基于SVM的灵敏度分析方法选取肿瘤特征基因

刘全金[1,2], 李颖新[2,3], 阮晓钢[2]
(1. 安庆师范学院 物理系, 安徽 安庆 246011; 2. 北京工业大学 电子信息与控制工程学院, 北京 100022;
3. 北京经纬纺机新技术有限公司 CCD部, 北京 100176)

摘 要: 提出基于支持向量机的灵敏度分析方法选取结肠癌特征基因. 用支持向量机分析基因对分类决策函数的灵敏度,递归去除灵敏度较低的若干基因,得到一组候选特征基因子集;以支持向量机为分类工具,检验候选特征基因子集对样本分类的贡献,选取具有最佳分类能力的候选特征基因子集作为结肠癌特征基因子集. 通过实验比较,该特征基因子集的分类能力优于文献给出的其他特征基因子集,表明了该方法的可行性和有效性.

关键词: 特征选取; 支持向量机; 基因表达谱; 灵敏度
中图分类号: TP 181; Q 16 **文献标识码**: A **文章编号**: 0254－0037(2007)09－0954－05

0 引言

DNA芯片可一次从实验中获得上万个基因的表达数据,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[1-4]. 利用计算机技术分析比较肿瘤组织与正常组织之间的基因差异,从中挖掘出在肿瘤组织中特异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. Golub等以"信噪比"作为衡量基因对样本分类贡献大小的量度,以加权投票法作为分类手段,就急性白血病的AML与ALL的识别,从7 129个基因中选出了50个分类特征基因[5]. Khan等结合主元分析法和人工神经网络,提出灵敏度方法,从2 308个基因中提取出96个分类特征基因用于儿童小圆蓝细胞瘤的亚型识别[6].

就结肠癌的基因表达谱,Alon等人用层次聚类法进行了分析,选出含有2 000个特征基因的数据集合[7]. Zhang等人在Alon的实验基础上,通过递归分割树归纳出2个基因子集[8];李霞等人对Alon的实验结果进行了分析,运用集成决策方法,得到3个特征基因子集[9];Guyon等人利用线性支持向量机也对Alon的实验结果做了分析,文献[8]给出了以所有样本为训练集得到的7个特征基因.

本文以支持向量机(support vector machine,记为SVM)为模型,计算结肠癌基因表达谱数据中基因对分类决策函数的灵敏度,去除灵敏度低的若干基因,得到候选特征基因子集,通过新的SVM模型,重新计算这个子集中各基因的灵敏度,去除灵敏度低的若干基因,作为新的候选特征基因子集. 如此进行下去,最终得到一组候选特征基因子集. 然后以支持向量机为分类器,检验这组候选特征基因子集中每个子集对样本的分类性能,选取最佳分类能力的候选特征基因子集作为结肠癌特征基因子集.

1 实验数据描述

肿瘤基因表达谱是指利用DNA芯片测定的数以千计的基因在肿瘤或正常组织样本中的表达水平值. 本文的实验数据来自Alon公布的结肠癌基因表达谱数据集[7],该数据集,有40个结肠癌组织样本和22个正常组织样本,每个样本包含2 000个基因的表达数据[10]. 先对样本数据进行归一化,然后将正常(Normal)样本和肿瘤(Tumor)样本按接近2:1的比例随机分配到训练集和测试集中. 如图1所示,训练集有40个样本,测试集有22个样本.

收稿日期: 2006-06-02.
基金项目: 国家自然科学基金重点资助项目(60234020); 安徽省教育厅科研项目(KJ2007B001).
作者简介: 刘全金(1971－), 男, 安徽寿县人, 讲师.

候选特征基因子集将根据训练集样本的基因表达数据选出. 就每个候选特征基因子集，利用基因在训练集样本中的表达谱数据构建“预测性”SVM 分类模型，并对测试集样本的类型进行识别，统计分类错误数，检验各个候选特征基因子集的分类性能.

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

图 1　基因表达谱实验数据集

Fig.1　Dataset of gene expression profile

2　支持向量机

支持向量机适合于处理基因表达谱样本少、维数高的数据集的分类和特征选取问题[8].

支持向量机是 Vapnik 等人基于统计学习理论，根据结构风险最小化原理提出的机器学习算法[11]. 通过调整判别函数使得它能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.

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

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

式中 i 为支持向量的个数，$k(x,x_i)$为核函数. 通过试验，径向基函数

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

鉴于实验数据的样本少，为了获得对分类错误率较为可靠的估计，在训练集和测试集上分别估计分类错误率

1) 在训练集上，采用“留一法”(leave-one-out cross validation，记为 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，记为 IT)的分类错误数.

3　基于支持向量机的基因灵敏度分析

基因间的调控和相互作用表现为“功能基因组合”形式，基因的功能与作用是基因集体作用的结果，而非单个基因单独作用的结果. 表现在分类特征对样本的分类能力方面就是以特征集合的形式以整体体现出来的. 为了考察特征基因子集的基因作为一个整体的分类能力，保留对分类贡献大的基因，去除对分类不重要的基因.

