基于一种集成的信息基因选择方法的乳腺肿瘤识别研究

杨晓慧[1],白欣宇[1],乔江华[2],陆寓非[2]
(1. 河南大学数据分析技术实验室,河南 开封 475004;
2. 河南省肿瘤医院/郑州大学附属肿瘤医院,河南 郑州 450000)

摘　要:[目的]探讨导致乳腺癌的可能致病基因及其生物学意义。[方法]基于国际上通用的乳腺癌公共测试集 Breast-2 (79) 数据库,提出了一种集成的决策信息因子(decision information factor,DIF)方法,以有效地选择出候选致病基因,并完成乳腺癌识别。基于 R 语言对原始基因数据做加权共表达网络分析以识别网络中的重要基因模块;使用 DAVID 软件对重要基因模块进行 Pathway 富集分析,验证是否具有统计学意义;使用 DIF 方法从具有统计学意义的重要基因模块中选择出 2 个候选致病基因;借助反空间稀疏表示分类模型完成乳腺癌识别。[结果]通过加权基因共表达网络得到 3 个基因模块,其中 2 个经 Pathway 富集分析检验具有统计学意义,在这两个模块上采用 DIF 基因选择方法选出的 2 个候选致病基因用于乳腺癌识别时,准确率达到 71.07%,比信噪比(signal noise ratio,SNR)、受试者工作特征曲线(receiver operating characteristic curve,ROC)、组内与组间平方和比率(the ratio of between-groups to within-groups sum of squares,BW)的方法分别高出 13.93%、11.19%和 8.57%。[结论]该文提出的集成 DIF 基因选择方法得到的候选致病基因能有效识别乳腺癌,并具有明确的生物学意义。
关键词:乳腺癌;微阵列基因表达数据;加权基因共表达网络;决策信息因子;反空间稀疏表示
中图分类号:Q-332　　**文献标识码**:A　　**文章编号**:1004-0242(2019)07-0557-06
doi:10.11735/j.issn.1004-0242.2019.07.A014

Identification of Breast Tumor Based on Integrated Information Gene Selection Method

YANG Xiao-hui[1],BAI Xin-yu[1],QIAO Jiang-hua[2],LU Yu-fei[2]
(1. Data Analysis Technology Lab of Henan University,Kaifeng 475004,China; 2. Henan Cancer Hospital/Affiliated Cancer Hospital of Zhengzhou University,Zhengzhou 450000,China)

Abstract:[Purpose] To explore the possible pathogenic genes and their biological significance in breast cancer. [Methods] Based on the standard public Breast-2(79) database,an integrated decision information factor(DIF) approach was proposed to select candidate pathogenic genes for identification of breast cancer. Firstly,based on the R language,the original gene data were analyzed by a weighted co-expression network analysis to select some important gene modules. Secondly,the pathway enrichment analysis was performed on these important gene modules using DAVID software to verify whether the genes were statistically significant. Thirdly,two candidate pathogenic genes were selected from the gene modules via the DIF. Finally,an inverse space sparse representation based classification was introduced to fulfill the breast tumor classification. [Results] Three gene modules were obtained by the weighted gene co-expression network,and two of them had statistically significant by pathway enrichment analysis. Two candidate pathogenic genes were selected by the integrated DIF gene selection method. Experiments showed that the classification accuracy reached 71.07%,which was higher than that of signal noise ratio (SNR,13.93%),receiver operating characteristic curve (ROC,11.19%),or the ratio of between-groups to within-groups sum of squares(BW,8.57%),respectively. [Conclusion] The two candidate genes selected by the integrated DIF gene selection method can be effectively used for identification of breast cancer.
Key words:breast cancer;microarray gene expression data;weighted gene co-expression network;decision information factor;inverse space sparse representation

乳腺癌严重威胁女性身心健康。近几年来,微阵列技术发展迅速,能够同时测量数千个基因的表达水平,并识别不同生物状态之间表达水平的变化,为高效分辨肿瘤类型提供了广阔前景,成为诊断肿瘤

