基于基因表达谱的结肠癌特征基因选取

刘全金[1,2],李颖新[1],阮晓钢[1]
(1. 北京工业大学 电子信息与控制工程学院,北京 100022;
2. 安庆师范学院 物理系,安徽 安庆 246011)

摘要: 在分析肿瘤基因表达谱的基础上,运用模式识别方法选取结肠癌特征基因. 利用浮动顺序搜索算法在结肠癌基因表达谱数据中生成若干个候选特征基因子集,再以 *RBF* 支持向量机作分类器,以其在训练集和测试集中的错误分类率为依据,从候选特征基因子集中选取结肠癌特征基因集合. 实验结果表明了该方法的可行性和有效性.

关键词: 支持向量机;基因表达谱; 肿瘤基因
中图分类号: Q786　　**文献标识码**: A　　**文章编号**: 1007 - 855X(2006)01 - 0089 - 04

Informative Genes Selection for Colon Tumor Based on Gene Expression Profiles

LIU Quan-jin[1,2], LI Ying-xin[1], RUAN Xiao-gang[1]
(1. School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100022, China;
2. Department of Physics, Anqing Teacher's College, Anqing, Anhui 246011, China)

Abstract: Gene expression profiles of tumor and select colon tumor feature subsets are analyzed by means of pattern recognition. Firstly, the "Floating Sequential Search Algorithm" is used to generate candidate feature subsets from gene expression profiles. Secondly, a support vector machine is employed to classify the samples. The feature subsets with a minimum error are chosen as a set of colon informative gene. The result proves that the tissue samples can be classified with high correctness and the proposed approach is effective and feasible.

Key words: support vector machine; gene expression profiles; tumor genes

0 引 言

DNA芯片技术是一种高通量的基因表达分析平台,一次实验中能获得上万个基因的表达数据,已被应用于生物医学研究、疾病诊断和药物筛选等领域[1~4]. 利用 DNA芯片技术分析,比较肿瘤组织与正常组织之间的基因表达差异,从中发掘出在肿瘤组织中特异表达的基因和药物治疗的靶序列,找出影响样本类别的特征基因,准确识别肿瘤类型,对肿瘤的诊断和治疗有重要的意义.

Alon等人用层次聚类等方法对结肠癌样本数据进行了分析研究,选出含有 2 000个特征基因的数据集合[5]. 对于 Alon选出的结肠癌数据集, Zhang等人通过递归分割树归纳出 2个特征基因集合[6];李霞等人运用集成决策方法,得到 3个特征基因集合[7]. 这些特征基因集合对结肠癌的临床诊断和医学研究有一定的参考价值,但这两种特征基因选取方法较复杂,选取出的特征基因集合在该结肠癌数据集中的分类准确率也不是很高.

本文将结肠癌基因数据集分成训练集和测试集,首先在训练集中用浮动顺序搜索算法[8~10]生成候选特征基因子集;再以支持向量机 (SVM)为分类器,以其在训练集和测试集中的错误分类率为依据,选取错分率最低的特征基因子集为结肠癌特征基因集合. 实验证明,选取的特征基因子集具有较高的分类能力.

收稿日期: 2005 - 05 - 15. **基金项目**:国家自然科学基金重点资助项目 (项目编号: 60234020)
第一作者简介:刘全金 (1971. 12～),男,硕士,讲师. 主要研究方向:信息处理、生物信息学.
E - mail: liuquanjing2002@yahoo. com. cn

1 实验数据描述

实验数据来自 Alon的实验结果[7],这套数据可从网站 http://microarray.Princefon.edu/oncology/affydata/index.html 获得,它包括 40个肿瘤组织样本和 22个正常组织样本,每个样本均有2 000基因表达谱数据.实验中,先对样本数据做归一化,然后将正常(Normal)和肿瘤(Tumor)样本按接近 2∶1 的比例随机地分配到训练集和测试集中.如图 1所示,训练集 40个样本中有肿瘤样本 26个、正常样本 14个,测试集 22个样本中有肿瘤样本 14个、正常样本 8个.

