DOI:10.16452/j.cnki.sdkjzk.2012.03.012

基于支持向量机的结肠癌信息基因提取

李 烨,王永丽,贺国平
(山东科技大学 信息科学与工程学院,山东 青岛 266590)

摘 要:基于结肠癌基因表达谱数据集,提出了一种信息基因提取的新方法。该方法结合了支持向量机(SVM)、Bhattacharyya距离、递归特征消除(RFE)和快速基于相关性过滤器(FCBF)方法。首先,利用Bhattacharyya距离与SVM-RFE方法结合去除无关基因,然后运用FCBF方法得到信息基因,最后以支持向量机作为分类器对结肠癌样本进行分类识别。实验结果表明,同现有的方法相比,该方法在提取基因数量和准确率上都有明显的优势。

关键词:结肠癌;支持向量机;信息基因;Bhattacharyya距离;递归特征消除;快速基于相关性过滤器

中图分类号:O224 **文献标志码**:A **文章编号**:1672-3767(2012)03-0084-06

The Extraction for Informative Genes of Colon Cancer Based on Support Vector Machine

LI Ye, WANG Yongli, HE Guoping
(College of Information Sciences and Engineering, Shandong University of Science and Technology, Qingdao, Shandong 266590, China)

Abstract: A new informative gene extracting method was developed based on the dataset of expressible spectrum in colon cancer genes and the method combined with the support vector machine, Bhattacharyya distance, recursive feature elimination (RFE) and fast correlation-based filter (FCBF). Firstly, a combination of Bhattacharyya distance and SVM-RFE was applied to remove the irrelevant genes. Secondly, FCBF was employed to generate the informative genes. Finally, SVM was used to classify the colon datasets. The experimental results demonstrated that compared to existing methods, the classification accuracy was improved and the number of informative genes was decreased.

Key words: colon cancer; support vector machine; informative genes; Bhattacharyya distance; recursive feature elimination; fast correlation-based filter

DNA微阵列(DNA microarray),也叫基因芯片,是近年发展起来的一种能快速、高效检测DNA片段序列和基因表达水平的新技术,该技术已被广泛应用于生物医学研究、疾病诊断和药物筛选等领域[1-3]。从DNA芯片所测量的成千上万个基因中,找出决定样本类别的一组基因"标签",即"信息基因"(informative genes)是正确识别癌症类型、给出可靠诊断和简化实验分析的关键,也为抗癌药物的研制提供了方便。

近年来,对于结肠癌识别与信息基因选取的研究取得了新的进展。Alon等[4]用层次聚类等方法对结肠癌样本数据进行了分类研究;Guyon等[5]首次提出SVM-RFE(support vector machine recursive feature elimination,支持向量机-递归特征消除)方法,该方法利用递归特征去除对分类影响小的基因,成为基因选择的经典算法。游伟等[6]在SVM-RFE算法的基础上提出了SVM-RFE-SFS方法,将SVM-RFE和序列向

收稿日期:2011-12-02
基金项目:国家自然科学基金项目(10971122);山东省优秀中青年科学家科研奖励基金项目(2010BSE06047);高等学校博士学科点专项科研基金项目(20093718110005);山东省科技攻关计划项目(2009GG10001012);山东省自然科学基金项目(Y2008A01)
作者简介:李 烨(1986—),女,山东泰安人,硕士研究生,主要从事支持向量机算法及应用的研究.
E-mail:happyxiaoye@hotmail.com
王永丽(1977—),女,山东烟台人,副教授,主要从事最优化理论与算法、支持向量机的研究.

前选择(sequential forward selection,SFS)算法结合选取结肠癌信息基因。张小丹等[7]提出一种将非相关线性判别式分析方法和支持向量机结合的分类算法对结肠癌进行分类。刘全金等[8]提出了基于 Boosting 算法构建 SVM 分类器对结肠癌进行分类识别的思想。

本研究基于生物信息学理论,对结肠癌基因表达谱数据集进行分析,研究结肠癌识别与信息基因的选取。首先,将 Bhattacharyya 距离与 SVM-RFE 方法相结合去除了大部分无关基因;然后,利用 FCBF 算法对信息基因进行进一步筛选,得到最佳的分类信息基因;最后,以支持向量机作为分类工具进行结肠癌的识别。实验结果表明,所提出的提取分类信息基因的方法不但提取的基因数较少,而且分类准确率有了较大提高。

