ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.7, July 2006, pp.1485−1493
DOI: 10.1360/jos171485

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

考虑样本不平衡的模型无关的基因选择方法*

李建中+, 杨 昆, 高 宏, 骆吉洲, 郭 政**

(哈尔滨工业大学 计算机科学与工程系,黑龙江 哈尔滨 150001)

Model-Free Gene Selection Method by Considering Unbalanced Samples

LI Jian-Zhong+, YANG Kun, GAO Hong, LUO Ji-Zhou, GUO Zheng**

(Department of Computer Science and Engineering, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: Phn: +86-451-86415827, Fax: +86-451-86415827, E-mail: lijzh@hit.edu.cn

Li JZ, Yang K, Gao H, Luo JZ, Guo Z. Model-Free gene selection method by considering unbalanced samples. *Journal of Software*, 2006,17(7):1485−1493. http://www.jos.org.cn/1000-9825/17/1485.htm

Abstract: In gene expression data analysis, discriminator genes are importantly informative genes for further research. Recently, a great deal of research has focused on the challenging task of identifying these informative genes from microarray data. However, the sizes of sample classes in microarray data are often unbalanced. The unbalance of samples has not been explicitly and correctly considered by the existing gene selection methods, especially nonparametric methods. Considering the unbalance of samples and the stability of the approach for identifying informative genes, a novel and model-free gene selection method is proposed in this paper. With considering within-class difference and between-class variation, as well as the homogeneities of the within-class difference and between-class variations, scoring functions of genes are constructed to select discriminator genes. This method is not only applicable in two-category case but also applicable in multi-category case. The experimental results on two publicly available microarray datasets, leukemia data and small round blue cell tumor data, show that the proposed method is very efficient and robust to select discriminator genes.
Key words: gene selection; gene expression; classification; microarray

摘 要: 在基因表达数据分析中,鉴别基因是后续研究中非常重要的信息基因.有很多研究致力于从基因表达数据中选出信息基因这一挑战性工作,并提出了一些基因选择方法.然而,这些方法(特别是非参数选择方法)都没有考虑不同样本类别中样本大小的不平衡性问题.考虑样本不平衡性和基因选择方法的稳定性,给出一个全新的与数据分布模型无关的基因选择方法.在类内变化小和类间差别大的策略下,选择敏感的度量函数提高方法的鉴别能力,同时,利用类内变化和类间差别的一致性来增加方法的稳定性和适用性.这一方法不但可以应用于两个类别的情况,也可以应用于多个类别的情况.最后,使用两组真实的基因表达数据对所提出的方法进行了验证.实验结果表明,这一方法比其他方法具有更高的有效性和稳健性.
关键词: 基因选择;基因表达;分类;微阵列
中图法分类号: TP391 文献标识码: A

* Supported by the National High-Tech Research and Development Plan of China under Grant Nos.863-317-01-04-99, 2004AA231071 (国家高技术研究发展计划(863))

** 目前工作单位是哈尔滨医科大学

Received 2005-04-19; Accepted 2005-12-13

基因芯片又称为 DNA 微阵列,是生物芯片中发展最成熟并最先实现商品化的产品.基因芯片是功能基因组、肿瘤、药物基因学研究中的重要监测手段.基因表达水平是衡量基因功能发挥作用的重要指标,通过基因表达水平的高低,可以揭示生物体的状态和基因在生物体内的活性.例如,利用基因芯片可以检测疾病样本的基因表达,并进行疾病的识别和分类.对基因表达量进行检测的基因芯片称为基因表达芯片.用基因表达芯片对样本进行检测所产生的数据称为这个样本的表达谱,多个表达谱组成一个基因表达数据.基因表达数据中的样本通常取自不同的类别(类别事先已知),当样本来自两个类别时称为两类问题;当样本来自大于等于 3 个类别时称为多类问题.

由于有些基因仅仅在特定的时间或特定的实验条件下被调控和表达,因此,一个基因在来自两个不同类别的样本中(例如正常样本和疾病样本),其表达值可能会不同.如果某基因在不同类别的样本中具有不同的表达值,那么该基因就具有了鉴别能力,可以作为特征对样本进行识别,进而可以用于疾病的分型和分类.有鉴别能力的基因称为鉴别基因.鉴别基因在生物学和医学中有着非常重要的应用.通过研究鉴别基因可以探索疾病的发生机制.在临床研究中,鉴别基因可以作为临床治疗的标记.对于基因芯片实验来说,大部分基因与具体的生理状态、实验条件无关,是没有鉴别能力的非鉴别基因.给定基因表达数据,人们希望从众多的基因中发现那些鉴别基因.从众多基因中选择鉴别基因的问题称为基因选择问题.由于样本的来源和实验经费的限制,使得基因表达数据具有一个独特的特点:小样本和超高维,即样本数目少(通常是数十个),而基因数目多(通常是数千到几万个)[1].小样本和超高维的特点使得基因选择成为十分困难的问题,引起了人们极大的关注.基因选择是目前基因表达数据处理和分析的热点研究问题[2,3].

Xiong 等人在特征 wrappers 的基础上,通过分类误差来选择鉴别基因[4].Guyon 等人提出了递归特征减少法(recursive feature elimination),借助支持向量机(support vector machines)递归去除分类函数中关联权重绝对值最小的基因,得到基因集合的排序来选择鉴别基因.然而,该方法需要很大的计算量[5].Ben-Dor 等人在一维空间为每个基因搜索一个最优判别点,然后根据此判别点进行样本分类,借助分类误差和 TNoM 数(threshold number of misclassification)来选择鉴别基因[6].这种方法的缺点是搜索最优判别点的计算复杂性大.上述方法均使用具体分类器或分类误差来选择鉴别基因,得到的基因选择结果容易受到具体分类方法的影响.