训练集		测试集
Tumor 26	+	Tumor 14
Normal 14		Normal 8

图 1 基因表达谱数据样本集

Fig 1 Sample sets of Gene expression profiles

2 浮动顺序搜索算法及候选特征基因子集的生成

模式识别中,特征选择是从一组数量为 D的特征中选择出数量为 $d(D>d)$的一组最优特征[10].从生物信息学角度看,基因间的调控和相互作用表现为"功能基因组合"的形式,基因的功能与作用是某些基因集体作用的结果,而非一些单个基因的作用结果.分类特征对样本的分类能力是以特征集合的形式以一个整体体现出来的.

训练集样本数据含有 2 000个基因,可以形成 $2^{2\,000}$个基因组合,每个基因组合就是一个特征基因子集.考虑到最优搜索算法的复杂度,实验中采用次优搜索算法——浮动顺序搜索算法(Floating Sequential Search Algorithm, FSSA)[8~10],又称增 l减 r算法,该算法避免了顺序前进法和后退法中特征被选入(或剔除)就不能再剔除(或选入)的缺点,在选择过程中加入了局部回溯过程.实验中以特征基因子集 F_k的 Battacharyya距离作为浮动顺序搜索算法的评价函数 $J(F_k)$,评估特征基因子集对样本分类的贡献,考察各特征基因子集整体的分类能力.维数为 k的特征基因子集是所有可能的"k维基因组合"中对分类贡献最大的基因集合.

$$J(F_k)=\frac{1}{8}(\vec{\mu}_2-\vec{\mu}_1)^T\left(\frac{\sum_1+\sum_2}{2}\right)^{-1}(\vec{\mu}_2-\vec{\mu}_1)+\frac{1}{2}\ln\frac{\left|\frac{\sum_1+\sum_2}{2}\right|}{\sqrt{|\sum_1||\sum_2|}} \tag{1}$$

式中 $J(F_k)$表示含有 k个基因的特征子集 F_k的 Battacharyya距离,$\vec{\mu}_2$和$\vec{\mu}_1$为特征子集中基因在正常和肿瘤样本中分布的均值向量,$\sum_1$和 $\sum_2$为对应的协方差矩阵.

令 $F_{k-\max}$为含有 k个基因的候选特征基因子集,它是所有种可能的由 k个基因构成的基因集合中评价函数值最大的那组基因集合.浮动顺序搜索算法在特征子集空间进行搜索,找出不同维数的候选特征子集 $F_{k-\max}$.

$FFSA(n,\{F_{k-max}\mid card(F)_k=k,k=1,2,\cdots,n\})$算法:

step1: $F_{2-\max}=\{g_1,g_2\}$, g_1, g_2为训练集中基因集合 G中 20 000个基因中 Battacharyya距离最大的两个基因;

step2: if $i=n$ then exit;

else (1) $G=G-F_{k-\max}$;

(2) seek $g\mid g\in G, F_{(k+1)-\max}=\{F_{k-\max},g\}$,使 $J(F_{(k+1)_\max})$最大;

step3: $F'_{k\_\max}=\underset{F\in\{card(F_k)=k,F_k\subset F_{(k+1)\_\max}\}}{\arg\max}J(F)$

step4: if $J(F'_{k\_\max})\leq J(F_{k\_\max})$ then $k=k+1$: goto tep2;

setp5: $F_{k\_\max}=F'_{k\_\max}$

if $k=2$ then goto step2;

if $k=k-1$ then goto step3.

训练集中有正常样本 14个、肿瘤样本 26个,为保证在计算式(1)的 $\sum_1$时不出现奇异,候选特征基因子集的最大维数不能超过 14.在训练集运行浮动顺序搜索算法得到 13个具有维数不同的候选特征基因子集 $F_{k-\max}$ $(k=2,3,\cdots,14)$.

3 支持向量机及特征基因集合的确定

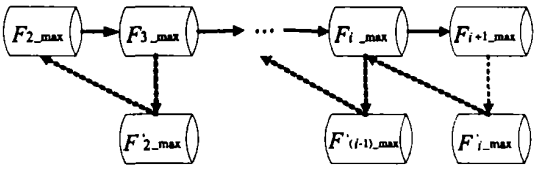

图2　浮动顺序搜索算法示意图

Fig.2　Flow chart of floating sequential

支持向量机(Support Vector Machine, SVM)是一种小样本学习理论,较适合于处理基因表达谱这种样本数少、维数高的数据集的分类和特征选取问题[12].本文所用基因芯片数据与其他基因芯片数据相比,在分类和特征选取方面属于较为困难的一个[12].实验中以 *RBF* 支持向量机为分类器识别样本,检验候选特征基因子集识别样本分类的能力.

