文章编号: 1000-2243(2008)S0-S140-03

# 基因表达谱数据的多分类问题研究

王 伟, 罗林开

(厦门大学自动化系, 福建 厦门 361005)

摘要: 针对基因表达谱微阵列的数据多分类问题, 给出一种在多病类情况下的基于信噪比和相关性的特征基因选择方法. 该方法一次性考虑基因区分所有病类的能力, 尽量避免基因的冗余性; 其次利用支持向量机, 构建了基因表达谱微阵列数据的多分类器; 最后通过实验表明了本方法的有效性.

关键词: 基因选择; 相关性; 信噪比; 支持向量机

中图分类号: TP311.13 文献标识码: A

## The research of multi-class problem based gene expression profile data

WANG Wei, LUO Lin-kai

(Department of Automation, Xiamen University, Xiamen, Fujian 361005, China)

**Abstract** Aiming at the multi-class problem of gene expression profile data, this paper proposes a gene selection method based on S2N and correlation for multiple diseases. This method takes the classification abilities of genes to separate all the diseases into consideration at a time and tries to avoid redundancy in selected genes. Secondly, we construct multi-classifier of gene expression profile data using SVM. Finally, we do experiment by this method, the result of which shows great effectiveness of the method.

**Keywords** gene selection; correlation; S2N; support vector machine

## 1 引言

DNA微阵列技术为生物医学带来了一场变革. 微阵列的应用, 使得生物学家可以大规模并行提取DNA或RNA信息, 从而能够在基因组水平上以系统的、全局的观念去研究生命现象. 通过对不同部位、不同阶段的基因表达对比, 正常和疾病状态下的基因表达对比, 可从分子水平上进行疾病的诊断、分类及治疗等. 通过微阵列实验, 得到的是包含成千上万个基因的表达数据-基因表达谱. 基因表达谱数据通常具有数据量大、维数高、样本小、非线性的特点[1], 这对传统的数据分析方法提出了挑战, 原因主要是: ①DNA微阵列数据的样本数相对于基因数极少, 造成了严重的维数灾难现象[2], 导致分类性能严重下降; ④样本数极少, 使得无法用传统的与估计概率密度有关的方法来做分类识别; ㈣高维使数据存在很多与分类无关的噪声.

为了解决上述问题, 进行基因选择就显得尤为重要. 基因选择, 即从输入特征集(原基因集合)中选择出与目标最相关的基因子集. 基因选择是DNA微阵列研究的一个非常重要的内容, 主要出于以下两个目的: ①数据集中许多被测基因的表达值与样本的区分没有很大关系, 在分类问题中引入这些不必要的基因, 将增加分类问题中样本的维数, 导致计算复杂度的增加, 同时这些基因的引入, 可能会产生一些不必要的噪声数据; ④如果存在能将两类区分开的较小基因子集, 将有利于生物医学工作者专门研究这些基因的功能, 了解其生物意义, 开发出基于这些基因的价格低廉的疾病诊断芯片. 因此, 在基因表达谱数据分析中, 进行特征提取找到足够少的能够进行有效分类的基因子集是非常必要和重要的.

目前, 人们在基因选择这方面已开展了大量的工作. 文献[3]采用Fisher线性判别函数与启发式逐步

收稿日期: 2008-06-13

作者简介: 王 伟(1984- ), 男, 硕士研究生; 通讯联系人: 罗林开, 副教授, 博士.

基金项目: 国家自然科学基金资助项目(60704042)

向前搜索结合的方法, 在两类中进行基因选择. 文献[4]将无监督的属性均值聚类网络中加入学习样本的类别信息, 形成堆近邻分类法, 进行基因选择. 文献[5]是关于两类的基因选择, 给每类设计一个能代表该类的理想基因, 使得这两个基因有最大的负相关性, 再用统计相关分析对每个基因与该理想基因的相关性进行计算, 选择基因. 然而, 这些方法基本上都是在两类情况下进行的基因选择, 在多病类情况下进行基因选择的方法还较少. 但现实世界中, 一般都是多病类的复杂情况.