基于 SVM 的基因灵敏度分析就是分析各基因影响 SVM 模型输出能力的大小[6,10]，即分析模型对各输入分量(基因)的敏感程度.

若给定样本集 $S_T=\{(x_i,y_i)\mid x_i\in R^d, y_i\in\{-1,+1\}, i=1,\cdots N\}$，则支持向量机的分类决策函数为

$$O(x)=\sum_{i=1}^{N} a_i y_i k(x,x_i)+b \tag{3}$$

对于各输入分量 $x=[x_1\quad x_2\quad\cdots\quad x_{2000}]^{\mathrm{T}}$ 中各分量对决策函数的影响，定义输入 x 中的第 j 个分量 x_j 对分类决策函数 $O(x)$的灵敏度函数为[12]

$$S(x_j)=\sum_{x\in S_T}\left|\frac{\partial O(x)}{\partial x_j}\right|=\frac{1}{\sigma^2}\sum_{x\in S_T}\left|\sum_{i=1}^{sv} a_i y_i k(x,x_i)(x_{ij}-x_j)\right| \tag{4}$$

其中，$S(x_j)$为分量 x_j 的灵敏度函数；S_T 为训练集.

输入分量 x 的第 j 个分量 x_j 对应于第 j 个分类特征基因. 基于训练集计算决策函数对每个基因的灵敏度. 分类特征基因的灵敏度可以看成该基因影响决策重要性的指标，据此，依次去除对决策影响最小的

若干个基因,将剩余的基因视为候选特征基因子集,考察它对样本的分类能力,从中找出具有最佳分类能力且所含基因最少的候选特征基因子集作为分类特征基因集合. 每去除一次基因,都要重新训练SVM模型,获取新的决策函数,并计算剩余基因对决策函数的灵敏度,分为4个步骤.

1) 在训练集中用候选特征基因子集 F 训练SVM模型,并记录 F 在"留一法"交叉校验和"独立测试实验"时分类错误数;

2) 由式(4),计算 F 中决策函数对各基因的灵敏度;

3) 找出 F 中灵敏度最小的若干个基因 f,,再从 F 中去除这些基因,得到新的候选特征基因子集: $F=F-\{f\}$;

4) 若 $F\neq\Phi$,则返回步骤(1)继续执行,否则退出.

通过试验,选取SVM模型的 $\sigma=10$,上界控制因子 $C=400$. 按上述过程分析训练数据集中的2 000个基因对决策函数的灵敏度,每次剔除基因子集的⌈15%⌉基因,从而得到一组维数单调下降的候选特征基因子集 $F_{1\,700}$、$F_{1\,445}$、…、F_1.

候选特征基因子集对样本的分类能力如图2所示. "LOOCV"曲线表示"留一法"交叉校验分类错误数,"IT"曲线则表示"独立测试实验"时的错分数. 在候选特征基因子集 F 的维数(基因数)从392下降到11的实验中,"留一法"错分数为0,"独立测试实验"的错分数为5,错分率最低. 当特征基因子集基因数继续下降时,错分率又会上升. 这表明,在候选特征基因子集中,基因数为11的特征基因子集含有最多的样本分类信息,称之为特征基因子集 F_g. F_g 中11个特征基因分别是X70326、M76378、R80427、R15447、H08393、T57619、T51558、M76378*、T41204、T71025和L07648. 其中,M76378在该生物芯片试验中被重复了2次,它俩在样本中的表达值相差不大,在选取的分类特征基因子集 F_g 中分别处于第2位和第8位.

如图3所示,特征基因子集 F_g 的11个特征基因在肿瘤和正常组织中的平均表达水平刚好相反. 有6个基因在肿瘤组织样本中呈上调表达,而在正常组织样本中下调表达;其余5个基因在肿瘤组织样本中下调表达,在正常组织样本中上调表达.

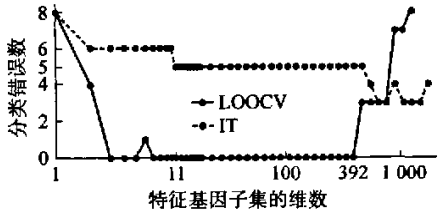

图2 各维数特征基因子集的样本分类能力

Fig.2 Classification performance when different subsets are applied

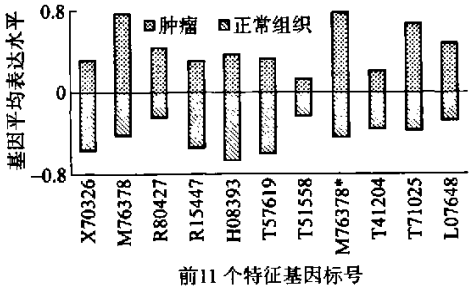

图3 F_g 中11个基因在肿瘤和正常组织中的表达水平均值

Fig.3 F_g's genes expression mean in Nomal and Tumor samples

4 特征基因选取的结果比较

以结肠癌基因表达谱为例,研究了基于SVM的基因灵敏度分析在肿瘤特征基因选取中的应用,选取出11个肿瘤特征基因. 文献[8]给出了Zhang等人通过递归分割树归纳出的2个结肠癌特征基因子集,文献[9]中李霞给出了他们运用集成决策方法得出的3个结肠癌特征基因子集. 为了综合分析各特征基因子集所含样本类别信息,本文分别做了聚类和分类实验. 利用各子集的特征基因数据对62个样本进行100次聚类,计算平均错聚率;利用各子集的特征基因数据对62个样本进行20次5重交叉校验:将62个

