计算机科学
COMPUTER SCIENCE

基于差异共表达邻接网络的癌症致病基因预测算法

李志杰, 廖旭红, 李青蓝, 刘丽

引用本文

李志杰, 廖旭红, 李青蓝, 刘丽. 基于差异共表达邻接网络的癌症致病基因预测算法[J]. 计算机科学, 2025, 52(5): 161-170.

LI Zhijie, LIAO Xuhong, LI Qinglan, LIU Li. Cancer Pathogenic Gene Prediction Based on Differential Co-expression Adjacent Network [J]. Computer Science, 2025, 52(5): 161-170.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240300110

基于差异共表达邻接网络的癌症致病基因预测算法

李志杰[1]　廖旭红[1]　李青蓝[2]　刘　丽[3]
1 湖南理工学院信息科学与工程学院　湖南 岳阳 414006
2 宾夕法尼亚大学医学院　费城 19019
3 弗吉尼亚联邦大学医学院　里士满 23284
(lzj0019@163.com)

摘　要　癌症是人类健康的第一杀手。随着测序技术的快速发展,积累了海量的癌症基因表达数据,利用计算方法进行致病基因预测成为癌症研究领域新的热点。然而,目前致病基因预测大多基于基因相互作用网络等,很少考虑网络局部连接与基因差异表达间的潜在联系。针对上述问题,首先利用患病前后的基因表达差异数据,通过互信息计算基因间的相关性并构建邻接网络,然后设计特征向量模型用于癌症致病基因预测。向量特征包括候选基因及其近邻的差异表达信息。从TCGA,OMIM和GEO等公共数据库获取癌症相关的致病与非致病基因以及患病前后基因差异表达数据进行实验,利用邻接网络中基因及其近邻的差异表达信息进行癌症致病基因预测(Differential Information of Gene and Nearest Neighbor for Cancer Pathogenic Gene Prediction,DICPG)。实验结果表明,DICPG癌症基因分类模型的生物学意义明显,分类精度和AUC等性能指标优于同类方法。

关键词:基因差异表达数据;邻接网络;候选基因;基因特征向量;癌症致病基因预测
中图分类号　TP181

Cancer Pathogenic Gene Prediction Based on Differential Co-expression Adjacent Network

LI Zhijie[1],LIAO Xuhong[1],LI Qinglan[2] and LIU Li[3]
1 School of Information Science and Engineering,Hunan Institute of Science and Technology,Yueyang,Hunan 414006,China
2 Medical College,University of Pennsylvania,Philadelphia 19019,USA
3 Medical College,Virginia Commonwealth University,Richmond 23284,USA

Abstract　Cancer is the first killer of human health. With the rapid development of sequencing technology,a massive amount of cancer gene expression data has been accumulated,and using computational methods to predict pathogenic genes has become a new hotspot in cancer research. However,currently,the prediction of pathogenic genes is mostly based on gene interaction networks,and little consideration is given to the potential connection between local network connections and differential gene expression. In response to the above issues,this paper first utilizes gene expression difference data before and after the disease,calculates the correlation between genes through mutual information,and constructs an adjacency network. Then,a feature vector model is designed for predicting cancer pathogenic genes. Vector features include differential expression information of candidate genes and their neighbors. Cancer-related pathogenic and non pathogenic genes are obtained from public databases such as TCGA,OMIM,and GEO,as well as differential expression data of genes before and after illness,for experiments. Differential expression information of genes and their neighbors in adjacency networks are used for cancer pathogenic gene prediction(DICPG). The experimental results show that the DICPG cancer gene classification model has significant biological significance,and its classification accuracy and AUC performance indicators are superior to similar methods.

Keywords　Gene differential expression data,Adjacent network,Candidate gene,Gene feature vector,Cancer pathogenic gene prediction

现有研究[1]表明,癌症主要是由患者的遗传、环境和生活方式等因素交互作用引起的,其中基因突变是诱发癌症的主要原因。癌症患者中突变的基因称为该疾病的致病基因。在生物分子网络中,单个基因或蛋白质很难引起表型的变化,

到稿日期:2024-03-18　返修日期:2024-07-29
基金项目:国家自然科学基金(62072475,61672391);湖南省自然科学基金(2019JJ40111)
This work was supported by the National Natural Science Foundation of China(62072475,61672391) and Hunan Provincial Natural Science Foundation,China(2019JJ40111).
通信作者:廖旭红(lxh2402@163.com)

通常是组成相互作用模块来发挥具体功能。识别致病基因与致病模块，成为癌症研究亟待解决的重要问题。

随着基因组学和生物信息学的发展，累积了海量的各种疾病相关的基因表达数据[2]。研究人员开始利用计算机技术从这些数据中挖掘出候选致病基因，提出了大量的基因对疾病贡献的排序算法。文献[3]采用高维大数据基因统计技术，提出了一种致病性模块信息识别方法。文献[4]提出了 LOTUS 单任务和多任务机器学习算法，用于预测癌症驱动基因。文献[5]提出一种基于体细胞突变预测癌症驱动基因的 deepDriver 方法，该方法采用了深度卷积神经网络技术。在文献[6]中，Liu 等将已知的少量禾谷镰刀菌(Fusarium Graminearum)致病基因视为种子基因，利用蛋白质-蛋白质相互作用网络识别种子基因邻域中的潜在致病基因，这些潜在致病基因在病原真菌入侵前后具有显著的差异表达。该预测致病基因方法是将蛋白质-蛋白质相互作用网络的邻域信息和禾谷镰刀菌患病前后基因差异表达信息作为指导，识别出一个由潜在致病基因组成的子网。

随机游走算法是生物分子网络中应用最广泛的识别致病基因的方法，例如目前流行的 PageRank 算法[7]和 HITS 算法[8]等。PageRank 最初是一个计算网页重要性的算法，其被应用于致病基因预测时，以 PR 值作为决策值区分正常基因和癌症致病基因的结果。HITS 算法(Hyperlink-Induced Topic Search)基于随机游走迭代搜索网络中的重要节点，与 Page-Rank不同，需计算 hub 和 authority 两个属性值。致病模块识别一般包括模块检测和模块排序两个部分，例如 Modularity[9]模块检测算法、Crank[10]模块排序算法。mRank[11]则在模块排序算法中结合了有指导的模块检测策略。

面向差异表达数据的差异分析方法[12]是常用的基因表达数据分析工具，可以帮助研究人员寻找不同条件下的基因差异，从而进一步了解基因的功能和作用。癌症基因表达数据是最重要的差异表达数据，由患癌前后样本形成分类数据集。在 TCGA，OMIM，GEO 等知名的癌症相关数据库和网站上，可以获得大量的癌症基因差异表达数据[13]。表 1 给出了一个癌症基因差异表达数据集示例，$\{\boldsymbol{g}_1,\boldsymbol{g}_2,\boldsymbol{g}_3,\boldsymbol{g}_4,\boldsymbol{g}_5,\boldsymbol{g}_6\}$ 是基因列集合，$\{\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_3,\boldsymbol{s}_4,\boldsymbol{s}_5\}$ 是患病前的样本，$\{\boldsymbol{s}_6,\boldsymbol{s}_7,\boldsymbol{s}_8,\boldsymbol{s}_9,\boldsymbol{s}_{10}\}$ 是对应的患病后样本。

表 1 癌症基因差异表达数据集示例

Table 1 Example of cancer gene differential expression dataset

样本(类别)	$\boldsymbol{g}_1$	$\boldsymbol{g}_2$	$\boldsymbol{g}_3$	$\boldsymbol{g}_4$	$\boldsymbol{g}_5$	$\boldsymbol{g}_6$
$\boldsymbol{s}_1$(−)	0.155	0.076	−0.201	0.254	0.013	−0.181
$\boldsymbol{s}_2$(−)	0.217	0.084	0.150	0.165	−0.159	0.132
$\boldsymbol{s}_3$(−)	0.375	0.115	0.284	0.076	−0.094	0.155
$\boldsymbol{s}_4$(−)	0.238	0	−0.159	0.129	−0.191	0.217
$\boldsymbol{s}_5$(−)	−0.073	−0.146	0.443	0.818	−0.341	0.227
$\boldsymbol{s}_6$(+)	0.394	0.909	0.426	0.768	1.070	0.226
$\boldsymbol{s}_7$(+)	0.385	0.822	0.244	0.550	1.013	0.327
$\boldsymbol{s}_8$(+)	0.329	0.690	0.066	0.529	0.790	0.313
$\boldsymbol{s}_9$(+)	0.384	0.730	0.066	0.529	0.852	0.313
$\boldsymbol{s}_{10}$(+)	−0.316	−0.191	0.202	−0.140	0.043	0.076