收稿日期:2019-01-21;修回日期:2019-03-10
通信作者,陆寓非,E-mail:lyf890@sina.com

的重要工具之一。若能从基因水平上探讨乳腺癌的发病机制，即在微阵列基因表达数据中找到致癌基因，并分析其生物学意义，将对提高肿瘤识别率提供很大帮助。然而，微阵列基因表达数据具有小样本（患者）、高维度（数千个基因）和高冗余[1]的特点，严重影响到其识别效果。因此，如何在微阵列基因表达数据中选择真正有利于肿瘤识别的信息基因至关重要。

作为小样本问题的降维方法，基因选择旨在去除不相关的冗余基因并获得少量的信息基因。常用的基因选择方法有：过滤法、缠绕法和嵌入法等。过滤法选择的基因具有高冗余和分类精度低；嵌入式方法[2]的实现和计算过程很复杂。Wixted 等基于受试者工作特征曲线（receiver operating characteristic curve，ROC）和曲线下面积（area under ROC curve，AUC），提出了一种不平衡基因数据集的差异表达基因选择算法[3]。ROC 虽然经常应用到医学领域[4]，然而它只考虑到分类的准确率，而没有考虑到临床上更加关注的漏诊率和误诊率。更重要的是以上所提及的基因选择方法均没有考虑到同一种疾病的致病基因倾向于紧密相关。决策曲线分析（decision curve analysis，DCA）[5] 是通过最大化利润的临床净收益（net benefit，NB）来评估治疗方案的一种方法，其目的选择在临床上使误诊率较低的治疗方案。然而，DCA 具有一定的主观性，没有类似于 ROC 对应的 AUC 那样的量化指标。基于此，Yang 等[6]构建了一种对应于 DCA 的统计量化指标，决策信息因子（decision information factor，DIF），并用于单个基因选择。然而，对于微阵列基因表达数据，直接计算每个基因的 DIF 值代价过高。

分类器的设计是影响肿瘤识别效果的另一个重要因素。常用的微阵列基因表达数据分类方法有随机森林[7]，神经网络[8]和支持向量机[9]等，这些方法大多是基于统计学习理论开发的，依赖于模型参数并可能产生“过拟合”。近年来，深度神经网络的发展在语音识别和图像识别中的应用初见成效，也尝试用于预测癌症亚型[10,11]。然而，深度神经网络模型复杂，通常需要大量的训练数据来训练模型[12]，因此，小样本仍是该方法的最大的困难和挑战之一。稀疏表示是基于过完备字典的稀疏编码技术，基于稀疏表示的分类（sparse representation based classification，SRC）被马毅等提出并成功用于鲁棒人脸识别[13]。然而，SRC 的成功依赖于每一类有足够多的训练样本。最近，SRC 及其改进方法已被应用于基于微阵列基因表达数据的肿瘤分类。Zheng 等 [14]基于 SRC 对肿瘤亚型的基因表达数据进行分类。然而，其识别效果仍然受限于在实际中难以获得用于肿瘤分类的足够且有效的训练样本。基于此，Yang 等[15]基于小样本的微阵列基因表达数据提出了一种反投影表示模型，并用于肿瘤分类，理论和应用两个方面都验证了其稳定性和有效性。接着 Yang 等[6]又针对微阵列基因表达数据的稀疏性特点，对反投影表示模型进行了改进，构建了一种反空间稀疏表示（inverse space sparse representation，ISSR）模型并用于肿瘤识别。

基于以上工作，本文提出了一种集成的 DIF 基因选择方法，该方法首先通过加权的基因共表达网络（weighted gene co-expression network analysis，WGCNA）将表达相关的基因聚成共表达模块，然后通过富集分析选择出具有统计学意义的基因模块，进而通过 DIF 从中选择出有代表性的候选致病基因。最后，由于常用的微阵列基因表达数据分类方法中随机森林[7]，神经网络[8]和支持向量机[9]等依赖于模型参数并可能产生“过拟合”；深度神经网络需要大量的训练样本，且网络中存在许多超参数，模型的分类性能很大程度上取决于调参的技巧，在实践中达到理想的分类效果并不可靠；SRC 高度依赖于充足的训练样本，且易忽略蕴含在测试样本中的信息，而基于反投影的反空间稀疏表示是将训练样本投影到测试样本空间，改善了训练样本不足的问题且充分挖掘了蕴含在测试样本中的信息，故在肿瘤分类阶段，通过反空间稀疏表示模型进行肿瘤分类。