1 基于 Bhattacharyya 距离的无关基因的删除

本研究采用 Alon 等[9]于 1999 年用层次聚类等方法收集的结肠癌基因表达谱数据集。该数据集由 62 个样本组成,其中包括 40 个结肠癌组织样本和 22 个正常组织样本,每个样本含有 2 000 个基因的表达数据。该数据集中的数值表示基因在癌症患者或者正常人样本中的表达水平。

在基因表达谱中,一些基因的表达水平在正常人和结肠癌患者中的分布无论均值还是方差都很接近,这些基因显然对癌症的识别不会提供有用的信息,因此将这些基因作为无关基因删除。为了衡量基因含有的分类信息量,Golub 等采用了"信噪比"(signal to noise ratio)指标[10],即

$$d(g)=\frac{|\mu_{g+}-\mu_{g-}|}{\sigma_{g+}+\sigma_{g-}}。\tag{1}$$

其中:$d(g)$—基因 g 的信噪比;μ_{g+},μ_{g-}—基因 g 在正常样本和癌症样本两个类别中表达水平的均值;σ_{g+},σ_{g-}—正常样本和癌症样本的标准差。

由式(1)可知,当基因 g 在正常样本和癌症样本中的均值相同或者相差很小时,则 $d(g)=0$ 或 $d(g)\leqslant\theta$($\theta>0$,但非常小)。于是,该基因就可以当作无关基因,当基因 g 在正常样本和癌症样本中的方差出现较大差异时,该基因很有可能是致病基因。

基因的 Bhattacharyya 距离[11]为信噪比的推广,即

$$d^*(g)=\frac{(\mu_{g+}-\mu_{g-})^2}{4(\sigma_{g+}^2+\sigma_{g-}^2)}+\frac{1}{2}\ln\frac{\sigma_{g+}^2+\sigma_{g-}^2}{2\sigma_{g+}\sigma_{g-}}。\tag{2}$$

$d^*(g)$显示了基因在正常人和结肠癌患者中分布均值和分布方差的差异,$d^*(g)$越大,基因 g 越有可能是致病基因。

通过对 2 000 个基因的 Bhattacharyya 距离进行计算,得到 Bhattacharyya 距离的详细分布情况,如表 1 所示,Bhattacharyya 距离小于 0.07 的基因占 86.9%。这些基因在正常人和结肠癌患者中分布均值和分布方差的差异不大,因此可以作为无关基因。取 0.07~0.35 区间内的 262 个基因作为与癌症关联大的信息基因,基本可以代表全部 2 000 个基因的表达水平。通过 Bhattacharyya 距离,从 2 000 个基因中筛选出 262 个基因,为进一步进行结肠癌相关基因的提取作准备。

表 1 2 000 个基因关于 Bhattacharyya 距离的分布情况表

Tab.1 Distribution of 2 000 genes about Bhattacharyya distance

Bhattacharyya 距离	基因总数	所占百分比/%
0 ~ 0.07	1 738	86.90
0.07 ~ 0.08	65	3.25
0.08 ~ 0.10	79	3.95
0.10 ~ 0.20	100	5.00
0.20~ 0.35	18	0.90

2 SVM-RFE 信息基因选择方法

SVM-RFE 算法是基于支持向量机的一种有效的特征选择方法。支持向量机(support vector machine,SVM)[12]是由 Vapnik 等人基于结构风险最小化原理提出的一种机器学习算法,该算法在有限样本条件下具有极好的泛化能力,被广泛应用于机器学习、模式识别和基因分类等领域。

给定训练样本 $T=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\mid i=1,2,\cdots,m\mid\}$,$\boldsymbol{x}_i\in R^n$,$\boldsymbol{y}_i\in\{1,-1\}$。支持向量分类机的目标就是寻找一

个最优分类函数，将两类样本分开。引入映射 $\Phi: x \to \Phi(\phi_1(x), \cdots, \phi_n(x)) \in H$，则 SVM 的一般模型为

$$\begin{aligned} & \min_{w,b,\xi} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{m} \xi_i \\ \text{s. t.} \quad & y_i (w^{\mathrm{T}} \Phi(x_i) - b) \geqslant 1 - \xi_i ; \\ & \xi_i \geqslant 0 \\ & i = 1, 2, \cdots, m \end{aligned} \tag{3}$$

其对偶问题为

$$\begin{aligned} & \min \left(\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{j=1}^{m} \alpha_i \right) \\ \text{s. t.} \quad & \sum_{i=1}^{m} \alpha_i y_i = 0 \\ & 0 \leqslant \alpha_i \leqslant C, i = 1, 2, \cdots, m \end{aligned} \quad 。 \tag{4}$$