Lee 等人基于分层 Bayesian 线性模型,结合马尔可夫链蒙特卡罗方法和 Gibbs 抽样来估计每个基因在模型中出现的概率,根据概率的大小来选择鉴别基因[7].基因表达数据通常含有噪音,而且不同数据间的差异很大[8],因此,很难构造一个适用于所有基因表达数据的数据模型,基于数据模型的基因选择方法的适用性较差.

Golub 等人使用均值和方差构造的统计量作为鉴别基因的选择测度[9].文献[10]直接使用两样本 t 统计量(two sample t-statistic)来选择基因.Bø 和 Jonassen 提出了一种成对基因选择方法,该方法把每个样本在一对基因上的表达数据(二维空间中的点)投影到样本点的线性判别界面的垂直线上,然后用投影点的两样本 t 统计量为测度来选择基因[11].这些方法仅适用于两种类别的问题,不适用于多类别的问题.

由于非参数方法不需要假定一个具体的数据分布,因此比较适于用来分析基因表达数据.针对两个类别的情况,Park 等人首先有序地组织样本,使得类别 1 的样本位于基因表达矩阵的左半面,类别 2 的样本在右边,并为类别 1 和类别 2 的样本分别分配标签 0 和 1;然后,在每个基因上对所有样本的表达值按大小排序,此时,有序的标签序列可能被破坏;接着两两交换样本,使得所有标签有序,以最小的交换数目作为这个基因的得分来排序和选择基因[12].Dudoit 等人用 between-group 和 within-group 的差别平方和之比 BW 来排序和选择基因[2].Cho 等人用样本到类质心距离的平均值和标准差来选择基因[13].虽然 Dudoit 和 Cho 提出的方法适用于多类别,但是它们都没有考虑样本数目的不平衡问题.

本文提出一种模型无关的稳健的基因选择方法,在类内变化小和类间差别大的策略下,通过比较来选择敏感的度量函数以提高方法的鉴别能力.同时,通过强调类内变化和类间差别的一致性来增加方法的稳定性和适用性.该方法不仅与具体应用的分类模型及数据模型无关,而且适用于任意类别的基因选择问题,同时克服了样本数目的不平衡现象.

1 基因选择方法

设有 n 个样本,来自 L 个不相交的类别.在样本 i ($1\le i\le n$)上,p 个基因的表达值称为样本 i 的基因表达谱(gene expression profile),简称表达谱,可以表示为向量 $\boldsymbol{A}_i=(a_{i1},a_{i2},\ldots,a_{ip})$,其中 a_{ij} 为样本 i 上基因 j 的表达值.表达谱与样本一一对应.矩阵 $\boldsymbol{A}=[\boldsymbol{A}_1^T,\boldsymbol{A}_2^T,...,\boldsymbol{A}_n^T]^T=[a_{ij}]_{n\times p}$ 称为 n 个样本 p 个基因的基因表达矩阵,简称基因表达矩阵.

把基因在属于相同类别的不同样本中表达值之间的大小差距称为类内变化;基因在不同类别的不同样本中表达值之间的大小差距称为类间差别.那么可以认为,一个理想的鉴别基因一定是类间差别比较大,而类内变化较小.把类间差别和类内变化分别记为 *scatter* 和 *compact*,用它们的比值 *compact*/*scatter* 来表示基因的鉴别能力.基于这种思想,提出本文的基因选择方法.

定义 **1**. 给定基因表达矩阵 $\boldsymbol{A}$,其样本的类别个数为 L,令 C_k 为属于第 k 个类别的样本标号的集合,$n_k=|C_k|$,那么,矩阵 $\overline{\boldsymbol{A}}=[\overline{a}_{kj}]_{L\times p}$ 称为表达谱的类质心矩阵(类均值矩阵),其中 $\overline{a}_{kj}=\frac{1}{n_k}\sum_{i\in C_k}a_{ij}$.矩阵 $\overline{\boldsymbol{A}}$ 中第 k 个行向量 $\overline{\boldsymbol{A}}_k$ 为类别 k 中所有样本表达谱的质心(即平均值).

定义 **2**. 给定基因表达矩阵 $\boldsymbol{A}$ 和类质心矩阵 $\overline{\boldsymbol{A}}$,设样本 i 属于第 k 个类别,向量 $\boldsymbol{X}_i=(x_{i1},x_{i2},\ldots,x_{ip})$称为样本 i 的类内变化谱,其中 $x_{ij}=\left|a_{ij}-\overline{a}_{kj}\right|$,即 $\boldsymbol{X}_i=|\boldsymbol{A}_i-\overline{\boldsymbol{A}}_k|$.矩阵 $\boldsymbol{X}=[x_{ij}]_{n\times p}$ 称为类内变化矩阵.

定义 **3**. 给定类内变化矩阵 $\boldsymbol{X}$,设样本的类别个数为 L,令 C_k 为属于第 k 个类别中样本标号的集合,$n_k=|C_k|$,那么,矩阵 $\overline{\boldsymbol{X}}=[\overline{x}_{kj}]_{L\times p}$ 称为类内变化谱的类质心矩阵,其中 $\overline{x}_{kj}=\frac{1}{n_k}\sum_{i\in C_k}x_{ij}$.