癌症基因差异表达数据可形式化表示为一个 $2m\times n$ 的矩阵 $\boldsymbol{D}=(\boldsymbol{S},\boldsymbol{G})$，其中，$\boldsymbol{S}=(\boldsymbol{S}_1,\boldsymbol{S}_2)$，$\boldsymbol{S}_1=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_m\}$ 表示患癌前样本行集合，$\boldsymbol{S}_2=\{\boldsymbol{s}_{m+1},\boldsymbol{s}_{m+2},\cdots,\boldsymbol{s}_{2m}\}$ 表示患癌后样本行集合，$\boldsymbol{G}=\{\boldsymbol{g}_1,\boldsymbol{g}_2,\cdots,\boldsymbol{g}_j,\boldsymbol{g}_{j+1},\cdots,\boldsymbol{g}_n\}$ 表示基因列集合。矩阵 $\boldsymbol{D}$ 的每一行代表一个组织样品 $\boldsymbol{s}_i(y_i,a_{i1},a_{i2},\cdots,a_{in})$，其中 y_i 是样本所属的类别标签，例如“+”“−”等。a_{ij} 是基因 $\boldsymbol{g}_j$ 在样本 $\boldsymbol{s}_i$ 中的表达量。矩阵 $\boldsymbol{D}$ 的每一列代表一个基因在不同组织样品的表达量 $\boldsymbol{g}_j(a_{1j},a_{2j},\cdots,a_{2mj})$。然而，癌症基因差异表达数据 $n\gg 2m$，即基因维度高，样本量少，且还有很多噪声，给癌症基因表达数据分析带来了很大的困难[14]。

从计算方法的角度，致病基因预测方法[15]可以划分为统计分析方法和机器学习方法两大类。例如，文献[16]应用 t 检验和 f 检验等统计分析方法选择基因，文献[17]提出信息熵增益(Information Entropy Gain，IEG)机器学习方法选择基因。

文献[18]为了挖掘基因表达数据中的差异共表达致病基因模块，提出了基于互信息和最大团相结合的方法(Finding Differentially Co-Expressed Disease-related Genes Based on Mutual Information，DCEG)，实验结果表明该方法能有效挖掘出差异共表达致病基因模块。

文献[19]提出一种基于网络和基因差异表达信息的癌症致病基因预测方法(Prediction of Cancerous Pathogenic Genes Based on Network and Gene Differential Expression Information，NGDE)。该方法采用了文献[6]中利用生物学网络邻域信息和患病前后基因差异表达信息的思想，并以此为指导设计基因节点特征表示向量。然后，癌症致病基因的预测就归结为基因特征表示向量的训练和测试问题。NGDE 选择的分类器是支持向量机。

本文受上述 3 类致病基因预测方法的启发，借鉴癌症基因表达数据差异分析的思想，提出了一种基于差异共表达邻接网络的癌症致病基因预测算法。本文的主要贡献有：

1)提出了一种新颖的癌症致病候选基因筛选方法，以候选基因作种子基因，通过计算互信息得到。首先，计算正常样本和疾病样本中各对基因的互信息值，分别求得互信息矩阵 $\boldsymbol{M}_1$ 和 $\boldsymbol{M}_2$。然后，设定 2 个不同阈值将 $\boldsymbol{M}_1$ 和 $\boldsymbol{M}_2$ 二值化，并利用元素逻辑“与”得到邻接矩阵。由邻接矩阵构建邻接网络，节点即为候选基因。

2)设计特征向量模型用于癌症致病基因预测，向量特征包括候选基因及其近邻的差异表达信息。癌症致病基因预测转变为特征表示向量的机器学习分类训练与测试问题。

3)在多个肿瘤或非肿瘤基因表达数据集上的实验结果验证了该方法的高效性和有效性，并从 Friedman 检验和 Nemenyi 后续检验等统计分析角度验证了所提方法相对现有基因选择方法的性能优势。

1 构建邻接网络筛选候选基因

本文通过构建基因邻接网络筛选癌症致病候选基因。为了构建如表 1 所列的癌症基因表达数据的邻接网络，先按样本类别挖掘基因相关性，得到基因的 2 个互信息矩阵，经过二值化和逻辑“与”运算后即得到网络图的邻接矩阵。

1.1 基因互信息

根据信息论，熵可用来度量随机变量的不确定性。对于复杂的基因关系，熵和互信息能有效描述癌症基因表达数据

的模式相似性[20-23]。

定义 1(熵，entropy) 熵 H(X)本质上是一个信息理论函数，使用概率分布 $P(X)$来度量离散型随机变量 X 的不确定性，具体计算公式如下：

$$H(X)=-\sum_{x}P(x)\log P(x) \tag{1}$$

定义 2(互信息，mutual information) 互信息 $I(X,Y)$表示一个随机变量 X 能提供给另一个随机变量 Y 的信息量，具体计算公式为：

$$I(X,Y)=\sum_{x\in X}\sum_{y\in Y}P(x,y)\log\frac{P(x,y)}{P(x)P(y)} \tag{2}$$

根据定义 2 计算互信息，要求随机变量 X 和 Y 离散取值。然而，按照基因表达取值的特点，某一个具体取值重复出现的概率很低，直接采用基因表达值计数显然不可行；使用 Weka 过滤器等工具，将连续的基因表达值离散化预处理，采用的相似性度量方式是欧几里得距离，并不符合基因表达数据模式相似性特征的要求。因此，将样本的基因表达值排序后，使用相关原子序列的计数作为基因随机变量的取值计算基因互信息。为了避免式(2)的分子或分母为 0，计数器 $c(X\rightarrow x)$，$c(Y\rightarrow y)$和 $c(X\rightarrow x,Y\rightarrow y)$的初始值均设为 1，式(2)中的概率由修正计算式得到：

$$P(x)=\frac{c(X\rightarrow x)+1}{n} \tag{3}$$

$$P(y)=\frac{c(Y\rightarrow y)+1}{n} \tag{4}$$

$$P(x,y)=\frac{c(X\rightarrow x,Y\rightarrow y)+1}{n} \tag{5}$$

其中，n 是样本个数。

式(3)－式(5)中，计数器 $c(X\rightarrow x)$，$c(Y\rightarrow y)$和 $c(X\rightarrow x,Y\rightarrow y)$的意义如下：首先对样本的基因表达值排序；然后用基因列下标置换形成基因列下标序列，将每个样本的基因序列分解为长度为 2 的原子序列集合；最后统计各个原子序列出现的次数。

定义 3(联合熵，joint entropy) 联合熵 $H(X,Y)$表示一对随机变量 X 和 Y 的不确定性，具体计算公式为：

$$H(X,Y)=-\sum_{x\in X}\sum_{y\in Y}P(x,y)\log P(x,y) \tag{6}$$

互信息和联合熵的关系为：

$$I(X,Y)=H(X)+H(Y)-H(X,Y) \tag{7}$$

1.2 正常样本和癌症样本的互信息矩阵

癌症基因表达数据基因间的互信息值由式(2)计算。对于任意两个基因 X 和 Y，若互信息 $I(X,Y)$值较大，则意味着 X 和 Y 存在较强的生物学意义上的相互关联；若互信息 $I(X,Y)=0$，通常认为 X 和 Y 相互独立，不存在任何生物学意义上的相互关联[24]。

以表 1 的癌症基因表达数据为例，存在正常和患病两种类别的样本。其中，$\boldsymbol{S}_1=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\boldsymbol{s}_3,\boldsymbol{s}_4,\boldsymbol{s}_5\}$是患病前的样本，类别为“－”；$\boldsymbol{S}_2=\{\boldsymbol{s}_6,\boldsymbol{s}_7,\boldsymbol{s}_8,\boldsymbol{s}_9,\boldsymbol{s}_{10}\}$是对应的患病后样本，类别为“＋”。