通过对问题(4)的求解，即可得到 SVM 的最优分类函数为

$$\sum_{i=1}^{m} \alpha_i y_i K(x_i, x) - b = 0。 \tag{5}$$

特别地，若映射 $\Phi(x) = x$，则问题(3)为线性可分问题，此时式(4)中的 $K(x_i, x_j) = x_i^{\mathrm{T}} x_j$；否则，为线性不可分问题。在线性不可分的情况下，$K(x_i, x_j) = \Phi(x_i)^{\mathrm{T}} \Phi(x_j)$ 就是所谓的核函数。目前常用且效果比较理想的核函数主要有多项式核函数(polynomial)、径向基核函数(radial basis function，RBF)和 Sigmoid 核函数等。径向基核函数是最常用的核函数，用来与本文方法比较的其他信息基因提取方法采用的是径向基核函数，为了方便对比，采用径向基核函数

$$K(x, y) = \exp(- \| x - y \|^2 / \sigma^2)。$$

SVM-RFE 算法是根据 SVM 在训练时生成的权向量 w 来构造排序系数，每次只删除一个特征，即每次求出分离超平面一个法向量 w^* 后，只删除对应于 w^* 的绝对值最小的那个分量的特征，然后利用其余特征组成新的训练集，重复上述步骤，最终得到特征排序表。该方法从基因集合的整体分类能力出发，得到的信息基因不仅具有较好的分类性能，也能更好地选出样本中蕴含的生物学信息。

算法 1　SVM-RFE 特征选择算法

Step 1　输入：训练样本矩阵 $X = [x_1, x_2, \cdots, x_m]^{\mathrm{T}}$，类别标签 $Y = [y_1, y_2, \cdots, y_m]^{\mathrm{T}}$

Step 2　确定保留的特征子集 $s = [1, 2, \cdots, k]$

Step 3　特征排序列表 $r = [\ \ ]$

Step 4　重复以下步骤直到 $s = [\ \ ]$：

{

a. 训练样本 $X_0 = X(:, s)$

b. 给定参数后训练分类器 $SVMtrain(y, X_0, c, g)$，其中 c 为惩罚参数，g 为 RBF 核参数

c. 得到权向量 w^*，删去 w^* 中绝对值最小的分量对应的特征 f

d. 更新特征排序列表：$r = [s(f), r]$

e. 排除具有最小判据的特征：$s = s(1: f-1, f+1: length(s))$

}

Step 5　输出：特征排序列表 r

采用以上 SVM-RFE 算法对 2 000 个基因进行了排序，为了与 Bhattacharyya 距离提取的信息基因数量一致，尽可能最大限度地保留信息基因，将 SVM-RFE 排序位于前 260 位的基因提出，与 Bhattacharyya 距离提取的 262 个信息基因共同组成癌症相关基因的集合，去除重复的基因，最终得到 352 个信息基因。

3 基于 FCBF 的信息基因的提取

在结肠癌研究领域，临床有以下生理学信息：大约 90%结肠癌在早期有 5 号染色体长臂 APC 基因失活，而只有 40%～50%的 ras 相关基因突变。依据已知致病基因的数据，利用与已知致病基因的相关性作为评价准则来筛选与结肠癌关系密切的基因，从而达到删除不相关和冗余基因的目的。本文采用 FCBF(fast correlation-based filter)[13]算法对 352 个信息基因进行相关性分析，删除不相关和冗余的基因，得到最佳的分类信息基因。

FCBF 算法是以对称不确定性 SU(symmetrical uncertainty)来度量两个特征之间相关性，表达式为

$$SU(\boldsymbol{X},\boldsymbol{Y})=2\left[\frac{M(\boldsymbol{X},\boldsymbol{Y})}{H(\boldsymbol{X})+H(\boldsymbol{Y})}\right]。\tag{6}$$

其中：M 为信息增益，$M(\boldsymbol{X},\boldsymbol{Y})=H(\boldsymbol{X})-H(\boldsymbol{X}\mid\boldsymbol{Y})$，$H(\boldsymbol{X})=-\sum_i p(x_i)\log_2 p(x_i)$，$H(\boldsymbol{X}\mid\boldsymbol{Y})=-\sum_j p(y_j)\sum_i p(x_i\mid y_i)\log_2 p(x_i\mid y_j)$。

