQ(表 3 算法 1、算法 2). 这说明：2.2 节提出的归一化互信息 *SU* 比斯皮尔曼等级相关系数 *RCC* 在度量基因相关(冗余)性时省时.

图 11 还显示：表 3～表 5 中算法 3、算法 4 的运行时间明显优于算法 1、算法 2，说明使用 *SU* 度量基因冗余性比使用 *RCC* 省时. 另外，表 3 算法 1、算法 2 与表 4 算法 3、算法 4(即算法 ARCD、ARCQ 和 AMID-DWSFS、AMIQ-DWSFS)的运行时间相当，表 4 算法 1、算法 2 与表 5 算法 3、算法 4(即 ARCD-DWSFS、ARCQ-DWSFS 与 AMID-DWSFFS、AMIQ-DWSFFS)的运行时间相当，再次说明 *SU* 度量基因冗余性比 *RCC* 省时.

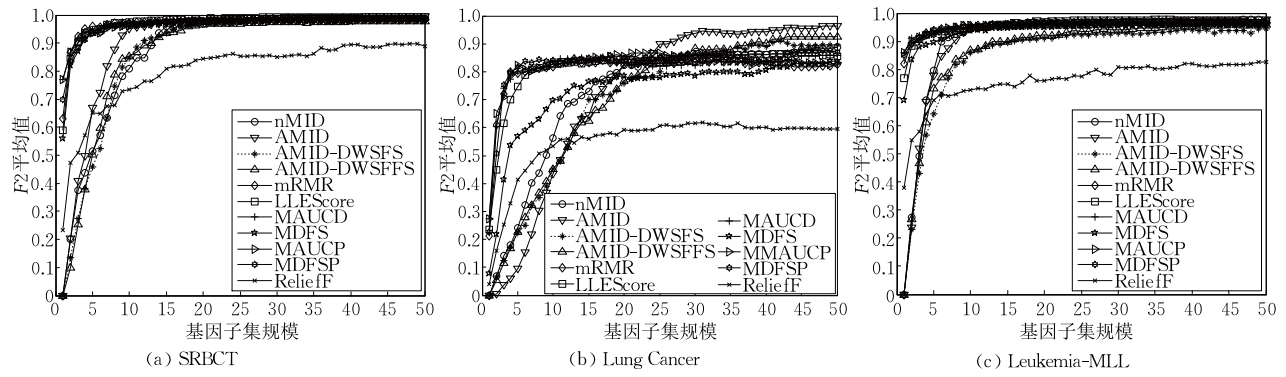
4.8　本文算法在多类数据集的性能测试

本小节将以表 7 所示 3 个分别含有 3、4、5 类的基因数据集测试本文算法在多类不平衡数据集的性能，其中的数据集 SRBCT 和 Lung Cancer 很不平衡，3 类白血病数据集 Leukemia-MLL 的不平衡性较弱. 选择该 3 个多类基因数据集不仅可以测试本文算法在多类基因数据集的性能，同时可以测试本文算法在多类不平衡或近似平衡的基因数据集的性能. 与 4.6 小节相同，本小节也是在 4.2～4.5 小节研究结论的基础上，以基于归一化互信息 *SU* 的 nMID 和 AMID 算法，以及分别结合 DWSFS、DWSFFS 的 AMID-DWSFS 和 AMID-DWSFFS 算法为例，即以表 2～表 5 的算法 3 为例，比较其与现有适用于多类问题的基因选择算法 mRMR[15]、LLEScore[38]、ReliefF[42]、MAUCD[32]、MDFS[32]、MAUCP[43]和 MDFSP[43]的性能. 需要说明的是，多类算法的 *AUC* 值就是 *MAUC* 值，*F*2 值就是式(10)所示的 *F*2 值. 图 12、图 13 分别给出了本文 4 种算



(a) SRBCT　　(b) Lung Cancer　　(c) Leukemia-MLL

图 12　11 个算法在表 7 各数据集所得基因子集的 NB 分类器的 *MAUC* 平均值曲线



(a) SRBCT　　(b) Lung Cancer　　(c) Leukemia-MLL

图 13　11 个算法在表 7 各数据集所得基因子集的 NB 分类器的 *F*2 平均值曲线

法与对比算法所选基因子集对应 NB 分类器的平均 *MAUC* 和平均 *F*2 值.