下面以患病前的样本 $\boldsymbol{S}_1$ 为例，说明计算 $\boldsymbol{S}_1$ 中每对基因互信息值的过程。

1)将每个样本的基因表达值按从大到小排序，具体如表 2 所列。

表 2 样本的基因表达值降序排序

Table 2 Descending order of gene expression values of samples

样本	基因表达值降序排序					
$\boldsymbol{s}_1$	0.254 (g_4)	0.155 (g_1)	0.076 (g_2)	0.013 (g_5)	－0.181 (g_6)	－0.201 (g_3)
$\boldsymbol{s}_2$	0.217 (g_1)	0.165 (g_4)	0.150 (g_3)	0.132 (g_6)	0.084 (g_2)	－0.159 (g_5)
$\boldsymbol{s}_3$	0.375 (g_1)	0.284 (g_3)	0.155 (g_6)	0.115 (g_2)	0.076 (g_4)	－0.094 (g_5)
$\boldsymbol{s}_4$	0.238 (g_1)	0.217 (g_6)	0.129 (g_4)	0 (g_2)	－0.159 (g_3)	－0.191 (g_5)
$\boldsymbol{s}_5$	0.818 (g_4)	0.443 (g_3)	0.227 (g_6)	－0.073 (g_1)	－0.146 (g_2)	－0.341 (g_5)

2)将基因表达值替换为基因列下标，具体如表 3 所列。

表 3 基因列下标序列

Table 3 Gene column index sequence

组织样本	基因列下标序列
$\boldsymbol{s}_1$	4→1→2→5→6→3
$\boldsymbol{s}_2$	1→4→3→6→2→5
$\boldsymbol{s}_3$	1→3→6→2→4→5
$\boldsymbol{s}_4$	1→6→4→2→3→5
$\boldsymbol{s}_5$	4→3→6→1→2→5

3)统计基因原子序列出现的次数。基因原子序列在表 3 中出现的次数统计如表 4 所列。

表 4 基因原子序列出现的次数

Table 4 Number of occurrences of gene atomic sequences

原子序列	次数	原子序列	次数	原子序列	次数
4→1	1	1→2	2	2→5	3
5→6	1	6→3	1	1→4	1
4→3	2	3→6	3	6→2	2
1→3	1	2→4	1	4→5	1
1→6	1	6→4	1	4→2	1
2→3	1	3→5	1	6→1	1

4)计算基因互信息。依据式(2)－式(5)，计算得到 $\boldsymbol{S}_1$ 的互信息值矩阵 $\boldsymbol{M}_1$，如表 5 所列。

表 5 互信息值矩阵 $\boldsymbol{M}_1$ 示例

Table 5 Example of mutual information value matrix $\boldsymbol{M}_1$

	$\boldsymbol{g}_1$	$\boldsymbol{g}_2$	$\boldsymbol{g}_3$	$\boldsymbol{g}_4$	$\boldsymbol{g}_5$	$\boldsymbol{g}_6$
$\boldsymbol{g}_1$	0	0.920	2.918	2.915	0.973	2.918
$\boldsymbol{g}_2$	0.920	0	1.198	0.865	1.103	1.287
$\boldsymbol{g}_3$	2.918	1.198	0	3.542	1.198	1.607
$\boldsymbol{g}_4$	2.915	0.865	3.542	0	1.250	1.929
$\boldsymbol{g}_5$	0.973	1.103	1.198	1.25	0	1.216
$\boldsymbol{g}_6$	2.918	1.287	1.607	1.929	1.216	0

患病后样本 $\boldsymbol{S}_2$ 的每对基因互信息值的计算过程与 $\boldsymbol{S}_1$ 类似，$\boldsymbol{S}_2$ 的互信息值矩阵 $\boldsymbol{M}_2$ 如表 6 所列。

表 6 互信息值矩阵 $\boldsymbol{M}_2$ 示例

Table 6 Example of mutual information value matrix $\boldsymbol{M}_2$

	$\boldsymbol{g}_1$	$\boldsymbol{g}_2$	$\boldsymbol{g}_3$	$\boldsymbol{g}_4$	$\boldsymbol{g}_5$	$\boldsymbol{g}_6$
$\boldsymbol{g}_1$	0	1.042	0.410	0.114	0.990	0.010
$\boldsymbol{g}_2$	1.042	0	1.172	1.777	0.984	1.790
$\boldsymbol{g}_3$	0.410	1.172	0	1.033	1.172	2.610
$\boldsymbol{g}_4$	0.114	1.777	1.033	0	2.265	3.509
$\boldsymbol{g}_5$	0.990	0.984	1.172	2.265	0	1.740
$\boldsymbol{g}_6$	0.010	1.790	2.610	3.509	1.740	0

1.3 构建基因差异共表达邻接网络

选定阈值 T_1 和 T_2($T_1>T_2$)，将 $\boldsymbol{M}_1$ 和 $\boldsymbol{M}_2$ 二值化后，就

可构建邻接矩阵 $\boldsymbol{M}$。$\boldsymbol{M}_1$ 和 $\boldsymbol{M}_2$ 中的互信息计算和邻接矩阵 $\boldsymbol{M}$ 的取值规则如算法 1 所示。

算法 1 $MIA(\boldsymbol{D}, T_1, T_2)$

输入：基因差异表达数据 $\mathbf{D}$，$\mathbf{M}_1$ 阈值 $\mathbf{T}_1$，$\mathbf{M}_2$ 阈值 T_2

输出：邻接矩阵 $\mathbf{M}$

$\mathbf{D}=(\mathbf{S},\mathbf{G})$，$\mathbf{S}=(\mathbf{S}_1,\mathbf{S}_2)$

$\mathbf{M}_1=\mathrm{MI}(\mathbf{S}_1,\mathbf{G})$// MI 由式(2)－式(5)计算

$\mathbf{M}_2=\mathrm{MI}(\mathbf{S}_2,\mathbf{G})$

for each i∈**G** do

  for each j∈**G** do

    if $M_1(i,j)\geqslant T_1$ then $M_1(i,j)=1$

    else $M_1(i,j)=0$

    if $M_2(i,j)\leqslant T_2$ then $M_2(i,j)=1$

    else $M_2(i,j)=0$

    $M(i,j)=M_1(i,j)$ & $M_2(i,j)$ //& 逻辑“与”

return **M**

以表 1 数据为例，设定 $T_1=1.19$ 和 $T_2=1.0$，利用 MIA 互信息矩阵算法得到邻接矩阵 $\boldsymbol{M}$，如表 7 所列。

表 7 邻接矩阵 $\boldsymbol{M}$ 示例

Table 7 Example of adjacency matrix $\boldsymbol{M}$

	g_1	g_2	g_3	g_4	g_5	g_6
g_1	0	0	1	1	0	1
g_2	0	0	0	0	0	0
g_3	1	0	0	1	0	0
g_4	1	0	1	0	0	0
g_5	0	0	0	0	0	0
g_6	1	0	0	0	0	0

图 1 给出了表 7 所列邻接矩阵 $\boldsymbol{M}$ 所构建的基因邻接网络，节点数字表示基因列下标。本文选定基因邻接网络节点为癌症致病候选基因。

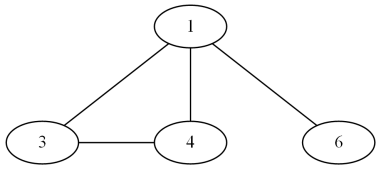


图 1 基因邻接网络图示例

Fig. 1 Example of gene adjacency network diagram

2 设计候选基因特征向量表示

2.1 癌症致病基因与邻居基因的生物特性

研究表明[19,25]，癌症致病基因及其邻居基因在患病前后基因表达的变化信息之间存在潜在的生物特性。

文献[19]从 TCGA 和 OMIM 数据库中获取 21 种癌症相关的基因表达数据及其致病基因数据，并以已知的致病基因作种子基因，分析生物学网络邻域信息和患病前后基因差异表达变化信息的潜在联系。结果表明：

