doi: 10.3969/j.issn.1006-1576.2011.07.015

生理学信息在基因标签提取中的应用

易波，文天柱，张原
（海军航空工程学院研究生管理大队，山东 烟台 264001）

摘要：根据临床研究所得生理学信息和结肠癌基因表达谱数据，对生理学信息在确定基因标签中的作用进行研究。在 ReCorre 算法的基础上，建立基于后验概率最大准则的基因标签提取模型，利用信息融合方法，将后验概率和 Bhattacharyya 距离相结合，建立融入临床知识的基因标签提取模型。实验结果证明了临床研究生理学信息在基因标签提取过程中的作用，该模型确定的基因标签进行留一交叉检验正确率可达 100%。

关键词：肿瘤；基因表达谱；信息融合
中图分类号：O236　　**文献标志码**：A

Application of Physiological Information in Selection of Informative Genes

Yi Bo, Wen Tianzhu, Zhang Yuan
(Administrant Brigade of Postgraduate, Naval Aeronautical & Astronautical University, Yantai 264001, China)

Abstract: For large number of gene expression profiles of the colon tumor, how to use the physiological information obtained in the clinical investigation was researched on during the process that the informative genes were selected in this paper. First, the ReCorre feature selection algorithm was applied, and a primary selection model of informative genes was built up which was based on the maximal criterion of the posterior probability. Then, the posterior probability and the Bhattacharyya distance were fused in the new selection model with the physiological information, so that the new informative genes could be selected. In the end, 2000 colon tumor gene expression samples were analyzed and the usage of physiological information turned out to be feasible and effective in the process of getting the best informative genes, which were checked out by the leave-one-out cross validation (LOOCV) method based on the sport vector machine (SVM) later in order to get the best informative genes.

Keywords: tumor; gene expression profiles; information fusion

0　引言

对基因标签选取问题的研究是发现肿瘤特异表达基因、研究肿瘤基因表达模式的重要手段。目前已对此问题进行了大量研究[1-4]。对于结肠癌的诊断，已有临床研究给出一定量的生理学信息，这些信息对于基因标签的确定也可以起到辅助作用。

对临床研究给出的生理学信息以及基因表达图谱，可考虑进行信息融合导出更为有效的信息，对样本数据进行诊断。笔者在 ReCorre 算法的基础上，利用信息融合的方法[5]，将后验概率和 Bhattacharyya 距离相结合，建立融入生理学信息的基因标签提取模型。并以实验证明了临床研究的生理学信息应用在基因标签提取过程中的可行性和有效性。

1　基于 Recorre 算法的特征选取

ReCorre 的核心思想是：首先运行 ReliefF 算法，得到每个基因的权重 w_i，将权值大于设定阈值 δ 的基因加入到初始状态为空的基因集 C，再求出基因集 C 中任意 2 个特征间的相关度(冗余度)，在相关度大于阈值 λ 的 2 个基因中，删除用 ReliefF 算法得到的权值小者，最后输出基因集 C。

1.1　ReliefF 算法

ReliefF 算法从同类和不同类中各选取 k 个最近邻样本，求其表达水平的均值，从而得到每个样本实例中各个基因的权值。ReliefF 算法具体计算过程如下[6]：

输入：选出的 m 个样本实例(每个样本有 n 个特征)及其所属类别；输出：权值向量 w，其维数与特征个数相同，表征特征与分类标志的相关性。

1) 初始化：$w=0$；

2) 从 m 个样本实例中随机抽取一个 R_j，判断其所属类别，并找出 R_j 的 k 个同类最近邻样本构成样本集 $P=\{R_p \mid p=1,2,\cdots,k\}$ 和 k 个非同类最近邻