1 资料与方法

1.1 微阵列基因表达数据

识别手术后肿瘤是否发生转移很有必要且有实际意义。下面对术后识别肿瘤是否发生转移的数据库 Breast-2 进行深入分析。

Breast-2 是来自 117 例年轻患者的 25 000 个基因的原发性乳腺肿瘤的数据集。在文献[16]中，选择了 79 例 55 岁以下患有原发性淋巴结阴性乳腺肿瘤的患者进行检测，其中 34 例患者在 5 年内发生了转

移,45 例患者在一段时间(至少 5 年)内无疾病。所有患者,肿瘤大小<5cm。为了后续研究方便,在 5 年内发生转移的患者被计为肿瘤样本,在至少 5 年的时间后续无病的患者被称为正常样本。

1.2 方 法

1.2.1 基于加权基因共表达网络的基因模块

加权基因共表达网络分析(WGCNA)是一种广泛使用的符合生物学原理的系统生物学算法[17],该算法基于高通量的基因信使 RNA(mRNA)表达芯片数据,被广泛应用于国际生物医学领域,并且 WGCNA 在 R(一种免费的开源统计编程语言)语言中已经实施,是 R 中的软件包。因此,这里考虑到基因与基因间的相关性,基于 WGCNA 进行层次聚类,进而得到表达相似的基因模块。

共表达相关矩阵:首先定义基因对$m=(x_1,\ x_2.....x_n)$与$n=(y_1,\ y_2.....y_n)$之间的相关系数:

$$S_{mn}=\left|cor(m,n)\right|=\frac{N\sum x_i y_i-\sum x_i\sum y_i}{\sqrt{N\sum {x_i}^2-(\sum x_i)^2}\sqrt{N\sum {y_i}^2-(\sum y_i)^2}}$$

其中 N 是基因 m,n 的维数。由任意基因对相关系数作为元素组成基因共表达相关矩阵$S=[S_{mn}]$。

邻接函数:通过指定基因之间的相关系数的阈值(R=0.85)将基因对划分为相关和不相关是最直接的邻接函数。这种分法信息量损失大,例如将阈值定为 0.85,即便是相关系数为 0.845 的基因对也将被划分为"不相关"的组中。因此,WGCNA 算法中应用基因对相关系数的幂指数$a_{mn}=power(S_{mn},\beta)=\left|S_{mn}\right|^{\beta}$作为邻接函数衡量基因对之间相关关系。

节点间的相异度衡量:通过上述步骤将相关矩阵$S=[S_{mn}]$转换成了邻接矩阵$A=[a_{mn}]$。通过考虑任意基因和除了被分析基因以外的其他所有基因之间的关系,邻接矩阵将被转换成拓扑矩阵$\Omega=[\omega_{mn}]$,

$$\omega_{mn}=\frac{l_{mn}+a_{mn}}{\min\{k_m,k_n\}+1-a_{mn}}$$

其中$l_{mn}=\sum_u a_{mu}a_{un}$表示基因 m,n 都与基因 μ 连接的邻接函数乘积和,$k_m=\sum_u a_{mu}$表示基因 m 单独连接点的邻接函数和。类似的,$k_n=\sum_u a_{un}$表示基因单独连接点的邻接函数和。当基因 m 与 n 之间不存在连接,且无任何其他的基因和这两个基因都连接的情况下$\omega_{mn}=0$。

聚类分析鉴定基因模块:定义节点的相异系数$d_{mn}^{w}=1-\omega_{mn}$,以相异系数d_{mn}^{w}作为网络构建的基础。以基因之间的相异系数代替原始基因构建分层聚类树。聚类树有静态剪切树和动态剪切树两种构建算法,简而言之,静态剪切树是通过定义一个固定的高度将群集分支,该方法识别群集的准确度不高。而动态剪切算法基于树状图的分支形状,可挖掘得到静态剪切无法检测到的基因模块,更重要的是,用动态剪切算法鉴定出的基因网络与以往的生物实验结果比较一致。因此,本文采用动态剪切算法进行层次聚类。