由上述定义可知，对称不确定性 $SU(\boldsymbol{X},\boldsymbol{Y})$取值范围在区间$[0,1]$上，$SU(\boldsymbol{X},\boldsymbol{Y})$值越大，表明两个基因之间相关性越高，可以认为在生物学意义上这两个基因之间存在较强的相关性。当 $SU(\boldsymbol{X},\boldsymbol{Y})=0$ 时，表明两个基因不相关，而当 $SU(\boldsymbol{X},\boldsymbol{Y})=1$ 时，则表明两个基因完全相关。

FCBF 是 Yu 等[13]基于相关关系度量给出的一种特征选择算法。该算法基本原理如下：设数据集有 m 个样本，每个样本由 n 个非目标基因和一个目标基因 C 组成，如果非目标基因 $i(i=1,2,\cdots,n)$和 C 之间 SU 较低，小于给定阈值，那么基因 i 就会作为不相关基因去除，如果两个非目标基因的 SU 值过大，超过这两个基因与目标基因 C 之间 SU 值，那么认为这两个基因存在冗余，也会被去除。

FCBF 算法的具体步骤如下：

Step 1 输入训练集 $S=$(Gene 1,Gene 2,…,Gene n,C)和阈值 δ

Step 2 for $i=1$ to n do

计算每个基因 i 与目标基因 C 的对称不确定性 SU_{ic}

end

Step 3 if $SU_{ic}\geqslant\delta$

将基因 i 添加至 S'_{list}

end

Step 4 将 S'_{list}中的基因按照 SU_{ic} 值降序排列

Step 5 for $i=1$ to n do

for $k=j+1$ to n do

计算 SU_{ik}

if $SU_{ik}\geqslant SU_{kc}$

将基因 k 从 S'_{list}中删除

end

end

重复上述步骤，直到 S'_{list}中所有冗余信息基因被删除，$S_{\text{best}}=S'_{\text{list}}$

Step 6 输出 S_{best}

选定 APC 相关基因为 L35545 (Has. 2238,869)，并将该基因设为目标基因 C，阈值设为 0.6，对 Bhattacharyya 距离和 SVM-RFE 方法提取的 352 个信息基因，执行 FCBF 算法，得到 12 个基因，如表 2 所示。

表 2 利用 FCBF 方法从 352 个信息基因中提取出的 12 个基因表

Tab. 2 12 genes extracted from 352 informative genes by using FCBF method

序号	基因编号	EST name	Gen Bank Acc No
1	105	Hsa. 1978	T72879
2	601	Hsa. 1132	D16431
3	869	Hsa. 2238	L35545
4	943	Hsa. 471	M29277
5	1042	Hsa. 549	R36977
6	1258	Hsa. 5211	R67358
7	1260	Hsa. 36655	X89986
8	1649	Hsa. 1240	M31994
9	1668	Hsa. 1454	M82919
10	1671	Hsa. 627	M26383
11	1674	Hsa. 14069	T67077
12	1675	Hsa. 3083	D00596

4 基于支持向量机的识别与分类实验

考虑到癌症基因表达谱数据集高维数、小样本的特殊性，以及支持向量机在处理小样本、非线性、高维数问题中的优越性，本节将采用支持向量机作为分类工具对结肠癌样本进行分类识别。

实验中，对包含选取的12个基因的共计62个样本数据[9]进行训练和测试。首先，将62个样本的数据集划分为训练集和测试集(表3)。

表3 结肠癌数据集

Tab.3 The dataset of colo

训练集	测试集
Normal 14	Normal 8
Cancer 26	Cancer 14

然后，采用非线性SVM作为分类器，并选用径向基核函数和默认参数，进行了20组实验。每次都在62个样本中随机抽取40个样本组成训练集(26 cancer，14 normal)，剩下的22个样本组成测试集(14 cancer，8 normal)。

由于实验数据的样本数较少，为了获得对信息基因分类错误率较为可靠的估计，在训练集和测试集上分别做分类错误率估计：

1)在训练集上，采用“留一交叉检验”(leave-one-out cross validation，LOOCV)方法进行样本分类，每次保留1个样本作为测试样本，其余39个样本作为训练样本，重复该过程，直到所有40个样本都被用作测试样本为止。40次检验的累计错误分类数和总训练数40之比，即为留一法的确认误差。

2)对于测试集，用SVM对训练集中的40个样本进行训练得到分类超平面，然后对测试集中的22个样本进行分类判别，被错误分类的样本数和总测试样本数之比即为独立测试集的误差率。

