

一种基于递归分类树的集成特征基因选择方法

李 霞^{1, 2)} 张田文¹⁾ 郭 政^{1, 2)}

¹⁾(哈尔滨工业大学计算机科学与技术系 哈尔滨 150001)

²⁾(哈尔滨医科大学生物医学工程与生物信息学系 哈尔滨 150086)

摘要 利用 DNA 芯片基因表达谱信息识别疾病相关基因, 对癌症等疾病分型、诊断及病理学研究有非常重要的实际意义。该文提出了一种基于递归分类树的特征基因选择的集成方法 EFST (Ensemble Feature Selection based on Recursive Partition-Tree)。EFST 可选择多组基于不同样本分布结构的特征基因, 结合有监督机器学习中的多分类器集成(ensemble)决策技术, 利用提出的衡量特征基因稳定性与显著性测度, 集成各特征基因组选择最终的特征基因。应用结肠癌 2000 个基因的表达谱实验数据分析结果显示: EFST 方法不仅具有寻找疾病相关基因的能力和较强的数据维数压缩能力, 而且由支持向量机(SVM)等 4 种模式分类方法证实 EFST 方法可以明显地提高疾病鉴别分类的准确率。

关键词 基因表达谱; 递归分类树; 特征选择; 集成决策

中图法分类号 TP391

An Novel Ensemble Method of Feature Gene Selection Based on Recursive Partition-Tree

LI Xia^{1,2)} ZHANG Tian-Wen¹⁾ GUO Zheng^{1,2)}

¹⁾(Department of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

²⁾(Department of Biomedical Engineering and Bioinformatics, Harbin Medical University, Harbin 150086)

Abstract To identify disease genes from these gene expression profiles is critically important for disease, such as cancer, subtype discovery, diagnosis and pathology study. This paper proposes a feature gene selection method named EFST(Ensemble Feature Selection Based on Recursive Partition-Tree) which can be applied to select multiple feature gene groups from one training sample set, and defines a significance and stability measure for each selected feature in a way similar to the ensemble decision method of supervised machine learning. Authors apply the EFST method to analyse the published 2,000 gene expression profile data. The results indicate that the EFST method can be used not only to select feature genes and reduce the dimension of feature space, but also to increase significantly the disease prediction accuracy of many classification methods including SVM, nearest neighbor classifier, Fisher linear and Logistic nonlinear discriminant analysis.

Keywords gene expression profile; recursive partition-tree; feature selection; ensemble decision

1 引言

DNA 芯片也叫基因芯片^[1, 2]或 DNA 微阵列,

是一种意义重大的新兴生物学技术。它可以在一次实验中同时检测成千上万个基因的表达量, 为从分子水平上研究疾病的发病机理和临床疾病诊断(疾病分型)提供了强有力的手段。然而, 在基因表达谱

收稿日期: 2002-12-17; 修改稿收到日期: 2004-01-20。本课题得到国家自然科学基金(30370798, 30170515)、国家“八六三”高技术研究发展计划项目基金(2003AA2Z2051, 2002AA2Z2052)、黑龙江省科技攻关重点基金(GB03C602-4)、黑龙江省自然科学基金(F0177)和 211 工程“十五”建设项目资助。李霞, 女, 1957 年生, 博士研究生, 教授, 主要研究方向为与生物信息学相关的知识发现、人工智能、机器学习等。E-mail: lixia@ems.hrbmu.edu.cn。张田文, 男, 1940 年生, 博士生导师, 教授, 主要研究方向为模式识别、图像处理和计算机视觉。郭政, 男, 1963 年生, 博士研究生, 教授, 主要研究方向为人工智能、知识发现。

数据获取过程中,由于非特异性杂交等原因,基因表达谱数据含有较大的实验误差。同时,由于实验成本较高,样本的数目一般为几十或上百例,而检测基因的数目往往高达几千甚至几万,其中含有大量无关的指标(检测基因),是典型的高维、高噪问题。另外一方面,由于功能相似的基因的表达高度相关,因此存在大量的在分类学意义上的冗余基因。如何利用这种具有高维、高相关(冗余)特点的有限样本基因表达谱数据,识别对疾病有鉴别意义的特征基因或疾病相关基因,为机器学习研究提出了新的课题。