图 12 实验结果显示:本文 AMID 算法性能最优,本文其他 3 种算法 nMID、AMID-DWSFS 和 AMID-DWSFFS 的性能紧随 AMID 算法,且性能优于其他 7 种对比算法;当基因子集规模较小时,对比算法的性能较好. ReliefF 算法的性能是 11 种算法中最差的.

图 13 实验结果显示:本文算法 AMID 选择的基因子集对应 NB 分类器的平均 *F*2 值最好,ReliefF 算法选择的基因子集对应 NB 分类器的 *F*2 值最差. 在 SRBCT 数据,除 ReliefF 外其他 10 种算法选择的基因子集对应 NB 分类器的平均 *F*2 基本一致;在 Lung Cancer 数据集,本文 AMID-DWSFFS 和 AMID-DWSFS 算法选择的基因子集的分类性能分别位居第二、第三,本文 nMID 算法和其他 6 种算法选择的基因子集的 NB 分类器的平均 *F*2 值相当;在 Leukemia-MLL 数据集,AMID-DWSFS 算法的基因子集的分类性能只优于 ReliefF 算法.

图 12、图 13 所示 11 种多类基因数据集特征选择算法的 50 次重复实验的平均 *MAUC* 值和平均 *F*2 值分析揭示:本文 AMID 算法的性能最优,能选择到分类性能非常好的基因子集. 这再次说明本文提出的归一化互信息 *SU* 非常有效,也验证了本文其他创新点的有效性,以及所得结论的正确性.

5 结 论

针对分类准确率在评价类别分布不平衡数据的特征选择算法所选特征子集分类性能时的缺陷,提出融合特征子集对正、负两类识别精度的新评价准则 *F*2. 提出归一化互信息 *SU*,避免经典互信息偏爱多值特征的问题. 提出兼顾基因 *SU* 和 *AUC* 值的基因预选择方法,剔除部分与分类无关基因,缩小特征搜索空间,提升特征选择算法效率. 提出了归一化特征权重方法,统一了特征区分能力和特征冗余性的取值区间. 在 mRMR 框架下,提出了采用特征区分能力减去特征冗余性和采用特征区分能力除以特征冗余性度量特征权值的多种特征选择算法. 提出了 DWSFS 和 DWSFFS 特征搜索,加快算法收敛速度,提升算法性能.

经典两类和多类样本分布不平衡基因数据集的 50 次重复实验验证:(1)提出的特征选择算法均能选择到不平衡癌症基因数据集的区分能力很好的差异基因;(2)提出的特征子集评价准则 *F*2 是非常有效的度量准则;(3)提出的归一化互信息 *SU* 能非常有效且高效地度量变量相关性;(4)提出的基因预选择方法降低了算法运行时间,剔除了部分无关基因;(5)提出的特征权重归一化方法有效避免了特征权值计算的量纲不一致问题;(6)提出的 DWSFS 和 DWSFFS 是非常有效的特征搜索策略,能搜索到规模小且分类性能好的基因子集;(7)mRMR 框架中基因(特征)权值计算采用特征与类标相关性减去特征之间相关(冗余)性的方案优于采用特征与类标相关性除以特征相关(冗余)性的方案.

提出的 *F*2 准则、特征预选择方法、*SU* 互信息、归一化特征权重方法以及特征搜索策略 DWSFS、DWSFFS 均有助于从类别分布不平衡基因数据中选择到规模小且能同时识别不同类样本的有效差异表达基因,是选择类别分布不平衡的高维生物医学大数据的区分特征的有效方法,为癌症等疾病的准确诊断提供了方法借鉴,也可用于其他高维大数据的特征降维.