根据定义 2,基因 j 在属于类别 k 的样本 i 上的表达值 a_{ij} 可以看成 $\overline{a}_{kj}$ 和 x_{ij} 两部分.它们可以分别用于量化类间差别 *scatter* 和类内变化 *compact*.在构造度量基因鉴别能力的函数时,需要考虑类别中样本数目差异(样本的不平衡)对度量函数的影响,给出受样本不平衡性的影响相对较小的度量函数.例如,对于基因 j 来说,平均值 $\sum_{k=1}^{L}\overline{a}_{kj}/L$ 和 $\sum_{i=1}^{n}a_{ij}/n$ 是不相同的,后者容易受到样本数多的那个类别的影响,因而是有偏倚的.我们给出一个简单策略来消除样本不平衡的影响.令 $F(j)$是关于基因 j 的函数,如果我们复制任意一个类别内的所有样本,使得这个类别拥有 $2n_k$ 个样本后,函数 $F(j)$的值保持不变,则称 $F(j)$关于样本偏斜是稳定的.

1.1 类间差别

若基因 j 是一个理想的鉴别基因,那么其类间差别较大.也就是说,关于基因 j 的表达谱的类质心 $\overline{a}_{kj}$ ($1\le k\le L$)间距离较大.对于任意的基因 j,直观地可以用下面定义的 $S(j)$来表示基因 j 上的样本的类间差别.

$$S(j)=\sqrt{\frac{1}{L}\sum_{k=1}^{L}(\overline{a}_{kj}-\hat{a}_j)^2}\ ,\text{其中}\ \hat{a}_j=\frac{1}{L}\sum_{k=1}^{L}\overline{a}_{kj}\ .$$

然而,$S(j)$并不能很好地体现样本的鉴别能力.

例 1:设某个基因表达数据的样本分成 3 个不同的类别,其基因 s 和基因 t 上样本表达谱的类质心分别是 $(\overline{a}_{1s},\overline{a}_{2s},\overline{a}_{3s})'=(0.1,0.5,0.6)'$ 和 $(\overline{a}_{1t},\overline{a}_{2t},\overline{a}_{3t})'=(0.2,0.4,0.6)'$.由公式 $S(j)$可以得到:$S(s)=0.216>S(t)=0.1633$,即基因 s 上的类间差别更大,$S(s)>S(t)$.然而,由于基因 s 上的类间差别的匀称性(一致性)不好,使得基因 s 上最小的类质心之间的距离(0.1)小于基因 t 上的最小值(0.2),因而实际上,基因 t 的鉴别能力更好.

由于类质心间的最小距离对基因的鉴别能力有重要的影响,因此为 $S(j)$增加一个惩罚项来强调类间差别的一致性,从而得到了一个新的类间差别计算函数,具体如下

$$scatter(j)=S(j)+\frac{1}{2}\min_{w\ne v}\left|\overline{a}_{vj}-\overline{a}_{wj}\right|\ .$$

对上例中的基因 s 和基因 t,有 $scatter(s)=0.316<scatter(t)=0.3633$.显然,由新公式得到的类间差别能够更好地反映样本的鉴别能力,从而有助于获得理想的鉴别基因.

进一步地,我们可以证明函数 $scatter(j)$的上下界.

引理 **1**. 设 $a_i>0,1\le i\le n$,那么有 $(a_1+...+a_n)/n\le\sqrt{(a_1^2+...+a_n^2)/n}$,等式成立当且仅当 $a_i=a_j,\forall i\ne j$.即算术平均数小于等于二次平均数.

证明:令 $\bar{a}=(a_1+...+a_n)/n$,把 $a_i=(a_i-\bar{a})+\bar{a}$ 代入不等式右边项,展开即可证明. □

引理 2. 对 $0\le a_1\le\ldots\le a_n(n\ge 2)$,令 $\bar{a}=\sum_{i=1}^{n}a_i/n$, $b=\min_{1\le k\le n-1}\{(a_{k+1}-a_k)/2\}$ 且 $S=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i-\bar{a})^2}$,则有 $S\ge b$ 且有 $S+b\le a_n-a_1$.

证明:如果 $b=0$,则两个不等式成立;否则,有 $a_1<\ldots<a_n$.

为证 $S\ge b$,只要证 $S^2\ge b^2$.不失一般性,假定 $b=(a_{p+1}-a_p)/2$.如果 $\bar{a}\in[a_p,a_{p+1}]$,根据引理 1,有 $(\bar{a}-a_p)^2+(a_{p+1}-\bar{a})^2\ge(a_{p+1}-a_p)^2/2=2b^2$.对 $\forall j\ne p$ 且 $j\ne p+1$,有 $(a_j-\bar{a})^2\ge b^2$.如果 $\bar{a}\in[a_k,a_{k+1}]$ 且 $k\ne p$,同样可得 $(a_k-\bar{a})^2+(a_{k+1}-\bar{a})^2\ge 2b^2$,且有对 $\forall j\ne p, j\ne p+1$,有 $(a_j-\bar{a})^2\ge b^2$.于是有 $nS^2\ge nb^2$.因此,不等式 $S\ge b$ 成立.