在有监督学习中,特征选择算法寻求的一个目标是从众多指标 $g_j (j=1, 2, \dots, p)$ 中选择一组最优的特征子集 G^* , 最大限度地提高模式分类的准确率。一类主要的特征选择方法是过滤(filter)法,如排秩(rank)、信息增益(information gain)和马尔可夫毯(Markov blanket)等方法^[3~5]。过滤方法的主要优点是计算复杂度较低、速度快。但由于过滤方法在特征选择过程中与分类器的决策机制脱离,一般难以确定由过滤方法选择的特征是否能使某一特定分类器的分类准确率达到最大。另一类特征选择方法是缠绕(wrapper)法^[6,7]。缠绕方法将分类算法嵌入特征选择过程中,是以达到最大分类准确率为引导的一类特征选择方法,由于缠绕方法选择出的特征与分类器的决策机制能够较好地耦合,不仅分类准确率高,而且在特定的决策机制下(或观察方式下),特征的分类学意义较为明确。所以,本文采用基于递归分类树的缠绕方法,选择有分类意义的特征基因。

在基因表达谱分析这类生物医学实际问题中,选择对疾病分型有鉴别意义的特征基因,除了对提高分类算法的分类准确率(即诊断预测准确率)有意义外,更重要的是选择出的特征基因提供了研究疾病病因的重要线索。由于在分类方法学意义上“冗余”的基因表达指标中,还有许多与疾病病因相关,因此,寻找这些特征基因,对研究疾病的复杂多基因病因具有重要的意义。

为此,我们借鉴有监督机器学习中的多分类器集成(ensemble)决策技术^[8],提出一种基于递归分类树的集成特征选择方法 EFST(Ensemble Feature Selection based on Recursive Partition-Tree)。EFST 方法以提高模式分类的准确率为引导,从一组训练样本中选择多组基于不同样本分布结构的特征,以实现挖掘有分类意义的特征基因和疾病病因。

相关的特征基因的目的。值得注意的是,多分类器集成决策技术在提高分类准确率方面取得了较大的进展,而应用集成技术进行特征选择则研究得较少。

本文第 2 节提出基于递归分类树的以分类准确率引导的特征基因选择集成方法 EFST 及特征基因选择的稳定性测度,并给出其实现算法;第 3 节将 EFST 应用于 40 例结肠癌组织与 22 例正常组织中 2000 个基因的 DNA 芯片表达谱实验数据,以交叉验证抽样策略构建具有不同分布结构的训练集,识别基于不同样本分布结构的特征基因,并由 4 种模式分类方法——支持向量机(SVM)、Fisher 线性判别、K 最邻近(K-NN)和 Logistic 非线性判别等证实 EFST 可以有效地提高分类正确率。第 4 节是结论及未来的工作。

2 特征基因选择 EFST 方法

用具有 p 个基因探针的 DNA 芯片检测 n 个 DNA 样品(样本)的表达谱数据可由 $n \times p$ 矩阵 $\mathbf{X} = (x_{ij})$ 表示^[1,2], 其中 x_{ij} 可代表第 j 个基因 g_j (属性变量)在第 i 个样品 X_i (观察个体)的表达水平。当 DNA 样品属于已知类别时,每一个样品观察值数据由基因表达谱 $\mathbf{X}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ 和类标签 y_i 组成,即预测变量 X_i 和应变量 y_i 。假设 DNA 样品属于 J 个类别 $\omega_1, \omega_2, \dots, \omega_J$, 对于 J 个类别,定义类标签 y_i 为从 $1 \sim J$ 的整数,以 n_j 表示第 j ($j=1, 2, \dots, J$) 类观察数。

我们提出的特征基因选择方法 EFST 是基于有监督的递归分类树学习方法^[9],根据不同的训练样本分布结构,以类纯度指标最大与分类错误率最小为目标引导,重复运行递归分层特征基因选择方法,然后,综合集成众多的特征选择器得到最终选择的特征基因子集。寻求最优特征子集是与建立纯化类别分类器同时进行的。图 1 给出 EFST 方法的基本框架。

EFST 过程的主要思路是(图 1):首先采用某种策略(见下文),由样本集构建不同分布结构的训练集 $L_d (d=1, 2, \dots, m)$ 和实验集 $T_d (d=1, 2, \dots, m)$,以类纯度为特征选择的评价指标,由训练集反复进行训练学习,递归分层识别基于训练集 $L_d (d=1, 2, \dots, m)$ 上的一系列特征基因组 $G_d = \{g_1^d, g_2^d, \dots, g_k^d\} (d=1, 2, \dots, m)$,由若干特征属性基因组 $\{G_d\}$ 综合集成最终的特征子集 $G^* = F(G_1, G_2, \dots, G_d)$ 。