另外,本文研究结果揭示,对二类不平衡基因数据,在时间因素优先,并考虑特征子集分类性能前提下,本文提出的 AMID 算法最佳;在基因子集分类性能优先,并考虑时间的前提下,本文提出的 AMID-DWSFS 算法是首选算法. 对多类不平衡基因数据,AMID 算法是最佳的基因选择算法,不仅高效,且能选择到区分能力最佳的基因子集. 该结论为同类研究或将本文算法应用于其他高维不平衡大数据的特征选择研究提供了借鉴和指导.

致 谢 我们非常感谢审稿人不辞辛劳对我们的稿件所做的非常认真地审读和提出的宝贵修改意见. 这些修改意见使得我们的稿件质量一次次不断提升. 另外,我们非常感谢稿件主编和编辑对我们稿件所做的一切审理和编辑工作. 真诚地说一声,谢谢大家! 最后,我们还要谢谢同行所提供的宝贵测试数据,以及同行的研究成果,正是这些已有工作才促进了我们的工作更加完善. 在此,一并谢过!

参 考 文 献

[1] Li Ying-Xin, Li Jian-Geng, Ruan Xiao-Gang. Study of information gene selection for tissue classification based on tumor gene expression profiles. Chinese Journal of Computers, 2006, 29(2): 324-330(in Chinese)

(李颖新，李建更，阮晓钢. 肿瘤基因表达谱分类特征基因选取问题及分析方法研究. 计算机学报，2006，29(2)：324-330)

[2] Wald R，Khoshgoftaar T M，Shanab A A. Comparison of two frameworks for measuring the stability of gene-selection techniques on noisy class-imbalanced data//Proceedings of the IEEE 25th International Conference on Tools with Artificial Intelligence (ICTAI). Virginia，USA，2013：881-888

[3] Li Xia，Zhang Tian-Wen，Guo Zheng. An novel ensemble method of feature gene selection based on recursive partition-tree. Chinese Journal of Computers，2004，27(5)：675-682(in Chinese)
(李霞，张田文，郭政. 一种基于递归分类树的集成特征基因选择方法. 计算机学报，2004，27(5)：675-682)

[4] Xie Juan-Ying，Gao Hong-Chao. Statistical correlation and *K*-means based distinguishable gene subset selection algorithms. Journal of Software，2014，25(9)：2050-2075(in Chinese)
(谢娟英，高红超. 基于统计相关性与*K*-means 的区分基因子集选择算法. 软件学报，2014，25(9)：2050-2075)

[5] Gulshan V，Peng L，Coram M，et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. The Journal of the American Medical Association，2016，316(22)：2402-2410

[6] Esteva A，Kuprel B，Novoa R A，et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature，2017，542(7639)：115-118

[7] Long E，Lin H，Liu Z，et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. Nature Biomedical Engineering，2017，1：0024

[8] Yin L，Ge Y，Xiao K，et al. Feature selection for high-dimensional imbalanced data. Neurocomputing，2013，105：3-11

[9] Wasikowski M，Chen X. Combating the small sample class imbalance problem using feature selection. IEEE Transactions on Knowledge and Data Engineering，2010，22(10)：1388-1400

[10] He H，Garcia E A. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering，2009，21(9)：1263-1284

[11] Xie Juan-Ying，Xie Wei-Xin. Several feature selection algorithms based on the discernibility of a feature subset and support vector machines. Chinese Journal of Computers，2014，37(8)：1704-1718(in Chinese)
(谢娟英，谢维信. 基于特征子集区分度与支持向量机的特征选择算法. 计算机学报，2014，37(8)：1704-1718)

[12] López V，Fernández A，García S，et al. An insight into classification with imbalanced data：Empirical results and current trends on using data intrinsic characteristics. Information Sciences，2013，250：113-141

[13] Zheng Z，Wu X，Srihari R. Feature selection for text categorization on imbalanced data. ACM SIGKDD Explorations Newsletter，2004，6(1)：80-89