这里采用 R 中的 WGCNA 软件包来对原始的微阵列基因表达数据进行层次聚类得到基因模块。

1.2.2 富集分析

为了检验 1.2.1 小节 WGCNA 得到的基因模块在生物学上是否具有统计学意义,下面采用富集分析,在给定的显著性水平下进行统计学检验。KEGG(the Kyoto Encyclopedia of Genes and Genomes)Pathway 富集分析旨在描述分子或基因的功能,是进行生物体内代谢分析、代谢网络研究的强有力工具,储存了基因功能的相关信息,通过通路表示细胞内的生物学过程。这里使用在线分析工具 DAVID(Database for Annotation,Visualization,and Integrated Discover)https://david.ncifcrf.gov/对共表达基因模块进行 KEGG Pathway 富集分析,在显著性水平 P<0.05 下的通路认为具有生物学统计意义,反之,无生物学统计意义。

1.2.3 基于 DIF 的基因选择

通过上面加权基因共表达网络和富集分析得到的具有统计学意义的基因模块,在每个模块内部基因表达模式相似,具有较强的相关性。因此,为了更进一步地去除冗余基因,下面将在每一个基因模块中来选择一个具有代表性的候选致病基因。

为了获得对应于最小的临床误诊率的信息基因,这里基于我们团队的工作[6],采用 DIF 进行候选致病基因选择。

假设阈值概率在有效概率区间[P,P_1]中变化,有效概率区间是 D_1、D_2 和 D_3 的交叉横坐标范围。每个治疗方案对应于曲线 D_3,最优的治疗方案对应于 NB 的最大值。类似地,每个基因对应于曲线 D_3,其中曲线中最大 NB 所对应的点被定义为基因的 DIF 指标。

$$DIF=\max_{p\le p_i\le p_1}\left(\frac{TP}{n}-\frac{FP}{n}\times\frac{P_t}{1-P_t}\right),$$
$$P_1=\{x\mid(x,y)=D_1\cap D_2\},$$
$$P_2=\max(\{x_1\mid(x_1,y_1)=D_1\cap D_3\}),\quad(1)$$

其中$DIF\in[0,1]$, $D_1:NB=0$, $D_2:NB=\frac{P}{n}-\frac{n-P}{n}\times\frac{P_t}{1-P_t}$, n 是样本总数, P 是真实患病的人数。

在每一个有统计学意义的基因模块中可通过公式(1)来计算基因的 DIF 值,并按照从大到小的顺序进行排列，最终选择值最大的基因作为该模块的代表候选致病基因。根据 DIF 的原理可以知道,此时既保证了选取的基因没有冗余信息,而且还保证了所选的基因恰好是临床上对患者影响最大的致病基因。

1.2.4 基于反投影稀疏表示模型的肿瘤分类

因为微阵列基因表达数据具有小样本的特性,带类别标签的训练样本个数极其有限，而当训练样本不足时,稀疏表示分类器的性能会降低。考虑到不带类别标签的测试样本在实际应用中相对容易获取,因此,这里采用反投影稀疏表示模型 (ISSR)[6]完成肿瘤分类。

假设 Y 是测试样本空间, $x_i\in X$, $i=1\cdots\cdots s_c$是训练样本。反投影表示意味着每个训练样本 x_i 由 Y 表示。

$$x_i=\alpha_{i,1}y_1+\mathrm{L}+\alpha_{i,l}y_l+\mathrm{L}+\alpha_{i,k}y_k=\sum_{l=1}^{k}\alpha_{i,l}y_l=Y\alpha_i\quad(2)$$

其中$\alpha_i=[\alpha_{i,1},\mathrm{L}\ \alpha_{i,l},\mathrm{L},\alpha_{i,k}]^T$是反投影表示的表示系数。考虑到微阵列基因表达数据存在明显的稀疏特征，稀疏性约束可以引入反投影表示并称为反投影稀疏表示(ISSR)。

$$\min_{\alpha_i}\|x_i-Y\alpha_i\|_2^2+\lambda\|\alpha_i\|_1\quad(3)$$

其中 λ 是正则化参数, a_i 是 x_i 的表示系数向量。

采用与反投影表示相匹配的类别贡献率作为分类准则。

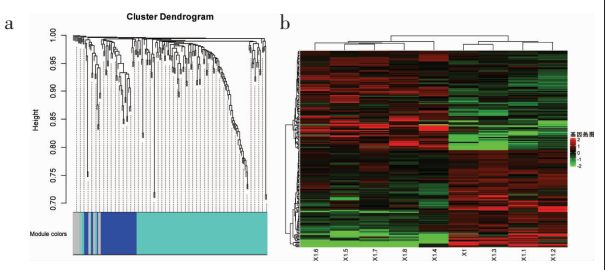
$$C_{j,l}=\frac{1}{s_j}\sum\left(\frac{\delta_j(\{|\gamma_{i,l}|\}_{i=1\ldots\ldots n})}{\sum_{i=1\ldots\ldots n}\{|\gamma_{i,l}|\}_{i=1\ldots\ldots n}}\right)\quad(4)$$

