Boosting 算法在基因表达谱样本分类中的应用

刘全金[1],李颖新[2]
LIU Quan-jin[1],LI Ying-xin[2]

1. 安庆师范学院 物理与电气工程学院,安徽 安庆 246011
2. 北京经纬纺机新技术有限公司 CCD 部,北京 100176
1. School of Physics & Electronic Engineering,Anqing Teacher's College,Anqing,Anhui 246011,China
2. CCD Item,Beijing Jingwei Textile Machinery New Technology Co.,LTD,Beijing 100176,China
E-mail:liuquanjing2002@yahoo.com.cn

LIU Quan-jin,LI Ying-xin. Application of Boosting algorithm to sample categorization of gene expression profiles. Computer Engineering and Applications,2008,44(14):228-230.

Abstract:In this paper an approach is proposed for sample categorization of gene expression profiles based on structure of gene expression profiles. Firstly,genes are removed as"noise genes"with small Bhattacharyya distance. Secondly,multi-edit-nearest-neighbor algorithm is modified to eliminate"noise samples". Then boosting-based support vector machines combination classifiers are constructed and employed to classify the samples. Finally,this methods is used to classify colon genes expression profiles samples. The results show that the means is feasible and effective.

Key words:Bhattacharyya distance;multi-edit-nearest-neighbor algorithm;Boosting algorithm

摘 要:基于基因表达谱结构提出一种基因表达谱的样本分类方法。首先用基因的 Bhattacharyya 距离衡量其所含样本类别的信息,过滤 Bhattacharyya 距离较小的噪声基因;然后修改重复剪辑近邻算法,剔除噪声样本;再基于 Boosting 算法构建支持向量机组合分类器;最后以结肠癌基因表达谱样本为例,进行了分类实验。实验结果表明该方法简单、有效,对基因表达谱样本的分类问题有强的实用性。

关键词:Bhattacharyya 距离;重复剪辑近邻法;Boosting 算法
DOI:10.3778/j.issn.1002-8331.2008.14.065 **文章编号**:1002-8331(2008)14-0228-03 **文献标识码**:A **中图分类号**:TP181

1 引言

利用计算机技术分析基因表达谱(Gene Expression Profiles)数据已经成为信息研究领域中的重要课题。基因表达谱数据分析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[1-4]。利用计算机术分析肿瘤组织与正常组织之间的基因表达差异,准确识别肿瘤类型,对肿瘤的诊断和治疗有重要的意义。

基因芯片技术是近年来分子生物学和医学诊断技术的重要进展。基因表达谱是在 DNA 芯片测定的组织样本中基因的表达水平值,它记录了组织样本所有被测基因在一定实验条件下的表达水平,全面反映了细胞中的基因表达情况。目前国内外已有不少基因芯片数据库。很多大学和研究机构还公布了肿瘤样本的基因表达谱数据,如 Golub 在文献[5]用的急性白血病两种亚型 AML 与 ALL 基因表达谱,Khan 在文献[6]中用的儿童小圆蓝细胞瘤的 4 个亚型基因表达谱数据,还有 Alon 用层次聚类等方法得到的结肠癌 2 000 个基因的表达数据[7]。

这些基因表达谱数据集有共同的特点,就是数据集呈扁平状,基因数成千甚至上万,但样本数量相对特别有少。对于类似高维数据集,首先要进行特征选取,降低特征数目,以提高分类识别效率。从高维数据中选取特征需要运算量大,特征选取算法也比较复杂。单基因过滤算法依据每个基因对样本类型的影响程度选取特征,算法简单,但由于该方法将基因作为孤立属性看待,没有考虑基因间相互作用对生物体的影响[8],很难得到好的实验结果;而从基因组合的角度将基因组合看作一个整体,考察其对样本的影响,计算量又特别大,且过程也比较复杂[9,10]。

针对上述问题,本文提出先通过 Bhattacharyya 距离过滤噪声基因;再套用重复剪辑近邻算法进行噪声样本过滤,剔除体现样本类别信息最少的噪声样本;然后基于 Boosting 算法构建支持向量机组合分类器,并将该方法运用于结肠癌基因表达谱样本[6]实验。实验表明,该方法与文献相比,过程简单,分类结果较好。