1)癌症致病基因患病前后表达倾向于无差异或者差异较小；

2)癌症致病基因的近邻在患病前后的表达值倾向于具有显著差异。

本文受文献[19]发现的上述 2 个生物特性启发，结合前面邻接网络筛选基因结果，以候选基因为种子基因，设计每个致病候选基因特征向量用于分类。

2.2 设计候选基因的特征向量表示

候选基因特征向量主要考虑候选基因与邻居在患癌前后的基因差异表达量，度量指标为 $\log_2 FC$，其中 $FC=(v_1-v_2)/v_2$，v_1 和 v_2 为患病前后的基因表达值($v_1>v_2$)。

候选基因特征向量包含 6 个特征，如表 8 所列。

表 8 基因特征向量表示

Table 8 Gene feature vector representation

特征 1	特征 2	特征 3	特征 4	特征 5	特征 6
候选基因 $\log_2 FC$ 值	候选基因邻居个数	候选基因前 N 邻居 $\log_2 FC$ 值	特征 3 均值	特征 3 方差	类标签

特征 1 候选基因的 $\log_2 FC$ 值，用来反映候选基因的差异表达程度；

特征 2 候选基因的邻居个数，表示候选基因子图密度；

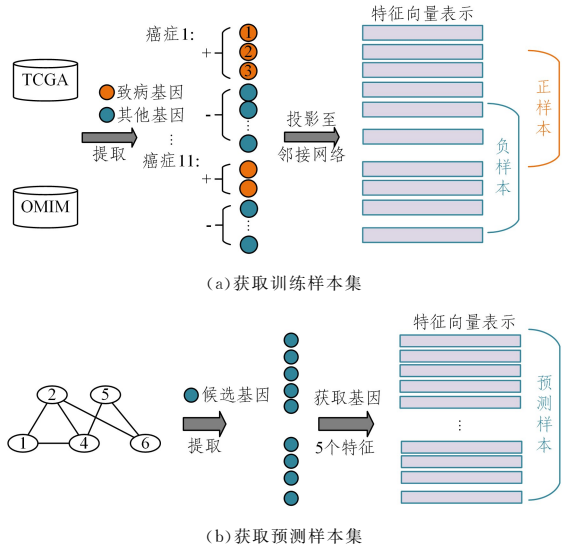
特征 3 候选基因的 N 个近邻的差异表达值 $\log_2 FC$，表示为 $u_1, u_2, \cdots, u_N$ ($u_1\geqslant u_2\geqslant\cdots\geqslant u_N$)，它们分别反映了候选基因各个邻居的差异表达程度；

特征 4 $avg(u_1, u_2, \cdots, u_N)$，候选基因 N 个邻居的 $\log_2 FC$ 均值；

特征 5 $var(u_1, u_2, \cdots, u_N)$，候选基因 N 个邻居的 $\log_2 FC$ 的方差；

特征 6 类标签，取值 0 或 1。类标签取值 0，意味着候选基因不是癌症致病基因，称为负样本；类标签取值 1，则表示候选基因是癌症致病基因，称为正样本。

图 2 给出了获取基因特征向量训练样本集和测试样本集的过程。



(a)获取训练样本集

(b)获取预测样本集

图 2 获取基因特征向量表示

Fig. 2 Obtaining gene feature vector representation

以表 1 所列癌症基因差异表达数据为例，通过差异共表达邻接网络筛选出候选基因 1,3,4 和 6，并挖掘出邻居关系 1-3,1-4,1-6,3-4，从而产生 4 条如表 8 所列的候选基因特征向量。利用支持向量机 SVM 对基因特征向量分类，类标签结果即表示候选基因是否被预测为癌症致病基因。

3 癌症致病基因预测

DICPG 是基于差异共表达邻接网络的癌症致病基因预测算法,其使用分类器对基因特征向量进行分类。差异共表达邻接网络使用 MIA 构建,如算法 1 所示。算法 1 所需的癌症基因差异表达数据来源于 TCGA(The Cancer Genome Atlas)数据库。分类训练过程提取有明确类标签的基因特征向量形成训练样本集,训练所需的特征向量类别值来源于 OMIM(Online Mendelian Inheritance in Man)数据库的癌症致病基因数据。同时,提取候选基因特征向量形成测试样本集用于分类。

算法 2 给出了基于差异共表达邻接网络的癌症致病基因预测的形式化描述。

算法 2 ***DICPG***$(\boldsymbol{D}_1,\boldsymbol{D}_2,T_1,T_2,CPG)$

输入:基因差异表达数据 $\mathbf{D}_1$ 和 $\mathbf{D}_2$,互信息矩阵阈值 T_1 和 T_2,致病基因集 CPG

输出:候选基因特征向量类标签

1. $\mathbf{AN}_1=\mathbf{MIA}(\mathbf{D}_1,T_1,T_2)$ //构建邻接网络
2. for each i$\in AN_1$ do
3. train[i]=fv(i,CPG) // fv 构建基因特征向量
4. CM=classify(train,classifior) //训练分类模型
5. $\mathbf{AN}_2=\mathbf{MIA}(\mathbf{D}_2,T_1,T_2)$
6. for each j$\in AN_2$ do
7. test[j]=fv(j)
8. **C**=**CM**(test) //测试样本使用模型分类
9. return **C** //**C** 为测试样本分类标签集

4 实验结果与分析

本文使用表 9 所列的 9 个数据集对所提算法的性能进行评价。实验在配置为 2.60 GHz、Intel(R) Core(TM) i7-6700HQ CPU、内存 16GB、操作系统 Windows 10 的计算机上进行。

表 9 基因表达数据集

Table 9 Gene expression datasets

数据集	基因个数	样本个数	类别数
Leukemia	7129	72	2
Colon	2000	62	2
SRBCT	2308	83	4
Brain	5920	90	5
Breast cancer	10	683	2
Duke_bc	7129	44	2
Heart	13	270	2
Mushrooms	112	8124	2
Protein	357	17766	3

实验将本文提出的 DICPG 分别与 t-test/f-test(t 检验和 f 检验)、IEG(信息熵增益)、DCEG(基于互信息和最大团相结合的方法)、NGDE(基于网络和基因差异表达信息的预测方法)、PageRank 和 HITS(随机游走)等其他致病基因预测方法进行了比较,同时对本文的 MIA 算法的致病基因模块挖掘功能进行了验证。

实验利用 SVM 和 NB 作为分类器,以分类精度 Acc,F-score 和 AUC 等性能指标评价基因选择方法。

4.1 数据集与评价指标

实验使用的 9 个基因表达数据包括 6 个肿瘤数据集和 3 个非肿瘤数据集,主要来自 libsvm 网站[1] 和 UCI 网站[2] 的基准数据集。肿瘤数据集包括 Leukemia(白血病)、Colon(结肠癌)、SRBCT(小蓝圆细胞)、Brain(脑瘤)、Breast cancer 和 Duke breast cancer(乳腺癌);非肿瘤数据集包括 Heart(心脏病)、Mushrooms(蘑菇菌)、Protein(蛋白质)。表 9 列出了 9 个数据集的相关参数,其中 SRBCT,brain 和 Protein 属于多分类问题,实验时将其处理为二分类问题,即指定其中的某一类为正样本,其余类别都作为负样本。9 个数据集均作为不平衡二分类问题实例,若正负样本数量不平衡程度用比例来衡量,小的相差 2 倍,大的相差近 70 倍。

二分类性能的评估主要基于表 10 所列的分类混淆矩阵。

表 10 分类混淆矩阵

Table 10 Classification confusion matrix

	预测为正样本	预测为负样本
正样本	*TP*	*FN*
负样本	*FP*	*TN*

基于混淆矩阵的二分类性能评价指标通常有:准确率 Acc(accuracy)、敏感性 Sen(sensibility)、特异性 Spe(specificity)、F-score 和 AUC(area under curve)等。由于少数类样本受偏态分布的影响,总体精度不能很好地评价不平衡二分类问题的性能,因此实验选用了敏感性 Sen,F-score,AUC 这 3 个性能指标来进行评价。

$$Sen=\frac{TP}{TP+FN} \tag{8}$$

敏感性 Sen 反映了分类器发现少数类的能力。由于在不平衡数据集分类中少数类的正确率往往受到更多重视,因此选择 Sen 为实验结果的性能度量指标。

F-score 是查准率 precision 与查全率 recall 的调和平均数,其值接近查准率 precision 与查全率 recall 中的较小者:

$$F\text{-}score=2\times\frac{precision\times recall}{precision+recall} \tag{9}$$