收稿日期：2011-02-25；修回日期：2011-04-02
作者简介：易波(1987—)，男，湖南人，硕士在读，从事海军航空兵作战运筹研究。

样本构成样本集 $Q=\{R_q \mid q=1,2,\cdots,k\}$；

3) 对于基因 g ($g=1,2,\cdots,n$)进行如下运算：

$$w[g]=w[g]-(\sum_{p=1}^{k}\mathrm{diff}(g,R_j,R_p))/mk+\sum_{c\neq \mathrm{class}(R_j)}\left[\frac{P(c)}{1-P(\mathrm{class}(R_j))}\sum_{l=1}^{k}\mathrm{diff}(g,R_j,R_q)\right]/mk \tag{1}$$

4) 再取下一个样本，转入 2)，当 m 个样本都参与计算转入 5)；

5) 输出最后的权值向量 w。

其中：$\mathrm{class}(R_j)$ 表示 R_j 与其同类所处的类别；$c\neq \mathrm{class}(R_j)$ 表示与 R_j 异类的所有类别；$P(c)$ 表示与 R_j 异类样本发生的概率；函数 $\mathrm{diff}(g,R_i,R_j)$ 用于计算 2 个不同样本 R_p 和 R_q 对应于基因 g 的差异。对于离散特征，

$$\mathrm{diff}(g,R_i,R_j)=\begin{cases}0, & \mathrm{value}(g,R_i)=\mathrm{value}(g,R_j)\\ 1, & \text{其他}\end{cases} \tag{2}$$

对于连续特征：

$$\mathrm{diff}(g,R_i,R_j)=\frac{\left|\mathrm{value}(g,R_i)-\mathrm{value}(g,R_j)\right|}{\max(g)-\min(g)} \tag{3}$$

其中：R_i，R_j 是 2 个样本，$\mathrm{value}(g,R_i)$ 表示第 i 个样本 R_i 的 g 基因水平表达值。

ReliefF 算法缺乏对基因之间的关联的处理，不能辨别冗余特征。参考文献[7]提出的组合式基因选择算法 ReCorre，该算法对 ReliefF 算法做进一步改进，能够有效去除冗余特征。

1.2　相关度计算

ReCorre 算法通过相关分析的方法来衡量基因之间冗余度。

1) 数值型特征间的相关度分析。对于任意 2 个数值型基因 $G_x=(x_1,x_2,\cdots,x_m)$ 和 $G_y=(y_1,y_2,\cdots,y_m)$，m 表示样本数量，则两者的相关泊松系数为：

$$r=\frac{\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i-\bar{x})^2\sum_{i=1}^{m}(y_i-\bar{y})^2}} \tag{4}$$

2 个数值型基因属性值的相关度 $correlate=r$。

2) 对于量化基因属性值，用熵和互信息相结合的方法能有效挖掘不同基因间的相似性。在信息论中，熵是随机变量的不确定度或自信息的平均值。设量化型基因 $G_x=(x_1,x_2,\cdots,x_m)$，m 表示样本数量，则其熵为：

$$H(G_x)=-\sum_{x\in G_x}p(x)\log_2 p(x) \tag{5}$$

其中 $p(x)$ 为对应基因属性值在总体中出现的概率，相应的 $p(x,y)$ 表示联合概率，则对应联合熵可表示为：

$$H(G_x,G_y)=-\sum_{x\in G_x}\sum_{y\in G_y}p(x,y)\log_2 p(x,y) \tag{6}$$

对于 2 个量化型基因 G_x 和 G_y，其互信息 $MI(G_x,G_y)$ 就是其中一个特征能提供给另一个特征的信息量，即：

$$MI(G_x,G_y)=\sum_{x\in G_x}\sum_{y\in G_y}p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)} \tag{7}$$

于是 2 个量化型基因的相关度为

$$correlate=2\left[\frac{MI(x,y)}{H(x)+H(y)}\right] \tag{8}$$

2　融合生理学信息的基因标签提取模型

临床研究生理学信息给出早期结肠癌样本中 5 号染色体长臂 APC 基因失活的概率 90%，以及 ras 相关基因突变概率 40%～50%。该信息和基因表达图谱的信息融合可以考虑从数据级融合、特征级融合和决策级融合[8]这 3 个不同的层次实施，如图 1。