基金项目:安徽省教育厅自然科研项目(No. KJ2007B001)。
作者简介:刘全金(1971-),男,副教授,主要研究方向:信息处理、机器学习、生物信息学;李颖新(1972-),男,博士,工程师,主要研究方向:模式识别、机器学习、生物信息学。
收稿日期:2007-08-24 **修回日期**:2007-10-24

2 噪声基因过滤

从生物学角度看,只有部分基因与样本某一特定的表型(生物类别)相关,其余基因是同该表型无关的"类别无关基因",或称为"噪声基因"。为有效选取样本的分类特征,在此,首先利用基因的 Bhattacharyya 距离[11]作为衡量基因含有样本分类信息多少的度量。以两类样本为例,Bhattacharyya 距离体现了属性在两个不同样本中分布的差异,这种差异既包含了属性在不同类别分布均值的差异,同时也考虑了样本分布方差不同对分类的贡献。其具体形式为:

$$B(g)=\frac{(\mu_+(g)-\mu_-(g))^2}{4(\sigma_+^2(g)+\sigma_-^2(g))}+\frac{1}{2}\ln\left(\frac{\sigma_+^2(g)+\sigma_-^2(g)}{2\sigma_+(g)\sigma_-(g)}\right) \quad (1)$$

式中 μ_+、μ_- 分别为基因 g 在两类不同样本中的表达水平的均值,σ_+、σ_- 为相应的标准差。基因的 Bhattacharyya 距离越大,该基因在两类样本中表达水平的分布差异也就越大,它对样本分类的能力也就越强。

然后,剔除 Bhattacharyya 距离小于某阈值的噪声基因,保留 Bhattacharyya 距离大于该该阈值的基因。

3 噪声样本过滤

基因表达谱数据是基因的表达水平值,因为实验环境改变或生物个体变异等情况,个别样本的基因表达数据就会带有噪声,如果仍与其他样本一起训练,就会影响分类实验结果。借鉴重复剪辑近邻法[12]的思想,设法将这些偏离样本类别特性的样本去除。

鉴于基因样本集数据结构的特点,本文修改重复剪辑近邻算法,对错分的样本采用"惩罚"式放回,即对于错判样本罚分,但不从训练集中去除该样本,如此进行若干次后,去除被罚分最多的几个"噪声"样本。具体算法如下:

算法 1

(1)将训练集样本随机划分为 S 个子集,即:$\chi^N=\{\chi_1\chi_2\chi_3\cdots\chi_s\}$,且 $s\geqslant 3$;

(2)用最近邻法,以 $\chi_{(i+1)Mod(s)}$ 为参考集,对 χ_i 中的样本进行分类;

(3)被错分的样本罚 1 分;

(4)回到第(1)步,重新随机划分子集;

(5)如此往复进行 N 次,统计所有错分样本被罚分数,将罚分最多的 M 个样本删除。

如果不同类别的样本个数相差较大,则第(1)步训练集样本的划分要尽量保证同类别样本个数在 S 个子集中的均等。这样可以避免子集内不同类别样本数失衡带来的偏差。

4 基于 Boosting 算法的组合分类器

Boosting 算法的目标是提高任何给定的学习算法的分类准确率[13,14]。基于 Boosting 算法有许多不同的变形,AdaBoost 算法就是其中具有代表性的一种。AdaBoost 算法的思想是将多个 Boosting 分类器级联起来以达到较高的分类准确度[15]。它允许设计都不断地加入新的"弱分类器",直到达到某个预定的足够小的误差率。每一个训练样本都被赋予一个权重,反映其被某个分类器选入训练集的概率:如果某训练集样本已经被准确分类,那么在构造下一个训练集时,该样本被选中的概率就被降低;相反,如果没有被准确分类,则其权重就得到提高。这样,AdaBoost 算法能够聚焦于那些困难的样本上,而这些样本所含的信息较其他样本更丰富。

以下以两类对像为例,说明 AdaBoost 算法。用 x^i 和 y_i 表示训练集的样本点和它们的类别,用 W_k 表示第 k 次迭代时训练集样本的权重分布,用 C_k 表示第 k 次迭代时的训练分类器。AdaBoost 算法表示如下:

算法 2

(1) begin initialize $D=\{x^1,y_1,\cdots,x^n,y_n\}$, $k_{\max}$, $W_1(i)=1/n$, $i=1,\cdots,n$