查准率 precision 与查全率 recall 的计算式如下:

$$precision=\frac{TP}{TP+FP} \tag{10}$$

$$recall=\frac{TP}{TP+FN} \tag{11}$$

F-score 也是评价分类器在不平衡数据上性能的重要指标之一。

二分类器以真阳性率(TP-rate)与假阳性率(FP-rate)为纵横二维坐标,形成性能空间分布曲线 ROC(Receiver Operating Characteristic)。ROC 曲线与纵轴与横轴所围成的面积即是 AUC 值。

4.2 MIA 互信息矩阵挖掘疾病相关基因与模块

4.2.1 Colon 结肠癌数据挖掘示例

本文利用 MIA 算法筛选候选基因,以候选基因作为与疾病相关的种子基因。为了验证 MIA 挖掘的候选基因是否与

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

[2] https://archive.ics.uci.edu/ml/datasets

疾病相关，采用表 9 中的 Colon 数据对 MIA 算法进行实验。

Colon 是由 Alon 等收集的结肠癌基因表达数据，共有 62 个样本，其中正常样本 22 个，疾病样本 40 个。Colon 基因数是 2000，数据标准化预处理后分别在正常样本和疾病样本中计算这 2000 条基因的互信息值。设定 $T_1=2.2$，$T_2=1.0$，得到的互信息矩阵 $\boldsymbol{M}_1$ 和 $\boldsymbol{M}_2$ 经过“逻辑与”运算，获得差异共表达网络邻接矩阵 $\boldsymbol{M}$。

按 MIA 算法处理得到的邻接矩阵 $\boldsymbol{M}$ 是仅有 0 和 1 两种元素值的对称矩阵，直接对应 Colon 结肠癌数据集的差异共表达邻接网络图 $\boldsymbol{G}$。$\boldsymbol{G}$ 中的每个节点基因是 MIA 算法筛选的候选基因，实验以 $\boldsymbol{G}$ 中挖掘的最大团(clique)中的基因作为种子基因，结合文献资料验证与致病基因的相关性。

在构建差异共表达网络邻接矩阵 $\boldsymbol{M}$ 的过程中，阈值 T_1 和 T_2 是两个关键参数。如果 T_1 太大或 T_2 太小，则 $\boldsymbol{M}$ 所对应的图 $\boldsymbol{G}$ 中边很少，邻接矩阵网络图密度很小，有可能挖掘不到所期望的 clique；如果 T_1 太小或 T_2 太大，则又会出现大量无实际意义的重叠 clique。因此，T_1 和 T_2 阈值需要根据邻接网络图的密度来合理选择，这个密度即是图中边的条数占最大可能边数的比例，取值为[0,1]。孤立点密度为 0，clique 的密度为 1。本实验设定 $T_1=2.2$，$T_2=1.0$，挖掘的最大 clique 为 4。

clique 中的任意两个基因都是连通的，密度为 1。Colon 共挖掘出 6 个分别有 4 条基因的 clique，如图 3 所示。6 个 clique 形成的图 3 模块中，共有 8 条种子基因，包含了 19 条边。由于最大可能边的数目是 C_8^2，图 3 模块的密度是 0.68。8 条种子基因登录号分别为：M63391，H64489，R87126，X74295，T92451，J02854，X86693，U19969 和 M63391。

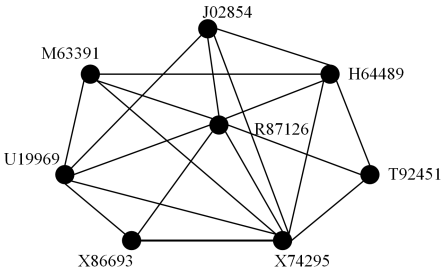


图 3 6 个相互重叠的 clique

Fig. 3 Six overlapping cliques

表 11 列出了图 3 所示模块的 8 个种子基因登录号与 Colon 数据致病基因研究的文献成果的相关性。其中，文献[26]采用 unified framework 挖掘与疾病相关的差异表达基因，文献[27]用集成决策方法挖掘复杂疾病相关基因，文献[28]基于贝叶斯网络方法选择最具可能性的致病基因。

表 11 MIA 算法挖掘的 8 个种子基因与文献致病研究的相关性

Table 11 Correlation of 8 seed genes mined by MIA algorithm and pathogenic in literatures

参考文献	8 条种子基因与致病基因研究文献成果相关性
[26]	8 条种子基因均包含在文献[26]提到的 66 个与疾病相关的差异表达基因中
[27]	基因 M63391、基因 J02854 和基因 R87126 包含在文献[27]的复杂疾病相关基因中
[28]	基因 M63391、基因 R87126、基因 T92451、基因 J02854 包含在文献[28]的 10 个最具可能的致病基因中
[29]	文献[29]的致病基因研究成果中也包含基因 J02854

从表 11 不难看出，差异共表达基因模块中的这 8 条与疾病相关的种子基因，与许多 Colon 数据的研究文献分析成果相吻合，说明 MIA 算法挖掘致病种子基因策略是有效的。

4.2.2 致病模块识别性能比较

本实验以肝细胞癌(Hepatocellular Carcinoma，HCC)致病基因模块的识别为例来分析比较 MIA 算法的致病模块识别性能。使用从 TCGA 库下载的 HCC RNA-seq 数据用于检测基因调控网络模块，其中正常样本 50 个，肿瘤样本 371 个。与 MIA 比较的致病基因模块识别算法为 mRank 与 Modularity+Crank。表 12 列出了这 3 种算法获得的排名前 10 的模块在 TCGA HCC 数据集上的分类指标平均值。

表 12 排名前 10 模块在 HCC 数据集的分类指标平均值

Table 12 Average classification indexes of the top 10 modules on HCC

algorithm	*Acc*	*Sen*	*Spe*	*F-score*	*AUC*
MIA	0.981	0.982	0.980	0.979	0.995
mRank	0.990	0.980	0.980	0.989	0.996
Modularity+Crank	0.970	0.980	0.960	0.970	0.992

从表 12 不难看出，MIA 有较好的致病基因模块识别功能，其分类性能优于 Modularity+Crank，与 mRank 算法大致相当。

4.3 DICPG 致病基因预测算法性能比较

为了评价 DICPG 的有效性和效率，采用了统计学习方法中一些常用的评估指标和方法，包括敏感性(Sen)、F-分数(F-Score)、ROC 曲线面积(Area Under Curve，AUC)、优劣记录(Win/Draw/Loss record，W/D/L)、散点图(scatter plot)、Friedman 和 Nemenyi 显著性检验方法(significance test)等。

4.3.1 分类性能比较

为验证 DICPG 致病基因预测方法的有效性，使用表 9 中的 9 个数据集进行实验。4 个癌症基因数据集均为高维数据，先使用 Emeditor 和 Excel 工具将从网站下载的数据集文本格式文件转变为 Arff 格式文件，然后通过 Weka 平台的过滤器进行降维预处理。leukemia，colon-cancer，SRBCT 和 brain 各选 143，198，128，186 个信息增益最大的基因构成关键基因。

实验中使用 NB 和 SVM 作为分类器，分类器参数均使用默认参数设置，以最大化突出采样方法自身的特点。评价指标主要选用 F-score，AUC 和 Sen。实验使用 5 折交叉验证的方法，将数据集随机分成 5 份，每次将其中 4 份作为训练集，剩下的 1 份作为测试集，重复 5 次。最后，将 5 次实验评价结果的平均值作为交叉验证的结果。所有的实验结果均为 5 次 5 折交叉验证结果，如表 13 所列。

由表 13 不难看到，当使用 SVM 分类器时，DICPG 算法在 SRBCT，brain，mushrooms 这 3 个数据集上，3 个性能指标值均最优；在 colon-cancer 数据集上，Sen 和 AUC 这 2 个性能指标值最优；在 leukemia 数据集上，1 个性能指标值 AUC 最优。综合来看，DICPG 算法的 SVM 分类性能指标最好。