其中$\delta_j(\{|\gamma_{i,l}|\}_{i=1\ldots\ldots n})$是测试样本 y_i 关于第 j 类的系数向量,$\{|\gamma_{i,l}|\}_{i=1\ldots\ldots n}$是测试样本 y_i 的系数向量,s_j 表示第 j 类的样本数。最后求平均是为了减轻不同类的训练样本数目不同的影响。令 $j=1,2\ldots\ldots c$, $l=1,2\ldots\ldots k$ 就可以计算出所有测试样本分别对所有类的类别贡献率矩阵 C,通过类别贡献率,可以比较每个测试样本和每一类的相关性。类别贡献率越大，相关性越高。最后把 y_i 分到 $C_{j,l}$ 中最大的贡献率所对应的类 m,即 $m=\arg\max_{j\in(1\ldots\ldots c)}(C_{j,l})$。

2 结 果

2.1 加权基因共表达网络分析和关键模块识别

首先根据组内与组间平方和比率 (the ratio of between-groups to within-groups sum of squares,BW)[18]初选，在原始的 5000 个基因中初选 200 个基因,然后根据 WGCNA 进行共表达基因模块的识别，并基于在线分析工具 OmicShare（http://www.omic share.com/)绘制共表达网络图。图 1a(Figure 1a)为基因共表达模块识别（模块被标注为不同的颜色)；图 1b(Figure 1b)为基因层次聚类热图,热图中每行对应一个基因,每列对应一个样本。在热图中用颜色由绿到红表示基因表达水平由低到高。图 2(Figure 2)为基因共表达网络图，可以展示基因与基因之间的调



Note:a:Gene co-expression module recognition;b:Gene hierarchical clustering.

Figure 1 Hierarchical clustering based on WGCNA

Figure 2 Gene co-expression network map

控关系。其中节点代表基因，线代表基因与基因之间的相互作用关系。

2.2 KEGG Pathway 富集分析

为了进一步研究所选择出来的候选致病基因的生物学功能，对 Breast-2 数据集的两个模块 185 个基因进行分析，发现其中有 123 个基因可以找到 ID 编号，故对这 123 个基因在网站 https://david.ncifcrf.gov/上进行 KEGG Pathway 富集分析，其 KEGG 通路富集结果显示有 3 个通路符合阈值要求被显著富集($P\leqslant0.05$)。Pathway 分析结果显示 123 个基因主要涉及细胞周期通路(cell cycle pathway)、脂肪酸降解通路(fatty acid degradation pathway)和卵母细胞减数分裂通路(oocyte meiosis pathway)。

进一步对 WGCNA 灰度模块所对应的 15 个离散基因也进行了 KEGG Pathway 富集分析，其 KEGG 通路富集结果表明这些离散基因并没有显著富集的通路。

2.3 集成 DIF 与其他基因选择方法的比较

通过 WGCNA 分析鉴定的两个基因模块，在每一个基因模块中计算基因的 DIF 值，并分别从这两个基因模块中选择一个有代表性的候选致病基因。为了观察集成 DIF 所选的候选致病基因对肿瘤分类的效果，在图 3(Figure 3)中分别展示了原始基因热图、BW 初选 200 基因热图及集成 DIF 所选两个候选致病基因热图。

为了进一步验证集成 DIF 基因选择的性能，和原始基因数据(无基因选择)、BW[18]，信噪比(signal noise ratio，SNR)[19]、ROC[3]等经典常用的基因选择方法进行了比较，识别结果显示，原始基因数据方法的精确度 51.6%，BW 方法[18]62.50%，信噪比[20]57.14%，ROC[3]59.88%，集成 DIF71.07%。并且还通过集成 DIF 在每一个基因模块中分别选择 5 个候选致病基因，来观察当在每个基因模块中选择多个候选致病基因时，其精确度 79.23%，与原始基因数据(无基因选择)(51.61%)、BW (69.82%)、SNR (64.29%)、ROC (62.43%) 等经典常用的基因选择方法的结果有差异。为了保证对比的公平性，分类部分基于相同的分类方法 ISSR 进行。