支持向量机由 Vapnik等人基于统计学习理论,采用结构风险最小化原理提出的一种机器学习算法[11].通过调整判别函数使得它最好地利用边界样本点的分类信息,构造出最佳分类超平面,该算法具有较强的泛化能力.

若给定的样本集为:$S_T = \{(x_i, y_i) \mid x_i \in R^d, y_i \in \{-1, +1\}, i = 1, \cdots, N\}$,则支持向量机的判别函数为:

$$g(x) = \operatorname{sign}\left(\sum_{i=1}^{sv} a_i y_i k(x, x_i) + b\right) \tag{2}$$

式中 sv 为支持向量的个数,$k(x, x_i)$ 为核函数,支持向量机核函数选用 *RBF* 函数:

$$k(x, x_i) = e^{\frac{-\|x - y\|^2}{2\sigma^2}} \tag{3}$$

由于实验数据的样本数量少,为了获得对候选特征基因子集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"留一法"(Leave - One - Out Cross Validation, LOOCV)进行样本类型识别:每次保留 1个样本为测试样本,其余 39个样本用作 SVM的训练样本.重复该过程,直到所有 40个样本都被用作测试样本为止.累计被错误分类的样本数为"留一法"分类错误数.

(2) 对于测试集,用训练集上的所有 40个样本训练 SVM,识别测试集中 22个样本类型,被错误分类的样本数为"独立测试实验"(Independent Test, IT)的分类错误数.

通过试验,选取 SVM模型的 $\sigma = 10$,上界控制因子 $C = 400$,以 *RBF* 支持向量机为分类器,让候选特征子集 $F_{i\text{-}max}$ ($i = 2, 3, \cdots, 14$)分别在训练集和测试集上做"留一法"和"独立测试"分类实验,图 3显示了 13个候选特征基因子集的样本分类能力,其中 $F_{3\text{-}max}$ 的分类能力最强:留一法错分数为 0,独立测试错分数为 3.表明这组由 3个基因组成的候选特征基因子集含有较多的样本分类信息.因此,选取 $F_{3\text{-}max}$ 为特征基因集合 F_{set}.

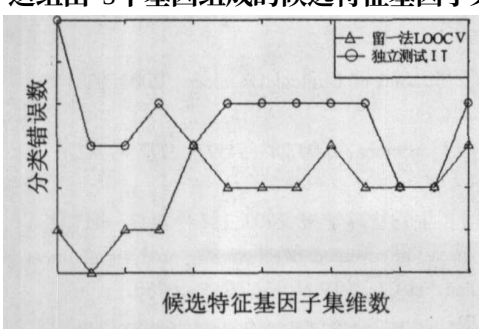

图3　候选特征基因子集的分类能力

Fig.3　Classifying capability of feature subsets

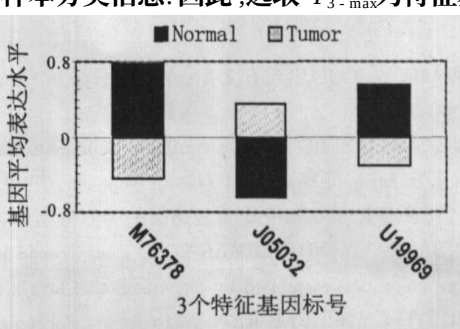

图4　Fset中3个基因在肿瘤和正常样本中的平均表达水平

Fig.4　Average expression level of 3 feature genes in tumor and normal samples

表 1列出了特征基因集合 F_{set} 的 3个基因标号及其描述.它们在肿瘤组织和正常组织中平均表达水平(归一化后的数值)刚好相反,如图 4所示,M76378和 U19969这 2个基因在肿瘤组织样本中呈下调表达,而在正常组织样本中相对上调表达;J05032

表 1　特征基因描述

Tab. 1　Description of feature genes

基因号	基因描述
M76378	Human cysteine - rich protein (CRP) gene, exons 5 and 6.
J05032	Human aspartyl - tRNA synthetase alpha - 2 subunit mRNA, complete cds
U19969	Human two - handed zinc finger protein ZEB mRNA, partial cds

在肿瘤组织样本中为上调表达,在正常组织样本中又相对下调表达.