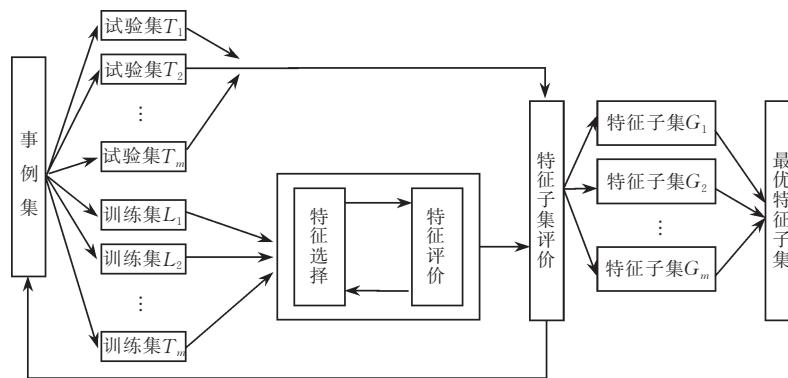


图 1 EFST 方法的基本框架

2.1 训练集的构建策略

构建训练集有许多方法,这里我们只介绍三种方法。构建训练集最直接的方法是 Bagging^[10](bootstrap aggregation 的缩写),在原训练集上采用有放回抽样,每次随机抽取与原训练集等同大小的集合(称这种集合为原训练集的副本),其特点是每一副训练集包含原训练集的 63.2%,由该副本作为训练集,余下的样本作为实验组;另一种构建训练集的方法是 n 倍(fold)交叉验证的方法^[10],我们可随机将样本集分为近似的 n 等份,选取每一份作为实验集,余下的 $n-1$ 份作为训练集,循环 n 次,这种方法产生不相重叠的训练集和实验集;第三种方法,可采用无放回随机抽样,每次抽取样本集的 $1/n$ 作为实验组,余下的样本集作为训练组。我们采用三种训练集构建方法进行了实验,得到了类似的结果,本文只报告由 n -倍(fold)交叉验证的方法所得的结果。

2.2 特征基因评价函数

按照上述训练集的构建策略,将样本集 $X = \{X_1, X_2, \dots, X_n\}$ 划分为训练集(L_d)和试验集(T_d)后,我们采用递归分类树方法^[9,11],为寻求一个最能改善分类正确率的特征属性基因,从包括所有训练集(L_d)的根结点开始,在特征空间做一次穷尽搜索,寻找一个特征属性基因和相应的阈值($cutoff$),使得最大程度地降低划分出的类别的杂质度,以此进行分叉。确定了一个最佳分叉后,当个体 X_i 的属性 $g > cutoff$ 或 $g \leq cutoff$ 时,数据集相应地划分成左右两个不相连的子集,称这些子集为源于根结点的两个子结点。然后,再在这两个子结点上,实施同样的特征空间穷尽搜索和子结点划分。

属性和相应阈值的选择准则是使得在一个结点 t 的划分最大程度地降低类别杂质度,我们采用 Gini 差异性指标(代价函数)为结点 t 的杂质函数:

$$E(t) = \phi(P(w_1 | t), P(w_2 | t), \dots, P(w_J | t))$$

$$= \sum_{i \neq j} P(w_i | t)P(w_j | t) = 1 - \sum_{j=1}^J P^2(w_j | t) \quad (1)$$

通常 $P(w_j | t) = p_j = n_j / n (j=1, 2, \dots, J)$, 参量 p_j 表示结点 t 中某一样品属于 j 类的概率(频率), $\sum_{j=1}^J p_j = 1$ 。因此,对给定结点,当所有类等同地混合于该结点时,其杂质函数最大;而当该结点只包含一个类时,其杂质函数最小。

选择特征基因的过程是试图找到一个最好的特征及阈值,在正在分叉的结点 t 这个分类层面上,确定一个相应的划分 s^* ,使得该划分有最大的杂质减少量,用符号表示为

$$\Delta E(s^*, t) = \max_{s \in S} \Delta E(s, t) \\ = \max_{s \in S} (E(t) - p_l E(t_l) - p_r E(t_r)) \quad (2)$$

其中, $E(t_l)$ 和 $E(t_r)$ 分别是左右分叉结点的杂质函数;而 p_l 和 p_r 分别是结点 t 中左右分叉事件的频率,式中 S 是结点 t 中分叉事件所有可能的方式集。

用最优分叉 s^* , t 被分为 t' 和 t'' ,并且在结点 t 这个分类层面上获得特征基因 g_t ,再对结点 t' 和 t'' 等都分别重复与结点 t 同样的搜索过程,最终得到基于训练集 L_d 的特征基因组 G_d 。