下面证明 $S+b\le a_n-a_1$.如果 $n=2$,则有 $S+b\le a_n-a_1$ 不等式成立.如果 $n>2$,可以证明 $S\le\max\{a_n-\bar{a},\bar{a}-a_1\}$ 成立.于是,只要证明 $\max\{a_n-\bar{a},\bar{a}-a_1\}+b\le a_n-a_1$ 即可完成证明.因为有 $a_n-\bar{a}+b\le a_n-\frac{a_2+a_1}{2}+\frac{a_2-a_1}{2}=a_n-a_1$,

$\bar{a}-a_1+b\le\frac{a_n+a_{n-1}}{2}-a_1+\frac{a_n-a_{n-1}}{2}=a_n-a_1$,所以有 $S+b\le a_n-a_1$. □

定理 1(函数 ***scatter***(*j*)的上下界限). 令 $\max(j)=\max_{w\ne v}|\bar{a}_{vj}-\bar{a}_{wj}|$ 且 $\min(j)=\min_{w\ne v}|\bar{a}_{vj}-\bar{a}_{wj}|$,其中 $1\le w,v\le L$.那么有 $\min(j)\le scatter(j)\le\max(j)$.

证明:由引理 2,易证.当只有两个类别时,不等式中的等号成立. □

1.2 类内变化

对一个理想的鉴别基因,它的类间差别较大,同时它的类内变化较小.给定一个基因表达矩阵 ***A***,通常可以采用下述两种方法之一来度量每个基因上的类内变化:

$$\tilde{\mu}(j)=\frac{1}{n}\sum_{i=1}^{n}x_{ij}\ 或\ \mu(j)=\frac{1}{L}\sum_{k=1}^{L}\bar{x}_{kj}=\hat{x}_j.$$

但是,$\tilde{\mu}(*)$ 对样本数量多的类别存在偏斜,而函数 $\mu(*)$ 的鉴别能力又较弱.

例 2:设(0.2,0.4,0.6)′和(0.4,0.4,0.4)′分别是对应 3 个类别时基因 *s* 和基因 *t* 上类内变化谱的类质心,那么有 $\mu(s)=\mu(t)=0.4$,函数 $\mu(*)$ 不能区分两个基因.因此,$\tilde{\mu}(*)$ 和 $\mu(*)$ 均不适合基因上的类内变化.

实际上,类内变化谱的类质心大意味着该类别中样本间的差距大.所以,一个基因的最大的类内变化谱的类质心越小,它的鉴别能力越好.例 2 中基因 *t* 上最大的类内变化谱的类质心数值(0.4)小于基因 *s*(0.6),因此基因 *t* 优于基因 *s*.

定义 4. 给定类内变化谱的类质心矩阵 $\bar{\boldsymbol{X}}$,则称如下定义的函数 $d_1(j)$ 为基因 *j* 上的类内变化因子 I.对于上面的例 2,有 $d_1(s)=0.4321>d_1(t)=0.4$,函数 $d_1(j)$ 比 $\mu(*)$ 有更强的鉴别力.

$$d_1(j)=\sqrt{\frac{1}{L}\sum_{k=1}^{L}\bar{x}_{kj}^2}\ .$$

定义 5. 给定类内变化矩阵 ***X***,令 *L* 为样本类别的个数,令 C_k 为属于第 *k* 个类别中样本标号的集合,$n_k=|C_k|$,那么称如下定义的函数 $d_2(j)$ 为基因 *j* 的类内变化因子 II.

$$d_2(j)=\sqrt{\frac{1}{L}\sum_{k=1}^{L}\left(\frac{1}{n_k}\sum_{i\in C_k}x_{ij}^2\right)}.$$

事实上,$d_1(j)$ 和 $d_2(j)$ 两者都是所有样本的类内变化谱的某种平均值.根据引理 1 有 $d_2(j)\ge d_1(j)$,这表明函数 $d_2(j)$ 在评价某个基因的类内变化时比函数 $d_1(j)$ 更敏感.因此在应用中,我们用 $d_2(j)$ 来度量基因 *j* 上的类内变化.

例 3:如图 1 所示,为两个类别 *A* 和 *B* 时,样本在基因 *s* 和基因 *t* 上的类内变化谱,其中:基因 *s* 和基因 *t* 上各自的类质心大小相等($\bar{x}_{ks}=\bar{x}_{kt}$,*k*=*A*,*B*);类别 *A* 有 9 个样本,类别 *B* 有 11 个样本.直观上可以看出,基因 *s* 的鉴别能力强于基因 *t*.对于基因 *s* 和 *t* 上的类内变化因子 I 和 II,有 $d_2(s)<d_2(t)$,但 $d_1(s)=d_1(t)$.

定义 6. 给定类内变化谱的类质心矩阵 $\bar{\boldsymbol{X}}$,则称如下定义的函数 $\delta_1(j)$ 为基因 *j* 上的类内变化匀质因子 I.