[14] Wang R，Tang K. Feature selection for maximizing the area under the ROC curve//Proceedings of the 2009 IEEE International Conference on Data Mining Workshops. Florida，USA，2009：400-405

[15] Peng H，Long F，Ding C. Feature selection based on mutual information criteria of max-dependency，max-relevance，and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence，2005，27(8)：1226-1238

[16] Zhang Yan-Xiang，Pan Hai-Xia. A feature selection method based on discriminative ability for multiclass text categorization on imbalanced data. Journal of Chinese Information Processing，2015，29(4)：111-119(in Chinese)
(张延祥，潘海侠. 一种基于区分能力的多类不平衡文本分类特征选择方法. 中文信息学报，2015，29(4)：111-119)

[17] Li Xia，Wang Lian-Xi，Jiang Sheng-Yi. Ensemble learning based feature selection for imbalanced problems. Journal of Shandong University (Engineering Science)，2011，41(3)：7-11+22(in Chinese)
(李霞，王连喜，蒋盛益. 面向不平衡问题的集成特征选择. 山东大学学报(工学版)，2011，41(3)：7-11+22)

[18] Han J W，Kamber M. Data Mining：Concepts and Techniques. 2nd Edition. San Francisco，USA：Morgan Kaufmann Publishers，2006

[19] Ding C，Peng H. Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology，2005，3(02)：185-205

[20] Yu L，Liu H. Efficient feature selection via analysis of relevance and redundancy. The Journal of Machine Learning Research，2004，5：1205-1224

[21] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters，2006，27(8)：861-874

[22] Lehmann E L，D'Abrera H J M. Nonparametrics：Statistical Methods Based on Ranks. New York，USA：Springer，2006

[23] Whitney A W. A direct method of nonparametric measurement selection. IEEE Transactions on Computers，1971，100(9)：1100-1103

[24] Pudil P，Novovičová J，Kittler J. Floating search methods in feature selection. Pattern Recognition Letters，1994，15(11)：1119-1125

[25] Alon U，Barkai N，Notterman D A，et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences，1999，96(12)：6745-6750

[26] Pomeroy S L，Tamayo P，Gaasenbeek M，et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature，2002，415(6870)：436-442

[27] Golub T R，Slonim D K，Tamayo P，et al. Molecular classification of cancer：Class discovery and class prediction by gene expression monitoring. Science，1999，286(5439)：531-537

[28] Freije W A, Castro-Vargas F E, Fang Z, et al. Gene expression profiling of gliomas strongly predicts survival. Cancer Research, 2004, 64(18): 6503-6510

[29] Shipp M A, Ross K N, Tamayo P, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine, 2002, 8(1): 68-74

[30] Gordon G J, Jensen R V, Hsiao L L, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Research, 2002, 62(17): 4963-4967

[31] Beer D G, Kardia S L R, Huang C C, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nature Medicine, 2002, 8(8): 816-824

[32] Wang R, Tang K. Feature selection for MAUC-oriented classification systems. Neurocomputing, 2012, 89: 39-54

[33] Khan J, Wei J S, Ringner M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine, 2001, 7(6): 673-679

[34] Bhattacharjee A, Richards W G, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proceedings of the National Academy of Sciences, 2001, 98(24): 13790-13795

[35] Armstrong S A, Staunton J E, Silverman L B, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature Genetics, 2002, 30(1): 41-47

[36] Chang C C, Lin C J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 27

[37] Kurgan L A, Cios K J. CAIM discretization algorithm. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(2): 145-153

[38] Li Jian-Geng, Pang Ze-Nan, Su Lei, et al. Feature selection method LLE score used for tumor gene expressive data. Journal of Beijing University of Technology, 2015, 41(8): 1145-1150(in Chinese)
(李建更，逄泽楠，苏磊等. 肿瘤基因选择方法 LLE Score. 北京工业大学学报，2015，41(8)：1145-1150)

[39] Kira K, Rendell L A. The feature selection problem: Traditional methods and a new algorithm//Proceedings of the 10th National Conference on Artificial Intelligence. California, USA, 1992: 129-134