2.3 特征基因集合 G_d 的显著性检验

为了检验每次选择出的一组特征基因集合 G_d 的分类能力是否显著,我们将统计方法引进 EFST 过程,我们基于被正确与错误分类样本的数目,构建显著性检验 χ^2 统计量:

$$\chi^2 = \frac{(|n_{00}n_{11} - n_{01}n_{10}| - n/2)^2 n}{(n_{00} + n_{01})(n_{10} + n_{11})(n_{00} + n_{10})(n_{01} + n_{11})} \quad (3)$$

其中, $n = n_{00} + n_{01} + n_{10} + n_{11}$, n_{00}, n_{01}, n_{10} 和 n_{11} 分别为真阴性、假阳性、假阴性和真阳性。

根据给定的显著性水平 α_0 , 可以检验每次选择出的一组特征基因集合 G_d 的分类能力是否显著. 当样本较少时, 我们选择使用精确的 P 检验. 很显然, 采用合适的显著性水平, 用该检验可以减少无关指标被选为特征的机会.

2.4 特征基因的集成与特征基因强度的显著性检验

采用 2.1 节的某种抽样策略, 由样本集合构建一系列不同分布结构的训练集, EFST 方法在构建的各训练集上重复运行同一特征选择方法, 构建众多的特征选择器, 可以得到一系列特征组集合 $\{G_1, G_2, \dots, G_m\}$. 然后, 综合集成众多的特征选择器得到的特征基因组^[8], 由各特征基因被选择的强度(或投票得分) FV 决定最终的特征子集.

对每个特征基因 g_k 可定义并计算各特征基因的被选择强度

$$FV(g_k) = F(G_1, G_2, \dots, G_m) = \frac{\sum_d w_d I(g_k, G_d)}{\sum_d w_d} \quad (4)$$

其中, $I(g_k, G_d)$ 是一个指示函数, 当 $g_k \in G_d$, $I(g_k, G_d) = 1$; 否则, $I(g_k, G_d) = 0$, FV 属于 $[0, 1]$, 权 w_d 可为与基于 G_d 所建立的分类器的分类效能相联系的指标, 例如, 可取 $w_d = \chi^2_d$ (基于 G_d 所建立分类器的 χ^2 值), 最简单的情况取等权 $w_d = 1$. 非等权 ($w_d = \chi^2_d$) 较等权 ($w_d = 1$) 更能反映所选特征的分类学意义, 特征基因选择更严格、更集中, 非等权 ($w_d = \chi^2_d$) 所选基因几乎被等权 ($w_d = 1$) 所选基因包含, 即在等权 ($w_d = 1$) 情况下, 除了可识别分类特征基因外, 不易遗漏疾病相关基因的识别. 因而, 可根据研究问题的精细程度, 选择等权或非等权.

为了决定最终被选择的特征基因组, 我们采用随机重排 (permutation) 技术, 将样本随机赋予类别标签, 按前面计算 $FV(g_k)$ 同样的过程构建 $FV(g_k)$ 的零分布 $FV^0(g_k)$. 根据零分布 $FV^0(g_k)$ 及选定的显著性水平 β (如 0.05 或 0.01), 可以确定临界值 FV_β^0 , 最终按 $FV \geq FV_\beta^0$ (单侧检验), 在给定的显著性水平上, 选择所有的特征基因.

按照 EFST 方法, 可以提供各特征被选择的强度, 据此可以给出各特征被选择到的显著性水平, 反映各特征被选择的稳定性与可靠性. FV_β^0 作为控制参数, 可进行适当选择, FV_β^0 选择过大, 特征基因选择标准过于严格, 最终特征基因选择结果可能只包含具有最强的模式分类能力的特征基因, 一些具有

弱模式分类能力的特征基因没有被选进, FV_β^0 选择过小, 特征基因选择标准较松, 可能将会选进一些由随机因素造成 $FV(g_k)$ 增大的无关特征基因. 这也是 EFST 方法的一个显著的优点.

EFST 方法的具体步骤如下:

1. 基于实例集构造训练集 L_d 与实验集 T_d , 检验一组特征显著性的水平 α_0 .

2. 对于给定的一个训练集 L_d , 可构建一棵树作为特征选择器. 确定特征基因选择搜索的终止条件:

叶子结点只包含相同一类样本, 或终结点仅包含最大允许数目 Max 的样本.

3. 在根结点 t 选择特征基因及 $cutoff$ 值, 确定最优分割.

3.1 在结点 t 计算先验概率 $P(\omega_j | t) = n_j / n$, 由公式(1)计算 $E(t)$.

3.2 对每个属性基因 g_j ($j=1, 2, \dots, p$) 的不同取值 g_{ij} ($i=1, 2, \dots, n$) 进行排序(从大到小).