定义 7. 给定类内变化矩阵 ***X*** 和类内变化谱的类质心矩阵 $\bar{\boldsymbol{X}}$,令 *L* 为样本类别的个数,C_k 为属于第 *k* 个类

别的样本标号的集合,$n_k=|C_k|$,则称函数$\delta_2(j)$为基因 j 上的类内变化匀质因子 II.

$$\delta_1(j)=\sqrt{\frac{1}{L}\sum_{k=1}^{L}\left(\bar{x}_{kj}-\mu(j)\right)^2}\ ,\ \delta_2(j)=\sqrt{\frac{1}{L}\sum_{k=1}^{L}\left(\frac{1}{n_k}\sum_{i\in C_k}\left(x_{ij}-\mu(j)\right)^2\right)}\ .$$

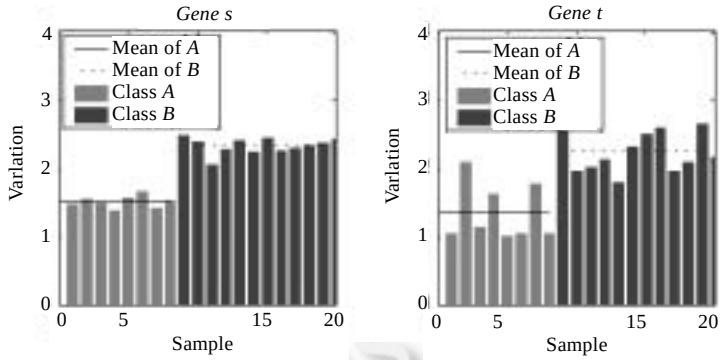


Fig.1 Within-Class variations of gene s and t

图 1 基因 s 和 t 上的类内变化

对于一个鉴别基因 j,我们希望它的类内变化的平均值要小;更进一步地,希望它在所有样本上的类内变化的匀质性要好.也就是说,对鉴别基因 j,希望其类内变化因子小,同时希望其类内变化匀质因子也小.

定理 **2**. 给定类内变化矩阵 $\boldsymbol{X}$ 和类内变化谱的类质心矩阵 $\overline{\boldsymbol{X}}$,对任意的 j,$1\le j\le p$,函数$\delta_1(j)$和$\delta_2(j)$满足 $\delta_1(j)=\sqrt{d_1(j)^2-\mu(j)^2}$ 和 $\delta_2(j)=\sqrt{d_2(j)^2-\mu(j)^2}$,且有$\delta_1(j)\le\delta_2(j)$.

证明:化简$\delta_1(j)$和$\delta_2(j)$即可. □

根据定理 2,可以简化$\delta_1(j)$和$\delta_2(j)$的计算,同时可以看到$\delta_2(j)$比$\delta_1(j)$更加敏感.我们选择函数$\delta_2(j)$来度量类内变化的匀质性.给定一个类内变化矩阵,组合函数$\delta_2(j)$和 $d_2(j)$来量化类内变化,于是得到度量类内变化的函数 $compact(j)$为

$$compact(j)=d_2(j)+\delta_2(j).$$

1.3 基因选择的方法

定义 **8**. 给定基因表达矩阵,其样本的类别个数为 L,令 C_k 为属于第 k 个类别的样本标号的集合,$n_k=|C_k|$,$compact(j)$和 $scatter(j)$分别为基因 j 上类内变化和类间差别的测度函数,则称 $score(j)=compact(j)/scatter(j)$为基因 j 的鉴别能力.

对于任意一个基因 j,可以计算它的鉴别能力 $score(j)$,其获得的分数值越小,表明这个基因与类别的相关程度越高,鉴别能力越强,以基因 j 为特征的基因表达数据的可分性越好.为每个基因计算一个分数之后,我们按照分数大小的递增序来排序每个基因.基于这种有序的基因序列或基因所获得的分数,有多种策略来选择鉴别基因.首先,可从排序的基因序列中选择前 x 个基因来进一步分析,如数据分类.其次,可以从排序的基因序列中选择前 x%基因来进一步分析.给定一个常数 C,选择其分数满足 $score(j)\le C$ 的基因来分析,例如,可令 $C=1$ 或 $C=1/2$.根据定理 1,还可以用一个更保守的函数 $S_{\max}(j)=compact(j)/\max(j)$来过滤鉴别能力差的基因,以降低基因表达数据的维度,减少数据分析时的复杂性.例如,给定常数 C,过滤满足 $S_{\max}(j)\ge C$ 的基因.

这一度量函数有下面的特点:(1) 函数 $compact(j)$是基因 j 上所有样本的类内变化的一种平均值;(2) 函数 $\delta_2(j)$是一个针对类内变化的匀质性的惩罚因子,当所有样本的类内变化相等时,即 $d_2(j)=\mu(\overline{\boldsymbol{X}},j)$,惩罚因子$\delta_2(j)$的值等于 0;(3) 由定理 1,有 $\max(j)\ge scatter(j)\ge\min(j)$.因此对某个基因 j 来说,函数 $scatter(j)$可以看成是两两类别质心之间“距离”的一个平均值;(4) 函数 $score(j)$的值是类内变化平均值和类质心间距的平均值的比值,这一比值具有某种统计上的意义,图 2 是一个示意图.

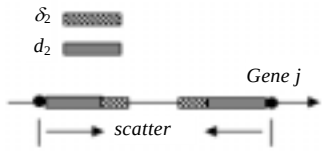


Fig.2 Conceptual illustration of the function *score*(*j*)

图 2 函数 *score*(*j*)的一个解释

2 复杂性分析

问题:给定一个基因表达数据矩阵 $\boldsymbol{A}$、类别数目 L、样本的类别信息 C_k 和 n_k.找出前 x 个鉴别能力好的鉴别基因.

输入:给定一个基因表达数据矩阵 $\boldsymbol{A}$、类别数目 L、样本的类别信息 C_k 和 n_k.

输出:前 x 个鉴别能力强的鉴别基因.