[40] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines. Machine Learning, 2002, 46(1): 389-422

[41] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 1996, 58(1): 267-288

[42] Kononenko I. Estimating attributes: Analysis and extensions of RELIEF//Proceedings of the European Conference on Machine Learning. Catania, Italy, 1994: 171-182

[43] Xie Juan-Ying, Wang Ming-Zhao, Hu Qiu-Feng. The differentially expressed gene selection algorithms for unbalanced gene datasets by maximizing the area under ROC. Journal of Shaanxi Normal University (Natural Science Edition), 2017, 45(1): 13-22(in Chinese)
(谢娟英，王明钊，胡秋锋. 最大化 ROC 曲线下面积的不平衡基因数据集差异表达基因选择算法. 陕西师范大学学报：自然科学版，2017，45(1)：13-22)

XIE Juan-Ying, Ph.D., professor. Her research interests include machine learning, data mining, and biomedical big data analysis.

WANG Ming-Zhao, Ph. D. candidate. His research interests include data mining and bioinformatics.

ZHOU Ying, M. S. Her research interests include data mining and bioinformatics.

GAO Hong-Chao, Ph. D. candidate. His research interests include deep learning and multimedia public opinion analysis.

XU Sheng-Quan, Ph. D., professor. His research interests include bioinformatics.

Background

With the development in biotechnology, there are more and more biomedical datasets with very high-dimensions and small samples, such as gene expression datasets of cancers usually with tens to thousands features but small numbers of samples. This kind of datasets usually contain a number of redundant and irrelevant genes for recognizing cancer patients. The redundant and irrelevant features in biomedical datasets bring great challenges to tell cancer patients from normal people. Furthermore, these datasets often comprise unbalanced samples from different classes, such as usually with small number of patients and the large number of normal people, which aggravates the challenges to discover the optimal feature subset with high capacity to recognize the samples from different classes. Therefore, there are more and more

experts focusing on the field to study the feature subset selection algorithms, so that to recognize those features highly related to labels and lowly redundant between each other, and to get the optimal feature subset to preserve or advance the capacity of the original system with spare features. In addition, recognizing the feature subset can make classification models built on it with more accurate and easier to understand, and have a better generalization capacity, higher efficiency, reduced curse of dimensionality, and with more intuitive visualization analysis. The feature subset selection has become the primary step to analyze the biomedical datasets.

There are three categories feature subset selection algorithms including filters, wrappers and embedded methods. All of them are to select a sparse and representative feature subset. However, most of them prefer assuming that the samples of each category are balance, which is not true especially in biomedical datasets. Therefore, there are many experts focusing on the study of feature subset selection algorithms for unbalanced datasets. We studied the available feature subset selection algorithms, and extensively got to know their limitations, after that we proposed our innovations and 16 feature subset selection algorithms to overcome the limitations of the available ones.

We tested all our proposed innovations and feature subset selection algorithms on 7 popular unbalanced binary gene expression datasets and on 3 multi-class unbalanced biomedical datasets, and compared the performance of our algorithms to that of the available feature subset selection algorithms. The experimental results of 50 runs demonstrate the power of our innovations and our proposed feature subset selection algorithms. All the experimental results disclose that the performance of our feature subset selection algorithms are superior to the available ones. In addition, we obtained some important conclusions. We can say that our work is very important in the field of feature subset selection, and in the field of analyzing the biomedical datasets.

This work is supported by the Natural Science Foundation of China under Grant No. 61673251, and the National Key Research and Development Program of China under Grant No. 2016YFC0901900, and at the same time supported by the Fundamental Research Funds for the Central Universities under Grant No. GK201701006, and supported by the Scientific and Technological Achievements Transformation and Cultivation Funds of Shaanxi Normal University under Grant No. GK201806013, and the Innovation Funds of Graduate Programs at Shaanxi Normal University under Grant Nos. 2015CXS028 and 2016CSY009 as well.