(2) $k=0$

(3) do $k=k+1$

(4) 选择 W_k 中权重大的样本 D 训练分类器 C_k

(5) 用分类器 C_k 测试 D 的训练误差 E_k

(6) $a_k=\frac{1}{2}\ln[(1-E_k)/E_k]$

(7) $W_{k+1}(i)=\frac{W_k(i)}{Z_k}\times\begin{cases}e^{-\alpha_k}\text{如果 } h_k(x^i)=y_i\\ e^{\alpha_k}\text{如果 } h_k(x^i)\neq y_i\end{cases}$

(8) until $k=k_{\max}$ or $E_k<$误差阈值

(9) return C_k 或 α_k, $k=1,\cdots,k_{\max}$

算法中的 Z_k 只是一个归一化的系数,$h_k(x^i)$ 是分量分类器 C_k 对样本点 x^i 的判断类别函数。最后对测试样本的总体分类判决可以使用各个分量分类器加权平均来得到:

$$g(x)=sign\left[\sum_{k=1}^{k_{\max}}\alpha_k h_k(x)\right] \quad (2)$$

基因表达谱数据的结构特性使得实验中分类器的选择受到限制。基因表达值为模拟量,而且样本数远远小于基因数,使得不少分类模型不能用作实验中的分类器。比如 fisher 分类器,如果样本数相对基因数过少,就不能保证类内总类内散度矩阵是对称半正定的,分类器也就无法正常工作。神经网络模型可以完成这个任务,但样本特征数量大会使训练计算量过大。支持向量机分类关键是求取支持向量,运算量小,运算速度较快,不过要注意相对其他分类器而言,以支持向量机为分类器迭代次数可能较少,分类器容易过拟合于训练样本,导致测试样本实验误差较大。

5 实验

5.1 实验数据

肿瘤基因表达谱是指利用 DNA 芯片所测定的肿瘤或正常组织样本中基因的表达水平值。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]。与其它基因芯片数据相比,该数据集是一个较难分析的数据集[16],它有 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。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 的比例随机地分配到训练集和测试集中:训练集有 40 个样本,正常样本 16 个,肿瘤样本 24 个;测试集有 22 个样本,正常样本 14 个,肿瘤样本 8 个。

5.2 噪声样本过滤

计算训练集样本基因的 Bhattacharyya 距离,用该距离度量基因所含的样本类别信息。将基因表达值代入式(1),并将按 Bhattacharyya 距离按从大到小排序。实验结果如图 1 所示。

如图 1 所示,以 Bhattacharyya 距离 0.2 为阈值,小于该值的基因在两类别中的分布,无论是均值还是方差均无明显差

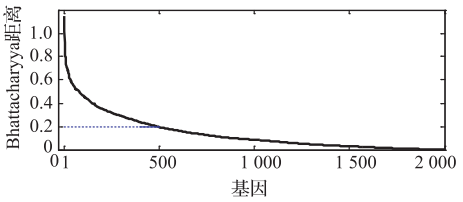

图1 基因 Bhattacharyya 距离曲线

异,可做作为噪声基因,将它们删除,保留大于阈值 0.2 的 495 基因作为后面实验的信息基因。

5.3 噪声样本过滤

去除噪声基因后,训练集每个样本有 495 基因。尝试了文献[12]中原始的重复剪辑近邻算法实验。多数情况下进行到第 6 次迭代时实验结束,但有近一半样本因错分被先后剪辑掉。剪辑比例过大,被剪辑的样本中仍可能含用丰富的样本类别信息,这样分类结果不会太好。这是因为,该方法只适于样本数足够多的样本集实验[12],基因表达谱数据集的样本相对特征数过少,而且正常样本数和肿瘤样本数也不均衡。

将训练集 40 个样本中的 495 基因数据代入算法 1,训练集分为 4 个子集,重复进行 1 000 次,即 $S=4$,$N=1\ 000$。通过实验,第 2 号肿瘤样本被罚 788 分,第 9 号肿瘤样本被罚 591 分,第 28 号正常样本被罚 612 分,其余样本被罚分数统计见图 2。罚分越高的样本被错分数的次数多,表明它们偏离本类样本中心也越远。对于罚分特别多的样本,如果仍作为训练样本将会影响分类结果,故应将它们剔除。从整个错分数的分布情况看,第 2 号肿瘤样本被错分次数远高于其他样本错分次数,另考虑训练集样本数不多,实验中将第 2 号肿瘤样本视为噪声样本去除。这样训练集就剩下 25 个肿瘤样本和 14 个正常样本。