将FCBF算法产生的12个信息基因分别在20组训练集和测试集上做留一交叉检验和独立测试集检验，得到平均分类准确率。为了比较实验效果，选用经典的信噪比方法[10]和SVM-RFE方法[5]在20组随机产生的训练集和测试集上进行对比实验，结果如表4所示。不难看出，SVM-RFE方法在训练集留一交叉检验中准确率较高，但是在独立测试集检验中，准确率却很低。此外，本文提出的方法较信噪比方法和SVM-RFE方法无论在留一交叉检验或者独立测试集检验上均显示出较高的准确率，并且需要的信息基因数较少。

表4 不同信息基因提取方法对结肠癌基因表达数据分类性能对比表

Tab.4 The contrast of classification performances of colon tumor dataset with different extracting methods of informative genes

信息基因提取方法	提取基因数	平均准确率/%	
		留一交叉检验	独立测试集检验
信噪比	10	85.50	88.41
	20	87.50	89.77
	30	85.25	87.27
SVM-RFE	10	84.63	57.73
	20	86.38	57.50
	30	86.13	57.50
本文方法	12	88.86	90.91

最后，将本文提出的方法与现有文献的方法[6-8]进行对比(表5)。结果表明，本文提出的方法无论是在提取的基因数量，还是检验准确率上都有明显的优势，因而具有较好的应用前景。

表5 不同信息基因提取方法对结肠癌基因表达数据分类的准确率对比表

Tab.5 The contrast of the accuracy rates of colon tumor dataset with different extracting methods of informative genes

信息基因提取方法	提取基因数	独立测试集检验准确率/%
SVM-RFE-SFS[6]	150	85.50
ULDA-SVM[7]	5	85.05
Boosting-SVM[8]	495	77.27
本文方法	12	90.91

5 结论

基于生物信息学理论，对结肠癌基因表达谱数据集进行分析，研究结肠癌识别与信息基因的选取，提出新的信息基因选取方法，该方法具有提取的信息基因数量少、分类性能好等优点，将结肠癌分类信息基因的

范围进一步缩小，从生物信息学的角度，为结肠癌诊断与研究提供了参考。

参考文献：

[1]RAMASWAMY S, GOLUB T R. DNA Microarrays in Clinical Oncology[J]. Journal of Clinical Oncology, 2002, 20(7): 1932-1941.

[2]KIM J, FISHER J W, et al. A nonparametric statistical method for image segmentation using information theory and curve evolution[J]. IEEE Transactions on Image Processing, 2005, 14(10): 1486-1502.

[3]ANDRADE P O, BITAR R A, et al. Study of normal colorectal tissue by FT-Raman spectroscopy[J]. Analytical and Bioanalytical Chemistry, 2007, 387(5): 2345-2352.

[4]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]. Proceedings of the National Academy of Sciences of the United States of America, 1999, 96(12): 6745-6750.

[5]GUYON I, WESTON J, BARNHILL S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2002, 46(1-3): 389-422.

[6]游伟，李树涛. 基 SVM-RFE-SFS 基因选择方法[J]. 中国生物医学工程学报，2010，29(1)：23-26.
YOU Wei, LI Shutao. Gene selection method based on SVM-RFE-SFS[J]. Chinese Journal of Biomedical Engineering, 2010, 29(1): 23-26.

[7]张小丹，吕建平. 基于支持向量机的特征提取方法研究[J]. 计算机工程与现代化，2008(8)：104-106.
ZHANG Xiaodan, LV Jianping. Research on uncorrelated linear discriminant analysis on support vector machine[J]. Computer and Modernization, 2008(8): 104-106.

[8]刘全金，李颖新. Boosting 算法在基因表达谱样本分类中的应用[J]. 计算机工程与应用，2008，44(14)：228-230.
LIU Quanjin, LI Yingxin. Application of boosting algorithm to sample categorization of gene expression profiles[J]. Computer Engineering and Applications, 2008, 44(14): 228-230.

[9]Cancer Program Data Sets[DB/OL]. [2011-11-20]. http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi.

[10]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.

[11]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.
LI Yingxin, LIU Quanjin, RUAN Xiaogang. Analysis of leukemia gene expression profiles and subtype informative genes identification[J]. Chinese Journal of Biomedical Engineering, 2005, 24(2): 240-244.

[12]李国正，王猛，曾华军. 支持向量机导论[M]. 北京：电子工业出版社，2004.

[13]YU L, LIU H. Feature selection for high-dimensional data: A fast correlation-based filter solution[C]// Proceedings of the 20th International Conference on Machine Learning. Washington D C, Aug. 21-24, 2003: 856-863.