3.3 对属性基因 g_j , 由 $s_{ij} = g_{ij} + (g_{ij} - g_{i+1,j}) / 2$ 得到对应 g_j 所有分划 s_{ij} .

3.4 由公式(2)计算所有 $\Delta E(s, t)$, 并记录满足条件(2)的对应的 s^* 和基因 g_j^d .

3.5 由 s^* 对 t 结点进行的二叉划分为 t_1, t_2 , 对 t_1, t_2 重复 3.1~3.5 过程, 直至满足树增大的终止规则, 即叶子结点只包含相同一类样本, 或在此结点有最大的允许量, 得到特征基因组 $G_d = \{g_1^d, g_2^d, \dots, g_d^d\}$.

4. 由式(3)计算 χ^2 值及评价指标: 一致性指标 $acc(G_d)$ 、预测真阳性率 $p(G_d)$ 和识别真阳性率 $r(G_d)$. 其中, $acc = (n_{00} + n_{11}) / n$, $p = n_{11} / (n_{01} + n_{11})$ 和 $r = n_{11} / (n_{10} + n_{11})$, $n, n_{00}, n_{01}, n_{10}$ 和 n_{11} 的意义同前.

5. 若 $\chi^2 > \chi^2_0$ (或 $\alpha < \alpha_0$), 记录特征基因组 $G_d = \{g_1^d, g_2^d, \dots, g_d^d\}$.

6. 重复步 1~步 5, 建立特征基因组群 $\{G_d\}$.

7. 对每个特征 g_k , 由式(4)计算特征选择强度 $FV(g_k)$.

8. 将样本随机赋予类别标签, 按以上步 1~步 7 的步骤, 计算各基因的选择强度, 重复 N 次. 根据选定的显著性水平 β , 确定临界值 FV_β^0 .

9. 输出特征选择强度不小于临界值 FV_β^0 的各特征基因及相应的选择强度.

整个算法是在 Matlab 6.0 平台上实现的, 其中部分程序采用美国 Infinity Technology Associates, Inc. 工具箱^[12] 的子模块^①.

3 应用实例与结果

我们将 EFST 应用于结肠癌基因表达谱实验数据^{[15]②}, 该实验采用 Affymetrix 公司的点有 2000

① <http://www.infinityassociates.com/>.

② www.sph.uth.tmc.edu/hgc.

个寡聚核苷酸探针组的 DNA 芯片,样本包括 40 例结肠癌和 22 例正常组织的样本,分别用类别标签 0,1 对正常结肠组织和结肠癌组织进行标定。对每个检测样本,获取 2000 个基因表达谱数据,其中每个基因都具有一定的生物学意义。

3.1 EFST 特征基因选择

采用 5-倍交叉验证方法建立基于不同分布的实验集,将样本集(正常组织,癌组织)分别随机近似地

等分成 5 组,即正常组分为组 $N_i (i=1, 2, \dots, 5)$, 结肠癌组织分为组 $D_i (i=1, 2, \dots, 5)$ 。分别取 $N_i (i=1, 2, \dots, 5)$ 和 $D_i (i=1, 2, \dots, 5)$ 中的一组 N_i 和 D_i 放在一起作为实验组,剩余 4/5 作为训练组。每一次随机分组可构建 25 对实验组和训练组,这样随机分组 20 次,产生了 500 对实验组和训练组。采用 EFST 特征基因选择算法,分析了等权 $w_d = 1$ (表 1)和非等权 $w_d = \chi_d^2$ (附表 1)两种情况所选特征基因组 G_d 。

表 1 基于不同训练集分布结构特征选择在 4 个水平上的相合性(等权 $w_d = 1$)

不限制 α_0		$\alpha_0 = 0.1 (\chi_0^2 = 2.71)$		$\alpha_0 = 0.05 (\chi_0^2 = 3.84)$		$\alpha_0 = 0.01 (\chi_0^2 = 6.63)$	
Gene ID	FV	Gene ID	FV	Gene ID	FV	Gene ID	FV
1671	0.4617	1671	0.5340	1671	0.5900	1671	0.6300
249	0.2715	249	0.3840	249	0.3800	47	0.4760
47	0.1934	47	0.3400	47	0.3680	249	0.4000
2	0.1903	576	0.3080	576	0.3460	576	0.3520
493	0.1895	2	0.2900	2	0.3080	2	0.3520
1	0.1462	682	0.1680	682	0.2180	201	0.2080
576	0.1315	1	0.1600	1	0.1800	72	0.2000
72	0.1176	3	0.1420	72	0.1500	1	0.1860
5	0.1083	72	0.1340	3	0.1400	682	0.1700
3	0.0982	737	0.0960	201	0.1240	1623	0.1480
1346	0.0982	201	0.0960	737	0.1140	491	0.1100
682	0.0905	9	0.0800	1504	0.0900	1346	0.0920
201	0.0603	1504	0.0780	9	0.0820	5	0.0920
9	0.0518	493	0.0760	1623	0.0620	3	0.0820
776	0.0510	5	0.0720	776	0.0600	737	0.0780
737	0.0503	1346	0.0640	5	0.0600	9	0.0620
1772	0.0487	491	0.0600	491	0.0580	1500	0.0440
1473	0.0441	776	0.0560	1500	0.0500	980	0.0440
11	0.0379	1500	0.0400	547	0.0420	550	0.0440
1423	0.0364	1623	0.0340	1346	0.0380	1952	0.0400