我们把加、减、乘、除、平方、开平方、绝对值等数值操作和比较大小等操作看成是消耗常数时间的单位操作,那么,我们的基因选择方法的计算复杂性如下:计算 $\overline{\boldsymbol{A}}$ 和 $\hat{\boldsymbol{A}}$ 的时间复杂性为 $O(p(L+n))$;计算 $\boldsymbol{X}$ 的时间复杂性为 $O(np)$;计算 $\overline{\boldsymbol{X}}$ 和 $\hat{\boldsymbol{X}}$ 的时间复杂性为 $O(p(L+n))$;计算每个基因的 *scatter*(*j*)共用时间为 $O(p(L+L\lg L))$;计算所有的 $d_2(j)^2$ 共用 $p(L+2n)$次单位操作;计算每个 *score*(*j*)用 $O(1)$次单位操作;排序 p 个基因用 $O(p\lg p)$次单位操作.所有操作的时间复杂性为 $O(np+pL+pL\lg L+p\lg p)$,由于 $L\leq n<<p$,故时间复杂性为 $O(p\lg p)$.

3 实验结果和讨论

在两个著名的数据集 leukemia[9]和 small round blue cell tumors (SRBCT)[14]上,我们从分类的精度、不同方法的差异性和方法的稳定性 3 个角度评价我们提出的方法.Leukemia 数据包括两个类别:ALL(47 样本)和 AML(25 样本),共 72 个样本.我们沿用文献[2]中提出的预处理方法处理 leukemia 数据,得到一个每个样本包含 3 571 个基因表达值的数据集.经 Han 等人处理,SRBCT 数据集包含 2 308 个基因,所有样本被分成一个训练集和一个测试集,分别拥有 63 和 25 个样本.为了与参考文献相一致,我们移去测试集中的 5 个非 SRBCT 样本,得到一个包含 20 个样本的测试集.SRBCT 数据集包括 4 个类别,分别为 EWS,BL,NB 和 RMS.

3.1 分类性能

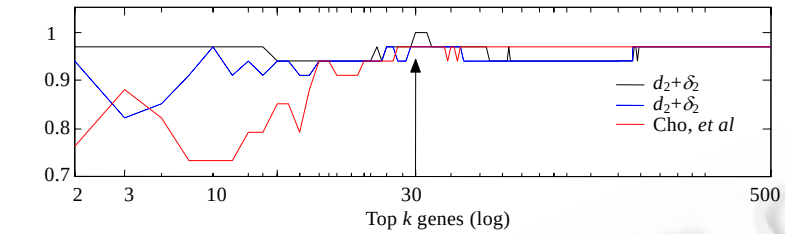
在分类时,我们选择支持向量机(SVMs)分类器[5]来分类.对于超过两个类别的情况,我们应用 One versus All (OVA)[15]策略来使用二元分类器 SVMs.我们所用的 Matlab 环境下的 SVMs 工具箱来自文献[16].对支持向量机,选择线性核函数,其他参数是默认的.

分类的具体过程如下:对训练样本上所有基因的表达值,用我们提出的两个基因选择方法(分别对应 *compact*(*j*)的两个计分函数$\delta_2(j)+d_2(j)$和$\delta_1(j)+d_1(j)$)和对照的方法(Cho 等[13])选出前 k 个鉴别基因,即 k 个特征基因;然后,用选出的 k 个基因在训练样本上的表达值来训练分类器;最后,分别用训练样本和测试样本在选出的 k 个基因上的表达值来测试,给出训练精度和测试精度.图 3 中分别给出了两个数据集 leukemia(如图 3(a)所示)和 SRBCT(如图 3(b)所示)上的测试精度(表示成用不同基因选择方法选出的前 k 个基因作为特征基因的函数,箭头表示达到 100%精度时所用最少的基因数).可以看到:在 3 种方法中,应用我们的基因选择方法(即函数 $\delta_2(j)+d_2(j)$),在最少的特征下能达到最好的分类精度.

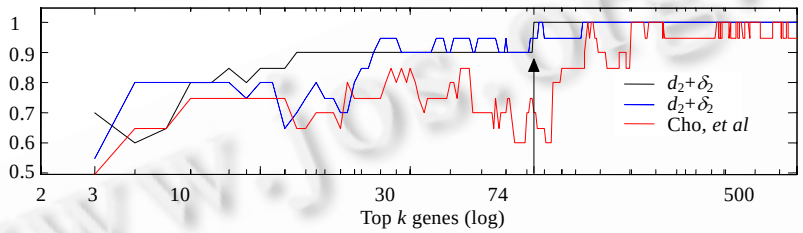
对 SRBCT 数据,当用我们的方法选出 74 个基因时,训练精度和测试精度都是 100%;用 SIMCA 方法[17]选出 60 个基因,对 20 个测试样本的测试精度是 95%;Khan 等人应用人工神经网络方法,选出 96 个基因,对训练样本正确分类,但是诊断精度只有 90%[14];选出 43 个基因,Tibshirani 等人正确分类测试样本[18];把 63 个训练样本和 20 个测试样本混合后,通过多次 5 倍交叉验证,Cho 等人选出 21 个基因,获得的平均最小测试误差为 0.96%[13].