当使用 SVM 分类器时，DCEG 算法在 leukemia 数据集上，Sen 和 F-score 这 2 个性能指标值最优；在 breast-cancer 数据集上，1 个性能指标值 F-score 最优。而 NGDE 算法在

breast-cancer 数据集上，Sen 和 AUC 这 2 个性能指标值最优；在 colon-cancer 数据集上，性能指标值 F-score 最优。

当使用 NB 分类器时，DICPG 算法在 SRBCT，brain，mushrooms 这 3 个数据集上，3 个性能指标值均最优；在 leukemia 和 breast-cancer 数据集上，F-score 和 AUC 这 2 个性能指标值最优。综合来看，DICPG 算法的 RF 分类性能最好。

当使用 NB 分类器时，DCEG 算法在 colon-cancer 数据集上，F-score 和 AUC 这 2 个性能指标值最优；而 NGDE 算法在 leukemia 和 colon-cancer 数据集上，性能指标值 Sen 最优。

DICPG 在 SVM 与 NB 分类器上对不平衡基因表达数据都取得了综合最好结果，说明 DICPG 的分类结果同样适用于多数分类器。

表 13 DICPG 与 t-test/f-test，IEG，DCEG，NGDE 方法的对比结果

Table 13 Comparison results between DICPG and t-test/f-test，IEG，DCEG，NGDE methods

数据集	评价指标	SVM					NB				
		DICPG	t-test/f-test	IEG	DCEG	NGDE	DICPG	t-test/f-test	IEG	DCEG	NGDE
leukemia	*Sen*	0.924	0.780	0.847	**0.933**	0.895	0.901	0.892	0.740	0.853	**0.915**
	F-score	0.772	0.835	0.580	**0.864**	0.761	**0.892**	0.865	0.539	0.880	0.557
	AUC	**0.933**	0.778	0.923	0.786	0.926	**0.980**	0.960	0.756	0.966	0.768
colon-cancer	*Sen*	**0.843**	0.691	0.533	0.726	0.775	0.819	0.752	0.558	0.845	**0.892**
	F-score	0.817	0.712	0.582	0.731	**0.820**	0.827	0.806	0.570	**0.909**	0.723
	AUC	**0.932**	0.837	0.793	0.848	0.926	0.930	0.935	0.793	**0.980**	0.844
SRBCT	*Sen*	**0.903**	0.726	0.651	0.735	0.703	**0.806**	0.721	0.710	0.740	0.697
	F-score	**0.945**	0.684	0.655	0.689	0.676	**0.941**	0.701	0.669	0.709	0.660
	AUC	**0.973**	0.850	0.840	0.850	0.861	**0.966**	0.868	0.842	0.876	0.836
brain	*Sen*	**0.737**	0.503	0.356	0.570	0.590	**0.855**	0.681	0.415	0.713	0.592
	F-score	**0.823**	0.562	0.431	0.602	0.748	**0.816**	0.730	0.445	0.736	0.617
	AUC	**0.905**	0.769	0.720	0.792	0.899	**0.912**	0.904	0.744	0.897	0.794
breast-cancer	*Sen*	0.638	0.605	0.619	0.510	**0.693**	0.810	**0.837**	0.597	0.517	0.518
	F-score	0.670	0.608	0.564	**0.694**	0.627	**0.860**	0.783	0.565	0.624	0.646
	AUC	0.868	0.831	0.834	0.820	**0.902**	**0.967**	0.902	0.833	0.884	0.872
mushrooms	*Sen*	**0.733**	0.644	0.644	0.610	0.661	**0.826**	0.715	0.616	0.720	0.573
	F-score	**0.756**	0.655	0.665	0.647	0.664	**0.828**	0.728	0.634	0.727	0.616
	AUC	**0.963**	0.936	0.949	0.935	0.951	**0.982**	0.958	0.923	0.957	0.930

4.3.2 可视化算法 AUC 平均值对比结果

AUC 是不平衡分类器综合性能公认的客观评价指标。由表 13 可获得 5 种算法在 6 个数据集上的平均 AUC 值，表 14 列出了 6 个数据集上各算法的 SVM 和 NB 分类 AUC 平均值，其中符号“*”表示算法 DICPG 和其他算法相比 AUC 提升显著。

表 14 算法在 6 个数据集上的平均 AUC 值

Table 14 Average AUC values of the algorithms on 6 datasets

数据集	DICPG	t-test/f-test	IEG	DCEG	NGDE
leukemia	0.9565	0.8690*	0.8395*	0.8760*	0.8470*
colon	0.9310	0.8860	0.7930*	0.9140	0.8850
SRBCT	0.9695	0.8590*	0.8410*	0.8630*	0.8485*
brain	0.9085	0.8365*	0.7320*	0.8445*	0.8465*
breast	0.9175	0.8665*	0.8335*	0.8520*	0.8870
mushrooms	0.9725	0.9470	0.9360	0.9460	0.9405

本文选取性能差异比例阈值为 0.05，差异比例小于 0.05 则认为两个算法质量相当(Draw)，否则比较结果存在输赢(Loss 或 Win)。W/D/L 为两个算法的比较结果提供了更直观的显示和解释，如表 15 所列。

表 15 DICPG 与其他算法在 AUC 上的 W/D/L 值

Table 15 W/D/L values of DICPG and other algorithms on AUC

W/D/L	t-test/f-test	IEG	DCEG	NGDE
DICPG	4\2\0	5\1\0	5\1\0	3\3\0

由表 15 看到，在 6 个数据集的 AUC 指标上，DICPG 相对于 t-test/f-test，IEG，DCEG，NGDE 算法分别赢了 4，5，5，3 次，其余的质量相当。

图 4 分别给出了 DICPG 与 t-test/f-test，IEG，DCEG，NGDE 这 4 种算法的 AUC 散点图。其中，横轴表示表 9 的 6 个数据集的索引数，纵轴表示 DICPG 与 t-test/f-test，IEG，DCEG，NGDE 关于 AUC 的比值 η。当 $\eta>1$ 时，说明 GBAF 比另一种算法表现更好。

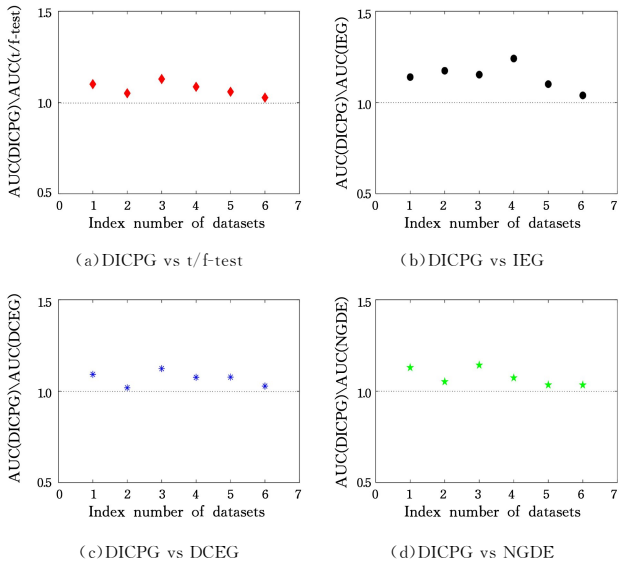


图 4 DICPG 与 t-test/f-test，IEG，DCEG，NGDE 算法的 AUC 散点图

Fig. 4 AUC scatter plots for DICPG and t-test/f-test，IEG，DCEG，NGDE algorithms

由图 4 可以清楚地看到，4 个散点图中，6 个 η 取值全都大于 1，意味着 DICPG 在 AUC 指标上从未输过其他 4 种

算法，只是在某些数据集上的值相对更接近而已，证明DICPG的不平衡分类性能优于其他4种基因选择方法。

4.3.3 分类时间比较

本文算法的另一个优势是训练时间相对较少。表16展示了不同选择方法在表3所列9个数据集上的时间消耗对比。

表16 不同基因选择方法的时间对比

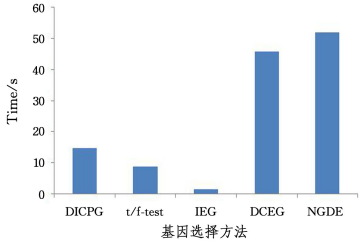
Table 16 Comparison of time for different gene selection methods