相应用于不同的权,给定 4 个显著性水平 α_0 (不限制 $\alpha_0; \alpha_0 = 0.1; \alpha_0 = 0.05$ 和 $\alpha_0 = 0.01$),用于检验每次选择出的一组特征基因集合 G_d 的分类能力是否显著,在每个水平各做 500 次特征选择,再分别做集成特征选择。我们分别得到 500 个特征属性基因组 $G_d (d=1, 2, \dots, 500)$,并获得特征基因选择强度的 FV 分布。同时,做 FV 的零分布 FV^0 ,取临界值 $FV_\beta^0 = 0.035 (\beta=0.01, w_d=1)$ 和 $FV_\beta^0 = 0.034 (\beta=0.01, w_d=\chi_d^2)$,选择所有 $FV \geq FV_\beta^0 (\beta=0.01)$ 的特征基因。结果显示:等权时所选的特征基因覆盖了非等权时所选的特征基因的 95%,表 1 也显示:在 $\alpha_0 = 0.1$ 的情况下,选择出了 19 个基因,在其它情况下都选择了 20 个有显著意义的基因,在 4 种显著性水平情况下,被共同选中的特征基因占 80%。这一结果提示用 EFST 方法选择特征基因具有高度的稳定性与一致性。特别值得注意的是,不限制 α_0 ,即对每次选择出的一组特征基因集合 G_d 的分类能力不做显著

性检验,只通过选择强度的临界值就能够获得很好的特征选择结果。所以,按照 EFST 方法的性质,在利用各个分类树对特征进行初选的阶段,特征入选的显著性标准可以较低,甚至不加限制,随机地选入的一些无关特征的选择强度可能会较小,可以通过选择强度被排除。这是 EFST 方法的一个显著的优点。

随着显著性水平增高,被 EFST 选到的特征基因被选择强度也增高。这是由于按照严格的显著性检验,在利用各个分类树对特征进行初选的阶段,可以入选的特征减少,造成入选的特征较为集中。这样,也可以使入选的特征较快地达到最终的显著性水平。

3.2 EFST 方法对 SVM 与统计模式分类学习算法的有效性

为证实 EFST 特征基因选择方法对提高不同模式分类算法的有效性,我们应用支持向量机(SVM)^[13]、Fisher 线性判别^[14]、K 最邻近(NN)^[10]和 Logistic

非线性判别^[14]4 种模式分类方法, 分别基于 3 组特征基因:(1)所有基因;(2)由 EFST 方法选择出的特征基因;(3)随机选择的结肠癌基因的表达谱实验数据进行分类分析.

我们选择表 1 的第 1 列 20 个特征基因参与分类. 表 2 给出了线性与非线性支持向量机在 6 种不同的核函数: 线性(linear), $d=1, 2, 3, 4$ 阶多项式核函数(Ploy1-Ploy4) 和 径向基(rbf)的情况下, 分别只用 EFST 方法选择的 20 个基因和用全部 2000 个基因参与结肠癌模式分类的结果, 每种支持向量机方法运行 100 次, 一致性指标 Acc 、预测真阳性率 p 和识别真阳性率 r 是运行 100 次的平均结果.

我们可以看到在 6 种不同核函数下, 用 EFST 方法选择的 20 个基因的指标(Acc , p 和 r)几乎全部大于 2000 个基因参与模式分类相应的指标, 其中径向基核函数 rbf 的指标大幅度提高.