在 leukemia 数据上应用我们的方法,当选出 30 个基因时,训练和测试精度达到 100%.如果只选前两个基因,那么训练精度是 100%并且其测试精度是 97.06%,也就是说,测试误差为 1/34;Golub 等人使用 50 个基因获得最

好的测试误差是 4/34[9];Tibshirani 等人利用他们提出的收缩质心方法,在使用 21 个基因时测试误差为 2/34[18];文献[17]应用 SIMCA 方法,最好的测试精度为 82.3%;Cho 等人[13]得到的平均交叉验证误差为 4.06%;同样,使用支持向量机分类器,Fu 等人的最好测试精度为 97.06%[19].



(a) Leukemia



(b) SRBCT

Fig.3 Test accuracy of classification on two datasets

图 3 在两个数据集上分类器的测试精度

从上面的分类精度对照可以看出:建立在我们的基因选择方法上的分类器对两个数据都能全部正确分类,分类器的性能优于目前报道的结果.

3.2 不同基因选择方法结果的差异性

为了验证不同基因选择方法结果间的差异性,我们在表 1 中列出了由不同基因选择方法选出的前 x 个基因间重叠的情况,其中对 leukemia 数据 x 为 30,对 SRBCT 数据 x 等于 74.非常明显,我们的方法和 Cho 的方法间的差异是很大的.

Table 1 Overlap of top k genes by different methods

表 1 不同基因选择方法选出的前 k 个基因间的重合度

	30 genes for leukemia		74 genes for SRBCT	
	δ_1+d_1	Cho	δ_1+d_1	Cho
δ_2+d_2	24 (80%)	6 (20%)	61 (82.43%)	21 (28.38%)

下面考察不同的基因选择方法的结果与已有文献的结果的一致性.在 SRBCT 数据的训练样本上,分别用 3 种基因选择方法(δ_2+d_2,δ_1+d_1 和 Cho,等)选出前 74 个基因.把这 3 个基因集合中的每一个分别与文献[14,17,18]中选出的 3 个特征基因集合进行比较,考察基因选择结果的一致性.比较的结果列于表 2 中,我们的方法一致性最好,Cho 的方法一致性最差.

Table 2 The consensus of top 74 genes by different methods and three reported sets of genes

表 2 不同基因选择方法选出的 74 个基因和文献中 3 个特征基因集合间的一致性

Method	96 genes, Khan, *et al.*	43 genes, Ribshirani, *et al.*	60 genes, Bicciato, *et al.*
δ_2+d_2	28 (29.17%)	12 (27.91%)	25 (41.67%)
δ_1+d_1	23 (23.96%)	9 (20.93%)	20 (33.33%)
Cho	17 (17.71%)	5 (11.63%)	8 (13.33%)

3.3 方法的稳定性

类别中样本的数目对基因选择结果的影响很大,特别是当样本数目出现偏斜时.我们希望基因选择方法受

样本数目的影响越小越好,也就是说,基因选择方法越稳定越好.在这一节中,我们通过两种方式来考察基因选择方法的稳定性.第一种方式的具体过程如下:对其中的某个类别 k,复制属于类别 k 的所有样本,使得这一类别有 $2\times n_k$ 个样本,其余类别的样本保持不变,所得到的这一新的数据称为"复制数据".然后,分别在复制数据和原始数据中选出前 x 个基因,比较两个结果间的重合情况.对 leukemia 数据的两个类别和 SRBCT 数据的 4 个类别的结果列于表 3 中,对每一个类别有两个结果:一个是从所有训练样本上计算的;另一个是从所有的训练和测试样本上计算的.实验结果表明:δ_2+d_2 和 δ_1+d_1 对样本的偏斜是稳定的,而 Cho 的函数是不稳定的.

Table 3 The consensus of genes selected by the same method from duplicated data and from original data

表 3 同一基因选择方法在复制数据和原始数据上选出的两个基因集合间的一致性

Method	Top x genes	Data of all training samples (%)						Data of all training and test samples (%)					
		EWS	BL	NB	RMS	ALL	AML	EWS	BL	NB	RMS	ALL	AML
δ_2+d_2	30/74/100	100	100	100	100	100	100	100	100	100	100	100	100
δ_1+d_1	30/74/100	100	100	100	100	100	100	100	100	100	100	100	100
Cho	30	83.3	83.3	96.7	86.7	83.3	86.7	90	93.3	90	86.7	96.7	83
	74	86.5	89.2	90.5	90.5	90.5	90.5	90.5	89.2	91.9	91.9	93.2	86.5
	100	93	88	92	91	90	93	90	94	92	91	92	90

对于第 2 种方式,对某个类别 k,我们把这一类的样本随机分割成大小近似相等的 3 份,每次移除其中的一份样本,剩下两份,其他类别保持不变,得到一个"移除数据".这一过程进行 3 次,使得类别 k 中的每个样本只被移出一次.我们分别在移除数据和原始数据上选出前 x 个基因各一组,计算重合度.与原始数据上结果的重合度越高,表明方法的稳定性越好.我们重复这一过程 20 次,计算平均重合度,实验结果列于表 4 中.对不同的类别和 x 的值,统计获得最高重合度的次数,我们方法的稳定性优于 Cho 的方法,其中 δ_2+d_2 的稳定性是最好的.