为了保证多病类情况下基因诊断的可靠性, 本文给出了一种多病类情况下的基因选择方法——基于信噪比和相关性的基因选择方法. 该方法一次性考虑了基因区分所有病类的能力, 我们通过该方法对DNA微阵列 lymphoma数据进行基因选择, 并利用 SVM (Support VectorMachine)对所选基因子集的分类推广能力进行验证, 结果表明该方法的有效性.

## 2 基因选择方法

### 2.1 信噪比方法

#### 2.1.1 两病类情况

针对两分类问题, 对于每一个基因, 根据其已知标签样本的数据计算以下统计量:

$$S(j) = \frac{\mu_+(j) - \mu_-(j)}{\sigma_+(j) + \sigma_-(j)} \tag{1}$$

式中: $\mu_+(j)$是指第 $j$个基因属于正类样本表达值的均值; $\mu_-(j)$是指第 $j$个基因属于负类样本表达值的均值. 同样, $\sigma_+(j)$指第 $j$个基因属于正类样本表达值的标准差; $\sigma_-(j)$指第 $j$个基因属于负类样本表达值的均值. 在 $S(j)$为正的情况下, $S(j)$越大, 说明该基因对于正类的影响越大; 在 $S(j)$为负的情况下, $S(j)$越小则说明该基因对于样本负类的影响越大.

#### 2.1.2 多病类情况

针对多分类问题, 考虑到本文多分类器构造方法是 SVM的 one-versus-rest方法, 因此对于每一个基因, 根据其已知标签样本的数据, 计算以下统计量:

$$S(j) = \sum_{i=1}^{n} \left| \frac{\mu_+^i(j) - \mu_-^i(j)}{\sigma_+^i(j) + \sigma_-^i(j)} \right| \tag{2}$$

式中: $n$是类别数; $\mu_+^i(j)$是指第 $j$个基因属于第 $i$类样本表达值的均值; $\mu_-^i(j)$是指第 $j$个基因不属于第 $i$类样本表达值的均值. 同样, $\sigma_+^i(j)$指第 $j$个基因属于第 $i$类样本表达值的标准差; $\sigma_-^i(j)$指第 $j$个基因不属于第 $i$类样本表达值的标准差. $S(j)$越大, 说明该基因对样本分类影响越大. 因此, 在这种情况下选取 $m$ 个绝对值最大的基因就达到了基因选择的目的.

### 2.2 基因间的相关性

相关的两个基因往往共同被表达, 比如基因$A$对分类贡献大, 与它相关性大的基因$B$对分类的贡献也很可能会比较大. 所以, 不考虑所选基因相关性的基因选择方法可能将这两个基因都选上. 但是, 这两个基因共同提供的对分类有用的信息和其中任何一个基因单独提供的差不多, 也就是说选择的基因中产生了冗余. 为了选择更有效的基因, 对所选基因间的相关程度加以约束. 本文用 Pearson线性相关系数[6]来度量基因间的相关性. 例如, 基因 $r$与基因 $r'$之间的线性相关性为:

$$P_{rr'} = \frac{\sum_S (g_{rs} - \bar{g}_r)(g_{r's} - \bar{g}_{r'})}{\sqrt{\sum_S (g_{rs} - \bar{g}_r)^2 \cdot \sum_S (g_{r's} - \bar{g}_{r'})^2}} \tag{3}$$

其中: $g_{rs}$是样本 $s$的 $r$基因的表达水平; $\bar{g}_r$是所有样本的 $r$基因表达水平的均值.

## 3 实验与结果分析

### 3.1 实验数据与来源

所用的实验数据是 lymphoma数据[7, 8]. 该数据有 6种病类, 共 96个样本、共 4026种基因组成. 从中选出 DLCL、Blood、FL、CLL四类共 78个样本. 由于数据中存在空值 (缺失), 所以某种基因若有空值, 即将该种基因删除, 不参与实验, 处理后只剩 1012维. 选取 52个样本作为训练样本, 26个样本作为测试样本. 训练样本和测试样本分布如表 1所示.