4 特征基因集合比较及分析

将浮动顺序搜索算法得到的特征基因集合 F_{set}与 Zhang等人通过递归分割树归纳出两个特征基因集合[6]及李霞等人运用集成决策方法得出的 3个特征基因集合[7]比较.选用层次聚类和 K均值方法在 62个样本中对各特征基因集合样本作 100次聚类,记录平均错聚率;用 Fisher分类器、K近邻和线性 SVM在62样本中对各特征基因集合样本作 20次 5倍交叉校验,得到平均错分率.比较表 2的实验结果,浮动顺序搜索算法得到的特征基因集合 F_{set}所含的样本类别信息要多于 Zhang和李霞的 5个分类特征基因集合所含的样本类别信息.

表 2 特征基因子集的样本聚类和分类结果比较

Tab. 2 Comparison of samples clustering and classifying in feature subsets

	特征基因集合	无监督学习		有监督学习		
		层次聚类(错聚率/%)	K均值(错聚率/%)	Fisher分类(错分率/%)	K近邻(错分率/%)	线性 SVM(错分率/%)
1	Zhang 1	35.48	40.32	36.75	36.43	24.84
2	Zhang 2	32.25	45.63	24.85	31.06	37.74
3	李 Tree1	30.64	41.94	23.23	30.79	21.29
4	李 Tree2	27.41	38.58	27.1	30.40	32.26
5	李 Tree3	17.74	41.94	11.29	19.03	15.48
6	本文 Fset	14.28	17.74	12.23	18.32	10.64

分析表 2的聚类和分类结果,前 5个特征基因集合中"李 Tree3"最好;比较"李 Tree3"和本文特征基因集合 F_{set}聚类和分类结果,特征基因集合 F_{set}总体上好于特征基因集合"李 Tree3".所以,从分类能力看,浮动顺序搜索算法得到的特征基因集合要优于其他 5个特征基因集合.

在特征基因选取方法方面,与 Zhang的递归分割树归纳方法和李霞等提出的集成决策方法相比,浮动顺序搜索算法简单易行,只是浮动顺序搜索算法所得特征基因子集的维数受制于参与搜索的样本数量.如果实验中特征基因子集需要更高的维数,可以采用增加伪样本办法解决,这是我们下一步的工作重点.

综上所述,结合支持向量机用浮动顺序搜索算法能较好地完成肿瘤特征基因的选取,对肿瘤的临床诊断和生物医学研究起到有益的参考作用.

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] LANDER E S, WEINBERG R A. GENOMICS: Journey to the Center of Biology [J]. Scince, 2000, 287(5459): 1777～1782.

[3] LANDER E S. Array of hope [J]. Nature Genetics, 1999, 21(supp1 1): 3～4.

[4] 李泽,包雷,黄英武,等.基于基因表达谱的肿瘤分型和特征基因的选取[J].生物物理学报, 2002, 18(4): 413～417.

[5] ALON U, BARKA I 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.

[6] ZHANG H, YU C Y, SINGER B, et al. Recursive Partioning for tumor classification with gene expression microarray data [J]. Proc Natl Acad Sci Usa, 2001, 98: 6730 - 67 - 35.

[7] 李霞,饶绍奇,张田文,等.应用 DNA芯片数据挖掘复杂疾病相关基因的集成决策方法[J].中国科学 C辑:生命科学, 2004, 34(2): 195～202.

[8] THEODORIDIS S, KOUTROUMBAS K. Pattern Recognition. second Edition[M]. Elsevier Science, 2003, 177～179.

[9] PADIL P, NOVOVICOVA J, KITTLER J. Floating Search Method in feature selection [J]. Pattern Recognition Letters, 1994, 15: 1119～1125.

[10] 边肇祺,张学工,等.模式识别(第 2版)[M].北京:清华大学出版社, 1998.

[11] VAPNIK V N. Statistical Leaning Theory [M]. New York: Wiley Interscience, 1998.

[12] GUYON I, WESTON J, BARNHILL S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2000, 46(13): 389～242.