表 2 同组数据特征基因选择前后 SVM 分类器性能的比较

Kernel	特征基金	Acc	p	r
linear	EFST20	0.8878	0.9066	0.9307
	2000	0.8241	0.8736	0.8658
Poly1	EFST20	0.8884	0.9064	0.9303
	2000	0.6462	0.6462	1.0000
Poly2	EFST20	0.8884	0.9007	0.9306
	2000	0.8882	0.9280	0.9023
Poly3	EFST20	0.8472	0.8788	0.9002
	2000	0.8227	0.8736	0.8631
Poly4	EFST20	0.8350	0.8747	0.8638
	2000	0.8214	0.8718	0.8630
rbf	EFST20	0.8992	0.9254	0.9246
	2000	0.6462	0.6462	1.0000

表 3 给出了在 Fisher 线性、 K 最近邻($K=3$) 和 Logistic 非线性判别分类器情况下, 分别用 EFST 方法选择的 20 个基因和随机选取 20 个基因参与模式分类的结果, 每种方法运行 100 次, 指标正确率 Acc , p 和 r 是运行 100 次的平均结果. 我们可以看到在 3 种不同分类器下 EFST 方法选择的 20 个基因的指标(Acc , p 和 r)全部大于随机选取 20 个基因参与模式分类相应的指标, 其中 Fisher 线性判别的指标大幅度提高.

表 3 同组数据 EFST 选择特征基因与随机选取基因不同分类器性能比较

方法	特征基金	Acc	p	r
Fisher 线性	EFST 20	0.882	0.912	0.910
	随机 20	0.592	0.751	0.600
K 最近邻	EFST 20	0.635	0.909	0.925
	随机 20	0.759	0.773	0.910
Logistic 非线性判别	EFST 20	0.747	0.834	0.765
	随机 20	0.660	0.729	0.775

4 讨论及未来的工作

对一般的特征选择方法, 一个重要的目标是排除高度相关的、冗余的指标, 追求识别单独一组特征集合. 这种策略在基因表达谱分析这类生物医学问题中, 存在着严重的缺陷, 不能够充分地发现与疾病病因关联的基因信息. 事实上, 在同一条代谢通路上的基因、共调控基因等功能相关的基因的表达倾向于高度相关, 在一般的特征选择的过程中, 这些基因通常就是所谓冗余的指标, 不会被选择到. 如果我们能够发现多个有分类意义的特征组, 就可以分析它们是否集中于某一条或几条代谢通路或功能类, 进而分析疾病的病因. 此时, 发现“冗余”的基因特征组特别有意义.

EFST 方法是基于分类树分类的一种特征基因选择的方法, 可以有效处理基因表达谱数据高相关(冗余)的特点, 通过选择多组特征基因, 最后集成获得一组分类意义明确的特征基因, 并提供特征基因选择结果的显著性与稳定性的测度. 同时, 我们发现 EFST 方法不仅有较强的降维能力, 而且对支持向量机(SVM)和传统的模式分类 Fisher 判别、最邻近(NN)和 Logistic 非线性判别法也有较强的适应性, 几乎在所有情况下提高了分类的准确率. 这两个特点在识别疾病相关靶基因、识别疾病亚型和疾病模式分类中很有实际意义.

我们将进一步研究 EFST 方法的特性, 研究由分类树选择的多组特征基因(未集成前)之间的关联关系, 构建基因相关网络, 并给出相应的生物学解释.

参 考 文 献

- 1 DeRisi J. L. et al.. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 1997, 278:680~685
- 2 Golub T. R. et al.. Molecular Classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 1999, 286:531~537
- 3 Cmll J. C. et al.. A new approach for filtering noise from high-density oligonucleotide microarray datasets. *Nucleic Acids Research*, 2001, 29(15):15~72
- 4 Hall M.. Correlation-based feature selection for machine learning[Ph. D. dissertation]. Department of Computer Science, University of Waikato, Hamilton, 1998
- 5 Blum A. L., Langley P.. Selection of relevant features and examples in machinelearning. *Artificial Intelligence*, 1997, 97(1)

- ~2):245~271
- 6 Kohavi R., John G. H.. Wrappers for feature subset selection. Artificial Intelligence, 1997, 97(1~2):273~324
- 7 Xing E. P., Jordan M. I., Karpy R. M.. Feature selection for high-dimensional genomic microarray data. In: Proceedings of International Conference on Machine Learning, Western Massachusetts, 2001, 601~608
- 8 Dietterich T. G.. Ensemble methods in machine learning. In: Proceedings of the 1st International Workshop on Multiple Classifier Systems. In: Roli F. ed.. Lecture Notes in Computer Science. New York: Springer, 2000, 1~15
- 9 Zhang H. P., Singer B.. Recursive Partitioning in the Health Sciences. New York: Springer, 1999
- 10 Breiman L.. Bagging predictors. Machine Learning, 1996, 24(2):123~140
- 11 Guo Zheng, Li Xia, Rao Shao-Qi. Analysis of Medical Data. Harbin: Harbin Publisher, 2002(in Chinese)
- (郭政, 李霞, 饶绍奇. 医学信息分析方法. 哈尔滨: 哈尔滨出版社, 2002)
- 12 Martinez W. L., Martinez A. R.. Computational Statistics Handbook with MATLAB. Chapman & Hall/CRC, Boca Raton, 2002
- 13 Brown M. P. S., Grundy W. N., Lin D. et al.. Support vector machine classification of microarray gene expression data. Department of Computer Sciences, University of California, Santa Cruz: Technical Report USCC-CRL-99-09, 1999
- 14 Li X., Rao S. Q. et al.. Genetic mapping of complex discrete human diseases by discriminant analysis. Progress in Natural Science, 2002, 12(6):27~33
- 15 Alon U., Barkai N., Notterman D. A., Gish K. et al.. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences, 1999, 96(12):6745~6750