将基于集成 DIF 选择的 2 个候选致病基因通过反投影稀疏表示模型进行肿瘤分类，图 4(Figure 4)展示了十折交叉验证中的部分识别结果，其中蓝色代表正常，黄色代表肿瘤，黑色字表示识别结果，红色圆圈表示识别错误。

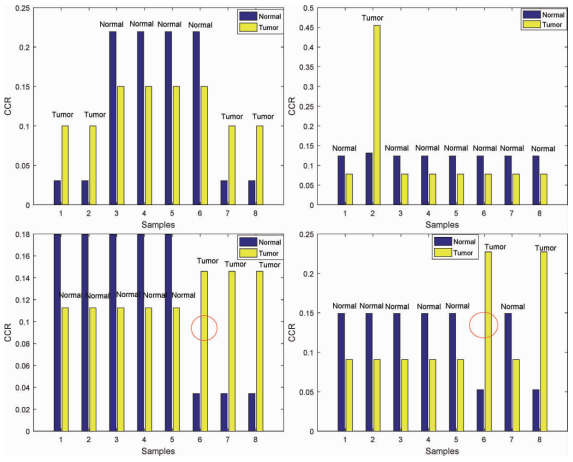
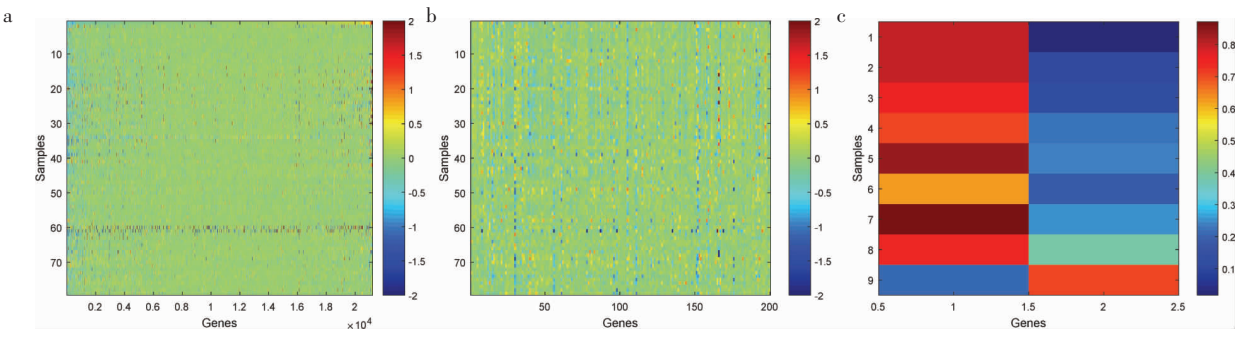


Figure 4 Partial classification results of 10-fold cross-validation (Dark blue represents normal and yellow represents tumor, black words indicate classification results, red circles indicate classification errors)



Note: a: original gene heat map; b: BW primary 200 gene heat map; c: integrated DIF selection 2 gene heat map.