Table 4 The average consensus of genes selected by the same method from removed data and from original data

表 4 同一基因选择方法在移除数据和原始数据上得出的两个基因集合间的平均一致性

Top genes	Method	Data of all training samples (%)						Data of all training and test samples (%)						Num. of max.
		EWS	BL	NB	RMS	ALL	AML	EWS	BL	NB	RMS	ALL	AML	
30	δ_2+d_2	84.8	77.8	81.9	87.4	75.4	65.3	87.5	81.3	85	83.8	87.4	84.9	7
	δ_1+d_1	87.6	75.0	84.8	89.1	75.1	69.0	82.9	73.9	80.4	81.8	85.5	80.4	4
	Cho	74.1	71.6	77.5	74.1	60.8	58.8	83.2	79.9	85.5	83.4	75.0	79.0	1
74	δ_2+d_2	89.0	87.1	89.8	88.7	77.3	71.0	85.5	82.6	86.4	83.9	84.5	80.8	5
	δ_1+d_1	85.7	80.9	87.0	84.6	78.3	73.3	84.6	80.2	84.3	85.2	83.0	80.9	4
	Cho	84.8	83.3	85.5	79.9	66.4	66.7	87.5	86.9	88.3	85	75.8	77.0	3
100	δ_2+d_2	87.3	86.1	90.1	87.3	80.0	74.1	86.6	83.9	87.6	86.3	83.7	80.8	5
	δ_1+d_1	87.4	87.4	87.9	86.5	77.5	74.9	84.9	79.2	84.1	82.9	83.6	80.9	4
	Cho	85.6	85.6	88.4	81.4	69.0	67.7	88.7	84.5	89.5	86	77.4	75.8	3

4 结 论

本文提出了一种全新的模型无关的基因选择方法.本文的方法考虑了如下的问题:首先,在基因表达数据中,类别中样本的不平衡是一个常见问题.我们的方法能够合理地处理这一常见问题.进一步地,我们强调了基因选择方法对样本数目的稳定性,使得方法更稳健,适用性更强.最后,我们提出的方法可以直接用于多种类别的问题.额外地,通过我们的方法分配给每个基因的分数有一个概念上的意义,并且作为一种单基因打分方法,我们提出的基因选择方法的计算复杂性是很低的.

References:

[1] Antoniadis A, Lambert-Lacroix S, Leblanc F. Effective dimension reduction methods for tumor classification using gene expression data. Bioinformatics, 2003,19(5):563−570.

[2] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association, 2002,97(457):77−87.

[3] Lu Y, Han J. Cancer classification using gene expression data. Information System, 2003,28(4):243−268.

[4] Xiong M, Fang X, Zhao J. Biomarker identification by feature wrappers. Genome Research, 2001,11(11):1878−1887.

[5] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning, 2002,46(3):389−422.

[6] Ben-Dor A, Bruhn L, Friedman N, Nachma I. Tissue classification with gene expression profiles. In: Shamir R, Miyano S, Istrail S, Pevzner P, Waterman M, eds. Proc. of the 4th Annual Int'l Conf. on Computational Molecular Biology (Recomb). Tokyo: ACM, 2000.

[7] Lee KE, Sha N, Dougherty ER, Vannucci M, Mallick BK. Gene selection: A Bayesian variable selection approach. Bioinformatics, 2003,19(1):90−97.

[8] Hunter L, Taylor RC, Leach SM, Simon R. GEST: A gene expression search tool based on a novel Bayesian similarity metric. Bioinformatics, 2001,17(Suppl. 1):S115−S122.

[9] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: Class discovery and class prediction gene expression monitoring. Science, 1999,286(5439):531−537.

[10] Varma S, Simon R. Iterative class discovery and feature selection using minimal spannig trees. BMC Bioinformatics, 2004,5:126.

[11] Bø TH, Jonassen I. New feature subset selection procedures for classification of expression profiles. Genome Biology, 2002, 3(4):research0017.

[12] Park PJ, Pagano M, Bonetti M. A nonparametric scoring algorithm for identifying informative genes form microarray data. In: Proc. of the Pacific Symp. on Biocomputing. 2001,6:52−63. http://psb.stanford.edu/psb-online/proceedings/psb01/

[13] Cho JH, Lee D, Park JH, Lee IB. New gene selection for classification of cancer subtype considering within-class variation. FEBS Letters, 2003,551(1):3−7.

[14] Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwarb M, Antonescu CR, Peterson C, Meltzer CR. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine, 2001,7(6):673−679.

[15] Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR. Multiclass cancer diagnosis using tumor gene expression signatures. Proc. of the National Academy of Sciences of the United States of America, 2001,98(26):15149−15154.

[16] Cawley GC. MATLAB support vector machine toolbox. 2004. http://theoval.sys.uea.ac.uk/~gcc/svm/toolbox/

[17] Bicciato S, Luchini A, Dibello C. PCA disnoint models for multiclass caner analysis using gene expression data. Bioinformatics, 2003,19(5):571−578.

[18] Tibshirani R, Hastie t Narasimhan B, Gilbert C. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. of the National Academy of Sciences of the United States of America, 2002,99(10):6567−6572.

[19] Fu LM, Casey SFL. Multi-Class cancer subtype classification based on gene expression signatures with reliability analysis. FEBS Letters, 2004,561(2):186−190.

李建中(1951－),男,黑龙江哈尔滨人,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库,数据流,传感器网络,生物信息学.

骆吉洲(1975－),男,博士生,主要研究领域为压缩数据库技术,计算生物学.

杨昆(1979－),男,博士生,主要研究领域为生物信息学.

郭政(1963－),男,博士,教授,主要研究领域为生物信息学.

高宏(1966－),女,博士,教授,主要研究领域为数据库,数据仓库,计算生物学.