表 1 lymphoma训练样本和测试样本分布情况

Tab. 1 Distribution of lymphoma train samples and test samples

| 类型 | DLCL | Blood | FL | CLL |
|---|---|---|---|---|
| trains | 28 | 11 | 6 | 7 |
| tests | 14 | 5 | 3 | 4 |

### 3.2 实验步骤

1) 将数据规范化;

2) 利用式(2)计算每个基因对所有病类分类的总贡献,并降序排列;

3) 利用式(3)计算出基因两两间的 Pearson线性相关系数;

4) 选出前 $n$个贡献最大的基因,并保证这 $n$个基因两两间的 Pearson线性相关系数小于阈值 $P$(判断第 $n$个基因是否选择的时候,分别计算它与前 $n-1$个基因的相关系数,如果均小于阈值 $P$,则选择该基因,否则不选择,判断下一个基因);

5) 对所选的基因子集用 SVM 构造 one-versus-one多分类器[9],选用 $K$ 折交叉验证方法($K=5$),求出准确率.

### 3.3 实验及结果分析

根据上面的实验步骤,对 lymphoma数据进行训练和测试. 核函数采用 RBF(采用网格算法寻找最优参数 $\gamma$ 和 $c$),实验结果如表 2所示.

从表 2看出,选择相同数目的特征时,Pearson线性相关系数的阈值 $P=0.8$的分类结果要优于对相关性不做要求的时候. 特别是在所选特征数在 30以外的时候,优势比较明显. 表明在基因选择的过程中,对所选基因间的线性相关性加以一定的约束是有意义的.

表 2 识别结果

Tab. 2 Identification results

| 特征数 | 准确率($P=1$) | 准确率($P=0.8$) |
|---|---|---|
| 5 | 0.8462 | 0.8462 |
| 10 | 0.9231 | 0.9231 |
| 20 | 0.9615 | 0.9615 |
| 30 | 0.9615 | 1 |
| 50 | 0.9615 | 1 |
| 100 | 0.9615 | 1 |

注: 当 $P=0.8$时,前 10个基因的相关系数均小于 0.8,所以所选基因与相关性不做要求时一样

## 4 结论

针对多病类的基因表达谱微阵列数据,利用基于信噪比和相关性的方法进行特征基因选择,用 SVM 的 one-versus-one方法构建多分类器. 该方法的优点是,加入了对所选基因线性相关性的约束机制,尽量避免选择冗余基因;一次性考虑了基因区分所有病类的能力,减少了计算时间. 用该方法对 DNA微阵列数据进行实验,在一定的线性相关性阈值下选出了性能良好的基因子集. 实验结果表明,该基因子集的分类结果好于没有相关性约束的情况,表明了该方法的有效性.

参考文献:

[1] 陈忠斌. 生物芯片技术[M]. 北京: 化学工业出版社, 2005: 1-10.

[2] Vladimir N Vapnik. The nature of statistical learning theory[M]. Berlin: Springer-Verlag, 1995: 98-114.

[3] Ioan Tabus, Jorma Rissanen, Jaakko Astola. Classification and feature gene selection using the normalized maximum likelihood model for discrete regression[J]. Signal Processing, 2003, 83: 713-722.

[4] Chu Feng, Wang Lipo. Gene expression data analysis using support vector machines[J]. Neural Networks, 2003, 3: 2268-2271.

[5] 刘申岭. 基于 SVM 的基因选择[D]. 西安: 西安电子科技大学, 2004.

[6] 李 云, 叶春晓, 李 季, 等. 基于特征关联性的特征选择算法研究[J]. 微型机与应用, 2004, 6: 58-60.

[7] http://llmpp.nih.gov/lymphoma

[8] Ash A Alizadeh. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling[J]. Nature, 2000, 403: 503-511.

[9] 李昆仑, 黄厚宽, 田盛丰. 模糊多类 SVM 模型[J]. 电子学报, 2004, 5: 830-832.

(责任编辑: 顾泉佩)