Figure 3 Heatmap analysis of gene selection

3 讨 论

乳腺癌是多个基因共同作用的结果，如何筛选出关键基因是乳腺癌研究的热点，而传统的单因素方法只考虑单个基因的异常表达对乳腺癌的影响，并没有考虑基因之间的相互作用。基于此，本文在Breast-2 基因数据库上进行加权基因共表达网络分析，将表达相似的基因聚到同一个基因模块中，图1a 和图 1b 表明总共聚了两个基因模块。为了从这两个基因模块中选择出可能的候选致病基因，本文应用 DIF 基因选择方法分别从这两个基因模块中选择出一个候选致病基因。一方面，本文对所选候选致病基因进行 Pathway 富集分析，其具有生物学意义，可以找到统计学上显著的途径。另一方面，图3c 的热图展示了通过集成 DIF 基因选择方法选择出的基因在两类样本中的表达是有显著差异的。而且，本文还将这两个候选致病基因通过反投影稀疏表示模型对测试样本进行分类，预测测试样本是否患病，从表 1 和图 4 中均能显示出识别率较高，即与BW[18]、SNR[19]、ROC[3]等基因选择方法相比，集成 DIF 基因选择方法选择出的基因更易于肿瘤分类。故本研究在方法层面和生物层面具有双重意义。方法层面意义在于相比于传统的单因素分析方法，本文提出的集成 DIF 基因选择方法利用了基因之间的交互信息，实现了更高效的基因选择。生物层面的意义在于，集成 DIF 基因选择方法在考虑基因表达数据的相关性以后，选择出的候选致病基因具有生物学意义。故本文提出的集成 DIF 基因选择和反投影稀疏表示模型的方法，对于识别乳腺癌术后是否发生转移具有重要的理论研究价值和临床实际意义。但值得注意的是，所筛选出的候选致病基因的异常表达可能导致乳腺癌的发生，但并非绝对发生。由于乳腺癌的形态结构十分复杂，类型有很多，不同类型的乳腺癌可能与不同的基因子集相关联。另外，基因之间的调控机制还尚未完全了解，基因异常表达与癌症之间的关系还尚不清楚。

参考文献：

[1] Veer LJV, Vijver MJVD, Dai H, et al. Expression profiling predicts poor outcome of disease in young breast cancer patients[J]. Eur J Cancer Care (Engl), 2001, 37: S271.

[2] Algamal ZY, Lee MH. An efficient gene selection method for high-dimensional microarray data based on sparse logistic regression [J]. EASA, 2017, 10: 242–256.

[3] Xie J, Wang M, Qiufeng HU. The differentially expressed gene selection algorithms for unbalanced gene datasets by maximize the area under ROC[J]. Journal of Shaanxi Normal University, 2017, 45: 13–22.[谢娟英，王明钊，胡秋峰. 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[J]. 陕西师范大学学报, 2017, 45: 13–22.]

[4] Wixted JT, Mickes L . Theoretical vs. empirical discriminability: the application of ROC methods to eyewitness identification[J]. Cogn Res Princ Implic, 2018, 3(1): 9.

[5] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models [J]. NIH-PA Author Manuscript, 2006, 26: 565 – 574.

[6] Yang XH, Wu WM, Chen YM, et al. An integrated inverse space sparse representation framework for tumor recognition[J]. Pattern Recognit, 2019, 93: 293–311.

[7] Breiman L. Random forest [J]. Mach Learn, 2001, 45: 5–32.

[8] Bishop CM. Neural networks for pattern recognition [M]. New York: Oxford University Press, 1996.

[9] Furey TS, Cristianini N, Duffy N, et al. Support vector machines classification and validation of cancer tissue samples using microarray expression data [J]. Bioinformatics, 2000, 16: 906–914.

[10] Liang M, Li Z, Chen T, et al. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach[J]. IEEE/ACM Trans Comput Biol Bioinform, 2015, 12(4): 928–937.

[11] Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis.[J]. Sci Rep, 2016, 6: 26286.

[12] Zhou ZH, Feng J . Deep Forest: towards an alternative to deep neural networks[A]. Proceedings of Twenty-Sixth International Joint Conference on Artificial Intelligence [C]. Melbourne: International Joint Conference on Artificial Intelligence, 2017.

[13] Wright J, Yang AY, Ganesh A, et al. Robust face recognition via sparse representation [J]. IEEE Trans Pattern Anal Mach Intell, 2009, 31(2): 210–227.

[14] Zheng CH, Zhang L, Ng TY, et al.Metasample-based sparse representation for tumor classification[J]. IEEE/ACM Trans Comput Biol Bioinform, 2011, 8(5): 1273–1282.

[15] Yang XH, Tian L, Chen YM, et al. Inverse projection representation and category contribution rate for robust tumor recognition [J]. IEEE/ACM Trans Comput Biol Bioinform, 2018 Dec 18. doi: 10.1109/TCBB.2018.2886334.[Epub ahead of print]

[16] Van LJ, Dai H, van MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer [J]. Nature, 2002, 415: 530–536.

[17] Liu J, Jing L, Tu X. Weighted gene co-expression network analysis identifies specific modules and hub genes related to coronary artery disease [J]. BMC Cardiovasc Disord, 2016, 16(1): 54.

[18] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data [J]. J Am Stat Assoc, 2002, 97: 77–87.

[19] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286: 531–537.