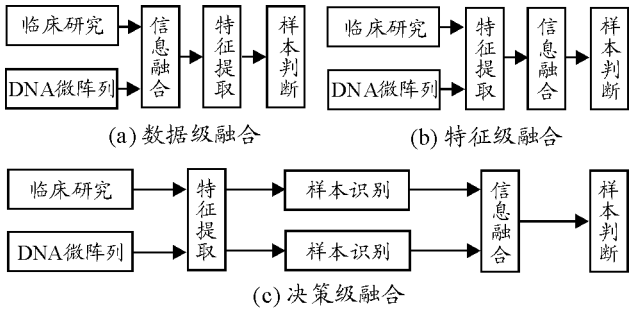

图 1　3 个层次信息融合

已知的生理学信息与 DNA 微阵列测出的基因表达谱信息是不同类的，不能进行数据级的融合。可将由 Bhattacharyya 距离得到的基因标签组和由最大后验概率得到的基因标签相融合，以实现特征级的融合并最终确定融入诊断信息的基因标签。

2.1　由最大后验概率确定可分性判据

根据最大后验概率准则，首先给出 2 个假设命题 H_0（假设检测样本是结肠癌患者）和 H_1（假设检测样本是正常人）。定义 c_{ij} 表示假设 H_j 为真却选择了 H_i 的代价，D_0 表示诊断为结肠癌患者，D_1 表示

诊断为正常人，r_0，r_1 分别表示假设 H_0，H_1 下的条件代价，因而总的平均代价等于各条件代价按其先验概率进行平均，即：

$$\overline{R} = P(H_0)r_0 + P(H_1)r_1 \tag{9}$$

贝叶斯准则要求确定区域 D_0 和 D_1 使得平均风险 $\overline{R}$ 最小，这个最小风险也称为贝叶斯风险[9]。

对于双择问题，有 $P(D_0 \mid H_1) = 1 - P(D_1 \mid H_1)$ 和 $P(D_0 \mid H_0) = 1 - P(D_1 \mid H_0)$。由式 (9) 得：

$$\begin{aligned}\overline{R} = {} & P(H_1)c_{01} + P(H_0)(c_{10} - c_{00})P(D_1 \mid H_0) - \\ & P(H_1)(c_{01} - c_{11})P(D_1 \mid H_1)\end{aligned} \tag{10}$$

虚警概率 $p(D_1 \mid H_0)$ 和检验概率 $p(D_1 \mid H_1)$ 可表示为：

$$\left.\begin{aligned} P(D_1 \mid H_0) &= \int_{D_1} p(G \mid H_0)\mathrm{d}g \\ P(D_1 \mid H_1) &= \int_{D_1} p(G \mid H_1)\mathrm{d}g \end{aligned}\right\} \tag{11}$$

其中 $p(G \mid H_0)$ 和 $p(G \mid H_1)$ 分别表示假设 H_0 和 H_1 条件下 n 个基因的联合概率密度函数，称为似然函数。

假设 H_1 和 H_0 下的似然函数之比称为似然比，记为 $\Lambda(G)$，则判决规则为：

$$\left.\begin{aligned} H_1\text{：}\ \Lambda(G) &= \frac{p(G \mid H_1)}{p(G \mid H_0)} > \frac{P(H_0)(c_{10} - c_{01})}{P(H_1)(c_{01} - c_{10})} = \Lambda_0 \\ H_0\text{：}\ \Lambda(G) &= \frac{p(G \mid H_1)}{p(G \mid H_0)} < \frac{P(H_0)(c_{10} - c_{01})}{P(H_1)(c_{01} - c_{10})} = \Lambda_0 \end{aligned}\right\}$$