(s)

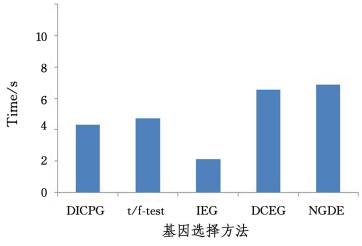
数据集	分类器	DICPG	t/f-test	IEG	DCEG	NGDE
Leukemia	SVM	0.53	0.66	**0.39**	0.67	0.71
	NB	1.77	1.84	**1.70**	1.91	1.98
Colon	SVM	27.92	22.04	**1.30**	80.22	80.74
	NB	7.32	9.45	**2.32**	13.76	13.91
SRBCT	SVM	0.31	0.34	**0.31**	0.36	0.39
	NB	**1.58**	1.56	1.54	1.56	1.58
Brain	SVM	75.68	20.02	**7.59**	189.22	238.91
	NB	13.48	16.36	**4.43**	18.02	19.92
Breast cancer	SVM	14.78	14.35	**3.08**	62.42	65.56
	NB	4.75	3.73	**2.86**	9.61	9.48
Duke_bc	SVM	0.73	0.58	**0.28**	1.32	1.29
	NB	1.67	1.33	**1.49**	1.74	1.76
Heart	SVM	3.46	2.31	**0.44**	9.25	9.30
	NB	2.14	2.57	**1.61**	2.59	2.66
Mushrooms	SVM	1.84	1.189	**0.44**	2.95	3.161
	NB	2.24	2.169	**1.65**	2.45	2.452
Protein	SVM	6.81	5.09	**0.44**	66.16	66.20
	NB	4.45	3.64	**1.65**	10.48	10.18

由表16可以看出，IEG，t-test/f-test，DICPG，DCEG，NGDE的总计用时分别为33.52 s，109.23 s，171.46 s，474.69 s，530.18 s。

图5是不同基因选择方法在9个数据集上SVM和NB分类的平均耗时柱形示意图。IEG，t-test/f-test，DICPG，DCEG，NGDE在9个数据集上SVM分类器的平均分类时间分别为1.59 s，7.40 s，14.67 s，45.84 s，51.81 s；5个算法的NB分类器平均分类时间分别为2.14 s，4.74 s，4.38 s，6.90 s，7.10 s。



(a)SVM 分类



(b)NB 分类

图5 在9个数据集上分类器的平均分类时间比较

Fig. 5 Comparison of average classification time of classifiers on nine datasets

4.3.4 统计分析和检验

本实验使用Friedman检验和Nemenyi后续检验方法，依据表13的实验结果统计分析5种致病基因预测方法的显著差异性，同时检验DICPG算法使用2种分类器SVM与NB的显著性差异。

1)Friedman 检验

表17列出了DICPG算法的SVM与NB分类性能比较序值，6个数据集，3个评价指标，共18个样本。

表17 SVM与NB分类性能比较序值

Table 17 Sorting values of performance comparison

SVM	NB	SVM	NB	SVM	NB
1	2	2	1	2	1
1	2	2	1	1	2
1	2	1	2	1	2
2	1	1	2	2	1
2	1	2	1	2	1
2	1	2	1	2	1

表17中，$r_1=1.611$，$r_2=1.389$，$k=2$，$N=18$，计算 $\tau_{\chi^2}=\frac{12N}{k(k+1)}\left(\sum_{i=1}^{k}r_i^2-\frac{k(k+1)^2}{4}\right)=0.887$，$\tau_F=\frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}}=0.881$。查$F$检验常用临界值表，$F_{0.05}=4.463$。由于$\tau_F<F_{0.05}$，因此2个分类器没有显著的差异性，DICPG算法不依赖于分类器。

表18列出了5种致病基因预测方法使用SVM与NB分类器的性能平均值，2个分类器，3个评价指标共，6个样本。

表18 5种致病基因预测方法的3个性能平均值

Table 18 Three performance averages of five pathogenic gene prediction methods

性能平均值	SVM			NB		
	Acc	*F-score*	*AUC*	*Acc*	*F-score*	*AUC*
DICPG	0.796	0.797	0.929	0.836	0.861	0.956
t/f-test	0.658	0.676	0.834	0.766	0.769	0.921
IEG	0.608	0.580	0.843	0.606	0.570	0.815
DCEG	0.681	0.705	0.839	0.731	0.764	0.927
NGDE	0.720	0.716	0.911	0.698	0.637	0.841

表18中，$k=5$，$N=6$。由表18计算得出5种基因选择方法的平均序值分别为$r_1=1$，$r_2=3.5$，$r_3=4.667$，$r_4=3.167$，$r_5=2.667$，$\tau_{\chi^2}=17.217$，$\tau_F=12.691$。查F检验常用临界值表$F_{0.05}=2.873$。由于$\tau_F>F_{0.05}$，“5种基因选择方法性能相同”假设被拒绝，因此对5种基因选择方法进行后续检验。

2)Nemenyi 后续检验

计算表18的Nemenyi平均序值差别临界值域$CD=q_\alpha\sqrt{\frac{k(k+1)}{6N}}$，查表$q_{0.05}=2.728$，得$CD=2.49$。由5种基因选择方法的平均序值以及CD值得出它们的检验图，如图6所示。

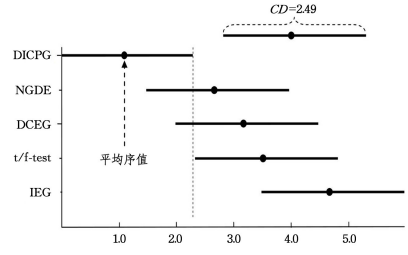


图6 Nemenyi 检验图

Fig. 6 Nemenyi test chart

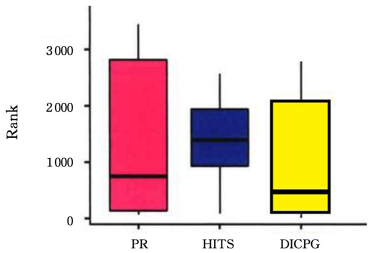
从图 6 中容易看出，DICPG 与 DCEG 和 NGDE 没有显著差别，因为它们的横线段有交叠区域。DICPG 算法相对最优；DICPG 算法显著优于 t/f-test 和 IEG 两种方法，因为它们的横线段没有交叠区域。以上统计分析结论的可信度为 95%。

4.3.5 DICPG 与随机游走致病基因识别性能对比

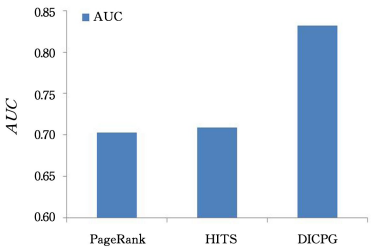
实验使用基因表达数据 GSE26125，该数据集与人类先天性心脏病（Congenital Heart Disease，CHD）相关，有 16 个疾病样本和 5 个对照样本，从 NCBI GEO 数据库下载使用。

生物分子网络数据使用人类蛋白质-蛋白质相互作用网络（Protein-Protein Interaction Network，PPIN），与 GSE26125 匹配后，获得包含 5 208 个节点和 19 588 条边的 CHD 特异性 PPIN。从 GWASdb 库筛选致病基因并与 GSE26125 匹配，获得 CHD 的 18 个致病基因，其余 PPIN 中的基因视为候选正常基因。

随机游走算法 PageRank 和 HITS 通过计算特征值作为区分 CHD 致病基因与正常基因的决策值，DICPG 算法则利用基因节点和近邻的差异表达信息识别候选致病基因。3 种算法的 CHD 致病基因排序结果和 30 次分类实验 AUC 值的比较如图 7 所示。



(a)CHD 致病基因排序结果箱线图



(b)3 种算法分类 30 次的 AUC 结果对比

图 7 DICPG 算法与随机游走算法的比较

Fig. 7 Comparison between DICPG and random walk algorithm

3 种算法的 CHD 致病基因分类性能指标如表 19 所列。

表 19 3 种算法的 CHD 致病基因分类性能指标

Table 19 CHD pathogenic gene classification indicators of three algorithms

algorithm	*Acc*	*Sen*	*Spe*	*F-score*	*AUC*
PageRank	0.782	0.783	0.781	0.777	0.703
HITS	0.790	0.781	0.780	0.788	0.709
DICPG	0.870	0.796	0.830	0.840	0.832