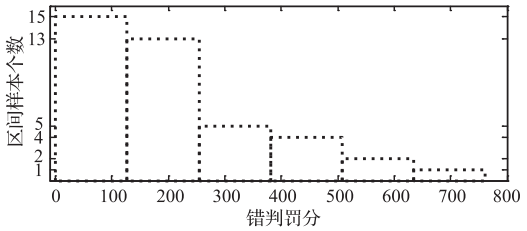

图2 样本错判的罚分统计直方图

5.4 基于 Boosting 算法的支持向量机分类实验

训练集每个样本有 495 个基因数据。上文算法 2 已经对 Boosting 算法进行了描述,实验中设置 $k_{\max}=11$,现在的任务是选择合适的分类器。

训练集基因数远大于样本数,显然 Boosting 算法分类器不能选用 fisher 分类模型。用单层神经网络模型构建 AdaBoost 算法中的组合分类模型,因为样本特征空间为 495 维,所以,神经网络计算量较大,运算速度较慢。实验中迭代 9 次后满足迭代终止条件,用这 9 个神经网络组合的分类器计算测试集样本输出,并代入式(2)识别样本类别,有 5 个样本被错识。

用核函数为线性的支持向量机做分类器训练 AdaBoost 算法中的组合分类模型。实验第 6 次迭代时,训练集错分数就为 0,迭代实验结束,将组合分类器代入式(2)测试测试集样本类别,分类错误数为 3。同一台机器,整个分类过程用时是神经网络组合分类器分类实验的 1/10,错分数也较少。所以,笔者认为对于基因表达谱这种扁平数据,可以选用线性支持向量机作为 AdaBoost 算法中的组合分类模型。

又将噪声样本(第 2 号肿瘤样本)放回到训练集,重新进行了基于 Boosting 算法的支持向量机分类实验,独立测试的错分数为 5,较放回前的错分数高 2 个(见表 1)。表明该样本的去除有益于改善后面独立测试实验的质量,提高了组合分类器的泛化能力。

表1 支持向量机组合分类器独立测试实验结果

训练集	错分数
原训练集	5
过滤噪声样本	3

综上所述,整个分类实验分三步进行,都简单易行。为了验证本文提出的分类方法的有效性,本文以支持向量机(核函数为径向基)为分类器,分别用文献[16-18]给出的特征子集,在同样的样本集中进行了独立测试实验。这些特征基因都是从 Alon 的实验[7]的 2 000 个基因中选取出的,Guyon 运用递归特征剪约算法选取基因子集[16],Zhang 通过递归分割树归纳出 2 个基因子集[17];李霞运用集成决策方法得到 3 个特征基因子集[18]。这些基因子集的特征基因个数均小于本文用 Bhattacharyya 距离过滤后得到的基因个数。表 2 给出了独立测试实验的结果,结果表明本文提出的分类方法分类结果要优于前 6 个分类结果。

表2 文献特征基因子集独立测试实验结果

	基因子集	独立测试实验错分数
1	Guyon	5
2	Zhang 1	6
3	Zhang 2	6
4	李 Tree1	8
5	李 Tree2	7
6	李 Tree3	6

从分类的特征维数看,本文得到的特征维数高,但从对样本的分类过程看本文的方法比较简单,不需要复杂的特征选取过程。另一方面,对于训练集新样本的加入,本文的分类方法更具灵活性,而且噪声样本的去除方法可以有效地过滤因实验环境变化等因素产生的变异样本,提高分类效率。

6 小结

本文针对肿瘤基因表达谱结构特性提出一种肿瘤基因表达谱样本的分类方法。首先用 Bhattacharyya 距离过滤噪声基因;然后修改重复剪辑近邻算法,剔除噪声样本;再基于 Boosting 算法构建支持向量机组合分类器。本文所用基因和样本过滤的方法简单,所选的组合分类器运算速度快,通过对结肠癌基因表达谱样本的分类实验,表明了该方法的有效性和可行性。

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.

(下转 238 页)