其中：Λ_0 表示判决门限。由此可见贝叶斯意义下的最佳判决系统变为计算似然比 $\Lambda(G)$ 的系统。

最大后验概率准则与贝叶斯准则不同。因为后验概率反映了获得观测矢量 G 后的信息，所以对应较大后验概率的假设更可能出现，于是最大后验概率准则表示为：

$$\left.\begin{aligned} H_1\text{：}\ & P(H_1 \mid G) > P(H_0 \mid G) \\ H_0\text{：}\ & P(H_1 \mid G) < P(H_0 \mid G) \end{aligned}\right\}$$

为引入临床经验知识，充分理解后验概率，对于含有 i 个基因的基因组 G_i，可分性判据 $J(G_i)$ 重新定义为：

$$J_1(G_i) = 1 - (1 - P_1)(1 - P_2)\cdots(1 - P_i) \tag{12}$$

其中：P_j 表示第 j 个基因的后验概率，$j = 1, 2, \cdots, i$，i 为 G_i 中基因的个数。

2.2 由 Bhattacharyya 距离确定可分性判据

基因组 G_i 的 Bhattacharyya 距离 $J(G_i)$ 定义为：

$$\begin{aligned} J_2(G_i) = {} & \frac{1}{8}(\mu_2 - \mu_1)^{\mathrm{T}}\left[\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}\right]^{\mathrm{T}}(\mu_2 - \mu_1) + \\ & \frac{1}{2}\ln\frac{\left|\frac{\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2}{2}\right|}{\sqrt{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}} \end{aligned} \tag{13}$$

其中：μ_1 表示基因组中包含的全部基因在正常人样本中的均值向量；μ_2 表示基因组中包含的全部基因在癌症患者样本中的均值向量；$\boldsymbol{\Sigma}_1$ 表示基因组中包含的全部基因在正常人样本中的协方差矩阵；$\boldsymbol{\Sigma}_2$ 表示基因组中包含的全部基因在癌症患者样本中的协方差矩阵。

2.3 信息融合可分性判据

由式 (12) 和式 (13) 定义新的可分性判据

$$J(G_i) = weight_1 J_2(G_i) + weight_2 J_1(G_i) \tag{14}$$

其中：$weight_1$ 表示 Bhattacharyya 距离在基因特征获取中所占权重；$weight_2$ 表示最大后验概率准则在基因特征获取中所占比重。此时 $J(G_i)$ 只是用来表示基因组属性的一个特征，将式 (14) 作为新的可分性判据，即可求得融入诊断信息的基因标签。

3 实验与结果分析

实验数据来自 Alon 等人选出的含有 2 000 个特征基因的结肠癌基因表达谱数据集[10]，包括 40 个结肠癌组织样本和 22 个正常组织样本。该实验数据仅出现正常样本和肿瘤样本 2 类，对于任意挑选的样本 R_j，它的非同类只有 1 个，则 $\frac{P(c)}{1 - \mathrm{P(class}(R))} = \frac{P(c)}{P(c)} = 1$。