从图 7 和表 19 可以看出，DICPG 算法在 CHD 致病基因重要值排序、致病基因分类预测性能等方面的实验结果都优于流行的随机游走算法 PageRank 和 HITS。

结束语 依据基因相关性挖掘基因邻接关系是筛选候选致病基因的关键问题。由于实际的基因表达数据是包含大量噪音的、连续的实数据值，而且通常样本数目不到 100，因此基因相关性度量既不符合统计度量要求，也不符合信息增益直接度量的要求。基因相关性度量目前尚无一个公认的、严格精确的定义。

对于复杂的基因相似性关系，传统的距离、相关系数等只能反映基因表达谱之间的线性相似性；而互信息等基于熵的度量方法能挖掘出基因表达数据的模式相似性，有效反映基因之间真实的非线性复杂关系。

本文基于患病前后的基因表达差异数据，首先利用互信息计算基因间的相关性并构建邻接网络，然后设计特征向量模型用于癌症致病基因预测，向量特征包括候选基因及其近邻的差异表达信息。实验结果验证了所提方法的有效性。

癌症的产生和发育受高度异质的多层面因素影响，往往需要整合基因表达谱数据、miRNA 和 CNV 等多源数据才能较全面地揭示癌症的复杂作用机理。因此，如何将基因表达谱数据与更多的源数据整合起来，更准确地预测癌症致病基因，是未来研究工作的重点。

参 考 文 献

[1] YANG S F，CHANG C W，WEI R J，et al. Involvement of DNA Damage Response Pathways in Hepatocellular Carcinoma[J]. BioMed Research International，2014，16：283-291.

[2] ZHANG X，ZOU Q，RODRIGUEZ-PATON A，et al. Meta-Path Methods for Prioritizing Candidate Disease MiRNAs[J]. IEEE ACM Trans. Comput. Biol. Bioinf.，2019，16：283-291.

[3] LI X，CHANG M，WANG L. Information Recognition of Pathogenic Modules in Gene Statistics of Big Data[J]. Nanomater Energy，2021，10：35-42.

[4] COLLIER O，STOVEN V，VERT J P. LOTUS：a Single and Multitask Machine Learning Algorithm for The Prediction of Cancer Driver Genes[J]. PLoS. Comput. Biol.，2019，15：100-108.

[5] LUO P，DING Y，LEI X，et al. deepDriveer：Predicting Cancer Driver Genes Based on Somatic Mutations Using Deep Convolutional Neural Networks[J]. Front Genet.，2019，15：12-19.

[6] LIU X，TANG W H，ZHAO X M，et al. A Network Approach to Predict Pathogenic Genes for Fusarium Graminearum[J]. PLoS ONE，2010，5：e13021.

[7] BOLDI P，SANTINI M，VIGNA S. PageRank as A Function of The Damping Factor[C]// Proceedings of The 14th International Conference on World Wide Web. 2005：557-566.

[8] CHAKRABARTI S，DOM B E，KUMAR S R，et al. Mining The Web's Link Structure[J]. Computer，1999，32(8)：60-67.

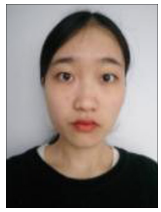
[9] NEWMAN M E J. Modularity and Community Structure in Networks[J]. National Academy of Sciences，2006，103(23)：8577-8582.

[10] ZITNIK M，SOSIC R，LESKOVEC J. Prioritizing Network Communities [J]. Nature Communications，2018，9(1)：1-9.

[11] SHANG H X，LIU Z P. Prioritizing Type 2 Diabetes Genes by Weighted PageRank on Bilayer Heterogeneous Networks[J]. IEEE/ACM Transactions on Computational Biology and Bioin-

formatics, 2021, 18(1): 336-346.
[12] PONTES B, GIRALDEZ R, AGUILAR-RUIZ J S. Biclustering on Expression Data: A Review[J]. Journal of Biomedical Informatics, 2015, 57(6): 163-180.
[13] CHENG L, YANG H, ZHAO H, et al. MetSigDis: A Mamually Curated Resource for The Metabolic Signatures of Diseases[J]. Briefings BioInf., 2019, 20: 203-209.
[14] POTTINGER T D, PUCKELWARTZ M J, PESCE L L, et al. Pathogenic and Uncertain Genetic Variants Have Clinical Cardiac Correlates in Diverse Biobank Participants[J]. J. Am. Heart. Assoc., 2020, 9: 26.
[15] ZOU Y, HUI R, SONG L. The Era of Clinical Application of Gene Diagnosis in Cardiovascular Diseases Is Coming[J]. Chronic. Dis. Transl. Med., 2019, 5: 214-220.
[16] TIMILSINA M, YANG H, SAHAY R, et al. Predicting Links Between Tumor Samples and Using 2-Layered Graph Based Diffusion Approach[J]. BMC Bioinf., 2019, 20: 1-20.
[17] XU B, LIU Y, YU S, et al. A Network Embedding Model for Pathogenic Genes Prediction by Multi-Path Random Walking on Heterogeneous Network[J]. BMC Med Genomics, 2019, 12: 188.
[18] ZHANG H P, WANG H N, LU G M, et al. Finding Differentially Co-Expressed Disease-Related Genes Based on Mutual Information[J]. Journal of Southeast University (Natural Science Edition), 2009, 39: 151-155.
[19] YU L, REN S J. Prediction of Cancerous Pathogenic Genes Based on Network and Gene Differential Expression Information [J]. Scientia Sinica Vitae, 2023, 53(1): 94-108.
[20] SHANNON C E. A Mathematical Theory of Communication [J]. The Bell System Technical Journal, 1948, 27: 379-423.
[21] WANG L, CHEN P, CHEN S, et al. A Novel Approach to Fully Representing The Diversity in Conditional Dependencies for Learning Bayesian Network Classifier[J]. Intelligent Data Analysis, 2021, 25(11): 35-55.
[22] DUAN Z, WANG L, CHEN S, et al. Instance-Based Weighting Filter for Superparent One-Dependence Estimators[J]. Knowledge-Based Systems, 2020, 203(8): 106-115.
[23] CABUZ S, ABREU G. Causal Inference for Multivariate Stochastic Process Prediction [J]. Information Sciences, 2018, 448(12): 134-148.
[24] SUN J, TAYLOR D, BOLLT E M. Causal Network Inference by Optimal Causation Entropy[J]. SIAM Journal on Applied Dynamical Systems, 2015, 14(3): 73-106.
[25] CHUA H N, SUNG W K, WONG L. Exploiting Indirect Neighbours and Topological Weight to Predict Protein Function from Protein—Protein Interactions[J]. Bioinformatics, 2006, 22(13): 1623-1630.
[26] SHAIK J S, YEASIN M. A Unified Framework for Finding Differentially Expressed Genes from Microarray Experiments[J]. BMC Bioinformatics, 2007, 8: 347.
[27] LI X, RAO S, WANG Y, et al. Gene Mining: A Novel And Powerful Ensemble Decision Approach to Hunting for Genes Using Microarray Expression Profiling[J]. Nucleic Acids Research, 2004, 32(9): 2685-2694.
[28] DIAO Q, HU W, ZHONG H, et al. Disease Gene Explorer: Display Disease Gene Dependency by Combining Bayesian Networks with Clustering[C]// Proceedings of The IEEE Computational Systems Bioinformatics Conference. Stanford, USA, 2004: 574-575.
[29] ZHANG X W, YAP Y L, WEI D, et al. Molecular Diagnosis of Human Cancer Type by Gene Expression Profiles and Independent Component Analysis[J]. European Journal of Human Genetics, 2005, 13(12): 1303-1311.

LI Zhijie, born in 1964, Ph.D, associate professor. His main research interests include computational biology, online learning of big data, and data mining.



LIAO Xuhong, born in 1997, master. Her main research interests include computational biology and data mining.

(责任编辑:何杨)