附表1 基于不同训练集分布结构特征选择在4个水平上的相合性($w_d = \chi_d^2$)

不限制 α_0		$\alpha_0 = 0.1 (\chi_0^2 = 2.71)$		$\alpha_0 = 0.05 (\chi_0^2 = 3.84)$		$\alpha_0 = 0.01 (\chi_0^2 = 6.63)$	
Gene ID	FV	Gene ID	FV	Gene ID	FV	Gene ID	FV
1671	0.5334	1671	0.5945	1671	0.5439	1671	0.6422
249	0.3684	47	0.4011	249	0.3994	47	0.4836
47	0.3289	249	0.3848	47	0.3848	249	0.3849
576	0.2826	576	0.3534	576	0.3409	2	0.3502
2	0.2718	2	0.3108	2	0.2908	576	0.3412
682	0.1620	682	0.2325	682	0.1932	201	0.2231
72	0.1530	1	0.1860	1	0.1772	72	0.2158
3	0.1431	72	0.1577	72	0.1461	682	0.1895
1	0.1248	3	0.1383	3	0.1425	1	0.1824
201	0.1191	201	0.1355	201	0.1103	1623	0.1632
493	0.0957	737	0.1074	737	0.0954	491	0.1121
737	0.0785	1504	0.0851	9	0.0818	5	0.0839
9	0.0755	9	0.0801	1504	0.0802	1346	0.0839
1346	0.0733	1623	0.0763	491	0.0758	3	0.0794
5	0.0642	491	0.0698	5	0.0637	737	0.0713
1504	0.0511	5	0.0563	1346	0.0563	9	0.0604
776	0.0497	776	0.0549	493	0.0563	550	0.0401
547	0.0438	1500	0.0456	776	0.0505	980	0.0401
1772	0.0378	547	0.0437	1623	0.0503	1500	0.0401
1623	0.0321	1346	0.0392	1500	0.0428	1952	0.0378
1473	0.0320	111	0.0261	547	0.0375	1466	0.0258



LI Xia, born in 1957, Ph. D. candidate, professor. Her current research interests include machine learning, data mining, artificial intelligence and knowledge discovery.

ZHANG Tian-Wen, born in 1940, professor, Ph. D. supervisor. His current research interests are in the areas of pattern recognition, image process and computer vision.

GUO Zheng, born in 1963, Ph. D. candidate, professor. His current interests include artificial intelligence, pattern recognition and data mining.

Background

This work is supported partly by the National Natural Science Foundation of China, “Pattern Recognition and Feature Extraction Techniques of Gene Mapping of Complex Human Diseases” (Grant No 30170515) and “Feature Gene Recognition Techniques of Gene Expression Profiles of Complex Human Diseases” (Grant No 30370798). This work is also supported partly by the National High Technology Research and Development Program (863 program), “Gene Chip Platform for the Analysis of the Anti-cancer Mechanism of Chinese Traditional Medicine” (Grant No. 2003AA2Z2051), and the 211 Project of the Tenth ‘Five-year’ Plan of Harbin Medical University, “Integrated Bioinformatics Analysis Platform of Gene Function and Drug Targets Identification”. All these projects involve in the identification of complex disease genes and drug targets.

This paper proposes a novel feature gene selection meth-

od named EFST (Ensemble Feature Selection Based on Recursive Partition-Tree) which can be applied to select multiple feature gene groups from one training sample set. Authors suggest that the proposed ensemble method is a powerful tool not only for prediction or classification of complex disease but also for complex disease gene discovery, which are one of the key problems in all the projects mentioned above.

The research focus of the group is on bioinformatics and computational biology. Since 1993, they have undertaken 22 projects, published over 90 papers and 7 books, and won 12 academic rewards in total. The PPAP(Population and Pedigree Analysis Package of human genetics) software developed by them has been applied successfully by 40 more research groups.