1) 根据分类基因的选择模型，求得各基因的权值，其分布如图 2。

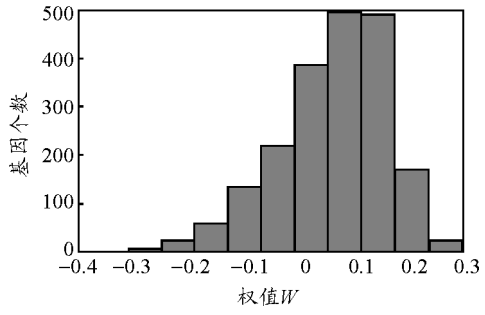

图 2 ReliefF 算法权值分布直方图

2) ReCorre 算法中求得各基因互信息的相关关系 SU 分布图如图 3。

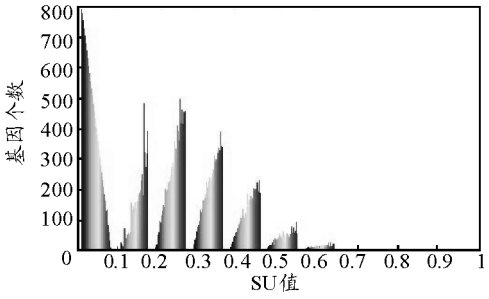

图 3　各基因互信息的 SU 分布图

3) 基于 Bhattacharyya 距离的权重分布见图 4。

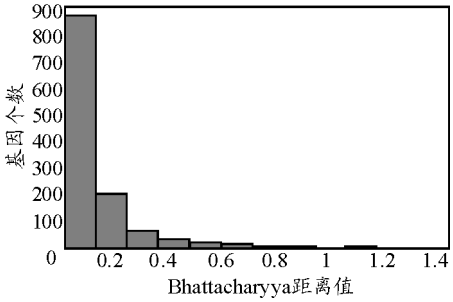

图 4　基于 Bhattacharyya 距离的权重分布

对于没有临床经验知识的基因，假设其后验概率为 50%，对于已知临床经验知识的染色体定义后验概率为：

$$p(H_i \mid g_j) = 1 - 0.5(1 - p_j)$$

其中：p_j 是根据临床经验已知的概率值。所以长臂 APC 基因对应的后验概率为 95%；ras 相关基因的后验概率为 75%。采用式 (14) 中信息融合后的 G_i 对由分类基因和后验概率已知基因组成的基因组应用交替选择算法确定了 3 组基因标签。再应用支持向量机进行留一交叉检验，得到检测正确率 100% 为最优，其包含的基因见表 1。

表 1　留一交叉检验确定的最优基因标签组

基因号	基因描述
U37673	Human neuron-specific vesicle coat protein and cerebellar degeneration antigen (beta-NAP) mRNA, complete cds.
Z50753	H.sapiens mRNA for GCAP-II/uroguanylin precursor.
R98410	S25641 HYPOTHETICAL PROTEIN - ;.
H04311	RAS GTPASE-ACTIVATING-LIKE PROTEIN IQGAP1 (Homo sapiens)

4　结束语

表 1 中 R98410 为与 ras 相关的基因，从仿真中不难发现：加入临床经验知识，即采用信息融合的方式进行基因标签的选取可以提高分类基因组的可分性，对算法实现进一步的优化，但与此同时算法复杂度也有所增加。采用 Alon 所得结肠癌数据集证明了对其噪声处理的可行性和有效性。笔者在信息融合模型中只研究了 2 种特征的融合，即后验概率和 Bhattacharyya 距离的融合，因此，在后续研究中可以考虑多种特征的融合。另外，也可以考虑决策级的信息融合算法。

参考文献：

[1] Ramaswamy S, Tamayo P, Rifkin R. et al.. Multiclass cancer diagnosis using tumor gene expression signatures[J]. Proceedings of the National Academy of Sciences of the United States of America, 2001, 98(26): 15149−15154.

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

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

[4] 李颖新, 刘全金, 阮晓钢. 急性白血病的基因表达谱分析与亚型分类特征的鉴别[J]. 中国生物医学工程学报, 2005, 24(2): 240−244.

[5] 苏思, 姜礼平, 邹明. 基于多分类支持向量机和证据合成方法的多传感器信息融合研究[J]. 兵工自动化, 2010, 29(1): 59−62.

[6] Marko Robnik-Šikonja. Igor Kononenko: Theoretical and Empirical Analysis of ReliefF and RReliefF[J]. Machine Learning Journal, 2003, 53(12): 23−69.

[7] 张丽新, 王家廞, 赵雁南, 等. 基于 Relief 的组合式特征选择[J]. 复旦学报: 自然科学版, 2004, 43(5): 893−898.

[8] 潘泉, 于昕, 程咏梅, 等. 信息融合理论的基本方法与进展[J]. 自动化学报, 2003, 29(4): 599−615.

[9] 张明友, 吕明. 信号检测与估计[M]. 北京: 电子工业出版社, 2006.

[10] 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[J]. Proc Natl Acad Sci USA，1999, 96: 6745−6750.