样本次序打乱后随机地分成 5 份，用其中的 4 份构建分类器，再用构建的分类器测试第 5 份样本的类别，记录分类错误数，如此依次做 5 次，使每份样本都被测试；该过程重复 20 次，统计平均错分率．从表 1 可看出，基于 SVM 的基因灵敏度分析方法得到的特征基因子集 F_g 的类型识别性能优于前 4 个特征基因子集．与"李 Tree3"特征基因子集相比，特征基因子集 F_g 除在 Fisher 分类器上的分类能力稍差外，在其他几个分类器上的分类能力均较强，聚类性能也好于"李 Tree3"的特征基因子集．这说明特征基因子集 F_g 含有较多的决定样本类别的信息．

表 1　特征基因子集的样本聚类和分类结果比较

Table 1　Comparison on Experimental Results

特征基因子集	错聚率/%		错类率/%		
	层次聚类	K 均值聚类	Fisher 分类	K 近邻	支持向量机
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　特征基因子集 F_g	8.06	8.06	16.06	14.11	9.35

从机器学习的角度看，将所有样本作为训练样本，存在过学习的问题．但从生物医学的角度出发，训练样本数的增加，有助于提高选取含有更多决定样本类别的信息的特征基因可能性．Guyon 等人以所有 62 个样本为学习样本，利用线性支持向量机找出了 7 个基因：H64807、T62947、R88740、H81558、T94579、M59040 和 H08393[10]．用本文提出的方法，仍以这 62 个样本为学习样本，选取出的 7 个基因分别为：H08393、H20709、M82919、T51849、T57619、K02268 和 R88740．通过实验发现，后一组特征基因的聚类能力和分类的能力均好于 Guyon 选取的特征基因．由此看出，至少就结肠癌基因表达谱数据集而言，基于 RBF 支持向量机的基因灵敏度分析方法是肿瘤特征基因选取的有效方法．

综上所述，基于 RBF 支持向量机分析基因对输出决策函数灵敏度的方法能有效地完成肿瘤分类特征基因的选取．从方法的实现上看，该方法简单易行，是依据基因表达谱对肿瘤类别进行可靠诊断，简化芯片实验的有效途径，对生物医学研究有重要参考价值．

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: 1777-1782.

[3] LANDER E S. Array of hope[J]. Nature Genetics, 1999, 21(supp): 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. LI Ze, BAO Lei, HUANG Ying-wu, et al. Cancer subtype discovery and informative gene identification with gene expression profiles[J]. Acta Biophysica Sinica, 2002, 18(4): 413-417. (in Chinese)

[5] GOLUB R R, SLONIM D K, TAMAYO P, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 289: 531-537.

[6] KHAN J, WEI J S, RINGNER M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural net works[J]. Nature Medicine, 2001, 7(6): 637-679.

[7] 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.

[8] 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: 6730-6735.

[9] 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.
LI Xia, RAO Shao-qi, ZHANG Tian-wen, et al. A ensemble decision approach to hunting for disease genes using microarray expression profiling[J]. Science in China Ser C Life Sciences, 2004, 34(2): 195-202. (in Chinese)
[10] 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.
[11] VAPNIK V N. Statistical leaning theory[M]. New York: Wiley Interscience, 1998.
[12] 李颖新，阮晓刚. 基于支持向量机的肿瘤分类特征基因选取[J]. 计算机研究与发展，2005，42(10)：1796-1801.
LI Ying-xin, RUAN Xiao-gang. Feature selection for cancer classification based on support vector machine[J]. Journal of Computer Research and Development, 2005, 42(10): 1796-1801. (in Chinese)

Analysis of Gene Sensitivity for Tumor Informative Genes Selection Based on SVM

LIU Quan-jin[1,2], LI Ying-xin[2,3], RUAN Xiao-gang[2]
(1. Department of Physics, Anqing Teacher's College, Anqing Anhui 246011, China;
2. College of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100022, China;
3. CCD Item, Beijing Jingwei Txtile Machinery New Technology Co., Ltd, Beijing 100176, China)

Abstract: In this paper we proposed an approach for tumor informative genes selection by analysis of gene sensitivity based on SVM. We analyzed the gene expression profiles of colon and recursively eliminated the genes which have lower sensitivity to SVM , then a set of candidate nested feature subsets were generated. Support Vector Machines were employed to classify the samples using these candidate feature subsets, and the feature subset with a minimum error was chosen as a set of colon informative genes. The results show that this feature subset contains more tumor classification information than other feature subsets identified in the literatures. The method proposed in this paper is feasible and effective.

Key words: feature selection; support vector machine; gene expression profile; sensitivity